

# Inference for Robust Canonical Variate Analysis

Stefan Van Aelst · Gert Willems

The date of receipt and acceptance will be inserted by the editor

**Abstract** We consider the problem of optimally separating two multivariate populations. Robust linear discriminant rules can be obtained by replacing the empirical means and covariance in the classical discriminant rules by S or MM-estimates of location and scatter. We propose to use a fast and robust bootstrap method to obtain inference for such a robust discriminant analysis. This is useful since classical bootstrap methods may be unstable as well as extremely time-consuming when robust estimates such as S or MM-estimates are involved. In particular, fast and robust bootstrap can be used to investigate which variables contribute significantly to the canonical variate, and thus the discrimination of the classes. Through bootstrap, we can also examine the stability of the canonical variate. We illustrate the method on some real data examples.

**Keywords** bootstrap · canonical variate · linear discriminant analysis · robustness

**Subject classification** AMS 62H30, AMS 62F35, AMS 62F40, JEL C63

## 1 Introduction

Linear discriminant rules are widely used to find representations of multivariate data that optimally separate the observations in two or more populations. We consider the situation with two  $p$ -dimensional populations,  $\Pi_1$  and  $\Pi_2$ , having respective population means  $\mu_1$  and  $\mu_2$ . It is assumed that the two populations share a common covariance matrix  $\Sigma$ . Furthermore, we also assume

---

Stefan Van Aelst and Gert Willems  
Department of Applied Mathematics and Computer Science, Ghent University, Krijgslaan  
281 S9, 9000 Gent, Belgium  
Phone: ++32-9-264 4908  
Fax: ++32-9-264 4995  
E-mail: Stefan.VanAelst@UGent.be, Gert.Willems@UGent.be

equal prior probabilities. The linear Bayes rule then classifies an observation  $\mathbf{x} \in \mathbb{R}^p$  into population  $\Pi_1$  if  $d_1^L(\mathbf{x}) > d_2^L(\mathbf{x})$ , where

$$d_j^L(\mathbf{x}) = \boldsymbol{\mu}_j^t \boldsymbol{\Sigma}^{-1} \mathbf{x} - \frac{1}{2} \boldsymbol{\mu}_j^t \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_j; \quad j = 1, 2, \quad (1)$$

and into population  $\Pi_2$  otherwise. The direction  $\mathbf{a}$  that best separates the two populations is given by  $(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \boldsymbol{\Sigma}^{-1}$ . The corresponding projection  $\mathbf{a}^t \mathbf{x}$  is called the canonical variate or discriminant coordinate.

Since  $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$  and  $\boldsymbol{\Sigma}$  are unknown, they need to be estimated from an available training sample of the form  $\mathcal{Z}_n = \{\mathbf{x}_{11}, \dots, \mathbf{x}_{1n_1}, \mathbf{x}_{21}, \dots, \mathbf{x}_{2n_2}\} \subset \mathbb{R}^p$  with  $n_1$  the number of samples from population 1,  $n_2$  the number of samples from population 2 and  $n = n_1 + n_2$ , the total sample size. Fisher's classical linear discriminant analysis is based on the empirical means and covariances of the training data  $\mathcal{Z}_n$ . Alternatively, more robust discriminant analysis methods are naturally obtained by using robust estimates of location and scatter instead (see e.g. He and Fung 2000, Croux and Dehon 2001, Hubert and Van Driessen 2004, Croux et al. 2008, Bianco et al. 2008).

Many robust estimators of multivariate location and scatter have been proposed in the literature. See e.g. Maronna et al. (2006) or Hubert et al. (2008) for a recent overview. In this paper we use the classes of S-estimators (Davies 1987, Rousseeuw and Leroy 1987, Lopuhaä 1989) and MM-estimators (Tatsuoka and Tyler 2000) to robustly estimate the centers of the populations and their common scatter matrix. Inference for these estimators can be derived from their asymptotic distribution. However, this asymptotic distribution is mainly known for elliptical model distributions, an assumption which is not appropriate in those cases where robust estimation is most recommended, i.e. for data with outliers. Inference based on the asymptotic variances derived at the central model may still yield reasonable results for large samples with a small fraction of contamination. The bootstrap (Efron 1979) is a computer-intensive alternative that can be more reliable for smaller sample sizes and for larger deviations from the central model. Moreover, because the bootstrap estimates the sampling distribution of the estimators, it has applications beyond the standard inference procedures of calculating standard errors, confidence intervals or p-values for hypothesis tests. For example, bootstrap allows us to assess the stability of the canonical variates by investigating the distribution of the angle between the estimated canonical variate and its population counterpart.

Applying the standard bootstrap on robust estimators raises a computational issue due to the high computation time of robust estimators as well as a robustness issue due to the varying amount of outliers in bootstrap samples. Both issues can be solved at once by the fast and robust bootstrap (FRB), introduced by Salibián-Barrera and Zamar (2002) in the context of robust regression based on MM-estimators. The FRB has later been extended to robust multivariate regression (Van Aelst and Willems 2005) and robust principal components analysis (Salibián-Barrera et al. 2006) based on S or MM-estimators. The FRB has also been used successfully for robust Wald tests in

linear models (Salibian-Barrera 2005), robust likelihood ratio type tests (Van Aelst and Willems 2009) and robust linear model selection (Salibian-Barrera and Van Aelst 2008). Here, we use the FRB to obtain many recalculations of the robust S or MM-estimates of the locations and common scatter matrix in a linear discriminant analysis. These FRB estimates can then be used for inference purposes. Moreover, in the context of discriminant analysis, we use the FRB to investigate which variables contribute significantly to the canonical variates. In this way we can investigate which variables carry discriminatory power.

The rest of this paper is organized as follows. In Section 2 we review multivariate S and MM-estimators. Section 3 explains the fast and robust bootstrap in this setting. In Section 4 we show some useful applications of the FRB for discriminant analysis and investigate the performance of the FRB through simulation. Section 5 illustrates the method on some real data examples and Section 6 concludes.

## 2 Multivariate S and MM-estimators

In the multivariate one-sample setting S-estimators are defined as follows. Suppose we have a sample  $\{\mathbf{x}_1, \dots, \mathbf{x}_n\} \subset \mathbb{R}^p$ . Then, for a function  $\rho_0 : [0, \infty[ \rightarrow [0, \infty[$  which is bounded, increasing and sufficiently smooth, the S-estimates of location and scatter  $(\tilde{\boldsymbol{\mu}}_n, \tilde{\boldsymbol{\Sigma}}_n)$  minimize  $|C|$  subject to

$$\frac{1}{n} \sum_{i=1}^n \rho_0 \left( [(\mathbf{x}_i - T)^t C^{-1} (\mathbf{x}_i - T)]^{\frac{1}{2}} \right) = b \quad (2)$$

