



UNIVERSITY OF LEEDS

This is a repository copy of *The use of a common location measure in the invariant coordinate selection and projection pursuit*.

White Rose Research Online URL for this paper:  
<http://eprints.whiterose.ac.uk/104150/>

Version: Accepted Version

---

**Article:**

Alashwali, F and Kent, JT [orcid.org/0000-0002-1861-8349](https://orcid.org/0000-0002-1861-8349) (2016) The use of a common location measure in the invariant coordinate selection and projection pursuit. *Journal of Multivariate Analysis*, 152. pp. 145-161. ISSN 0047-259X

<https://doi.org/10.1016/j.jmva.2016.08.007>

---

© 2016, Elsevier. Licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

**Reuse**

Unless indicated otherwise, fulltext items are protected by copyright with all rights reserved. The copyright exception in section 29 of the Copyright, Designs and Patents Act 1988 allows the making of a single copy solely for the purpose of non-commercial research or private study within the limits of fair dealing. The publisher or other rights-holder may allow further reproduction and re-use of this version - refer to the White Rose Research Online record for this item. Where records identify the publisher as the copyright holder, users can verify any specific terms of use on the publisher's website.

**Takedown**

If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing [eprints@whiterose.ac.uk](mailto:eprints@whiterose.ac.uk) including the URL of the record and the reason for the withdrawal request.



[eprints@whiterose.ac.uk](mailto:eprints@whiterose.ac.uk)  
<https://eprints.whiterose.ac.uk/>

# The use of a common location measure in the invariant coordinate selection and projection pursuit

Fatimah Alashwali<sup>a,\*</sup>, John T. Kent<sup>b</sup>

<sup>a</sup>Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia

<sup>b</sup>University of Leeds, Leeds, United Kingdom

---

## Abstract

Invariant coordinate selection (ICS) and projection pursuit (PP) are two methods that can be used to detect clustering directions in multivariate data by optimizing criteria sensitive to non-normality. In particular, ICS finds clustering directions using a relative eigen-decomposition of two scatter matrices with different levels of robustness; PP is a one-dimensional variant of ICS. Each of the two scatter matrices includes an implicit or explicit choice of location. However, when different measures of location are used, ICS and PP can behave counter-intuitively. In this paper we explore this behavior in a variety of examples and propose a simple and natural solution: use the same measure of location for both scatter matrices.

*Keywords:* Cluster analysis, Invariant coordinate selection, Projection pursuit, Robust scatter matrices, Location measures, Multivariate mixture model.

---

## 1. Introduction

Consider a multivariate dataset, given as an  $n \times p$  data matrix  $X$ , and suppose we want to explore the existence of any clusters. One way to detect clusters is by projecting the data onto a lower dimensional subspace for which the data are maximally non-normal. Hence, methods that are sensitive to non-normality can be used to detect clusters.

One set of methods based on this principle is invariant coordinate selection (ICS), introduced by Tyler et al. [17], together with a one-dimensional variant called projection pursuit (PP), introduced by Friedman and Tukey [5]. ICS involves the use of two scatter matrices,  $S_1 = S_1(X)$  and  $S_2 = S_2(X)$  with  $S_2$  chosen to be more robust than  $S_1$ . An eigen-decomposition of  $S_2^{-1}S_1$  is carried out. If the data can be partitioned into two clusters, then typically the eigenvector corresponding to the smallest eigenvalue is a good estimate of the clustering direction. The main choice for the user when carrying out ICS is the choice of the two scatter matrices.

However, in numerical experiments based on a simple mixture of two bivariate normal distributions, some strange behavior was noticed. In certain circumstances, ICS, and its variant PP, badly failed to pick out the right clustering direction. Eventually, it was discovered that the cause was the use of different location measures in the two scatter matrices. The purpose of this paper is to explore the reasons for this strange behavior in detail and to demonstrate the benefits of using common location measures.

Section 2 gives some examples of scatter matrices and reviews the use of ICS and PP as clustering methods. Section 3 sets out the multivariate normal mixture model with two useful standardizations of the coordinate system. Section 4 demonstrates in the population setting an ideal situation where ICS and PP work as expected and where an analytic solution is available — the two-group normal mixture model where the two scatter matrices are given by the covariance matrix and a kurtosis-based matrix. Some examples with other robust estimators are given in Sections 5–6, which show how ICS and PP can go wrong when different location measures are used and how the problem is fixed by using a common location measure. Further issues, including unbalanced mixtures and heteroscedasticity, are discussed in Section 7.

---

\*Corresponding author

Email addresses: fsalashwali@pnu.edu.sa (Fatimah Alashwali), j.t.kent@leeds.ac.uk (John T. Kent)

24 *Notation.* Univariate random variables, and their realizations, are denoted by lowercase letters,  $x$ , say. Multivariate  
 25 random vectors, and their realizations, are denoted by lowercase bold letters,  $\mathbf{x}$ , say. A capital letter,  $X$ , say is used for  
 26  $n \times p$  data matrix containing  $p$  variables or measurements on  $n$  observations;  $X$  can be written in terms of its rows as

$$X = (\mathbf{x}_1^\top, \dots, \mathbf{x}_n^\top)^\top,$$

27 with  $i$ th row  $\mathbf{x}_i^\top = (x_{i1}, \dots, x_{ip})$ ,  $i = 1, \dots, n$ .

## 28 **2. Background**

### 29 *2.1. Scatter matrices*

30 A scatter matrix  $S(X)$ , as a function of an  $n \times p$  data matrix  $X$ , is a  $p \times p$  affine equivariant positive definite matrix.  
 31 Following Tyler et al. [17], it is convenient to classify scatter matrices into three classes depending on their robustness.

- 32 (1) Class I: is the class of non-robust scatter matrices with zero breakdown point and unbounded influence function.  
 33 Examples include the covariance matrix defined below in (1) and the kurtosis-based matrix in (2).
- 34 (2) Class II: is the class of scatter matrices that are locally robust, in the sense that they have bounded influence  
 35 function and positive breakdown points not greater than  $1/(p+1)$ . An example from this class is the class of  
 36 multivariate  $M$ -estimators, such as the  $M$ -estimate for the  $t$ -distribution, e.g., [4, 8].
- 37 (3) Class III: is the class of scatter matrices with high breakdown points such as the Stahel-Donoho estimate, the  
 38 minimum volume ellipsoid (mve) and the constrained  $M$ -estimates, e.g., [7, 18].

39 Each scatter matrix has an implicit location measure. Let us look at the main examples in more detail, and note what  
 40 happens in  $p = 1$  dimension. The labels in parentheses are used as part of the notation later in the paper.

41 The sample covariance matrix (var) is defined by

$$S = \frac{1}{n} \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^\top, \quad (1)$$

42 where for convenience here a divisor of  $1/n$  is used, and where  $\bar{\mathbf{x}}$  is the sample mean vector. The implicit measure of  
 43 location is just the sample mean.

44 The kurtosis-based matrix (kmat) is defined by

$$K = \frac{1}{n} \sum_{i=1}^n \{(\mathbf{x}_i - \bar{\mathbf{x}})^\top S^{-1}(\mathbf{x}_i - \bar{\mathbf{x}})\}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^\top. \quad (2)$$

45 Note that outlying observations are given higher weight than for the covariance matrix, so that  $K$  is less robust than  
 46  $S$ . Again the implicit measure of location is just the sample mean. When  $p = 1$ , the scatter matrix  $S^{-1}K$  reduces to 3  
 47 plus the usual univariate kurtosis.

48 The  $M$ -estimator of scatter based on the multivariate  $t_\nu$ -distribution for fixed  $\nu$  is the maximum likelihood estimate  
 49 obtained by maximizing the likelihood jointly over scatter matrix  $\Sigma$  and location vector  $\boldsymbol{\mu}$ . If both parameters are  
 50 unknown and  $\nu \geq 1$ , then under mild conditions on the data, the mle of  $(\boldsymbol{\mu}, \Sigma)$ , is the unique stationary point of the  
 51 likelihood. Similarly, if  $\nu \geq 0$  and  $\boldsymbol{\mu}$  is known, the mle of  $\Sigma$  is the unique stationary point of the likelihood; see Kent  
 52 et al. [8]. In either case, an iterative numerical algorithm is needed. Note that when  $\boldsymbol{\mu}$  is to be estimated as well as  $\Sigma$ ,  
 53 the mle of  $\boldsymbol{\mu}$  is the implicit measure of location for this scatter matrix. For this paper we limit attention to the choice  
 54  $\nu = 2$  (and label it below by  $t_2$ ).

55 The minimum volume ellipsoid (mve) estimate of scatter  $S_{\text{mve}}$ , introduced by Rousseeuw [14], is the ellipsoid  
 56 that has the minimum volume among all ellipsoids containing at least half of observations, and its implicit estimate  
 57 of location,  $\bar{\mathbf{x}}_{\text{mve}}$ , say, is the center of that ellipsoid. Calculating the exact mve requires extensive computation.  
 58 In practice, it is calculated approximately by considering only a subset of all subsamples that contain 50% of the  
 59 observations, e.g., [9, 18]. If the location vector is specified, the search is limited to ellipsoids centered at this location  
 60 measure.

When  $p = 1$ , the mve reduces to the lshorth, defined as the length of the shortest interval that contains at least half of observations. The corresponding estimate of location,  $\bar{x}_{\text{lshorth}}$ , say, is the midpoint of this interval. Calculating the lshorth around a known measure of location is trivial; just find the length of the interval that contains half of observations centered at this location measure. The lshorth was introduced by Grubel [6], building on an earlier suggestion of Andrews et al. [3] to use  $\bar{x}_{\text{lshorth}}$ , which they called the shorth, as a location measure.

