Nearest-Neighbor Variance Estimation (NNVE): Robust Covariance Estimation via Nearest-Neighbor Cleaning

Naisyin WANG and Adrian E. RAFTERY

Robust covariance estimation commonly proceeds by downweighting outliers. In this article we measure the "outlyingness" of a data point by the standardized distance between the point and its Kth nearest neighbor. The appropriate weights for robust estimation are found by a model-based mixture modeling approach that follows from considering the data cloud as a realization of a high-dimensional point process. To correct a potential bias when there are no outliers, we introduce a boundary correction procedure that artificially adds in extra outlying points; the resulting methodology is called nearest-neighbor variance estimation (NNVE). The strength of NNVE is its robustness against a large proportion of noise points and against deviation from normality of the signal. A consistency result for the method is established. Under some reasonable assumptions, it is shown that the covariance estimate is bounded and that each point has only bounded influence on the final estimates. NNVE outperformed the popular minimum volume ellipsoid (MVE) estimator in simulation studies, with a big improvement when the proportion of outliers was very large (>50%). In our simulation study, when the proportion of outliers was ≥50%, the mean squared error of the NNVE estimator of variance was at least 100 times smaller than that of the MVE estimator. The proposed estimator also outperformed MVE in cases where the underlying data distribution was not normal. Good performance of NNVE in several real examples is demonstrated. A potential drawback of NNVE is that data points condensed in moderate-sized clusters would be classified as signal. Even though we do not support approaches discarding moderate-sized clusters as outliers without checking, this feature of NNVE could be problematic, particularly when only the main data cloud is of interest. A simple diagnostic tool built on existing model-based clustering procedure is proposed. This procedure allows us to check whether there is more than one separate data cloud in the data after cleaning. It also supplies the central locations of the separated moderate-sized clusters, which allows further investigation. Finally, because NNVE reduces the problem of finding the robustness weights to a one-dimensional problem, it may be useful in high-dimensional problems, such as those encountered in data mining.

KEY WORDS: EM algorithm; M estimator; Minimum volume ellipsoid estimator; Mixture model.

1. INTRODUCTION

Outliers are observations that do not follow the pattern of the majority of the data. It is well known that the presence of outliers can damage the quality of estimates. In this article we address the problem of robustly estimating a covariance matrix. There is a substantial literature on this problem (see, e.g., Campbell 1980; Huber 1981; Lopuhaä 1989; Maronna 1976; Maronna and Yohai 1995; Poston et al. 1997; Rousseeuw and Leroy 1987; Ruppert 1992; Tyler 1994; Woodruff and Rocke 1994). In this article we consider a "weighted" covariance matrix for robust covariance estimation in which the weight for each data point depends on a measure of "outlyingness." However, the nature of this estimator and the way in which we define outlyingness are very different from the previous approaches.

After ensuring scale standardization of each variable, we measure the outlyingness of a point by the distance from the point to its Kth nearest neighbor. If a point, or a small group of points, with a number less than K is far from the rest of the data, then it tends to have a large Kth nearest-neighbor (NN) distance and is viewed as an outlier. On the other hand, a tight cluster containing a moderate number of data points would not be automatically classified as outliers because of our belief that a moderate size cluster may not arise simply by chance and could instead be an interesting feature of the data. A simple procedure is then used to locate these clusters, after which further action can be taken. This idea was first discussed by Byers and Raftery (1998), hereafter referred to as BR, who proposed a nearest-neighbor cleaning (NNC) procedure to remove clutter from a dataset comprising both "main" data and irrelevant clutter. Their work was motivated by the problem of detecting minefields from noisy images collected by an aircraft in which many objects are identified as possible mines; some are mines, whereas many others are just clutter. In such cases, the proportion of outliers is unknown and can be much greater than 50%—a scenario under which traditional covariance estimators are not applicable.