among all  $T \in \mathbb{R}^p$  and  $C \in \text{PDS}(p)$ . Here,  $\text{PDS}(p)$  denotes the set of positive definite symmetric  $p \times p$  matrices and  $|C|$  denotes the determinant of the square matrix  $C$ . In this paper, the loss function  $\rho_0$  is taken from the common class of Tukey biweight functions, given by  $\rho_c(t) = \min(t^2/2 - t^4/(2c^2) + t^6/(6c^4), c^2/6)$ . The constant  $b$  is usually chosen such that  $b = E_{\Phi}[\rho(\|\mathbf{x}\|)]$ , where  $\Phi$  is the multivariate standard normal distribution. This ensures consistency of the S-estimator at the normal model. The breakdown point of an estimator is the smallest fraction of contamination that can have an arbitrarily large effect on the estimator. The asymptotic breakdown point of the S-estimators  $\tilde{\boldsymbol{\mu}}_n$  and  $\tilde{\boldsymbol{\Sigma}}_n$  equals  $\min(b/\rho_c(\infty), 1 - b/\rho_c(\infty))$  (Lopuhaä and Rousseeuw 1991). Hence, for any given dimension  $p$ , the Tukey biweight loss function  $\rho_c$  can be tuned in order to achieve a 50% breakdown point, by choosing the constant  $c$  appropriately. However, the choice of  $c$  does not only determine the breakdown point of the S-estimator, but at the same time affects the efficiency of the S-estimators. Therefore, high-breakdown point S-estimators can have quite low efficiency at normal distributions, especially in lower dimensions (see e.g. Salibian-Barrera et al. 2006).

MM-estimators have been introduced as a class of robust estimators that can attain high breakdown point and high Gaussian efficiency at the same time.

MM-estimators were first introduced in regression by Yohai (1987). Multivariate (one-sample) MM-estimators of location and shape have been introduced by Tatsuoka and Tyler (2000) as follows. Let  $\tilde{\Sigma}_n$  be the S-estimate of scatter and denote  $\hat{\sigma}_n := |\tilde{\Sigma}_n|^{1/(2p)}$  the corresponding S-estimate of multivariate scale. Let  $\rho_1$  be a loss function from the same class as  $\rho_0$ . Then, the multivariate MM-estimates of location and shape  $(\hat{\boldsymbol{\mu}}_n, \hat{\Gamma}_n)$  minimize

$$\frac{1}{n} \sum_{i=1}^n \rho_1 \left( [(\mathbf{x}_i - T)^t G^{-1} (\mathbf{x}_i - T)]^{\frac{1}{2}} / \hat{\sigma}_n \right)$$

among all  $(T, G) \in \mathbb{R}^p \times \text{PDS}(p)$  for which  $|G|=1$ . The corresponding MM-estimator for the scatter matrix is given by  $\hat{\Sigma}_n = \hat{\sigma}_n^2 \hat{\Gamma}_n$ .

The MM-estimator thus starts from the highly robust S-estimate of multivariate scale and then estimates the location and shape using a different loss function  $\rho_1$ . In this way, the location and shape estimates inherit the breakdown point of the initial S-estimate of multivariate scale as determined by the loss function  $\rho_0$  (Tyler 2002, Salibián-Barrera et al. 2006). Hence, the loss function  $\rho_1$  can be tuned to obtain a high efficiency, e.g. 95%, at the normal model.

To obtain the linear discriminant scores, we need a robust estimate of the common covariance matrix  $\Sigma$  of the two populations involved. Similarly as for the classical estimates, we can start from the robust scatter estimates  $\hat{\Sigma}_{1n_1}$  and  $\hat{\Sigma}_{2n_2}$  for the individual groups and then calculate a pooled scatter estimate  $\hat{\Sigma}_n$  as

$$\hat{\Sigma}_n = \frac{n_1 \hat{\Sigma}_{1n_1} + n_2 \hat{\Sigma}_{2n_2}}{n_1 + n_2}.$$

This is the approach taken by Croux and Dehon (2001), Hubert and Van Driessen (2004), Croux et al. (2008), and Bianco et al. (2008) among others.

Alternatively, the definition of the S and MM-estimators can be adjusted to the multigroup setting, as proposed by He and Fung (2000). For example, simultaneous S-estimates of the two locations and the common scatter matrix can be defined as the solution  $\hat{\boldsymbol{\mu}}_{1n}$ ,  $\hat{\boldsymbol{\mu}}_{2n}$  and  $\hat{\Sigma}_n$  that minimizes  $|C|$  subject to

$$\frac{1}{n_1 + n_2} \sum_{j=1}^2 \sum_{i=1}^{n_j} \rho \left( [(\mathbf{x}_{ji} - T_j)^t C^{-1} (\mathbf{x}_{ji} - T_j)]^{\frac{1}{2}} \right) = b \quad (3)$$

among all  $T_1, T_2 \in \mathbb{R}^p$  and  $C \in \text{PDS}(p)$ . Similarly, simultaneous MM-estimates for the two locations and common shape/scatter can be defined (see Van Aelst and Willems 2009 for details).

### 3 Fast and robust bootstrap

The fast and robust bootstrap procedure assumes that the robust estimates can be written as a solution of a set of sufficiently smooth fixed point equations.

This is indeed the case for the S and MM-estimates defined in the previous Section. For example, the multivariate one-sample S and MM-estimates can be written in the following way (see e.g. Salibián-Barrera et al. (2006)):

$$\tilde{\boldsymbol{\mu}}_n = \left( \sum_{i=1}^n \frac{\rho'_0(\tilde{d}_i)}{\tilde{d}_i} \right)^{-1} \left( \sum_{i=1}^n \frac{\rho'_0(\tilde{d}_i)}{\tilde{d}_i} \mathbf{x}_i \right) \quad (4)$$

$$\tilde{\Sigma}_n = \frac{1}{nb} \left( \sum_{i=1}^n p \frac{\rho'_0(\tilde{d}_i)}{\tilde{d}_i} (\mathbf{x}_i - \tilde{\boldsymbol{\mu}}_n)(\mathbf{x}_i - \tilde{\boldsymbol{\mu}}_n)^t + \left( \sum_{i=1}^n \tilde{w}_i \right) \tilde{\Sigma}_n \right) \quad (5)$$

$$\hat{\boldsymbol{\mu}}_n = \left( \sum_{i=1}^n \frac{\rho'_1(d_i/|\tilde{\Sigma}_n|^{1/(2p)})}{d_i} \right)^{-1} \left( \sum_{i=1}^n \frac{\rho'_1(d_i/|\tilde{\Sigma}_n|^{1/(2p)})}{d_i} \mathbf{x}_i \right) \quad (6)$$

$$\hat{I}_n = H \left( \sum_{i=1}^n \frac{\rho'_1(d_i/|\tilde{\Sigma}_n|^{1/(2p)})}{d_i} (\mathbf{x}_i - \hat{\boldsymbol{\mu}}_n)(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_n)^t \right) \quad (7)$$

where the function  $H$  is defined as  $H(A) = |A|^{-1/p} A$  with  $A$  a square matrix of size  $p$ . Moreover,  $d_i = [(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_n)^t \hat{I}_n^{-1} (\mathbf{x}_i - \hat{\boldsymbol{\mu}}_n)]^{1/2}$ ,  $\tilde{d}_i = [(\mathbf{x}_i - \tilde{\boldsymbol{\mu}}_n)^t \tilde{\Sigma}_n^{-1} (\mathbf{x}_i - \tilde{\boldsymbol{\mu}}_n)]^{1/2}$  and  $\tilde{w}_i = \rho_0(\tilde{d}_i) - \rho'_0(\tilde{d}_i)\tilde{d}_i$ . Similarly, the simultaneous S and MM-estimators of the two locations and common scatter can be written as a solution of fixed point equations (Van Aelst and Willems 2009).