The minimum covariance determinant estimate of scatter (mcd),  $S_{\text{mcd}}$ , say, is defined as the covariance matrix of half of observations with the smallest determinant. The mcd location measure,  $\bar{x}_{\text{mcd}}$ , say, is the sample mean of those observations. The mcd can be calculated approximately by considering only a subset of all subsamples that contain at least half of observations, e.g., Rousseeuw and Driessen [15]. The mcd estimate of scatter with respect to a known location measure  $\boldsymbol{\mu}$  is defined as the covariance matrix about  $\boldsymbol{\mu}$  of half of observations with the smallest determinant. Recall that the covariance matrix about  $\boldsymbol{\mu}$  for a dataset is given by  $S + (\boldsymbol{\mu} - \bar{\boldsymbol{x}})(\boldsymbol{\mu} - \bar{\boldsymbol{x}})^\top$ , where  $S$  and  $\bar{\boldsymbol{x}}$  are the sample covariance matrix and mean vector of the dataset.

When  $p = 1$ , the mcd reduces to a truncated variance,  $v_{\text{trunc}}$ , say, defined as the smallest variance of half the observations. Its implicit measure of location,  $\bar{x}_{\text{trunc}}$ , say, is the sample mean of that interval. Also, a modified definition of  $v_{\text{trunc}}$  using a known location measure is trivial and does not require any search; just find the interval that contains half of observations centered at the given location measure and calculate the variance.

Routines are available in R [13] to compute (at least approximately) these robust covariance matrices and their implicit location measures, in particular, `tM` from the package `ICS` [10] for the multivariate  $t$ -distribution, `cov.rob` from the package `MASS` [19] for mve, and `CovMcd` from the package `rrcov` [16] for mcd. Modified versions of these routines have been written by us to deal with the case of known location measures.

## 2.2. Invariant coordinate selection and projection pursuit

Given an  $n \times p$  data matrix  $X$ , the ICS objective function is given by the ratio of quadratic forms

$$\kappa_{\text{ICS}}(\boldsymbol{a}) = \frac{\boldsymbol{a}^\top S_1 \boldsymbol{a}}{\boldsymbol{a}^\top S_2 \boldsymbol{a}}, \quad \boldsymbol{a} \in \mathbb{R}^p, \quad (3)$$

where  $S_1 = S_1(X)$  and  $S_2 = S_2(X)$  are two scatter matrices. By convention,  $S_2$  is chosen to be more robust than  $S_1$ . The intuition behind this convention is as follows. Under a balanced elliptically symmetric model, the population center is always uniquely defined. In the clustering direction the data will appear to have shorter tails, for the same reason that kurtosis is negative in this direction (see Section 4) than in the perpendicular directions, and hence we expect a more robust estimator to give a larger estimate of scatter, relative to a less robust estimator, in this direction than in the perpendicular direction.

For exploratory statistical analysis, attention is focused on the choices for  $\boldsymbol{a}$  maximizing or minimizing  $\kappa_{\text{ICS}}(\boldsymbol{a})$ . These values can be calculated analytically as the eigenvectors of  $S_2^{-1}S_1$  corresponding to the maximum/minimum eigenvalues.

The original ICS method did not make a strong distinction between the largest and the smallest eigenvalues. However for clustering purposes between two groups, when the mixing proportion is not too far from  $1/2$ , it is the minimum eigenvalue which is of interest; see Section 4.

The method of PP can be regarded as a one-dimensional version of ICS. It looks for a linear projection  $\boldsymbol{a}$  to maximize or minimize the criterion,

$$\kappa_{\text{PP}}(\boldsymbol{a}) = \frac{s_1(X\boldsymbol{a})}{s_2(X\boldsymbol{a})}, \quad (4)$$

where  $s_1 = s_1(X\boldsymbol{a})$  and  $s_2 = s_2(X\boldsymbol{a})$  are two one-dimensional measures of spread. In general, optimizing  $\kappa_{\text{PP}}(\boldsymbol{a})$  must be carried out numerically. Searching for a global optimum is computationally expensive, and the complexity of the search increases as the dimension  $p$  increases. Alternatively, we can search for a local optimum starting from a sensible initial solution, such as the ICS optimum direction.

Both ICS and PP are equivariant under affine transformations. That is, if  $X$  is transformed to  $U = \mathbf{1}_n \boldsymbol{h}^\top + XQ^\top$ , where  $Q(p \times p)$  is nonsingular and  $\boldsymbol{h}$  is a translation vector in  $\mathbb{R}^p$ , then for both ICS and PP the new optimal vector  $\boldsymbol{b}$ , say, for  $U$  is related to the corresponding optimal vector  $\boldsymbol{a}$  for  $X$  by

$$\boldsymbol{b} \propto Q^{-\top} \boldsymbol{a}. \quad (5)$$

104 For numerical work it is convenient to have an explicit notation for the different choices in ICS and PP. If Scat1  
 105 and Scat2 are the names of two types of multivariate scatter matrix, each computed with its own implicit location  
 106 measure, then the corresponding versions of ICS and PP will be denoted

$$\text{ICS : Scat1 : Scat2, \quad and \quad PP : Scat1 : Scat2.}$$

107 Note that PP is based on the univariate versions of Scat1 and Scat2. For example, ICS based on the covariance matrix  
 108 and the minimum volume ellipsoid will be denoted by ICS:var:mve. Other choices for scatter matrices have been  
 109 summarized in Section 2.

110 When a common location measure is imposed on Scat1 and Scat2, then this restriction will be indicated by the  
 111 augmented notation

$$\text{ICS : Scat1 : Scat2 : Loc,}$$

112 and similarly for PP. In this paper the only choice used for the location measure is the sample mean (mean). For  
 113 example, ICS based on the covariance matrix and the minimum volume ellipsoid, both computed with respect to the  
 114 mean vector, is denoted

$$\text{ICS : var : mve : mean.}$$

### 115 3. The two-group multivariate normal mixture model

116 The simple model used to demonstrate the main points of this paper is the two group multivariate normal mixture  
 117 model, with density

$$f(\mathbf{x}) = q\phi_p(\mathbf{x}, \boldsymbol{\mu}_1, \Omega) + (1 - q)\phi_p(\mathbf{x}, \boldsymbol{\mu}_2, \Omega),$$

118 where  $\phi_p$  is the multivariate normal density,  $\boldsymbol{\mu}_1$  and  $\boldsymbol{\mu}_2$  are two mean vectors,  $\Omega$  is a common covariance matrix, and  
 119  $0 < q < 1$  is the mixing proportion. Even in this simple case, major problems with ICS and PP can arise.

120 Since ICS and PP are affine equivariant, we may without loss of generality choose the coordinate system so that

$$\boldsymbol{\mu}_1 = \alpha \mathbf{e}_1, \quad \boldsymbol{\mu}_2 = -\alpha \mathbf{e}_1, \quad \Omega = I_p,$$

121 where  $\mathbf{e}_1 = (1, 0, \dots, 0)^\top$  is a unit vector along the first coordinate axis, and  $\alpha > 0$ . That is,  $\boldsymbol{\mu}_1$  and  $\boldsymbol{\mu}_2$  lie equally  
 122 spaced about the origin along the first coordinate axis, and the covariance matrix of each component equals the identity  
 123 matrix.

124 A random vector  $\mathbf{x}$  from the mixture model can also be given a stochastic representation,

$$\mathbf{x} = \alpha s \mathbf{e}_1 + \boldsymbol{\epsilon},$$

125 where  $\boldsymbol{\epsilon} \sim \mathcal{N}_p(0, I_p)$  independently of an indicator variable  $s$ ,

$$s = \begin{cases} 1 & \text{with probability } q \\ -1 & \text{with probability } (1 - q) \end{cases}.$$

126 Moments under the mixture model are calculated most simply in terms of this stochastic representation. In particular,

$$\boldsymbol{\mu}_x = E(\mathbf{x}) = q\boldsymbol{\mu}_1 + (1 - q)\boldsymbol{\mu}_2 = (2q - 1)\alpha \mathbf{e}_1, \quad E(\mathbf{x}\mathbf{x}^\top) = \alpha^2 \mathbf{e}_1 \mathbf{e}_1^\top + I_p,$$

127 so that the covariance matrix is

$$\Sigma_x = \text{var}(\mathbf{x}) = E(\mathbf{x}\mathbf{x}^\top) - \boldsymbol{\mu}_x \boldsymbol{\mu}_x^\top = 4q(1 - q)\alpha^2 \mathbf{e}_1 \mathbf{e}_1^\top + I_p. \quad (6)$$

128 For practical work it is also convenient to consider a standardization for which the overall covariance matrix is the  
 129 identity matrix. That is, define a new random vector

$$\mathbf{y} = C^{-1} \mathbf{x}, \quad (7)$$

130 where  $C^{-1} = \text{diag}(1/c_1, \dots, 1/c_p)$ , where  $c_1 = \{1 + 4q(1 - q)\alpha^2\}^{1/2}$ , and  $c_2 = \dots = c_p = 1$ . Then  $\mathbf{y}$  has a stochastic  
 131 representation

$$\mathbf{y} = \delta \mathbf{s} \mathbf{e}_1 + \boldsymbol{\eta},$$

132 where

$$\delta = \alpha / \{1 + 4q(1 - q)\alpha^2\}^{1/2}, \quad (8)$$

133 and

$$\boldsymbol{\eta} \sim \mathcal{N}_p(0, \text{diag}(\sigma_\eta^2, 1, \dots, 1))$$

134 where the first diagonal term  $\sigma_\eta^2$  has two equivalent formulas,

$$\sigma_\eta^2 = \{1 + 4\alpha^2 q(1 - q)\}^{-1} \quad \text{or} \quad \sigma_\eta^2 = 1 - 4q(1 - q)\delta^2$$

135 The first two moments of  $\mathbf{y}$  are

$$\boldsymbol{\mu}_y = (2q - 1)\delta \mathbf{e}_1, \quad \boldsymbol{\Sigma}_y = I_p.$$

#### 136 4. A population example: PP based on the kurtosis and ICS based on the kurtosis-based matrix and the co- 137 variance matrix

138 In this section we look at ICS:kmat:var and PP:kmat:var in the population case. In this setting it is possible to  
 139 derive analytic results. Note that since kmat is based on fourth moments it is less robust than the variance matrix;  
 140 hence kmat is listed first.