BR mentioned the possibility of using their method to produce robust covariance estimates, and carried out an intriguing simulation study (BR, table 3). This compared the performance of NNC with that of the minimum volume ellipsoid estimator (MVE) (see Rousseeuw and Leroy 1987), probably the most commonly used robust covariance estimator in practice. In BR's example, signal and noise points are both normally distributed, and the noise points are assumed to have a larger variance. BR found that the performance of MVE started to deteriorate markedly when the proportion of outliers came close to 50%, whereas NNC remained steady. However, when there are no outliers, the NNC covariance estimator

Naisyin Wang is Professor of Statistics and Toxicology, Department of Statistics, Texas A&M University, College Station, TX 77843. Adrian E. Raftery is Professor of Statistics and Sociology, University of Washington, Seattle, WA 98195. Wang's research was funded by National Cancer Institute grant CA-74552 and Texas Advanced Research Program grant TARP-010366-0194 and was conducted while Wang was visiting the Statistics Department, University of Washington, Seattle. Raftery's research was supported by Office of Naval Research grant N-00014-96-1-0192. The authors are grateful to Simon Byers, Ray Carroll, and Doug Martin for helpful conversations and to Geoffrey McLachlan for pointing out the reference of Day (1969). They also gratefully acknowledge the editor, associate editor, and the referees for their helpful comments and suggestions that substantially improved the article.

Journal of the American Statistical Association
December 2002, Vol. 97, No. 460, Theory and Methods
DOI 10.1198/016214502388618780

tends to underestimate the true variances. The good performance of NNC with very large numbers of outliers is interesting, but the bias shown in BR's study makes it impossible to recommend it as a general procedure.

In this article we investigate the potential of using NNC as a general method for robust covariance estimation. Examples in Section 4 suggest that traditional ellipsoid-based methods, such as MVE, could lead to dubious conclusions when the underlying distribution of the signal points deviates from normality. Exploring an alternative concept of "outlyingness" beyond the usual approach of covering the central 50% of the data with an ellipsoid is a main theme of this article. In light of BR's study and our own findings in Sections 3 and 4, our goal is to provide an estimator that performs as well as the original NNC when there are outliers, but does not suffer much from bias when there are no outliers. In Section 2 we describe our approach, which is based on artificially adding extra outliers to the data. We refer to the resulting methodology as nearest-neighbor variance estimation (NNVE). We give theoretical findings to support the use of NNVE in Section 3. In Section 4 we report on several simulation studies and data analyses that compare the performance of various estimators. We provide concluding remarks in Section 5.

2. METHODOLOGY

2.1 Nearest-Neighbor Cleaning

The original NNC procedure was proposed by Byers and Raftery (1998). They suggested viewing the data as a mixture of two components, one arising from the outliers or clutter and other arising from the data that are not outliers. We refer to these two types of data as noise points and signal points. BR proposed basing the decision as to whether to view a data point as signal or noise on the distance from it to the Kth nearest data point, that is, the Kth nearest-neighbor distance, D_K. They then pointed out that under certain assumptions, the distribution of D_K is approximately a mixture of two generalized gamma distributions,

$$D_K \sim \tau\Gamma^{1/p}(D_K; K, \lambda_1\alpha_p) + (1-\tau)\Gamma^{1/p}(D_K; K, \lambda_2\alpha_p), \quad (1)$$

where τ indicates the mixture proportion; λ_1 and λ_2 characterize the distributions corresponding to signal and noise points; α_p is the volume of the p-dimensional unit hypersphere, with p being the dimension of the data; and $D \sim \Gamma^{1/p}(\cdot; a, b)$ indicates that $D^p \sim \Gamma(\cdot; a, b)$. The intuition is that the density of points is lower outside the region of p space in which the signal points are concentrated, and hence that outliers have a less dense distribution, and thus a large Kth NN distance, on average. The distribution in (1) was derived by BR based on the assumption that the original data, X, could be approximately modeled as a mixture of two Poisson processes.