Let  $\hat{\boldsymbol{\theta}}_n$  be a vector of length  $d$  that collects all parameter estimates of interest. In our case  $\hat{\boldsymbol{\theta}}_n$  contains the location estimates as well as the scatter estimates in vectorized form. Then, a set of fixed-point equations such as (4)-(7) can be written as

$$\hat{\boldsymbol{\theta}}_n = \mathbf{g}_n(\hat{\boldsymbol{\theta}}_n) \quad (8)$$

where the function  $g_n : \mathbb{R}^d \rightarrow \mathbb{R}^d$  depends on the training sample  $\mathcal{Z}_n$ . Given a bootstrap sample  $\mathcal{Z}_n^*$  (i.e. a sample of size  $n$  drawn with replacement from  $\mathcal{Z}_n$ ), the recalculated estimate  $\hat{\boldsymbol{\theta}}_n^*$  then is the solution of the corresponding fixed point equation  $\hat{\boldsymbol{\theta}}_n^* = \mathbf{g}_n^*(\hat{\boldsymbol{\theta}}_n^*)$ , where the function  $\mathbf{g}_n^*$  now depends on  $\mathcal{Z}_n^*$ . However, in case of high-breakdown robust estimators such as S or MM-estimators, computing  $\hat{\boldsymbol{\theta}}_n^*$  for every bootstrap sample  $\mathcal{Z}_n^*$  becomes a computationally expensive task. This makes it infeasible to obtain a large number of recalculations in a reasonable amount of time. Moreover, even if the robust estimates for the original sample (corresponding to the solution of (8)) were able to resist the effect of the outliers in  $\mathcal{Z}_n$ , this does not guarantee that  $\hat{\boldsymbol{\theta}}_n^*$  will be equally resistant. Indeed, due to the resampling with replacement, bootstrap samples may contain a larger fraction of outliers than the original sample. Hence,  $\mathbf{g}_n^*$  is potentially more severely affected by outliers than  $\mathbf{g}_n$ .

An intuitive and cheap way to obtain an approximation for the recalculated estimates  $\hat{\boldsymbol{\theta}}_n^*$  corresponding to each bootstrap sample would be to calculate

$$\hat{\boldsymbol{\theta}}_n^{1*} := \mathbf{g}_n^*(\hat{\boldsymbol{\theta}}_n). \quad (9)$$

For example, for the multivariate one-sample S and MM-estimates corresponding to (4)-(7) this leads to the following equations:

$$\tilde{\boldsymbol{\mu}}_n^{1*} = \left( \sum_{i=1}^n \frac{\rho'_0(\tilde{d}_i^*)}{\tilde{d}_i^*} \right)^{-1} \left( \sum_{i=1}^n \frac{\rho'_0(\tilde{d}_i^*)}{\tilde{d}_i^*} \mathbf{x}_i^* \right) \quad (10)$$

$$\tilde{\Sigma}_n^{1*} = \frac{1}{nb} \left( \sum_{i=1}^n p \frac{\rho'_0(\tilde{d}_i^*)}{\tilde{d}_i^*} (\mathbf{x}_i^* - \tilde{\boldsymbol{\mu}}_n)(\mathbf{x}_i^* - \tilde{\boldsymbol{\mu}}_n)^t + \left( \sum_{i=1}^n \tilde{w}_i^* \right) \tilde{\Sigma}_n \right) \quad (11)$$

$$\hat{\boldsymbol{\mu}}_n^{1*} = \left( \sum_{i=1}^n \frac{\rho'_1(d_i^*/|\tilde{\Sigma}_n|^{1/(2p)})}{d_i^*} \right)^{-1} \left( \sum_{i=1}^n \frac{\rho'_1(d_i^*/|\tilde{\Sigma}_n|^{1/(2p)})}{d_i^*} \mathbf{x}_i^* \right) \quad (12)$$

$$\hat{\Gamma}_n^{1*} = H \left( \sum_{i=1}^n \frac{\rho'_1(d_i^*/|\tilde{\Sigma}_n|^{1/(2p)})}{d_i^*} (\mathbf{x}_i^* - \hat{\boldsymbol{\mu}}_n)(\mathbf{x}_i^* - \hat{\boldsymbol{\mu}}_n)^t \right) \quad (13)$$

$d_i^* = [(\mathbf{x}_i^* - \hat{\boldsymbol{\mu}}_n)^t \hat{\Gamma}_n^{-1} (\mathbf{x}_i^* - \hat{\boldsymbol{\mu}}_n)]^{1/2}$ ,  $\tilde{d}_i^* = [(\mathbf{x}_i^* - \tilde{\boldsymbol{\mu}}_n)^t \tilde{\Sigma}_n^{-1} (\mathbf{x}_i^* - \tilde{\boldsymbol{\mu}}_n)]^{1/2}$  and  $\tilde{w}_i^* = \rho_0(\tilde{d}_i^*) - \rho'_0(\tilde{d}_i^*) \tilde{d}_i^*$ . Note that the right-hand side of expressions (10)-(13) only depends on the robust estimates for the original sample. Hence, calculating these approximations only involves calculating weighted means and covariances, which is very easy and computationally efficient. A similar result holds for simultaneous S or MM-estimates for the two locations and common scatter matrix.

The approximation  $\hat{\boldsymbol{\theta}}_n^{1*}$  in (9) can be viewed as one-step estimation of  $\hat{\boldsymbol{\theta}}_n^*$  starting from the initial value  $\hat{\boldsymbol{\theta}}_n$ . However, since we are keeping the estimates  $\hat{\boldsymbol{\theta}}_n$  fixed on the right-hand side of (9), these approximations will likely underestimate the variability of the MM-estimator. To remedy this, a linear correction can be applied as follows. Using the smoothness of  $\mathbf{g}_n$ , we can calculate a Taylor expansion about  $\hat{\boldsymbol{\theta}}_n$ 's limiting value  $\boldsymbol{\theta}$ ,

$$\hat{\boldsymbol{\theta}}_n = \mathbf{g}_n(\boldsymbol{\theta}) + \nabla \mathbf{g}_n(\boldsymbol{\theta})(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) + R_n, \quad (14)$$

where  $R_n$  is the remainder term and  $\nabla \mathbf{g}_n(\cdot) \in \mathbb{R}^{m \times m}$  is the matrix of partial derivatives. When the remainder term is negligible ( $R_n = o_p(1)$ ), equation (14) can be rewritten as

