Table 1. Relative Mean Squared Errors of Estimated Correlation Using RMCD with $\alpha = .75$ With Respect to NNVE for Various Levels of Contamination in the Bivariate Normal and the Bivariate Skewed-Normal Scenarios

%MSE	Bivariate normal scenario; correlation = 0	Bivariate skew-normal scenario; correlation = −.5
0% outliers	8.37	6.39
5% outliers	8.82	6.02
33% outliers	2.27	1.46
50% outliers	1.61	1.26
67% outliers	.64	.29

comparing it to the RMCD estimator. The choice of the trimming constant of the MCD is nonstandard: $\alpha = .75$, meaning that the MCD will be based on that subset containing 25% of the data having the smallest value for the determinant of its covariance matrix. This choice of α will give us more protection against percentages of scattered outliers >50%. But on the other hand, we will be less well protected against clusters of outliers having little or no dispersion. This behavior under contamination is therefore similar to that of the NNVE. Table 1 reports the relative MSEs (%MSE) of the RMCD procedure with $\alpha = .75$ with respect to NNVE, using the same simulation setup as in Tables 2 and 3 of Wang and Raftery. We only present here the relative MSE for the estimator of the correlation coefficient. Indeed, we think that it is far more important to have an accurate estimate of the shape of the covariance matrix than of its size. Many procedures in multivariate analysis are even size invariant and use only the correlation structure of the data.

From Table 1 we see that the NNVE is outperforming the RMCD in practically all situations. An exception is the case with >50% of outliers, for which the RMCD with $\alpha = .75$ performs better. This modest simulation study confirms that the NNVE has good properties at finite samples, but is also shows us that NNVE is not the only robust covariance matrix estimator that can cope with a large amount of scattered outliers.

ADDITIONAL REFERENCES

Butler, R. W., Davies, P. L., and Jhun, M. (1993), "Asymptotics for the Minimum Covariance Determinant Estimator," *The Annals of Statistics*, 21, 1385–1400.

Croux, C., and Dehon, C. (2001), "Robust Linear Discriminant Analysis Using S-Estimators," *The Canadian Journal of Statistics*, 29, 473–492.

——— (2002), "Analyse Canonique Basée sur des Estimateurs Robustes de la Matrice de Covariance," *La Revue de Statistique Apliquée*, 2, 5–26.

Croux, C., and Haesbroeck, G. (1999), "Influence Function and Efficiency of the Minimum Covariance Determinant Scatter Matrix Estimator," *The Journal of Multivariate Analysis*, 71, 161–190.

——— (2000), "Principal Component Analysis Based on Robust Estimators of the Covariance or Correlation Matrix: Influence Function and Efficiencies," *Biometrika*, 87, 603–618.

Lopuhaä, H. P. (1999), "Asymptotics of Reweighted Estimators of Multivariate Location and Scatter," *The Annals of Statistics*, 27, 1638–1665.

Maronna, R. A., and Yohai, V. J. (1998), "Robust Estimation of Multivariate Location and Scatter," in *Encyclopedia of Statistical Sciences Update*, Vol. 2, eds. S. Kotz, C. Read, and D. Banks, New York: Wiley, pp. 589–596.

Pison, G., Rousseeuw, P. J., Filzmoser, P., and Croux, C. (2002), Robust Factor Analysis, *Journal of Multivariate Analysis*, to appear.

Rocke, D. M., and Woodruff, D. L. (2000), A Synthesis of Outlier Detection and Cluster Identification," unpublished manuscript.

Rousseeuw, P. J., and Van Driessen, K. (1999), "A Fast Algorithm for the Minimum Covariance Determinant Estimator," *Technometrics*, 41, 212–223.

Rousseeuw, P. J., Van Aelst, S., and Van Driessen, K. (2000), "Robust Multivariate Regression," unpublished manuscript.

Comment

Geoffrey J. McLachlan and Karyn L. Hamaty

We congratulate the authors on their interesting article and their new NNVE procedure for the robust estimation of a covariance matrix by exploiting the NNC cleaning method of Byers and Raftery (1998). The detection of outliers in multivariate data is a difficult but very important problem. The NNC method for removing much clutter from a dataset as in, for instance, the linear minefield example, is very impressive. Its adaptation to robust estimation by the artificial introduction of extra outlying points is novel. In this discussion, we focus on the performance of NNVE relative to the approach based on a mixture model analysis using normal and t components via the EMMIX software (McLachlan, Peel, Basford, and Adams 1999).