141 Recall the kurtosis of a univariate random variable  $u$ , say, with mean  $\mu_u$ , is defined by

$$\text{kurt}(u) = \frac{E\{(u - \mu_u)^4\}}{[E\{(u - \mu_u)^2\}]^2} - 3.$$

142 The univariate kurtosis is zero when the random variable has normal distribution. For non-normal distributions the  
 143 kurtosis lies in the interval  $[-2, \infty]$  and is often nonzero. In particular, the kurtosis takes the following possible values:

- 144 (1)  $\text{kurt}(u) = 0$ ; satisfied under normality.
- 145 (2)  $\text{kurt}(u) < 0$ ; this case is called sub-Gaussian.
- 146 (3)  $\text{kurt}(u) > 0$ ; this case is called super-Gaussian.

147 The sub-Gaussian case appears in distributions flatter than the normal and have thinner tails; one example is the  
 148 uniform distribution. On the other hand, the super-Gaussian case appears in distributions that are more peaked than  
 149 the normal distribution and have longer tails; examples include  $t$ , and Laplace distributions.

150 Define a balance parameter  $\psi(q) = |q - 1/2|$ . Peña and Prieto [11] studied the population version of PP:kmat:var and  
 151 showed that when the mixing proportion is not too far from  $1/2$ , more precisely, if  $q(1 - q) > 1/6$ , i.e.,  $\psi(q) < 1/\sqrt{12}$ ,  
 152 then minimizing the PP objective function picks out the correct clustering direction. Similarly, if  $q$  is far from half,  
 153 i.e.,  $\psi(q) > 1/\sqrt{12}$ , then maximizing the objective function picks out the correct clustering direction.

154 Their result can be derived simply as follows. Let  $\mathbf{a} \in \mathbb{R}^p$  be a unit vector. Write  $\mathbf{a}^\top \mathbf{x} = \alpha a_1 s + v$ , where  
 155  $v = \mathbf{a}^\top \boldsymbol{\epsilon} \sim \mathcal{N}(0, 1)$  is independent of  $s$ . The moments of  $s$  are  $E(s) = E(s^3) = m$ , say, where

$$m = 2q - 1, \quad (9)$$

156 and  $E(s^2) = E(s^4) = 1$ . Hence,  $\text{var}(s) = \sigma^2$ , say, where

$$\sigma^2 = 4q(1 - q). \quad (10)$$

157 Then

$$\text{kurt}(s) = -6 + 4/\sigma^2.$$

158 It can be checked that  $\text{kurt}(s) < 0$  provided  $\phi(q) < 1/\sqrt{12}$ .

159 Next, we use the property that if  $u_1, u_2$  are independent random variables with the same variance, and if  $\delta_1, \delta_2$  are  
 160 coefficients satisfying  $\delta_1^2 + \delta_2^2 = 1$ , then

$$\text{kurt}(\delta_1 u_1 + \delta_2 u_2) = \delta_1^4 \text{kurt}(u_1) + \delta_2^4 \text{kurt}(u_2).$$

161 Applying this result to  $\mathbf{a}^\top \mathbf{x}$  yields

$$\text{kurt}(\mathbf{a}^\top \mathbf{x}) = \frac{a_1^4 \alpha^4 \sigma^4}{(\alpha^2 a_1^2 \sigma^2 + 1)^2} \text{kurt}(s). \quad (11)$$

162 Provided  $\text{kurt}(s) < 0$ , (11) is minimized when  $a_1^2$  is maximized, that is, if  $a_1^2 = 1$ , so that  $\mathbf{a} = \pm e_1$  picks out the first  
 163 coordinate axis.

164 The ICS calculations proceed similarly. First note that  $E(x_1) = \alpha m$ , and the first diagonal term in  $\Sigma_x$ , defined in  
 165 (6), can be expressed in terms of  $\sigma^2$ , defined in (10), as  $\alpha^2 \sigma^2 + 1$ .

166 The first factor in the population version of  $K$  defined in (2),  $K_x$ , say, is given by

$$(\mathbf{x} - \boldsymbol{\mu}_x)^\top \Sigma_x^{-1} (\mathbf{x} - \boldsymbol{\mu}_x) = \frac{(x_1 - \alpha m)^2}{1 + \alpha^2 \sigma^2} + x_2^2 + \dots + x_p^2 = D^2, \text{ say,}$$

167 where  $m$  is defined in (9). Note that  $D^2$  is an even function in  $x_2, \dots, x_p$ . Hence by symmetry all the off-diagonal  
 168 terms in  $K_x$  vanish. The first diagonal term is given by

$$E\{D^2(x_1 - \alpha m)^2\} = (1 + \alpha^2 \sigma^2)(p + 2) + \frac{\alpha^4 \sigma^4 \text{kurt}(s)}{(1 + \alpha^2 \sigma^2)}.$$

169 The remaining diagonal terms,  $j = 2, \dots, p$  are given by

$$E(D^2 x_j^2) = p + 2.$$

170 Hence  $\Sigma_x^{-1} K_x$  reduces to

$$\text{diag} \left\{ p + 2 + \frac{\text{kurt}(s) \alpha^4 \sigma^4}{(1 + \alpha^2 \sigma^2)}, p + 2, \dots, p + 2 \right\}.$$

171 These diagonal values are the eigenvalues. Hence provided  $\text{kurt}(s) < 0$ ,  $\kappa_{\text{ICS}}$  is minimized when  $\mathbf{a} = \mathbf{e}_1$ , that is, when  
 172  $\mathbf{a}$  picks out the clustering direction.

173 If  $p = 2$ , we can write a unit vector as  $\mathbf{a} = (\cos \theta, \sin \theta)^\top$ , and since  $\mathbf{a}$  and  $-\mathbf{a}$  define the same axis, we can  
 174 parameterize the ICS and PP objective functions in terms of  $\theta$ ,  $-\pi/2 \leq \theta \leq \pi/2$ . Plots of  $\kappa_{\text{ICS}}(\theta)$  and  $\kappa_{\text{PP}}(\theta)$  for  $\alpha = 3$   
 175 and  $q = 1/2, 0.85$  and  $1/2 + 1/\sqrt{12}$  are shown in Figure 1.

176 For numerical work, especially when the underlying mixture model is unknown, the only feasible standardization  
 177 is to ensure the overall variance matrix  $\Sigma_y$  is the identity rather than the within-group variance matrix. In terms of the  
 178 population model of this section, it means working with  $\mathbf{y}$  from (7) rather than  $\mathbf{x}$ . If  $p = 2$  and  $\mathbf{b} \propto (\cos \phi, \sin \phi)^\top$ , say,  
 179 is also written in polar coordinates, then from (5) and (7)  $\mathbf{a}$  and  $\mathbf{b}$  are related by

$$\mathbf{b} \propto C \mathbf{a};$$

180 hence,  $\phi$  and  $\theta$  are related by

$$\begin{pmatrix} \cos \phi \\ \sin \phi \end{pmatrix} \propto \begin{pmatrix} c_1 & 0 \\ 0 & c_2 \end{pmatrix} \begin{pmatrix} \cos \theta \\ \sin \theta \end{pmatrix}.$$

181 Thus,

$$\tan \phi = c \tan \theta,$$

182 where  $c = c_2/c_1$ .

183 The plot of the ICS and PP objective functions in Figure 2 shows that there is a sharper minimum in  $\phi$  coordinates  
 184 than in  $\theta$  coordinates because under our mixture model  $c$  is less than 1. If  $\mathbf{x}$  is scaled as in (7) with  $c_1 > c_2$ , i.e.,  $c > 1$ ,  
 185 then there will be a wider minimum in  $\phi$ .

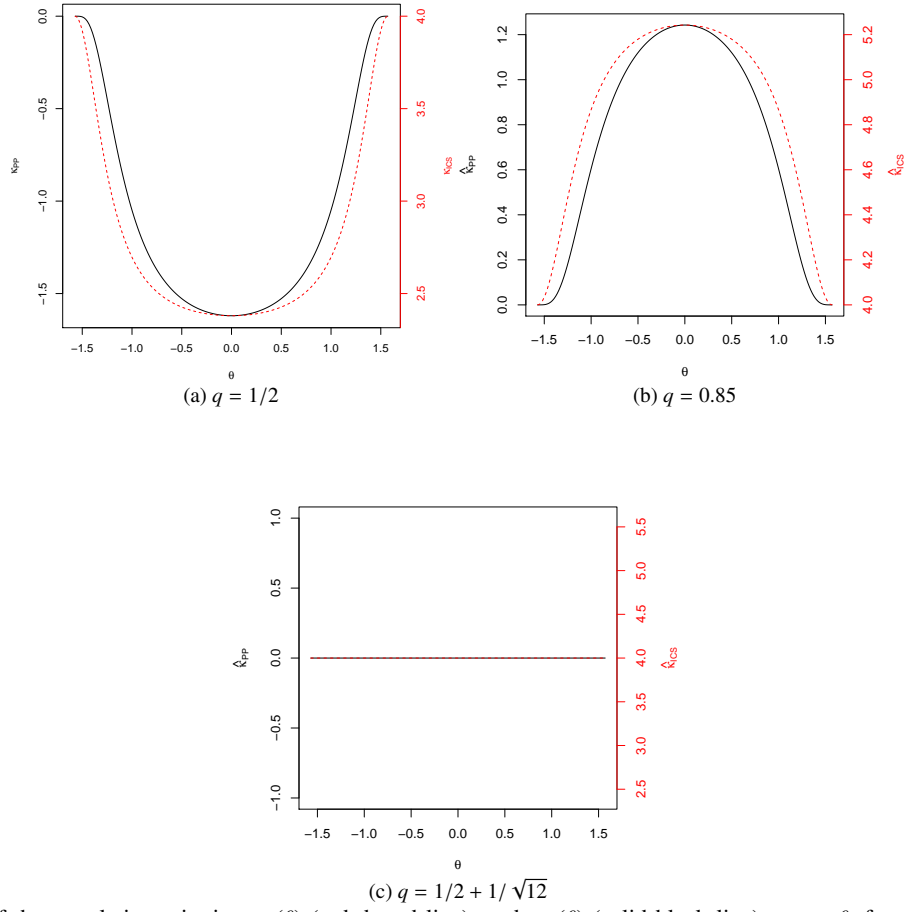


Figure 1: Plot of the population criteria  $\kappa_{CS}(\theta)$  (red dotted line), and  $\kappa_{PP}(\theta)$  (solid black line) versus  $\theta$ , for  $q = 1/2, 0.85$  and  $1/2 + 1/\sqrt{12}$ , and  $\alpha = 3$ .