By modeling the distribution of D_K, BR reduced the dimensionality of the mixture modeling problem from p to 1. Denote the density of $\Gamma^{1/p}(\cdot; K, \lambda\alpha_p)$ by $f_{D_K}(\cdot, \lambda)$, let $\theta = (\tau, \lambda_1, \lambda_2)$, and define

$$h(d, \theta) = \frac{\tau f_{D_K}(d, \lambda_1)}{\tau f_{D_K}(d, \lambda_1) + (1-\tau) f_{D_K}(d, \lambda_2)}, \quad (2)$$

which is the probability of a datum being a signal point given that its D_K is equal to d. A simple EM algorithm can be used to obtain the estimates of the parameters λ_1, λ_2, and τ. The "missing data" in this problem is the binary 0-1 indicator variable, Z_i, where $Z_i = 1$ if the ith datum is a signal point. On convergence, we obtain

$$\hat{\lambda}_1 = \frac{K\sum_i h(D_{Ki}, \hat{\theta})}{\alpha_p \sum_i \{h(D_{Ki}, \hat{\theta}) D_{Ki}^p\}}, \quad (3)$$

$$\hat{\lambda}_2 = \frac{K\sum_i \{1 - h(D_{Ki}, \hat{\theta})\}}{\alpha_p \sum_i \{1 - h(D_{Ki}, \hat{\theta})\} D_{Ki}^p}, \quad (4)$$

and

$$\hat{\tau} = n^{-1} \sum_i h(D_{Ki}, \hat{\theta}). \quad (5)$$

Let $\hat{h}_i = h_i(\hat{\theta}) = h(D_{Ki}, \hat{\theta})$ and $\hat{w}_i = \hat{h}_i / \sum_j \hat{h}_j$, where $\hat{\theta}$ denotes the maximum likelihood estimator, obtained by the EM algorithm. The mean, μ_1, and covariance, Σ_1, of the distribution of the signal points can be estimated by $\hat{\mu}_{1,NN} = \sum \hat{w}_i X_i$ and

$$\hat{\Sigma}_{1,NN} = \sum \hat{w}_i (X_i - \hat{\mu}_{1,NN})(X_i - \hat{\mu}_{1,NN})'. \quad (6)$$

2.2 Boundary Correction and Estimation Algorithm

It is well known that when a two-component 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 (see Day 1969; Bryant and Williamson 1978). For example, a simulated study that we conducted with normal data suggested that the misclassification rates varied from 5% to 25% when the data were assumed to be from a mixture of two normals. When there is only one component, we have $\tau = 1$, which is on the boundary of the parameter space, $\tau \in (0, 1)$. It is well known that the MLE tends to perform less well in such situations than when the true parameter value is inside the parameter space (e.g., Feng and McCulloch 1992). In our setup, this misclassification causes underestimation of the covariance matrix of the signal distribution when there are no outliers. When there are outliers, the parameter of the mixture model is in the interior of the parameter space, and the original NNC procedure works well. Our goal is to find a method that works similarly to the original NNC when $\tau < 1$, but corrects the underestimation when $\tau = 1$.

Our idea is quite simple. Because NNC tends to misclassify signal points as noise when there is no noise, we propose to artificially add extra noise to the original dataset, so that the true τ can be moved away from the boundary. We first give some numerical illustrations to show how the method works, and then give a more systematic description of it.

A conventional measurement of distance from a data point to the center is the Mahalanobis distance, $d_M(x, \mu, \Sigma) = \{(x-\mu)'\Sigma^{-1}(x-\mu)\}^{1/2}$. Let $\hat{\mu}_{all}$ and $\hat{\Sigma}_{all}$ be the sample mean and covariance of the entire dataset. Then one way to obtain new outliers is to generate data points with squared Mahalanobis distances, $d_M^2(x, \hat{\mu}_{all}, \hat{\Sigma}_{all})$, equal to some prespecified

value, q. Here we simply choose q to be the ζ quantile of the chi-square distribution with p degrees of freedom; that is, $q = \chi^2_{p,\zeta}$, with ζ being a prespecified probability close to 1. Various other choices could be made; this particular choice worked well in all of our numerical studies.

To illustrate our method, we use the two-dimensional situation considered in BR where both signal and noise points have bivariate normal distributions with mean 0, the signal covariance matrix is diag(4, 25), and the noise points have the same distribution, only multiplied by 10. We consider two scenarios: data with no outliers, and data with 10% outliers. Six different values of q were used; these were equally spaced between $\chi^2_{2,.99} = 9.2$ and $\chi^2_{2,.9999} = 18.4$. For each q, M data points with the desired squared Mahananobis distance, q, were added to each of the two datasets, where $M = 5\%$ of the original sample size. These were chosen by selecting the M points in the dataset with the largest Kth NN distances, and then projecting each of them to a randomly chosen direction such that the squared Mahalanobis distance between the new location and $\hat{\mu}_{all}$ is equal to q. In each case, a new data point is added at that location. The NNC procedure was then used to estimate the signal covariance in each case. Our numerical experiments indicate that the results are relatively insensitive to the choice of the M artificial data points. We refer to this overall methodology as NNVE.