1. SPURIOUS CLUSTERS

In Section 2.2 the authors state that "when a mixture model is fit to data that have only one component in reality, the maximum likelihood estimator (MLE), when it exists, tends to falsely indicate that there are two components." It is true that bimodality in histograms of linear combinations of multivariate observations does not always imply that the data have been sampled from a mixture distribution. This point was illustrated in the seminal paper of Day (1969) on normal mixture models in which he demonstrated the presence of spurious clusters in a dataset. Following his approach, McLachlan and Peel (2002, sec. 1.8) generated a random sample of size $n = 50$ from a spherically symmetric $p = 10$-dimensional normal distribution.

Geoffrey J. McLachlan is Professor and Karyn L. Hamaty is Research Assistant, Department of Mathematics, University of Queensland, Brisbane, 4072 Australia (E-mail: *gjm@maths.uq.edu.au*).

Journal of the American Statistical Association
December 2002, Vol. 97, No. 460, Theory and Methods
DOI 10.1198/016214502388618807

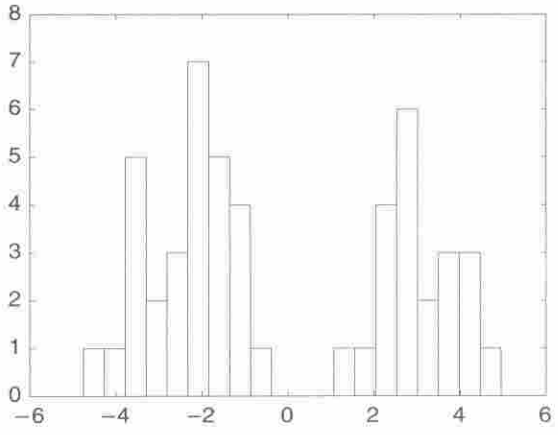

Figure 1. Histogram of First Canonical Variate for 10-Dimensional Simulated Normal Dataset of Size $n = 50$.

They then plotted the histogram of the univariate projections $\hat{a}^T x_1, \ldots, \hat{a}^T x_n$, where

$$\hat{a} = \widehat{\Sigma}^{-1}(\hat{\mu}_1 - \hat{\mu}_2),$$

and $\hat{\mu}_1, \hat{\mu}_2$, and $\widehat{\Sigma}$ are the estimates obtained from fitting a mixture of two 10-dimensional normal components with means μ_1 and μ_2 and common covariance matrix Σ. That is, these univariate projections are the first canonical variates when two multivariate normal groups with means $\hat{\mu}_1$ and $\hat{\mu}_2$ and common covariance matrix $\widehat{\Sigma}$ are imposed on the data. Their plot is given in Figure 1. The bimodal nature of the histogram suggests that the data have not come from a single normal distribution.

However, this spurious clustering can be detected in practice. For example, the likelihood ratio test statistic λ can be applied to the simulated data represented in Figure 1 to test the null hypothesis H_0 of a single normal component against the alternative of a two-component normal mixture with equal covariance matrices. The value of $-2\log\lambda$ was found to be 31.41. As is well known, regularity conditions do not hold for the likelihood ratio test statistic for this test to have its usual null chi-squared distribution. However, the resampling approach advocated by McLachlan (1987) can be used to the assess the p value. Using this approach with $B = 199$ replications, McLachlan and Peel (2002) assessed the p value to be approximately 47%. Hence the null hypothesis of a single normal component would be retained at any conventional level of significance. Note that as $g = 1$ under H_0, the null distribution of λ does not depend on any unknown parameters, and so the B replications of $-2\log\lambda$ generated here are actual, not bootstrap, replications. Thus if we were to reject the null hypothesis H_0 if the test value of $-2\log\lambda$ were greater than, say, the bth largest replicated value of this statistic, then this test would be of exact size $\alpha = 1 - b/(B+1)$.