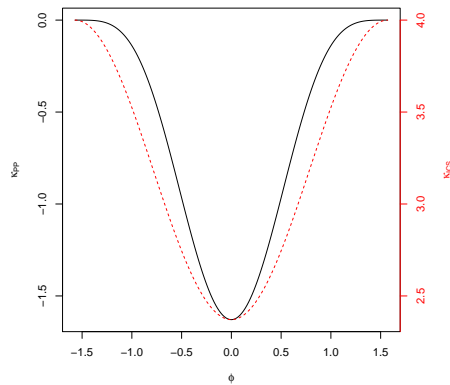


Figure 2: Plot of the population criteria  $\kappa_{CS}(\phi)$  (red dotted line), and  $\kappa_{PP}(\phi)$  (solid black line) versus  $\phi$ , for  $q = 1/2$ , and  $\delta = 0.95$ .



186 **5. The effect of using a common location measure on ICS and PP**

187 As mentioned earlier in Section 2.2, the ICS and PP criteria are expected to have similar behavior to the kurtosis-  
 188 based criteria in Section 4. Namely, they are expected to be minimized in the clustering direction when the mixing  
 189 proportion is not too far from 1/2.

190 However, when applying ICS with at least one robust estimate of scatter (mainly from Class III), some peculiar  
 191 behavior was observed on many datasets. In particular, the ICS criterion was often maximized in the clustering  
 192 direction rather than minimized.

193 Here is an explanation. Under the two-group mixture model with one group slightly bigger than the other, a class  
 194 III scatter matrix will typically home in on the larger group, with its corresponding location measure at the center of  
 195 this group and its estimate of the scatter matrix capturing the spread of this group. The other scatter matrix (Class I  
 196 or II) will measure the overall scatter of the data with its corresponding location measure at the overall center of the  
 197 data. The result is erratic behavior in  $\kappa_{ICS}$  and  $\kappa_{PP}$ .

198 Imposing a common location measure on the two scatter matrices fixes this problem. Here is a population example  
 199 in  $p = 2$  dimensions to illustrate the issues in greater detail.

200 In this example we look at ICS:var:mve for the population bivariate normal mixture model in Section 3, with  
 201  $q = 1/2$  and any value of  $\alpha > 0$ , i.e.,  $0 \leq \delta \leq 1$ , where  $\delta$  is given in (8). Standardize the coordinate system so that the  
 202 overall covariance matrix is the identity,  $\Sigma_y = I_2$ . Let  $\Sigma_{mve}$  denote the population minimum volume ellipsoid scatter  
 203 matrix.

204 Then it turns out that  $\Sigma_{mve}$  is the within-group covariance matrix for (either) one of the groups,

$$\Sigma_{mve} = \begin{pmatrix} 1 - \delta^2 & 0 \\ 0 & 1 \end{pmatrix}, \quad (12)$$

205 where  $0 \leq \delta \leq 1$  is given in (8). The implicit estimate of the center of the data will be given by the center of either  
 206 group,  $\pm\delta\mathbf{e}_1$ ; both values fit equally well.

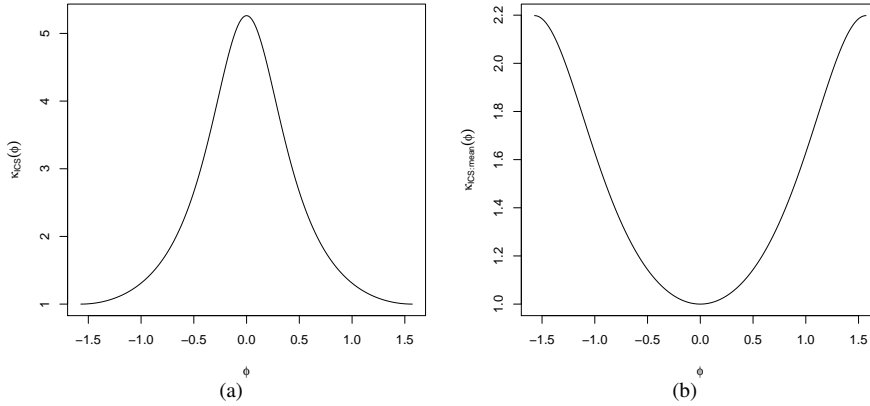


Figure 3: For  $\delta = 0.9$ , plots of the population criterion of: (a) ICS:var:mve vs.  $\phi$ , and (b) ICS:var:mve:mean .

207 Figure 3 (a) shows that the ICS:var:mve estimate of clustering direction is  $(0, 1)^\top$ , i.e.,  $\phi = \pm\pi/2$ . However, the  
 208 true direction of group separation direction is  $(1, 0)^\top$ , i.e.,  $\phi = 0$ .

209 Next consider ICS:var:mve:mean, i.e., the common mean version of the previous example. The overall mean of  
 210 the data is at the origin. When  $\Sigma_{mve}$  is constrained to have its location measure at the origin, then the ICS criterion  
 211 now picks out the true clustering direction. In order to give an analytic proof of this result, we restrict attention to the  
 212 limiting case of the balanced mixture model, i.e., when  $\delta = 1$ ,  $q = 1/2$ . Hence, the group components will lie on  
 213 two parallel vertical lines with means

$$\mu_1 = (1, 0)^\top, \quad \mu_2 = (-1, 0)^\top,$$

214 and within-group covariance matrix

$$\begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}.$$

215 In this setting, the following theorem gives population version of the MVE matrix.

216 **Theorem 1.** Consider the limiting balanced bivariate normal mixture model,

$$\mathbf{y} = s\mathbf{e}_1 + z\mathbf{e}_2,$$

217 where  $s = \pm 1$ , each with probability  $1/2$ , independent of  $z \sim \mathcal{N}(0, 1)$ , and  $\mathbf{e}_1 = (1, 0)^\top$ ,  $\mathbf{e}_2 = (0, 1)^\top$ . This model is  
 218 standardized with respect to the “total” coordinates; i.e.,  $\mathbf{E}(\mathbf{y}) = \mathbf{0}$  and  $\text{var}(\mathbf{y}) = \mathbf{I}_2$ . The model can also be described  
 219 in terms of a mixture of two normal distributions, concentrated on the vertical lines  $y_1 = 1$  and  $y_1 = -1$  as shown in  
 220 Figure 4.

221 The minimum volume ellipsoid mve of  $\mathbf{y}$ ,  $\Sigma_{mve}$ , say, takes the form

$$\Sigma_{mve} = c_t \Sigma_t = \begin{pmatrix} 2 & 0 \\ 0 & 2d^2 \end{pmatrix},$$

222 where  $d = \Phi^{-1}(.75) = 0.674$ , the 75th quantile of the standard normal distribution. Hence the dominant eigenvector  
 223 is  $\mathbf{e}_1$ .

224 Theorem 1 is proved in the Appendix. The ellipse of  $\Sigma_{mve}$  is plotted in Figure 4. Figure 3 (b) shows that the  
 criterion of ICS:var:mve:mean,  $\kappa_{ICS;\mu}(\phi)$  picks out the correct clustering direction  $\mathbf{e}_1$ .

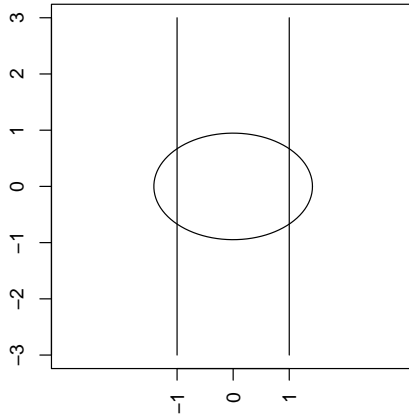


Figure 4: Plot of the ellipse of  $\Sigma_{mve}$  with its location measure forced at the origin superimposed on a mixture of two normal distributions concentrated on the vertical lines  $y_1 = 1$  and  $y_1 = -1$ .

225 Like ICS, PP can fail to detect the clustering direction if applied using different location measures. If the projection  
 226 direction separates the data into two groups with one slightly bigger than the other, then the more robust measure of  
 227 spread will measure the spread of the larger group. In Section 6, we give a detailed numerical example of the problem  
 228 arising from using two different location measures in PP:var:mcd, and how the problem is fixed by using a common  
 229 location measure.  
 230

231 **6. Numerical examples**

232 *Overview*

233 In this section, we give numerical examples that demonstrate different ways in which ICS and/or PP can go wrong.  
 234 We also show the beneficial effect of using common location measures in these examples. We use one simulated data  
 235 set and apply different ICS and PP methods, with and without imposing a common location measure (the mean).