Figure 1 plots the resulting variances of the two variables against the corresponding q values. In each plot, the original variance without the added noise is indicated by the y coordinate of the triangle. When there were no outliers, the original NNC underestimated the true variances. The estimated variance of the first variable was 2.3, whereas the truth is 4.0. The top two plots of Figure 1 show that NNVE reduces the underestimation; when $q = \chi^2_{2,.99}$, the estimate is 3.1. The results improve as q increases, reaching an estimate of about 3.5; the improvements become negligible after about $q = 15$. On the other hand, with 10% outliers, the added noise appears to have little or no effect on the estimated variances, and NNVE

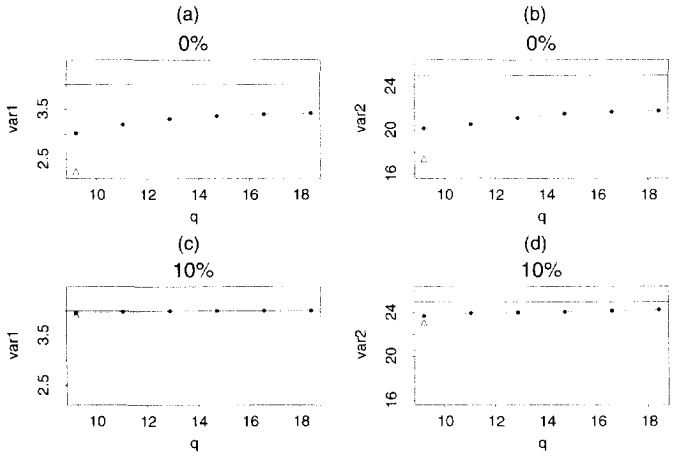

Figure 1. Plots of Changing Variance Estimates for Various Values, q, of the Artificial Added Data. (a) and (b) are for data with no outliers, and (c) and (d) are for data with 10% outliers. The true variance for the first variable [(a) and (c)] is 4, and that for the second variable [(b) and (d)] is 25. The y coordinate of the triangle shows the original NNC variance estimate. The solid horizontal lines indicate the true variances.

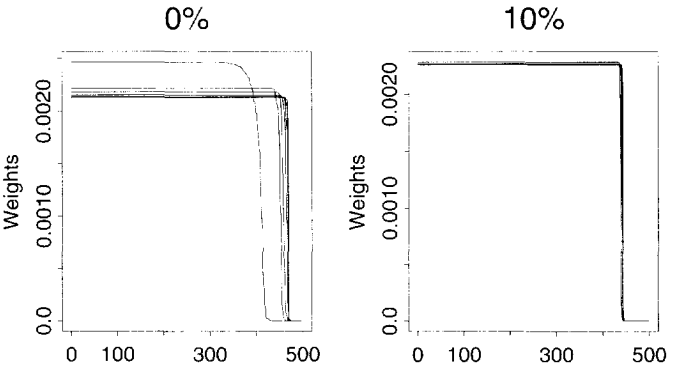

Figure 2. Estimated Weights versus the Ranks of Kth NN Distance for the Original Data and Six Augmented Versions Corresponding to Different Values of q Described in Figure 1.

gave results very similar to the original NNC method. This is exactly the kind of adjustment to NNC that we sought.

The explanation for this behavior is shown in Figure 2, where the estimated weights, w_i, in (6) for each data point against the rank of its Kth NN distance are given for both scenarios. In each plot there are seven curves, corresponding to the original dataset and the six augmented datasets described in Figure 1. For the augmented datasets, only the w_i and the Kth NN distance of the original 500 data points were used to construct the curves. The weights for the added outliers are practically zero. By design, they have large NN distances.