2. MIXTURE ANALYSIS VIA NORMAL AND t COMPONENTS

Concerning the application of NNVE to the Hertzsprung–Russell and the Australian Athletes datasets, we now consider the analysis of these two sets using mixtures of normal and t components. Normal mixture models provide a model-based approach to clustering; (see, e.g., McLachlan and Basford 1988; McLachlan and Peel 2000). However, a single outlier can break down the parameter estimation for at least one of the components. McLachlan and Peel (1998) and Peel and McLachlan (2000) suggested using mixtures of t components as an alternative, because the t components are less sensitive to outliers, having longer tails than the normal. The t density with location parameter μ, positive definite matrix Σ, and ν degrees of freedom is given by

$$f(x;\mu,\Sigma,\nu) = \frac{\Gamma(\frac{\nu+p}{2})|\Sigma|^{-1/2}}{(\pi\nu)^{\frac{1}{2}p}\Gamma(\frac{\nu}{2})\{1+\delta(x,\mu;\Sigma)/\nu\}^{\frac{1}{2}(\nu+p)}}, \quad (1)$$

where

$$\delta(x,\mu;\Sigma) = (x-\mu)^T\Sigma^{-1}(x-\mu) \quad (2)$$

denotes the Mahalanobis squared distance between x and μ (with Σ as the covariance matrix). If $\nu > 1$, μ is the mean of X, and if $\nu > 2$, $\nu(\nu-2)^{-1}\Sigma$ is its covariance matrix. As ν tends to infinity, X becomes marginally multivariate normal with mean μ and covariance matrix Σ.

The t distribution does not have substantially better breakdown behavior than the normal (Tyler, 1994). The advantage of the t mixture model is that, although the number of outliers needed for breakdown is almost the same as with the normal mixture model, the outliers have to be much larger. This point is made more precise in Hennig (2002) who has provided an excellent account of breakdown points for maximum likelihood estimation of location-scale mixtures with a fixed number of components g. Of course as explained in Hennig (2002), mixture models can be made more robust by allowing the number of components g to grow with the number of outliers.

Hertzsprung–Russell Data

We first consider the Hertzsprung–Russell dataset. Fitting a single t component to it via the expectation-maximization algorithm reveals the presence of six outliers, as indicated by six observations having very small weights in the iterative computation of the estimates. Table 1, reports the results on the next stage of fitting a mixture of $g = 2$ normal components with unrestricted covariance matrices and a mixture of $g = 2$ t components with unrestricted scale matrices and degrees of freedom ν_1 and ν_2. It can be seen from Table 1 that these two mixture models lead to estimates of the covariance matrix similar to that given by NNVE.

Table 1. Covariance Estimates for the Star Data

NNVE		*t mixture*		*Normal mixture*	
.0115	.0343	.0116	.0348	.0116	.0345
.0343	.2390	.0348	.2403	.0345	.2392

NOTE: Mixture model estimates are those for the covariances of the component corresponding to the main body of the data.

Australian Athletes Data

The authors also analyzed the Australian Athletes dataset to illustrate the relative performance of NNVE where there is a lack of elliptical symmetry in the data. We use this dataset to demonstrate how we might use a mixture analysis based on t and normal component distributions to assess whether the data consist of one major cloud. This approach can be viewed as complementary or even as an alternative to the procedure proposed by the authors for determining whether the signal consists of more than one data cloud.

Because the fitting of a single t component to these $n = 202$ five-dimensional observations clearly revealed many outliers and a bad fit, we proceeded to fit a mixture of $g = 2\,t$ components. It gave a clustering of the data into 2 clusters of almost the same size (105 and 97), with the first 100 observations and 5 of the last 102 observations comprising the first cluster. This clustering has (almost) recovered the sex of the athletes, as it is known that the first 100 observations are on females and the last 102 are on males. The estimated degrees of freedom $\hat{\nu}_1$ and $\hat{\nu}_2$ for the two components are 17.62 and 5.59. The small value of $\hat{\nu}_2$ suggests that the data on the males have longer tails than the normal distribution. A subsequent inspection showed that several observations $\boldsymbol{x}_j$ on the males had very small values for their weights with respect to the second component, suggesting that there are several outliers among the male data. The fitting of $g = 3$ normal components (t components were assessed as not being necessary in the case of three components) produced a clustering in which the males were partitioned into the second and third clusters of size 69 and 32 (with another male being put in the first cluster corresponding to the females). The estimates of the mean and covariance matrix for the second and third components showed that the smaller cluster of males has a greater mean for all five variables than for the larger cluster of males and a greater variance for all but the third variable. The differences are appreciable for the fourth and fifth variables (LBB and FERR). The p value obtained via resampling for the likelihood ratio test of $g = 2$ versus $g = 3$ normal components was found to be significant at the 5% level. The subsequent test of $g = 3$ versus $g = 4$ normal components was not significant ($p = .45$). This mixture model analysis has thus revealed that this set is comprised of data from essentially three normal populations and so the estimation of a single covariance matrix in the sense that the signal consists of one major cloud would be inappropriate.