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \sim [\mathbf{I} - \nabla \mathbf{g}_n(\boldsymbol{\theta})]^{-1} \sqrt{n}(\mathbf{g}_n(\boldsymbol{\theta}) - \boldsymbol{\theta}),$$

where  $\sim$  denotes that both sides have the same limiting distribution. Under certain conditions (see Salibián-Barrera et al. 2006, Section 4.2 for details) we will have that  $\sqrt{n}(\hat{\boldsymbol{\theta}}_n^* - \hat{\boldsymbol{\theta}}_n) \sim \sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})$  and  $\sqrt{n}(\mathbf{g}_n^*(\boldsymbol{\theta}) - \boldsymbol{\theta}) \sim \sqrt{n}(\mathbf{g}_n(\hat{\boldsymbol{\theta}}_n) - \hat{\boldsymbol{\theta}}_n)$ . If we furthermore approximate  $[\mathbf{I} - \nabla \mathbf{g}_n(\boldsymbol{\theta})]^{-1}$  by  $[\mathbf{I} - \nabla \mathbf{g}_n(\hat{\boldsymbol{\theta}}_n)]^{-1}$  we obtain

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n^* - \hat{\boldsymbol{\theta}}_n) \sim [\mathbf{I} - \nabla \mathbf{g}_n(\hat{\boldsymbol{\theta}}_n)]^{-1} \sqrt{n}(\mathbf{g}_n^*(\hat{\boldsymbol{\theta}}_n) - \hat{\boldsymbol{\theta}}_n). \quad (15)$$

We now define the corrected version of the one-step approximation as

$$\hat{\boldsymbol{\theta}}_n^{R*} := \hat{\boldsymbol{\theta}}_n + [\mathbf{I} - \nabla \mathbf{g}_n(\hat{\boldsymbol{\theta}}_n)]^{-1}(\hat{\boldsymbol{\theta}}_n^{1*} - \hat{\boldsymbol{\theta}}_n), \quad (16)$$

which is a better approximation to  $\hat{\theta}_n^*$  than the initial approximation  $\hat{\theta}_n^{1*}$ . Moreover, for one-sample multivariate S and MM-estimators it has been shown that the fast and robust bootstrap approximations  $\hat{\theta}_n^{R*}$  given by (16) are consistent in the sense that they estimate the same limiting distribution as  $\hat{\theta}_n^*$  does (Salibian-Barrera et al. 2006, Theorem 2). The same result can be shown for the simultaneous two-sample S and MM-estimators.

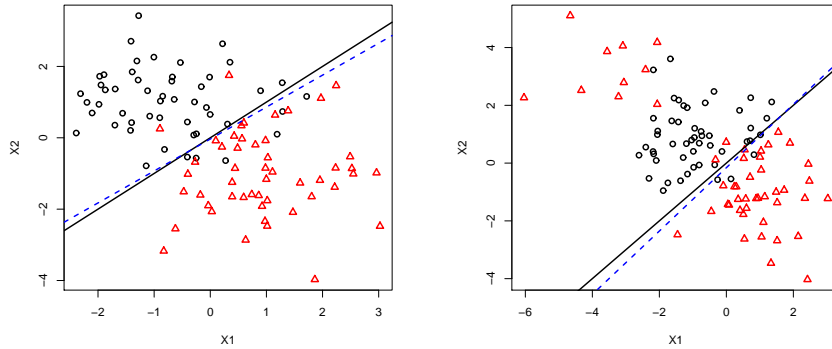
As explained before, calculating the approximation  $\hat{\theta}_n^{1*}$  for each bootstrap sample is easy in our setting. Moreover, note that the correction matrix  $[\mathbf{I} - \nabla \mathbf{g}_n(\hat{\theta}_n)]^{-1}$  needs to be calculated only once, based on the original sample, so also calculating the approximations  $\hat{\theta}_n^{R*}$  requires little effort. The FRB approximations  $\hat{\theta}_n^{R*}$  are also more robust than the completely recalculated bootstrap estimates  $\hat{\theta}_n^*$ . The reason is that any observation that was found to be outlying in the original sample  $\mathcal{Z}_n$ , will be associated with a small weight in the estimating equations. Consequently, this observation will receive the same small weight in the computation of the initial approximation  $\hat{\theta}_n^{1*}$  for any bootstrap sample, no matter how many copies of it were drawn into  $\mathcal{Z}_n^*$ , and hence will be harmless. Clearly, also the final FRB approximations  $\hat{\theta}_n^{R*}$  will thus be little affected by the outliers, as has been shown through simulation in e.g. Salibian-Barrera and Zamar 2002, Salibian-Barrera et al. 2006, Salibian-Barrera et al. 2008). This is also confirmed by the fact that quantiles of the FRB distribution achieve the maximal possible breakdown point (Salibian-Barrera and Zamar 2002, Theorem 2).

#### 4 Applications in discriminant analysis

Bickel and Freedman (1981) have shown that the bootstrap commutes with smooth functions. It has been shown by Salibian-Barrera et al. (2006, Theorem 3) that this property carries over to the FRB. A useful implication of this result for the current setting is the consistency of bootstrapping the coefficients of the canonical variate  $\mathbf{a} = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\boldsymbol{\Sigma}^{-1}$  in a discriminant analysis.

In discriminant analysis, part of the interest can lie in the canonical variate, which is the univariate direction that best separates the two groups according to Fisher's criterion. We can then consider the angle of the estimated canonical variate with respect to the population canonical variate as a performance measure of an estimator. Indeed, if an estimated canonical variate is relatively aligned with its population counterpart, then it provides valuable information regarding the discriminant coordinates of the underlying distribution. On the other hand, canonical variate estimates that can be almost orthogonal to the true canonical variate are far less reliable.

We can assess the variability of the canonical variate estimate by looking at the bootstrap distribution of the angles that the recalculated canonical variates have with the estimated canonical variate of the original data. The angle between the normalized canonical variates  $\hat{\mathbf{a}}^*$  and  $\hat{\mathbf{a}}$  is given by



**Fig. 1** Example of data sets in simulation. The scatterplot of the first two variables is shown. Left: clean data. Right: data contaminated with 20% outliers in the second group. Both groups are of size 50.

$\text{acos}(|\hat{\mathbf{a}}^t \hat{\mathbf{a}}^*|) \in [0, \pi/2]$ . The bootstrap distribution of these angles is then an estimate of the distribution of the angles  $\text{acos}(|\mathbf{a}^t \hat{\mathbf{a}}|)$  between the canonical variate estimator and the population canonical variate.