236 A two-dimensional data set of size  $n = 500$  is generated from the balanced mixture model, defined in Section 3,  
 237 with  $q = 1/2$ , and  $\alpha = 3$ , so that  $\delta = 0.95$ . Thus the two groups are well-separated and no sensible statistical method  
 238 should have any problem finding the two clusters. All calculations are done after standardization with respect to the  
 239 “total” coordinates. That is, the data matrix  $Y(500 \times 2)$  is standardized to have sample mean  $\mathbf{0}$  and sample covariance  
 240 matrix  $I_2$ .

241 The ICS and PP methods used are:

- 242 (1) (PP,ICS):var:t2 with corresponding criteria  $\kappa_{\text{ICS}}^1$ , and  $\kappa_{\text{PP}}^1$ .
- 243 (2) (PP,ICS):var:mcd with corresponding criteria  $\kappa_{\text{ICS}}^2$ , and  $\kappa_{\text{PP}}^2$ .
- 244 (3) (PP,ICS):var:mve with corresponding criteria  $\kappa_{\text{ICS}}^3$ , and  $\kappa_{\text{PP}}^3$ .
- 245 (4) (PP,ICS):t2:mcd with corresponding criteria  $\kappa_{\text{ICS}}^4$ , and  $\kappa_{\text{PP}}^4$ .
- 246 (5) (PP,ICS):t2:mve with corresponding criteria  $\kappa_{\text{ICS}}^5$ , and  $\kappa_{\text{PP}}^5$ .

247 When imposing the mean as the common location measure, the ICS and PP criteria will be denoted by  $\kappa_{\text{ICS:mean}}^j$  and  
 248  $\kappa_{\text{PP:mean}}^j$ , where  $j = 1, \dots, 5$ .

249 To understand the behavior of ICS and PP, their criteria are plotted against  $-\pi/2 \leq \phi \leq \pi/2$ . The plots are shown  
 250 in Figure 5. From the panels in Figure 5, we make the following remarks based on the simulated data set:

- 251 (1) Panel (a) shows that ICS:var:t2 and PP:var:t2 work well since  $\bar{y}$  and  $y_{i2}^-$  are approximately equal. Hence,  
 252 imposing a common location measure has little effect, as shown in (b).
- 253 (2) Panels (c), (e), (g), (i) show examples when ICS and/or PP go wrong because of the difference in the location  
 254 measures.
- 255 (3) Using a common location measure fixes the problem in panel (d) for (PP, ICS):var:mcd, panel (f) for (PP,  
 256 ICS):var:mve, and panel(h) for (PP, ICS):t2:mcd.
- 257 (4) From panel (j), using a common location measure in PP:t2:mve:mean does not seem to work well. The reason  
 258 might be due to the unstable behavior of the mve and lshorth.
- 259 (5) The plots generally suggest that PP will be more accurate than ICS, since the PP plots are narrower at the  
 260 clustering direction than the ICS plot. This property has been confirmed empirically in Alashwali [2] for certain  
 261 multivariate normal mixture models and choices of scatter matrix.
- 262 (6) Similar patterns are seen with most simulated data sets from this model.

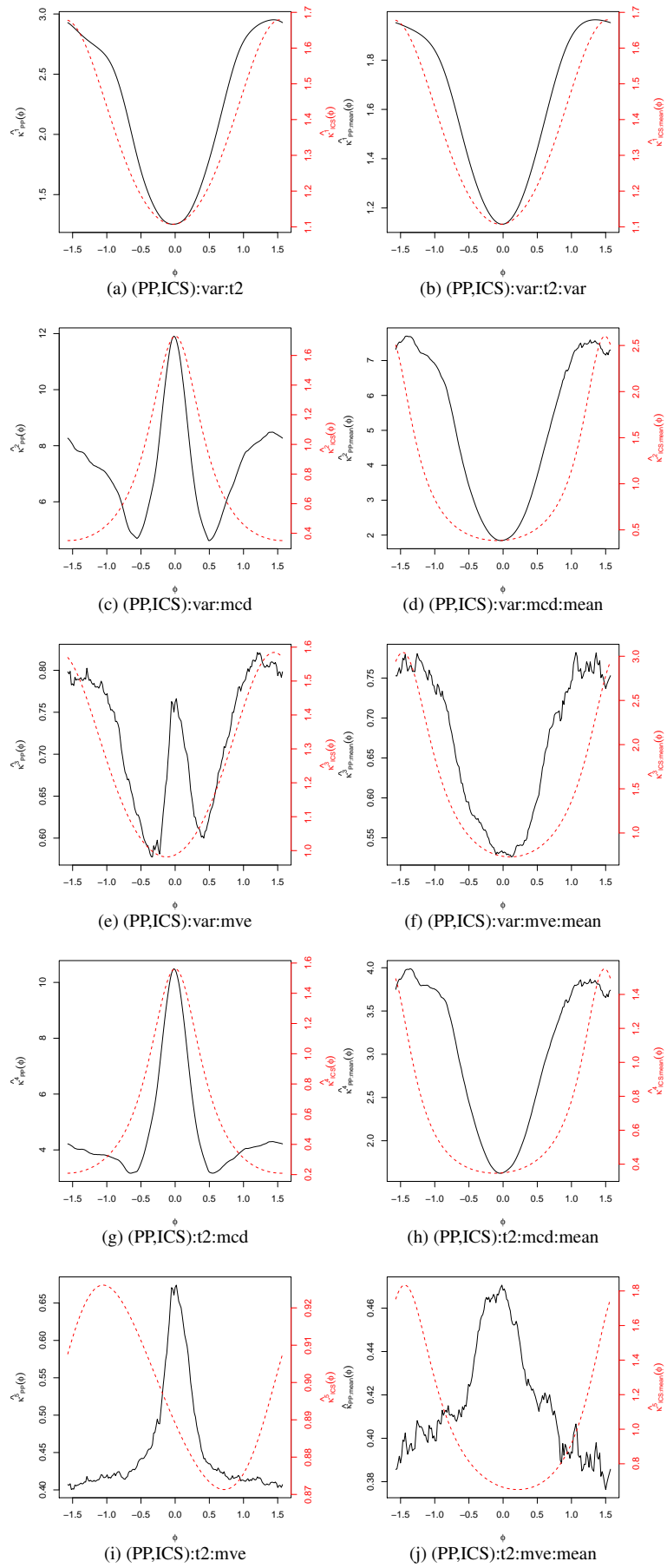


Figure 5: For  $\delta = 0.95$  and  $q = 1/2$ , plots of different ICS (red dashed curve) and PP (black solid curve) criteria without (left) and with imposing a common location measure (right).

263 *Behavior of ICS:var:mcd*

264 To gain a deeper understanding of the behavior of ICS:var:mcd in panel 5 (c) and the effect of forcing a common  
 265 location measure on mcd in panel (d), we plot the ellipse of  $S_{\text{mcd}}$  (both with and without imposing a common  
 266 location measure) and superimpose it on the data points of our example. The plots are shown in panels 6 (a) and (b). The  
 behavior in this example agrees with the interpretation given for the population example in Section 5.

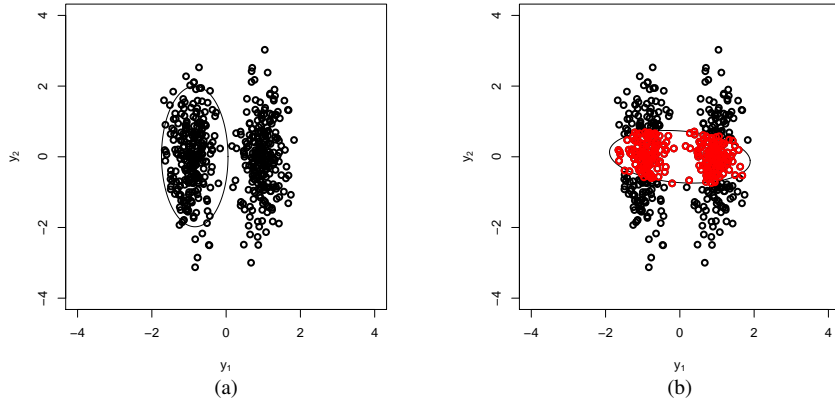


Figure 6: Plots of the ellipses of mcd scatter matrix based on (a) mcd location measure, and (b) the sample mean, superimposed on data of size  $n = 500$ , distributed as mixtures of two normal distributions.

267

268 *Behavior of PP:var:mcd*

269 The objective function for PP:var:mcd, has a similar problem to ICS; it is maximized rather than minimized near  
 270 the correct clustering direction.

271 To understand this behavior in more detail, we plot in Figure 7 one-dimensional histograms after projections by  
 272 the following choices for the angle  $\phi$ :  $0^\circ$ ,  $15^\circ$ ,  $30^\circ$ , and  $90^\circ$ . For each histogram, we plot the 50% of the data that  
 273 has the smallest variance, and the corresponding location measure  $\bar{x}_{\text{trunc}}$ . The plots are repeated where the location  
 274 measure is constrained at the sample mean  $\bar{x} = 0$ . Note that the shape of the histograms depends on of the projection  
 275 directions. Also, as  $v_{\text{trunc}}$  gets smaller, the PP criterion  $\kappa_{\text{pp}}$  gets larger. From the panels of Figure 7, we make the  
 276 following remarks:

- 277 (1) The  $0^\circ$  projection produces two widely separated groups with one group is slightly bigger than the other. In this  
 278 case,  $\bar{x}_{\text{trunc}}$  is at the larger group and  $v_{\text{trunc}}$  is essentially the variance of this group. Hence  $v_{\text{trunc}}$  takes its smallest  
 279 value and  $\kappa_{\text{pp}}$  is largest.
- 280 (2) The  $15^\circ$  projection produces two slightly separated groups with within-group variance is larger than in the  $0^\circ$   
 281 projection. The value of  $v_{\text{trunc}}$  is larger than for  $0^\circ$ .
- 282 (3) The  $30^\circ$  projection produces one group, with a pseudo-uniform distribution. The value of  $v_{\text{trunc}}$  is larger than  
 283 for  $15^\circ$ .
- 284 (4) The  $90^\circ$  projection produces one normally distributed group. The value for  $v_{\text{trunc}}$  becomes small again.