3. RELATIVE EFFICIENCY OF NEAREST-NEIGHBOR VARIANCE ESTIMATION

Wang and Raftery (2002) conducted a simulation study to evaluate the relative performance of their NNVE method in estimating the covariance matrix on the basis of a sample of $n = 500$ bivariate observations drawn from a mixture in proportions π_1 and $\pi_2 = 1 - \pi_1$ of two normals with mean 0 and covariance matrix $\boldsymbol{\Sigma}$ and $10\boldsymbol{\Sigma}$, where $\boldsymbol{\Sigma} = \text{diag}(4, 25)$ for $\pi_2 = 0\%$ (no outliers), 5%, 33%, 50%, and 67%. Five hundred simulation trials were performed for each level of the proportion of outliers π_2. To illustrate the efficiency of the NNVE in estimating $\boldsymbol{\Sigma}$, we performed a simulation experiment with the same number of trials for the same population configurations with $\pi_2 = 33\%, 50\%$, and 67%, but with $\boldsymbol{\Sigma}$ estimated by fitting by maximum likelihood a mixture of $g = 2$ normal components with unrestricted means and covariance matrices. The estimate $\widehat{\boldsymbol{\Sigma}}$ of $\boldsymbol{\Sigma}$ was taken to be the estimate of the covariance matrix for the component corresponding to the population with $\boldsymbol{\Sigma}$ as its covariance matrix. The results are displayed in Table 2. Comparing the MSEs of the estimates for each of the three distinct elements of $\boldsymbol{\Sigma}$, it can be seen in the cases of 33% and 50% outliers that the (simulated) relative efficiency of NNVE ranges between 52% and 92% for the estimation of the two variances and is >100% for the covariance. However, in the case of 67% outliers, the relative efficiency is extremely low, only 1% for the two variances.

Table 2. Monte Carlo Averages, Standard Errors, MSEs, and Relative MSEs of Estimated Covariance for Various Levels of Contamination in the Bivariate Simulated Example

	NNVE			Normal mixture		
	4.0	25.0	0	4.0	25.0	0
33% outliers						
Mean	3.86	23.08	−.01	3.99	25.08	.02
SE	.35	2.17	.53	.35	2.09	.66
MSE	.14	8.37	.28	.12	4.37	.44
%MSE	1.00	1.00	1.00	.86	.52	1.57
50% outliers						
Mean	4.06	23.72	−.01	4.00	24.89	.02
SE	.48	2.63	.71	.43	2.80	.73
MSE	.23	8.53	.50	.18	7.84	.53
%MSE	1.00	1.00	1.00	.78	.92	1.06
67% outliers						
Mean	4.76	26.44	.11	4.01	25.39	−.03
SE	7.38	40.39	2.23	.58	3.42	.94
MSE	54.87	1,630.42	4.99	.34	11.82	.88
% MSE	1.00	1.00	1.00	.01	.01	.18

NOTE: Each dataset has 500 observations. Each relative MSE was calculated by dividing the MSE by that of NNVE as given in Table 2 of the article.

ADDITIONAL REFERENCES

Hennig, C. (2002), "Breakdown Points of Maximum Likelihood-Estimators of Location-Scale Mixtures," Private communication.

McLachlan, G. J. (1987), "On Bootstrapping the Likelihood Ratio Test Statistic for the Number of Components in a Normal Mixture," *Applied Statistics*, 36, 318–324.

McLachlan, G. J., and Basford, K. E. (1988), *Mixture Models: Inference and Applications to Clustering*, New York: Marcel Dekker.

McLachlan, G. J., and Peel, D. (1998), "Robust Cluster Analysis via Mixtures of Multivariate t-Distributions," in *Lecture Notes in Computer Science*, Vol. 1451, eds. A. Amin, D. Dori, P. Pudil, and H. Freeman, Berlin: Springer-Verlag, pp. 658–666.

McLachlan, G. J., and Peel, D. (2000), *Finite Mixture Models*, New York: Wiley.

McLachlan, G. J., Peel, D., Basford, K. E., and Adams, P. (1999), "The EMMIX Algorithm for the Fitting of Mixtures of Normal and t-Components," *Journal of Statistical Software*, 4 (*http://www.stat.ucla.edu/journals/jss/*).

Peel, D., and McLachlan, G. J. (2000), "Robust Mixture Modelling Using the t Distribution," *Statistics and Computing*, 10, 335–344.