To investigate how well the FRB can estimate the variability of the canonical variate, we ran a small simulation study. We considered two samples of the same size in  $p = 4$  dimensions where the sample sizes were 25, 50 and 100. Both samples were drawn from a multivariate normal distribution with identity covariance matrix. For the first group, the center was  $\boldsymbol{\mu}_1 = (-1, 1, 0, 0)^t$  while the center of the second group was  $\boldsymbol{\mu}_2 = (1, -1, 0, 0)^t$ . The normalized canonical variate at the population level then equals  $(-1/\sqrt{2}, 1/\sqrt{2}, 0, 0)^t$ . Hence, the first two components carry discriminatory power, while the last two variables do not aid in separating the two populations. To examine the robustness of the FRB, we also consider contaminated data sets with 20% of outliers in the second group. The outliers were generated from the same multivariate normal distribution, but the center was shifted to  $\boldsymbol{\mu}_{\text{out}} = (-3, 3, -3, 3)^t$ . Figure 1 shows the scatterplot of the first two components for examples of clean data (left plot) and contaminated data (right plot) when both samples have size 50. As can be seen, the outliers in group two are close to the observations of the first group. Hence, the outliers can be considered as spurious outliers or as misclassified observations of the first population. The solid line in these plots is the population canonical variate and the dashed line represents the robustly estimated canonical variate using simultaneous two-sample MM-estimates. For each setting, we generated  $m = 500$  data sets and computed the simultaneous two-sample MM-estimates with 50% breakdown and 95% location efficiency. Subsequently, we performed the FRB with  $B = 999$  recalculations.

For each simulated data set we computed the mean angle between the  $B$  bootstrap estimates  $\hat{\mathbf{a}}^{R*}$  of the canonical variate and the original MM-estimate  $\hat{\mathbf{a}}$  of that canonical variate. The average and standard deviation of the  $m = 500$

mean angles are displayed in Table 1 for both the clean data and the data with 20% outliers. Each average is compared to the corresponding Monte Carlo estimate of the mean angle between  $\hat{\mathbf{a}}$  and the population canonical variate  $\mathbf{a}$ , based on the same  $m$  simulated samples. From Table 1 we can see that

**Table 1** Average bootstrap estimates (with standard deviations) of the mean angle between MM canonical variate estimate and population canonical variate for data sets with samples of 25, 50 and 100. Results are shown for both clean data (Eps=0%) and contaminated data (Eps=20%).

Eps		25	50	100
0%	Monte Carlo	0.292	0.207	0.145
	Bootstrap	0.298	0.201	0.144
	(SD)	(0.05)	(0.02)	(0.01)
20%	Monte Carlo	0.325	0.226	0.166
	Bootstrap	0.344	0.227	0.161
	(SD)	(0.14)	(0.07)	(0.04)

the FRB mean angle accurately estimates the angle between the estimated and population canonical variate. The outliers clearly affect the precision of the canonical variate estimate (due to the loss of useful information), but also in the presence of outliers the FRB still accurately estimates the (now larger) angle between the robustly estimated and population canonical variate. Moreover, as expected the FRB estimates become more accurate as the sample size grows.

For inference concerning the individual variables, we can construct for instance confidence intervals for each of the coefficients in the canonical variate. In Tables 2 and 3 we show respectively the observed coverage levels and average length of FRB confidence intervals with respectively 95% and 99% nominal confidence level. From these results we can see that the FRB coverage levels correspond quite well to their respective nominal level. The average length of the confidence intervals decreases with increasing sample size, as expected. Introducing 20% of contamination in the second group affects the average length of the FRB confidence intervals, but the coverage level is maintained well. Hence, we can conclude that the FRB confidence intervals reflect well the imprecision on the estimates of the coefficients, even in the presence of contamination.

To investigate which variables contribute significantly to the canonical variate and thus to the discrimination of the two populations, we can use the duality between confidence intervals and hypothesis tests. We thus check whether the confidence interval for each variable's contribution to the canonical variate contains zero or not. Table 4 shows the results for the test at 5% significance level. The results for the first two variables show that the FRB test has high power to identify the first two variables as relevant for discriminating the two groups. This power is maintained well in the presence of contamination. The

**Table 2** Coverage (average length) of 95% FRB confidence intervals for the coefficients of each variable in the standardized canonical variate. Results are shown for both clean data (Eps=0%) and contaminated data (Eps=20%) with samples of sizes 25, 50 and 100.

Eps		25	50	100
0%	1	93.8 (0.455)	93.8 (0.305)	96.4 (0.248)
	2	94.6 (0.517)	94.2 (0.313)	94.8 (0.219)
	3	94.8 (0.776)	94.4 (0.564)	95.2 (0.291)
	4	94.6 (0.645)	91.0 (0.385)	93.8 (0.312)
20%	1	93.4 (0.393)	94.0 (0.368)	96.0 (0.281)
	2	93.0 (0.385)	93.8 (0.386)	94.8 (0.247)
	3	96.2 (0.601)	94.8 (0.625)	93.4 (0.319)
	4	96.0 (0.681)	96.0 (0.518)	95.8 (0.322)

**Table 3** Coverage (average length) of 99% FRB confidence intervals for the coefficients of each variable in the standardized canonical variate. Results are shown for both clean data (Eps=0%) and contaminated data (Eps=20%) with samples of sizes 25, 50 and 100.

Eps		25	50	100
0%	1	98.2 (0.626)	98.6 (0.400)	99.8 (0.343)
	2	98.4 (0.722)	98.0 (0.414)	98.0 (0.299)
	3	99.0 (0.948)	98.2 (0.740)	98.6 (0.368)
	4	98.0 (0.771)	98.0 (0.495)	98.8 (0.391)
20%	1	97.6 (0.540)	97.8 (0.495)	97.6 (0.540)
	2	97.8 (0.474)	98.6 (0.505)	97.8 (0.474)
	3	99.0 (0.783)	99.0 (0.802)	99.0 (0.783)
	4	99.4 (0.829)	98.6 (0.669)	99.4 (0.829)

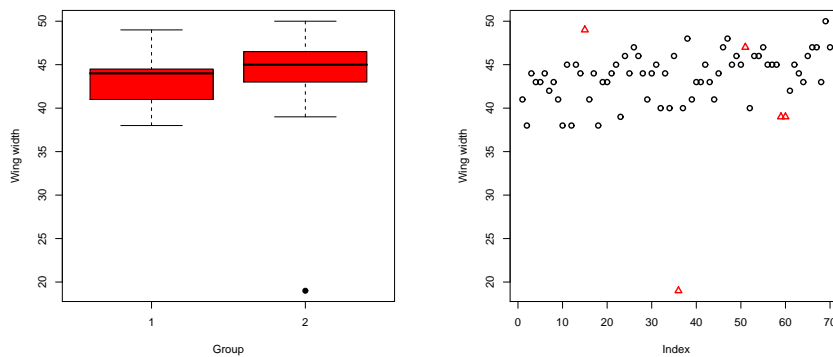
results for the last two variables show that the observed significance level of the FRB test corresponds well to its nominal significance level.

**Table 4** Observed probability of rejecting the null hypothesis that the coefficient of each variable in the standardized canonical variate equals zero when the FRB test is performed at 5% significance level. Results are shown for both clean data (Eps=0%) and contaminated data (Eps=20%) with samples of sizes 25, 50 and 100.