285 Constraining the mean to be at the origin fixes the problem. The value of  $v_{\text{trunc}}$  steadily decreases from  $0^\circ$  to  $90^\circ$ .

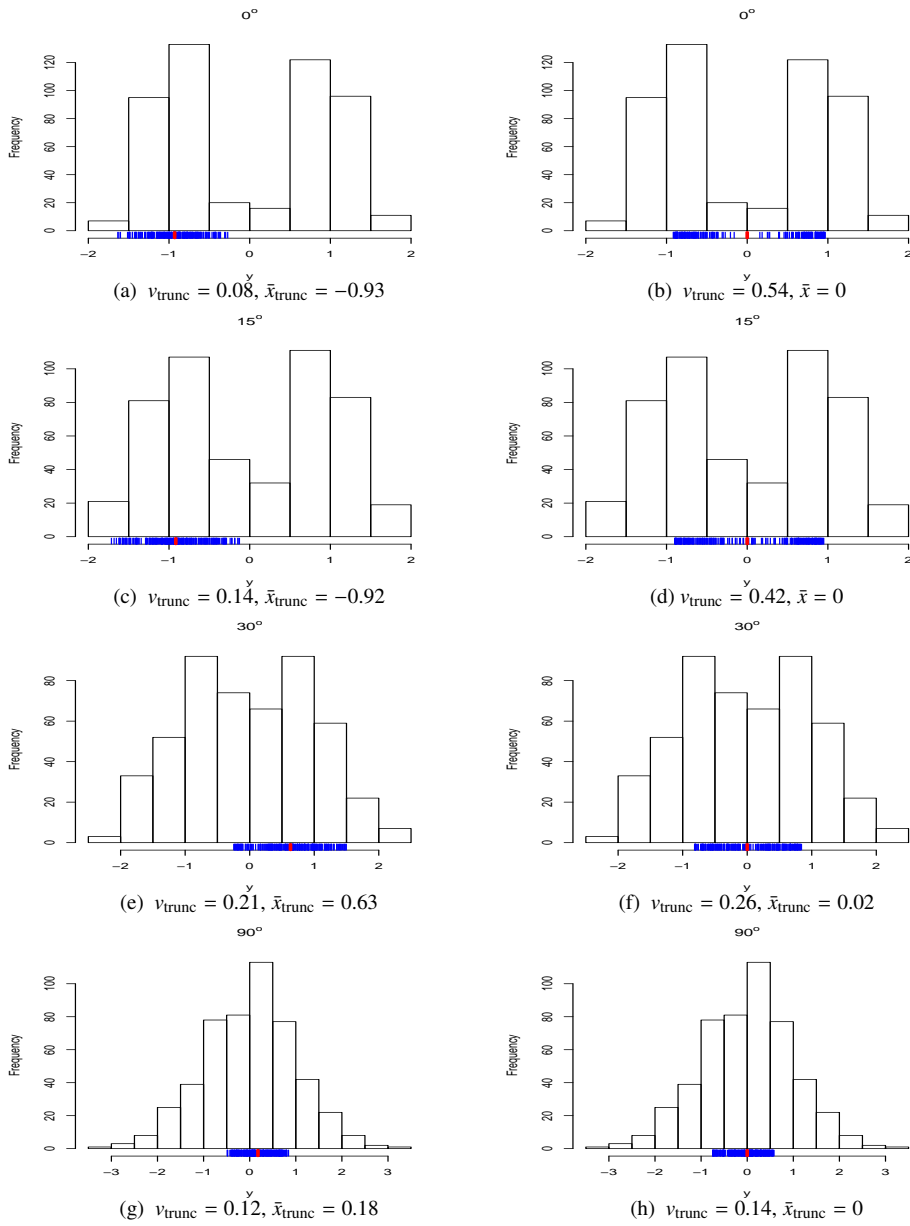


Figure 7: Histograms of  $0^\circ$ ,  $15^\circ$ ,  $30^\circ$  and  $90^\circ$  projections. Left panels show the vectors of 50% of data with the smallest variance (the blue lines), and its location measure (the red lines), right panels show the 50% of data with the smallest variance computed around the mean 0.

Table 1: Estimates of  $\hat{\theta}_{ICS}$  in degrees for simulated data sets with  $n = 500$ ,  $q = 0.6$  and  $0.85$ ,  $\alpha = 3$ , and equal covariance matrices.

Method	$q = 0.6$ (min)	$q = 0.85$ (max)
ICS:var:t2	1.02	-2.05
ICS:var:t2:mean	1.09	-0.05
ICS:var:mcd	88.87	-1.29
ICS:var:mcd:mean	2.90	19.31
ICS:var:mve	33.00	0.13
ICS:var:mve:mean	0.89	0.98
ICS:t2:mcd	89.44	-1.04
ICS:t2:mcd:mean	3.77	23.58
ICS:t2:mve	88.53	0.65
ICS:t2:mve:mean	0.77	1.26

## 286 7. Further issues

287 So far, we have investigated the importance of using a common location measure in the performance of ICS based  
 288 on robust estimates of scatter under mixtures of two balanced normal distributions. In this section, we discuss some  
 289 further issues regarding ICS performance including lack of balance, heteroscedasticity, and the importance of robust  
 290 estimates.

### 291 *Lack of balance*

292 Recall from Section 4 that under mixtures of two normal distributions with  $S_1 = K$  and  $S_2 = S$ , if  $q$  is close to  
 293 half, then  $\kappa_{ICS}$  is minimized in the clustering direction, whereas if  $q$  is far from half then  $\kappa_{ICS}$  is maximized in the  
 294 clustering direction. In this section, we want to explore the extent to which this behavior continues to hold for other  
 295 choices of  $S_1$  and  $S_2$ .

296 Several data sets were simulated from the mixture model defined in Section 3 with  $n = 500$ ,  $\alpha = 3$ , and dif-  
 297 ferent choices of  $q$ . After standardizing the data as in (7), the following ICS methods are applied: ICS:var:t2,  
 298 ICS:var:t2:mean, ICS:var:mcd, ICS:var:mcd:mean, ICS:var:mve, ICS:var:mve:mean, ICS:t2:mcd, ICS:t2:mcd:mean,  
 299 ICS:t2:mve, and ICS:t2:mve:mean for  $q = 0.6$  (near half),  $0.85$  (far from half),. Table 1 shows a comparison of the  
 300 clustering direction estimates of the ICS methods. The simulation results can be summarized as follows:

- 301 (1) if  $q$  is close enough to  $1/2$ , then minimization is still appropriate.
- 302 (2) if  $q$  is far enough from  $1/2$ , then maximization is appropriate. In this case, forcing a common location measure  
 303 is unnecessary, because the Class II and III estimates of location will be at the center of the larger group, and  
 304 the Class I estimate of location will be close to the center of the larger group.
- 305 (3) several simulations for different values of  $q$  suggest that robust ICS methods have the same balance parameter  
 306  $\phi(q) = 1/\sqrt{12}$  as discussed in Section 4.

### 307 *Heteroscedasticity*

308 Following Peña et al. [12], consider the heteroscedstic model:

$$q\mathcal{N}(\mu_1, \Omega) + (1 - q)\mathcal{N}(\mu_2, \Omega + \Delta\Omega),$$

309 where  $\Delta\Omega$  is the added perturbation. Without loss of generality assume  $\mu_1 = (\alpha, 0)^\top$ ,  $\mu_2 = (-\alpha, 0)^\top$ ,  $\Omega = I_2$ .

310 To investigate the effect of heteroscedasticity, restrict attention to the balanced case ( $q = 1/2$ ) in  $p = 2$  dimensions,  
 311 with three different scenarios for  $\Delta_j$ ,  $j = 1, \dots, 4$ ,

$$\Delta_1 = \text{diag}(0.5, 1.5), \Delta_2 = \text{diag}(1, 1.5), \Delta_3 = \text{diag}(1, 3), \text{ and } \Delta_4 = \text{diag}(2, 1.5).$$

312 In the simulation study,  $N = 500$  datasets of size  $n = 500$  were simulated under each scenario for  $\alpha = 1, 2$ , and 3. All  
 313 data set are standardized as in (7) to have the identity matrix as the total covariance matrix. The following methods  
 314 are applied: ICS:kmat:var, ICS:var:t2:mean, ICS:var:mcd:mean, ICS:var:mve:mean. Note that the different location  
 315 measures version of the used ICS methods have the same problems that appear under equal covariance mixture model  
 316 (see Section 6). Each method gives a set of estimates of the clustering direction as follows:  $\hat{\theta}_1, \dots, \hat{\theta}_{500}$ , with the true  
 317 clustering direction at  $\theta_0 = 0$ . To compare the performances of the four methods, we use the following measure of  
 318 spread:

$$\hat{\nu}(\hat{\theta}) = \frac{1}{N} \sum_{k=1}^N \sin^2(\hat{\theta}_k - \theta_0), \quad (13)$$