The curve for the no outlier case is clearly separated from the curves for the augmented data cases. This no-outlier curve indicates that signal points with large Kth NN distances (about the top 16% in ranking) were classified as outliers and contributed close to zero weights in (6). With the original signal points being normally distributed (unimodal), a point far away from the mean (mode) usually belongs to a low-density area and tends to have a larger Kth NN distance compared to a point close to the mode. Consequently, such a point is more likely to be downweighted by the original NNC and to cause underestimation of the variances. As shown by the other curves, the added noise helps the signal points that were originally misclassified as outliers to regain substantial weight. On the other hand, when there are 10% outliers, these outliers are correctly downweighted. The weights used to calculate the estimated covariance hardly change, as desired, even with the added points. We observed similar graphs when the proportion of outliers was >10%. For an easy graphical presentation, Figures 1 and 2 were simply obtained based on two original datasets. Nonetheless, our simulation study in Section 4 under the same structure indicates that similar behavior is to be expected in general.

Some further aspects of the practical implementation of NNVE are as follows. When calculating the NN distances, each variable in the data was standardized by subtracting its sample median and then divided by its median absolute deviation (MAD). This step provides scale invariance when calculating the NN distances. As a result, NNVE is equivariant for scale multiplication of each variable, due to, for example, changes of scale in the coordinates.

When q is increased past a certain point, eventually some original outliers get misclassified as signal points. In that case,

we observe a surge in the estimated variance as q increases. To avoid misclassifying influential outliers as signal points, if any estimated variance increases by a factor of κ or more between two consecutive values of q, we stop the procedure and use the results obtained with the previous value of q as the final answer. The variance ratio of a normal random variable and a truncated normal with the central 80% and 85% are 2.39 and 1.92. Motivated by this, we use $\kappa = 2$, which lies between the two values. Note that this surge can happen only when there are outliers with moderate influence, a situation in which even the original NNC provides a reasonable answer. Therefore, the choice of κ is not very crucial.

We let Q be the set of values of q used. Extending q outside the range of Q improves the resulting estimates when there are no outliers. Therefore, we can consider extending q using the same consecutive addition as for the q's in Q and stop the procedure either when a variance increases by a factor of more than κ as before or when none of the estimated variances changes much (we define this as being a change of $<1\%$).

Our NNVE procedure can be summarized as follows:

1. Standardize each variable by subtracting its sample median and dividing by its MAD.
2. Use the Kth NN distance and the EM algorithm to obtain $\hat{w}_i$ and calculate the resulting NNC weighted covariance estimate.
3. Choose Q, as the set of six chi-square quantiles described earlier, by themselves or augmented with the extension points.
4. For each q_i in Q, generate M extra outliers, add them to the dataset, and obtain the resulting new weighted covariance estimate.
5. Stop incrementing q when the maximum value in Q has been reached, or when one of the estimated variances increases by a factor of 2 or more for one increment of q, or when none of the estimated variances changes by much. Otherwise, increment q and repeat the procedure.

In the numerical examples throughout this article, we used $M = \max(2, .05 * n)$, where n is the sample size. The 5% was chosen to represent a small proportion; thus in each augmented dataset, the original data comprise the vast majority. Numerical studies were conducted to compare the effects of using different proportions ranging from 3% to 8%. The results are insensitive to changes in M, indicating that there is little to be gained by developing a method for choosing M optimally.

2.3 Remarks on the Estimation Procedure

Remark 1. The reason for using more than one value of q is simply to improve the estimation when there are no influential outliers. A simpler version of the method would just use a single value of q, such as $q = \chi^2_{p,.99}$; this would provide a reasonable and fast option. Nevertheless, using several values of q leads to better performance when there are no outliers, at the cost of a modest increase in complexity.