Eps		25	50	100
0%	1	1.000	1.000	1.000
	2	0.966	1.000	1.000
	3	0.052	0.056	0.048
	4	0.054	0.090	0.062
20%	1	1.000	1.000	1.000
	2	0.868	0.996	1.000
	3	0.038	0.052	0.066
	4	0.040	0.040	0.042

## 5 Examples

As a first illustration, we consider the Biting Flies data from Johnson and Wichern (2002). The data set consists of two groups of 35 flies (*Leptoconops torrens* and *Leptoconops carteri*) and we consider the measurements `wing length`, `wing width`, `third palp length`, `third palp width`, and `fourth palp length`. The variable `wing width` contains a clear outlier in the second group as can be seen from the left panel of Figure 2. Hence, a robust discriminant analysis is advisable to reduce the possible effect of outliers. The right panel of Figure 2 shows that the simultaneous two-sample MM-estimates of locations and scatter indeed identify this observation as an outlier and appropriately downweight it in the corresponding robust discriminant analysis.



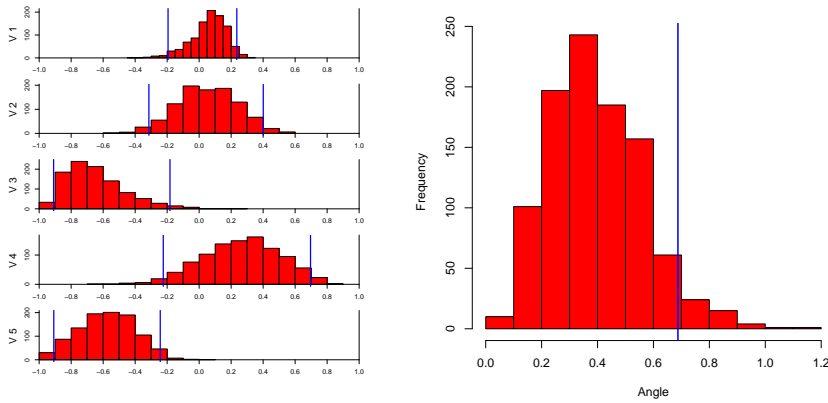
**Fig. 2** Biting Flies data. Left: boxplot of wing width for both groups. Right: Observations flagged as outliers according to the simultaneous two-sample MM-estimates of locations and scatter.

Table 5 shows the standardized coefficients of the canonical variate for both the classical linear discriminant analysis and its robust counterpart based on simultaneous two-sample MM-estimates. From the results in this Table it seems that the effect of the outliers is largest for the coefficients of variable 2 (wing width), variable 3 (third palp length) and variable 5 (fourth palp length). However, if we look at the FRB distribution of the robust coefficient

**Table 5** Biting Flies data. Standardized coefficients of the canonical variate for classical and robust linear discriminant analysis.

Method	V1	V2	V3	V4	V5
Classical	0.084	0.122	-0.825	0.278	-0.469
Robust MM	0.076	0.043	-0.730	0.283	-0.616

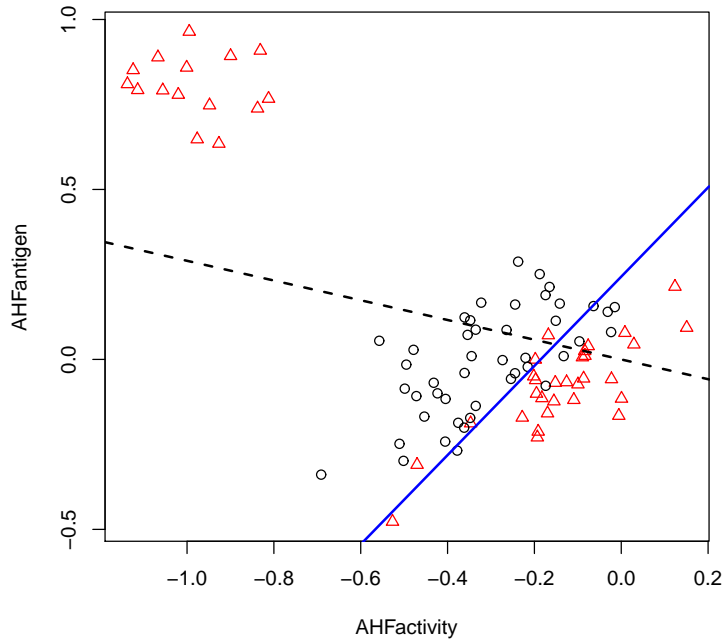
estimates in the left panel of Figure 3 and the corresponding 95% confidence intervals (indicated by the vertical lines in the plot), then we see that the classical coefficient estimates do not differ significantly from the robust estimates. This indicates that the outliers were not very influential on the discriminant analysis. Moreover, from the same plot we can also derive that variables 3 and 5 contain the most discriminatory power. To further investigate the stability of the discriminant analysis, the right panel of Figure 3 shows the FRB distribution of the angle between the estimated and population canonical variate. The vertical line indicates the upper limit (at value 0.69) of a one-sided 95% confidence interval for this angle. This upper limit corresponds with an angle of about 40 degrees, showing that the variability of the canonical variate is quite high.



**Fig. 3** Biting Flies data. Left: FRB distribution of the standardized coefficients of the robustly estimated canonical variate. Right: FRB distribution of the angle between the robustly estimated and population canonical variate.

As a second example, we consider the Hemophilia data (Habbema *et al.* 1974), which consists of  $n_1 = 30$  observations of normal women and  $n_2 = 45$  of hemophilia A carriers, with  $p = 2$  variables (AHF activity and AHF antigen). Robust discriminant analysis methods were already applied to these data by Hawkins and McLachlan (1997) and by Hubert and Van Driessen (2004). Salibian-Barrera *et al.* (2008) used this data set to illustrate that FRB can be used to estimate the error rate of robust classification rules. The data are shown in the lower right corner of Figure 4. Group 1 observations are plotted by circles and group 2 observations by triangles. The original data set does not contain outliers and hence robust procedures show similar results as classical canonical variate analysis, which is indicated by the solid line in Figure 4. Similarly as in Salibian-Barrera *et al.* (2008), we added 15 outliers to group 2 (the points in the upper left corner of Figure 4) to illustrate the robustness of the robust canonical variate analysis and the FRB. The outliers

strongly affect the classical canonical variate as can be seen from the dashed line in Figure 4.



**Fig. 4** Hemophilia data. The original data are in the lower right corner. The 15 outliers that have been added to group 2 are in the upper left corner.