319 If the distribution of  $\hat{\theta}$  is concentrated around  $\theta_0$ , then  $\nu(\hat{\theta}) = 0$ . If the distribution of  $\hat{\theta}$  is concentrated around  $\theta_0 + \pi/2$   
 or  $\theta_0 + 3\pi/2$ , then  $\nu(\hat{\theta}) = 1$ . If  $\hat{\theta}$  is uniformly distributed, then  $\nu(\hat{\theta}) = 1/2$ . Figure 8 shows plots of  $\hat{\nu}(\hat{\theta})$  for the

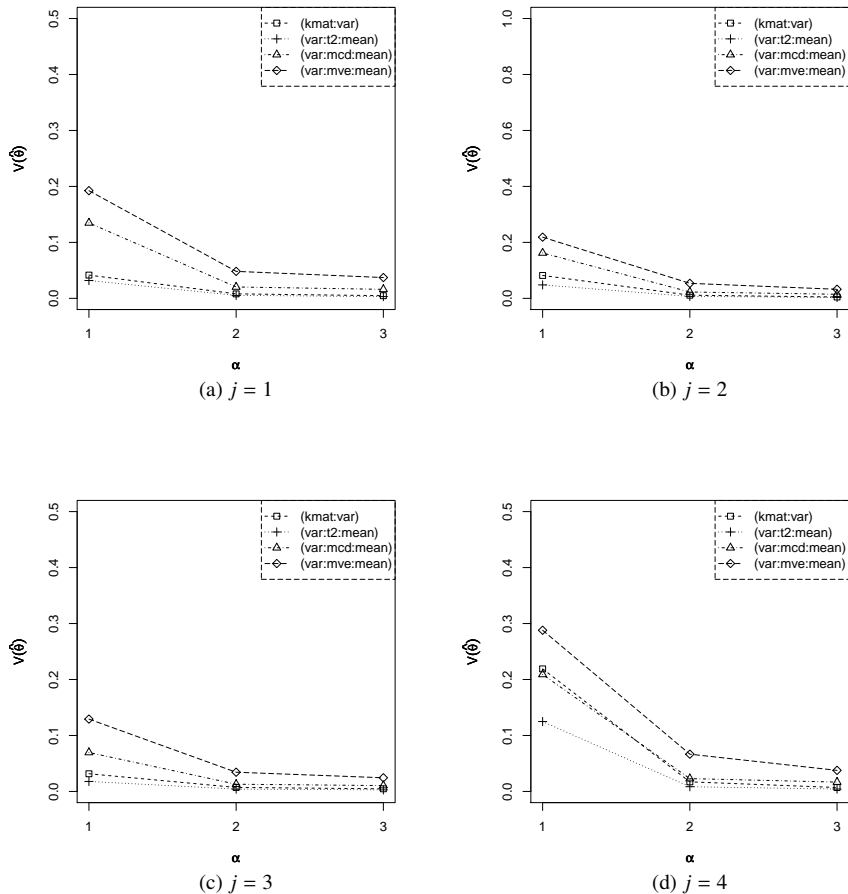


Figure 8: Plot of  $\hat{\nu}(\hat{\theta})$  for estimates of clustering directions estimated by the methods ICS:kmat:var:mean, ICS:var:t2:mean, ICS:var:mcd:mean, and ICS:var:mve:mean versus  $\alpha = 1, 2$  and 3 for four different heteroscedastic models labeled by  $j = 1, \dots, 4$ .

320  
 321 four different methods. The plots show that forcing a common location measure works well under the heteroscedastic



322 model. Also, the methods ICS:kmat:var and ICS:var:t2 have the best performance for all  $\Delta_j$  among all other methods  
 323 used in this study.

### 324 *Importance of robust estimators*

325 In this section, we compare the performance of different ICS methods using robust estimates of scatter versus  
 326 ICS:kmat:var under mixtures of long-tailed distributions.

327 The data sets used in this section are simulated from the following model. Suppose that the clustering direction is  
 328 along the first coordinate axis. Let  $\mathbf{x} = (x_1, x_2)^\top$  be a bivariate random vector, where  $x_1$  follows a balanced mixture of  
 329 two  $t$  distributions with  $\nu$  degrees of freedom, and  $x_2$  follows a standard normal distribution. The random variable  $x_1$   
 330 can be written as:

$$x_1 = \alpha s + z,$$

331 where  $\alpha$ , and  $s$  are defined in Section 3, and  $z$  is a  $t$  random variable with  $\nu$  degrees of freedom. The first and third  
 332 moments of  $z$  are equal to zero, the second and fourth moments are given by, e.g., Ahsanullah et al. [1],

$$E(z^2) = \frac{\nu}{\nu - 2}, \quad E(z^4) = \frac{3\nu^2}{(\nu - 2)(\nu - 4)}.$$

333 The kurtosis of  $z$  is  $6/(\nu - 4)$  for  $\nu > 4$ . Following our model in Section 3, we first standardize with respect to the  
 334 within-group variance, i.e.,  $x_1$  can be written as:

$$x_1 = \alpha s + u,$$

335 where  $u = z \sqrt{(\nu - 2)/\nu}$ . The second moment of  $u$  is 1 and its fourth moment is  $3(\nu - 2)/(\nu - 4)$ . The kurtosis of  $u$  is  
 336  $6/(\nu - 4)$ .

The kurtosis of  $x_1$  is given by:

$$\begin{aligned} \text{kurt}(x_1) &= \frac{\alpha^4}{(\alpha^2 + 1)^2} \text{kurt}(s) + \frac{1}{(\alpha^2 + 1)^2} \text{kurt}(u) \\ &= -\frac{2\alpha^4}{(\alpha^2 + 1)^2} + \frac{1}{(\alpha^2 + 1)^2} \left( \frac{6}{\nu - 4} \right). \end{aligned}$$

337 We want to explore settings in which each mixture component has positive kurtosis and the mixture has zero or  
 338 negative kurtosis. Let  $\nu = 7$ ; then the kurtosis of each mixture component is 9.8 and the kurtosis of  $x_1$  is

$$\text{kurt}(x_1) = \frac{1}{(\alpha^2 + 1)^2} (-2\alpha^4 + 2).$$

339 For  $\alpha = 1$ , the kurtosis equals to zero, and as  $\alpha$  increases the kurtosis decreases (takes negative values).

340 The simulation is repeated  $N = 500$  times for each  $\alpha = 1, 2$ , and 3, and sample size  $n = 500$ . The following  
 341 ICS methods are applied: ICS:kmat:var; ICS:var:t2:mean; ICS:var:mcd:mean; ICS:var:mve:mean. To compare the  
 342 performances of the ICS methods, we use (13). Figure 9 shows plots of  $\hat{v}(\hat{\theta})$  versus  $\alpha = 1, 2$ , and 3. The plots show  
 343 that for small  $\alpha$ , robust ICS methods especially ICS:var:t2:mean are more accurate than ICS:kmat:var.

## 344 **8. Conclusion**

345 This paper has clarified several issues about role of the location measure when ICS and PP are used for two-group  
 346 cluster analysis. The key observation is that if the mixing proportion  $q$  is near 1/2 (the balanced case) and the two  
 347 scatter measures use different location measures, then ICS and PP are prone to erratic behavior. This problem is most  
 348 severe when one scatter matrix comes from Class I and the other comes from Class II or III. The solution is to modify  
 349 the definition of the scatter matrices to ensure they both use the same measure of location. The clustering direction  
 350 can be found by minimizing the ICS and PP criteria, respectively.

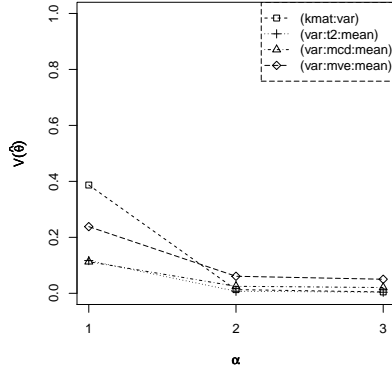


Figure 9: Plot of  $\hat{v}(\hat{\theta})$  for estimates of clustering directions estimated by the methods ICS:kmat:var, ICS:var:t2:mean, ICS:var:mcd:mean, and ICS:var:mve:mean versus  $\alpha = 1, 2$  and  $3$  for balanced mixtures of two  $t_7$  distributions.

351 In the unbalanced case when  $q$  is far from  $1/2$ , the situation is simpler. The clustering direction is found by  
 352 maximizing the ICS or PP criteria, respectively, and in this case it does not matter whether or not a common location  
 353 parameter is used.

354 Most of paper focuses on the use of normal distributions for the mixture components. It is also possible to reach  
 355 some conclusions when the mixture components have longer tails. In this setting it is beneficial for one of the scatter  
 356 matrices to be robust. In particular, if  $q = 1/2$  then ICS:var:t2:mean outperforms ICS:kmat:var.

## 357 Appendix

358 In this appendix we shall prove Theorem 1. In particular, we show that the population version of the mve, con-  
 359 strained to be centered at the origin, is given by

$$\Sigma_{\text{mve}} = \begin{bmatrix} 2 & 0 \\ 0 & d^2 \end{bmatrix},$$

360 where  $d = \Phi^{-1}(.75)$  in terms of the cumulative distribution function of the  $\mathcal{N}(0, 1)$  distribution.