Remark 2. BR proposed an entropy-based method for choosing K. They suggested calculating a negative entropy-based measurement of the signal-outlier classification $\mathcal{E}(K) = \sum h_i \log(h_i)$ for each K, where $h\log(h)$ is defined to be 0 when $h = 0$. It is motivated by the fact that the classification of the ith data as signal or noise can be described by the binary indicator variable Z_i as in Section 2.1. The negative entropy, $h_i \log(h_i)$, of the binary distribution of Z_i provides information on classification certainty and will reach a supremum 0 in the extreme case where there is no classification uncertainty and where h_i is either 0 or 1. $\mathcal{E}(K)$ therefore summarizes the amount of classification certainty inherent in the classification probabilities for all data points. BR proposed plotting $\mathcal{E}(K)$ against K, and choosing the smallest value of K beyond which $\mathcal{E}(K)$ increases little. That is, to choose the smallest value of K which approximately provides the lowest classification uncertainty. We have found that our results do not change much when K is varied within a reasonable range; BR reported the same thing. As a result, visual inspection of the $\mathcal{E}(K)$ plot may well be sufficient to choose a reasonable K; BR also proposed a more formal method based on change-point estimation.

For the first two examples in Section 4, where the estimated proportion of outliers is smaller than 50%, we used $K = 8$. For both examples, the entropy approach suggests that one use a value of K in the range 8 to 15, and similar results were obtained for all the values of K in this range. An entropy plot is also given in the first example of Section 4 below.

Remark 3. In classifying signal and noise, the choice of K has certain practical implications. For example, using $K = 8$ would tend to classify a cluster of size smaller than 8 as noise, but points in a tight cluster of a larger size as signal. This is in spite of the signal being from one major data cloud. In practice, however, it would seem wise to investigate further before classifying any tight cluster of moderate size as noise, since such groupings may indicate some feature of interest in the data and are unlikely to arise by chance. One way of identifying such moderate-sized tight clusters is to apply a clustering procedure to the initally identified signal points after cleaning. The cleaning step is important here because it makes the clustering procedure feasible. In the last example of Section 4, where the proportion of noise points is extremely high, we illustrate how to implement such a procedure using a publicly available software package, MCLUST (Fraley and Raftery 1999), to check whether there is more than one data cloud. If prior knowledge indicates that the signal comprises one major data cloud (a common assumption taken by many robust methods, e.g., MVE), then one either calculates the covariance from the largest data cloud or reapplies NNVE to the data after the first cleaning. More details are given in Section 4.

3. THEORETICAL JUSTIFICATIONS

In this section we derive some theoretical results that support the use of NNVE. We start by reformulating NNVE as the solution to a set of estimating equations. We then provide a consistency result for the NNVE estimator. We also show that under certain reasonable assumptions, each data point has only bounded influence on the final estimates. Finally, we show that the NNVE estimates remain bounded, even when the proportion of outliers is >50%.

Without loss of generality, we assume that $\lambda_1 > \lambda_2 > 0$, with λ_2 corresponding to the noise population. More precisely, we

assume that $\theta = (\tau, \lambda_1, \lambda_2)$ is an interior point of Θ, where Θ is $\{(\tau, \lambda_1, \lambda_2); \tau \in (0, 1)$ and $\lambda_1 > \lambda_2 > 0\}$. Let $A_i(\lambda) = f_{D_K}(D_{Ki}, \lambda)$ and define

$$\Psi_{Ni} = \Psi_N(D_{Ki}, \theta) = \begin{bmatrix} \dfrac{A_i(\lambda_1) - A_i(\lambda_2)}{\tau A_i(\lambda_1) + (1-\tau)A_i(\lambda_2)} \\ h_i(\theta)(-\alpha_p D_{Ki}^p + K/\lambda_1) \\ \{1 - h_i(\theta)\}(-\alpha_p D_{Ki}^p + K/\lambda_2) \end{bmatrix} \tag{7}$$

and

$$\Psi_{Pi} = \Psi_P(X_i, D_{Ki}, \mu, \Sigma, \theta) = \begin{bmatrix} h_i(\theta)(X_i - \mu) \\ h_i(\theta)\mathrm{vec}\{\Sigma - (X_i - \mu)(X_i - \mu)'\} \end{bmatrix},$$

where μ and Σ are the mean and covariance of the signal points and "P" and "N" correspond to "primary" and "nuisance" (parameters). Then NNVE can be reformulated as the solution to a set of estimating equations by noting that $\hat{\theta}$, $\hat{\mu}_{1,NN}$, and $\widehat{\Sigma}_{1,NN}$ can be written as a solution to $\sum_i \Psi_i = 0$, where $\Psi_i = (\Psi_{Pi}^T, \Psi_{Ni}^T)^T$. The basic idea here is to use (7) to obtain $\hat{\theta}$. Weights based on the $\hat{\theta}$ are then used to obtain the estimates of primary interest, $\hat{\mu}_{1,NN}$ and $\widehat{\Sigma}_{1,NN}$.