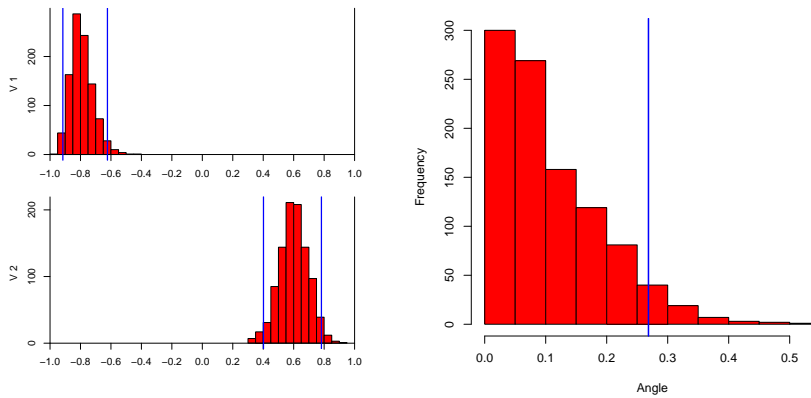
**Table 6** Hemophilia data. Standardized coefficients of the canonical variate for classical and robust linear discriminant analysis. Results for both the original and contaminated data set are shown

Data	Method	V1	V2
Clean	Classical	-0.748	0.663
	Robust MM	-0.834	0.551
Contaminated	Classical	-0.279	-0.960
	Robust MM	-0.795	0.606

Table 6 shows the standardized coefficients of the estimated canonical variate for both the original and contaminated hemophilia data. The canonical variate estimates based on both classical linear discriminant analysis and robust discriminant analysis using simultaneous two-sample S-estimates are shown. From this table it can be seen that the outliers strongly affect the

classical estimates of the canonical variate. The robust estimates on the other hand are quite stable and resemble well the classical estimates based on the original data without outliers.

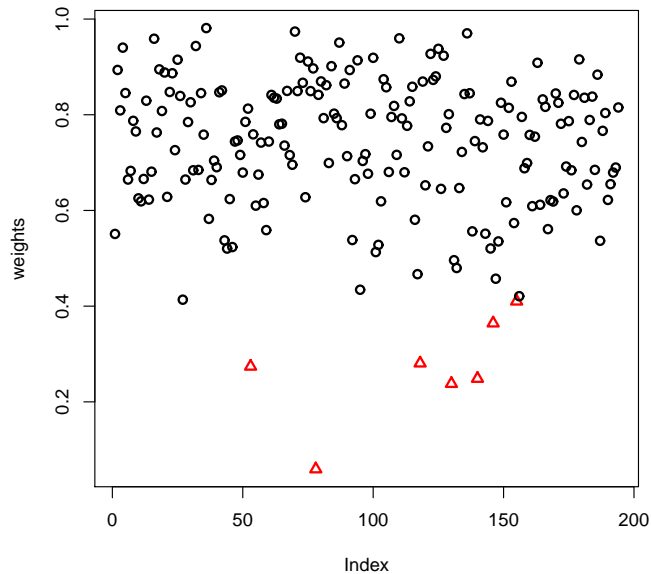
From the left panel of Figure 5 we can see that both variables are important to discriminate the two populations. This plot also confirms that the outliers largely influenced the classical linear discriminant analysis. Comparing the classical coefficient estimates in the presence of outliers with the 95% confidence intervals in the plot, we see that the classical estimates are significantly different from the robust estimates, where the effect of the outliers has been appropriately reduced in the latter. The right panel of Figure 5 shows the stability of the canonical variate in this example, even in the presence of the large amount of contamination. The upper limit of the one-sided 95% confidence interval for the angle between the estimated and population canonical variate has value 0.27, which corresponds with an angle of less than 16 degrees.



**Fig. 5** Hemophilia data. Left: FRB distribution of the standardized coefficients of the robustly estimated canonical variate. Right: FRB distribution of the angle between the robustly estimated and population canonical variate.

As a final example, we consider the Duchenne Muscular Dystrophy (DMD) data set of Andrews and Herzberg (1985). This data set contains measurements of  $n_1 = 127$  DMD carriers and  $n_2 = 67$  noncarriers. The first two measured variables are age ( $X_1$ ) and month of year ( $X_2$ ), and the other four variables are serum marker levels. A detailed description of the data is given in Riani and Atkinson (2001) which provides an extensive robust analysis of the data based on the forward search (see also Atkinson et al., 2004). We considered the transformed data set, using the transformation advocated in Riani and Atkinson (2001). Figure 6 shows the weights that are given to the observations by the simultaneous two-sample MM-estimates. A standard outlier identification rule is to identify observations as outliers if their squared robust distance exceeds the 97.5% quantile of the  $\chi^2$  distribution with 6 degrees of freedom. In

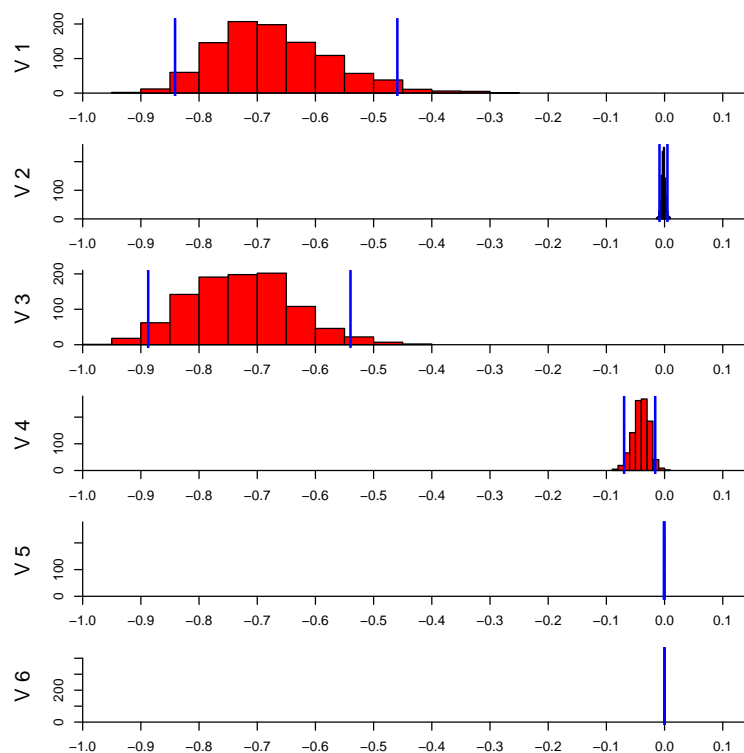
Figure 6 these outlying cases are plotted as triangles. We see that there are five clear outliers and two boundary cases (cases 146 and 155). The five most outlying observations are the cases 53, 78, 118, 130, and 140, which are also the five observations flagged as outliers in the forward search analysis of Riani and Atkinson (2001). Hence, a robust discriminant analysis is needed to avoid a potentially damaging effect of the outliers on the analysis. Figure 7 shows



**Fig. 6** Duchenne Muscular Dystrophy (DMD) data. The weight that the observations receive in the calculation of the simultaneous two-sample MM-estimates are shown. Observations with squared robust distance exceeding the 97.5% quantile of the  $\chi^2$  distribution with 6 degrees of freedom are considered outliers and are indicated by triangles.

the FRB distribution of the robust coefficient estimates of the canonical variate. From this plot we can see that the coefficients of variables 2, 4 and 5 are always negligibly small. Variables 3 and 1 carry the most discriminatory power while the contribution of variable 4 is significant but smaller. These findings correspond with the conclusion of Riani and Atkinson (2001) and is especially useful in this application because the serum marker levels of variables 3 and 4 are inexpensive to measure while the levels in variables 5 and 6 are far more expensive.