361 First let  $u_1 < u_2$  be two possible values for  $y_2$  and consider an ellipse based on a matrix  $\Sigma$  with inverse  $\Sigma^{-1} = \Omega$ ,

$$\mathbf{y}^\top \Omega \mathbf{y} = 1, \tag{A.1}$$

362 which intersects the vertical line passing through  $(1, 0)^\top$ , at these points,

$$\begin{bmatrix} 1 & u_1 \end{bmatrix} \Omega \begin{bmatrix} 1 \\ u_1 \end{bmatrix} = 1, \quad \begin{bmatrix} 1 & u_2 \end{bmatrix} \Omega \begin{bmatrix} 1 \\ u_2 \end{bmatrix} = 1. \tag{A.2}$$

363 By symmetry the ellipse also intersects the points  $(-1, -u_1)^\top$  and  $(-1, -u_2)^\top$ . Note that  $\Sigma$  will be a candidate for the  
 364 mve matrix if the interior of the ellipse covers 50% of the probability mass, that is,

$$\Phi(u_2) = \Phi(u_1) + 1/2. \tag{A.3}$$

365 If  $u_1$  and  $u_2$  are finite, then necessarily  $u_1 < 0$  and  $u_2 > 0$ .

366 The proof will proceed in two stages. First, for fixed  $u_1, u_2$  satisfying (A.3), we choose  $\Sigma$  to minimize  $\det(\Sigma)$  (or  
 367 equivalently maximize  $\det(\Omega)$ ). Secondly, we optimize over the choice of  $u_1, u_2$ .

368 Thus, start with a fixed pair of values  $u_1, u_2$  satisfying (A.3). If  $\mathbf{y} = (1, u)^\top$  represents a point on one of the vertical  
 369 lines, then the intersection with the ellipse (A.1) can be written

$$\omega_{11} + 2\omega_{12}u + \omega_{22}u^2 = 1,$$

370 or equivalently as the quadratic equation in  $u$ ,

$$Au^2 + Bu + C = 0,$$

371 where  $A = \omega_{22}$ ,  $B = 2\omega_{12}$ ,  $C = \omega_{11} - 1$ . If this ellipse passes through  $(1, u_1)^\top$  and  $(1, u_2)^\top$ , then  $u_1, u_2$  are roots  
372 of the quadratic equation, so

$$u_1, u_2 = \frac{-B \pm \sqrt{B^2 - 4AC}}{2A}. \quad (\text{A.4})$$

373 In particular, setting  $M = (u_1 + u_2)/2$  to be the mean of the roots, and  $P = u_1 u_2$  to be the product of the roots, we have

$$M = -\frac{B}{2A} = -\frac{\omega_{12}}{\omega_{22}}, \quad P = \frac{C}{A} = \frac{\omega_{11} - 1}{\omega_{22}}. \quad (\text{A.5})$$

374 Let us try to maximize  $\det(\Omega)$  subject to the ellipse satisfying (A.2). Start with an arbitrary  $\omega_{22} > 0$ . Then (A.5)  
375 determines the remaining elements of  $\Omega$ ,

$$\omega_{12} = -M\omega_{22}, \quad \omega_{11} = 1 + P\omega_{22}.$$

376 Hence

$$\det(\Omega) = \omega_{11}\omega_{22} - \omega_{12}^2 = \omega_{22} - Q\omega_{22}^2,$$

377 where

$$Q = M^2 - P = \frac{1}{4}(u_1 - u_2)^2 > 0. \quad (\text{A.6})$$

378 Maximizing  $\det(\Omega)$  with respect to the choice of  $\omega_{22}$  leads to  $\omega_{22} = 1/(2Q)$  and

$$\det(\Omega) = 1/(4Q).$$

379 The remaining task is to choose  $u_1 < 0$  (which determines  $u_2 > 0$  by (A.3)) to maximize  $\det(\Omega)$ , or equivalently,  
380 to minimize  $Q$  in (A.6).

381 Recall a basic result from calculus. If  $t = f(u)$  and  $u = g(t)$  are monotone functions which are inverse to one  
382 another, then  $g(f(u)) = u$ . Differentiating two times yields the relation between the derivatives,

$$g' = 1/f', \quad g'' = -f''/\{f'\}^3.$$

383 In particular, consider  $f(u) = \Phi(u)$ , with derivatives  $f'(u) = \phi(u)$  and  $f''(u) = -u\phi(u)$ , where  $\phi(u)$  is the probability  
384 density function of  $\mathcal{N}(0, 1)$ . Then  $g(t) = \Phi^{-1}(t)$  with derivatives  $g'(t) = 1/\phi(u)$  and  $g''(t) = u/\{\phi(u)\}^2$ , where  $u =$   
385  $\Phi^{-1}(t)$ .

With this notation, write  $u_1 = g(t)$  for  $0 < t < 1/2$ . Then  $u_2 = g(t + 1/2)$ . Write  $\phi_1 = \phi(u_1)$ ,  $\phi_2 = \phi(u_2)$ . The  
quantity  $Q$  in (A.6), treated as a function of  $t$ , has derivatives

$$\begin{aligned} Q' &= \frac{1}{2} \{u_1 u_1' - u_1 u_2' - u_1' u_2 + u_2 u_2'\} \\ &= \frac{1}{2} \{u_1(1/\phi_1 - 1/\phi_2) + u_2(1/\phi_2 - 1/\phi_1)\}, \\ Q'' &= \frac{1}{2} \{u_1 u_1'' + (u_1')^2 - u_1 u_2'' - 2u_1' u_2' - u_1'' u_2 + u_2 u_2'' + (u_2')^2\} \\ &= \frac{1}{2} \{u_1^2/\phi_1^2 + 1/\phi_1^2 - u_1 u_2/\phi_2^2 - 2/(\phi_1 \phi_2) - u_1 u_2/\phi_1^2 + u_2^2/\phi_2^2 + 1/\phi_2^2\} \\ &= \frac{1}{2} \{(1/\phi_1 - 1/\phi_2)^2 + u_1^2/\phi_1^2 - u_1 u_2/(1/\phi_1^2 + 1/\phi_2^2) + u_2^2/\phi_2^2\}. \end{aligned}$$

386 If  $u_1 = -d$ , then  $u_2 = d$  and  $\phi_1 = \phi_2$  so that the first derivative vanishes. For all  $(0 < t < 1/2)$ , the second derivative  
387 is positive, so the function is convex. Hence  $Q$  is minimized for  $u_1 = -d$ ,  $u_2 = d$ . Then  $M = 0$ ,  $Q = -P = d^2$  and the  
388 optimal  $\Sigma$  becomes

$$\Sigma = \Omega^{-1} = \begin{bmatrix} 2 & 0 \\ 0 & 2d^2 \end{bmatrix},$$

389 as required.

## 390 References

- 391 [1] M. Ahsanullah, B. M. Golam Kibria, and M. Shakil. *Normal and Student's  $t$  Distributions and Their Applications*. Atlantis Press, Paris,  
392 2014.
- 393 [2] F. S. Alashwali. *Robustness and Multivariate Analysis*. PhD thesis, University of Leeds, November 2013.
- 394 [3] D. F. Andrews, P. J. Bickel, F. R. Hampel, P. J. Huber, W. H. Rogers, and J. W. Tukey. *Robust Estimates of Location: Survey and Advances*.  
395 Princeton University Press, 1972.
- 396 [4] O. Arslan, P. D. Constable, and J. T. Kent. Convergence behavior of the EM algorithm for the multivariate  $t$ -distribution. *Comm. Statist.*  
397 *Theor. Meth.*, 24:2981–3000, 1995.
- 398 [5] J. H. Friedman and J. W. Tukey. A projection pursuit algorithm for exploratory data analysis. *IEEE Transactions on Computers*, 100:881–890,  
399 1974.
- 400 [6] R. Grubel. The length of the shorth. *Ann. Statist.*, 16:619–628, 1988.
- 401 [7] J. T. Kent and D. E. Tyler. Constrained M-estimation for multivariate location and scatter. *Ann. Statist.*, 24:1346–1370, 1996.
- 402 [8] J. T. Kent, D. E. Tyler, and Y. Vardi. A curious likelihood identity for the multivariate  $t$ -distribution. *Comm. Statist. Sim. Comp.*, 23:441–453,  
403 1994.
- 404 [9] R. A. Maronna, R. D. Martin, and V. J. Yohai. *Robust Statistics*. Wiley, Chichester, 2006.
- 405 [10] K. Nordhausen, H. Oja, and D. E. Tyler. Tools for exploring multivariate data: the package ICS. *Journal of Statistical Software*, 28:1–31,  
406 2008.
- 407 [11] D. Peña and F. J. Prieto. Cluster identification using projections. *J. Am. Statist. Ass.*, 96:1433–1445, 2001.
- 408 [12] D. Peña, F. J. Prieto, and J. Viladomat. Eigenvectors of a kurtosis matrix as interesting directions to reveal cluster structure. *Journal of*  
409 *Multivariate Analysis*, 101:1995–2007, 2010.
- 410 [13] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014.  
411 <http://www.R-project.org/>.
- 412 [14] P. Rousseeuw. Multivariate estimation with high breakdown point. In W. Grossman, G. Pflug, I. Vincze, , and Wertz W., editors, *Mathematical*  
413 *Statistics and its Applications*, volume B, Dordrecht, 1985. Reidel.
- 414 [15] P. J. Rousseeuw and K. V. Driessen. A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 41:212–223, 1999.
- 415 [16] Valentin Todorov and Peter Filzmoser. An object-oriented framework for robust multivariate analysis. *Journal of Statistical Software*, 32:  
416 1–47, 2009. <http://www.jstatsoft.org/v32/i03/>.
- 417 [17] D. E. Tyler, F. Critchly, L. Dumbgen, and H. Oja. Invariant co-ordinate selection. *J. R. Statist. Soc. B*, 71:549–592, 2009.
- 418 [18] S. Van Aelst and P. Rousseeuw. Minimum volume ellipsoid. *Wiley Interdisciplinary Reviews: Computational Statistics*, 1:71–82, 2009.
- 419 [19] W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Springer, New York, fourth edition, 2002.  
420 <http://www.stats.ox.ac.uk/pub/MASS4>.