To avoid extraordinary issues, we assume the following:

(A1) There exists a $\theta^* = (\tau^*, \lambda_1^*, \lambda_2^*)'$ in the interior of the parameter space and an $\eta > 0$, such that $n^\eta(\hat{\theta} - \theta^*) = O_p(1)$.

Proposition 1. Let $\mu_1^* = E\{(\tau^*)^{-1}h(D_K, \theta^*)X\}$ and $\Sigma_1^* = E\{(\tau^*)^{-1}h(D_K, \theta^*)(X - \mu_1^*)(X - \mu_1^*)'\}$. If (A1) holds, then $\hat{\mu}_{1,NN}$ and $\widehat{\Sigma}_{1,NN}$ are consistent estimators of μ_1^* and Σ_1^*.

Proposition 1 can be obtained from straightforward asymptotic derivation and the fact that by (5), $n^{-1}\sum h_i(\theta^*)$ converges to τ^*. The exact form of $h(D, \theta)$ is given in (8). Even more general theoretical derivations for M estimators with this structure have been provided in appendix 3.6 of Carroll, Ruppert, and Stefanski (1995). Note that condition (A1) does not require the true underlying distribution of D_K to be as specified in (1). The whole procedure of obtaining $\hat{\theta}$ is simply a process of establishing reasonable and robust weights. When the probability of the ith point being a signal point given the data is in fact $h_i(\theta^*)$, we have the following.

Corollary 1. If (A1) holds and $\Pr(Z_i = 1 \mid \text{data})$ equals $h_i(\theta^*)$, where $Z_i = 1$ indicates that X_i is a signal point, then Proposition 1 applies with $\mu_1^* = \mu_1$ and $\Sigma_1^* = \Sigma_1$, where μ_1 and Σ_1 are the mean and covariance of the signal distribution.

A sketch of the proof of Corollary 1 is given in the Appendix. Corollary 1 simply states that when $h_i(\theta^*)$ correctly specifies the conditional probability of the ith data value being a signal point given the data, then the given covariance estimator is consistent. In general, this can be at best approximately correct. Nonetheless, what we found was that, as with many other robust procedures, $\widehat{\Sigma}_{1,NN}$ achieves its robustness by downweighting outliers. In practice, $h_i(\theta^*)$ provides a convenient weight function that works very well in applications. To appreciate this, we note that $h(D, \theta)$ in (2) can be written as

$$h(D, \theta) = \left[1 + \frac{(1-\tau)\lambda_2^K}{\tau\lambda_1^K}\exp\{(\lambda_1 - \lambda_2)\alpha_p D^p\}\right]^{-1}. \tag{8}$$

The quality of the proposed estimator is determined mainly by $h(D, \theta)$. In the Appendix, we show the following result.

Lemma 1. For any θ in the interior of the parameter space, $h(D, \theta)$ is a monotone decreasing function of nonnegative D. Furthermore, for any positive integer ℓ and a bounded density $f_d(\cdot)$ with positive support, (a) $D^\ell h(D, \theta)$ is bounded for $D \geq 0$, (b) $\lim_{D\to\infty} D^\ell h(D, \theta) = 0$, and (c) $\int_0^\infty D^\ell h(D, \theta) df_d(D)$ is bounded.

Also note that, on convergence, eqs. (3)–(5) hold. As a result, data points with large D_K are more likely to be grouped into the noise population and to contribute to $\hat{\lambda}_2$. Such data points are also associated with small values of $h(D_K, \theta)$. Thus the NNVE procedure downweights the points with far-away neighbors and views them as potential outliers. Property (b) of Lemma 1 indicates that for a data point with extremely large D_K, the weight $h(D_K, \theta)$ not only goes to 0, but also reduces the contribution of this point toward the estimate to 0. Consequently, property (c) of Lemma 1 further indicates that $n^{-1}\sum_i D_{Ki}^\ell h(D_{Ki}, \theta)$ is always bounded.