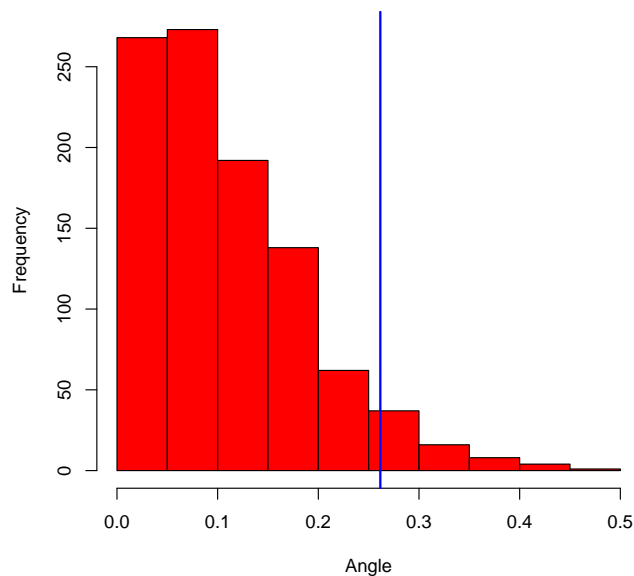
Finally, Figure 8 shows the stability of the canonical variate. The upper limit of the one-sided 95% confidence interval for the angle between the estimated and population canonical variate has value 0.26, which corresponds with an angle of about 15 degrees.



**Fig. 7** Duchenne Muscular Dystrophy (DMD) data. FRB distribution of the standardized coefficients of the robustly estimated canonical variate.

## 6 Conclusions

We considered robust canonical variate analysis based on robust estimates of the group centers and joint scatter matrix. We used  $S$  and  $MM$ -estimates of multivariate location and scatter. One-sample robust estimates can be applied on the data of each group separately, and the joint scatter estimate can then be obtained by pooling the individual group estimates. Alternatively, simultaneous robust estimators for the locations and joint scatter estimator can be defined directly. In both cases the fast and robust bootstrap method can be used to obtain inference for the robustly estimated canonical variate. More particularly, we showed that the FRB can be used to construct confidence intervals for the contribution of each variable to the canonical variate and thus to investigate which variables contribute significantly to the canonical variate. Moreover, the stability of the robust discriminant analysis can be examined further through the FRB distribution of the angles between the bootstrapped and original canonical variate estimates. This distribution estimates the distribution between the original canonical variate estimate and its population counterpart. We considered the two-group discriminant analysis problem in



**Fig. 8** Duchenne Muscular Dystrophy (DMD) data. FRB distribution of the angle between the robustly estimated and population canonical variate.

this paper, but the method can straightforwardly be extended to discrimination problems with more than two groups.

**Acknowledgements** The research of Stefan Van Aelst was supported by a grant of the Fund for Scientific Research-Flanders (FWO- Vlaanderen) and by IAP research network grant nr. P6/03 of the Belgian government (Belgian Science Policy).

## References

- Andrews D F, Herzberg A M (1985) *Data*. Springer Verlag, New York
- Atkinson A C, Riani M, Cerioli A (2004) *Exploring Multivariate Data with the Forward Search*. Springer Verlag, New York
- Bianco A, Boente G, Pires A M, Rodrigues I M (2008) Robust discrimination under a hierarchy on the scatter matrices. *Journal of Multivariate Analysis* 99: 1332-1357
- Bickel P J, Freedman D A (1981) Some asymptotic theory for the bootstrap. *The Annals of Statistics* 9: 1196-1217
- Croux, C, Dehon, C (2001) Robust linear discriminant analysis using S-estimators. *Canadian Journal of Statistics* 29: 473-492
- Croux C, Filzmoser P, Joossens K (2008) Classification efficiencies for robust linear discriminant analysis. *Statistica Sinica* 18: 581-599
- Davies P L (1987) Asymptotic behavior of S-estimates of multivariate location parameters and dispersion matrices. *Annals of Statistics*, 15: 1269-1292
- Efron B (1979) Bootstrap methods: another look at the jackknife. *Annals of Statistics* 7: 1-26

- Habbema J D F, Hermans J, Van den Broeck K (1974) A stepwise discriminant analysis program using density estimation. In: Bruckmann G, Ferschl F, Schmetterer L (eds) *Proceedings in Computational Statistics*. Physica-Verlag, Vienna pp 101–110
- He X, Fung W K (2000) High breakdown estimation for multiple populations with applications to discriminant analysis. *Journal of Multivariate Analysis* 72: 151–162
- Hubert M, Rousseeuw P J, Van Aelst S (2008) High-breakdown robust multivariate methods. *Statistical Science* 23: 92–119
- Hubert M, Van Driessen K (2004) Fast and robust discriminant analysis. *Computational Statistics and Data Analysis* 45: 301–320
- Johnson R A, Wichern D W (2002) *Applied multivariate statistical analysis*. Prentice Hall, Upper Saddle River
- Lopuhaä H (1989) On the relation between S-estimators and M-estimators of multivariate location and covariance. *Annals of Statistics* 17: 1662–1683
- Lopuhaä H, Rousseeuw P J (1991) Breakdown points of affine equivariant estimators of multivariate location and covariance matrices. *Annals of Statistics* 19: 229–248
- Maronna R A, Martin D R, Yohai V J (2006) *Robust statistics: theory and methods*. John Wiley and Sons, England
- Riani M, Atkinson A C (2001) A Unified approach to outliers, influence, and transformations in discriminant analysis. *Journal of Computational and Graphical Statistics* 10: 513–544
- Rousseeuw P J, Leroy A M (1987) *Robust regression and outlier detection*. John Wiley and Sons, New York
- Salibian-Barrera M (2005) Estimating the p-values of robust tests for the linear model. *Journal of Statistical Planning and Inference* 128: 241–257
- Salibian-Barrera M, Van Aelst S (2008) Robust model selection using fast and robust bootstrap. *Computational Statistics and Data Analysis* 52: 5121–5135
- Salibian-Barrera M, Van Aelst S, Willems G (2006) PCA based on multivariate MM-estimators with fast and robust bootstrap. *Journal of the American Statistical Association* 101: 1198–1211
- Salibian-Barrera M, Van Aelst S, Willems G (2008) Fast and robust bootstrap. *Statistical Methods and Applications* 17: 41–71
- Salibian-Barrera M, Zamar, R H (2002) Bootstrapping robust estimates of regression. *Annals of Statistics* 30: 556–582
- Tatsuoka K S, Tyler D E (2000) The uniqueness of S and M-functionals under non-elliptical distributions. *Annals of Statistics* 28: 1219–1243
- Tyler D E (2002) High-Breakdown point multivariate M-estimation. *Estadística* 54: 213–247
- Van Aelst S, Willems G (2005) Multivariate regression S-estimators for robust estimation and inference. *Statistica Sinica* 15: 981–1001
- Van Aelst S, Willems G (2009) Robust and efficient one-way MANOVA tests. Ghent University, Technical report, submitted for publication
- Yohai V J (1987) High breakdown-point and high efficiency robust estimates for regression. *Annals of Statistics* 15: 642–656