Recall the definition of $\widehat{\Sigma}_{1,NN}$ in (6). In light of the results in Lemma 1, the following assumption is needed to link a point with a relatively large scale of $|X - \mu_1|$ with a small $h(D_K, \theta)$ so that the point will be downweighted:

(A2) For any $C > 0$, there exists a positive constant η such that for all $|X_i - \mu_1| > C$, $D_{Ki}^\eta / |X_i - \mu_1| > C$.

Condition (A2) simply requires that D_K be large when $|X - \mu_1|$ is. The purpose is to rule out a situation that is unlikely in real applications in which a tight cluster of more than K noise points go to infinity together. If such a situation does occur, then a direct application of NNVE would not be appropriate, and the cluster of outliers should be identified and analyzed separately. Note that (A2) is simply a theoretical condition to ensure that the estimates are bounded. The implementation of NNVE does not use this assumption. Because μ_1 is unknown, an assumption-checking procedure that does not require the location of μ_1 is provided in the last example of Section 4. The basic idea is that after the "cleaning" procedure, we identify any data point that potentially could be a violator of the assumption. For a data point with a large D_K, a simple process is used to decide whether this point is "close" enough to the main data cloud, and thus is analogously "close" enough to μ_1. Details are given in Section 4.

Proposition 2. As long as (A1) and the relationship between X and D as described in (A2) hold for all of the data, we have the following robustness properties of $\hat{\mu}_{1,NN}$ and $\widehat{\Sigma}_{1,NN}$:

$$(\hat{\mu}_{1,NN} - \mu_1^*) = \left\{\frac{1}{n}\sum_{i=1}^n \psi_\mu(D_{Ki}, X_i, \mu_1^*, \tau^*, \hat{\theta})\right\}\{1 + o_p(1)\}$$

and

$$vec(\hat{\Sigma}_{1,NN} - \Sigma_1^*) = \left\{ \frac{1}{n} \sum_{i=1}^{n} \psi_\Sigma(D_{Ki}, X_i, \mu_1^*, \Sigma_1^*, \tau^*, \hat{\theta}) \right\} \{1 + o_p(1)\},$$

where

$$\psi_\mu(D, X, \mu, \epsilon, \theta) = \epsilon^{-1} h(D, \theta)(X - \mu)$$

and

$$\psi_\Sigma(D, X, \mu, \Sigma, \epsilon, \theta) = \epsilon^{-1} vec[h(D, \theta)\{(X - \mu)(X - \mu)' - \Sigma\}].$$

Also, let $\mathcal{T}$ in Θ be a neighborhood of θ^* where all $\theta \in \mathcal{T}$ are bounded away from the boundary of Θ. Then the following conditions hold:

(A) For any ϵ bounded away from 0, $sup_{\theta \in \mathcal{T}} \psi_\mu(D, X, \mu, \epsilon, \theta)$ and $sup_{\theta \in \mathcal{T}} \psi_\Sigma(D, X, \mu, \Sigma, \epsilon, \theta)$ are bounded functions of D and X for $D \geq 0$.

(B) Both $\hat{\mu}_{1,NN}$ and $\hat{\Sigma}_{1,NN}$ are bounded.

Proposition 2 is a straightforward application of direct asymptotic derivations, conditions (A1) and (A2), and Lemma 1. Under certain conditions that can be reasonably expected to hold in applications, part (A) indicates that, locally, each data point can have only bounded influence on the final estimates. Part (B) implies that even when there is more than 50% noise, $\hat{\mu}_{1,NN}$ and $\hat{\Sigma}_{1,NN}$ would still be bounded.

4. EXAMPLES

Here we report analyses of several observed and simulated datasets to investigate the numerical performance of the NNVE estimator. The main approach with which we compare it is the popular MVE estimator (Rousseeuw and Leroy 1987). The MVE covariance estimate is based on the minimum volume ellipsoid covering at least $[n/2 + 1]$ data points. A suitable factor is used to achieve consistency in the multivariate normal case. We used the default version of `cov.mve` in S-PLUS. The differences between NNVE and BR's NNC procedure (NN-BR) are also illustrated.

4.1 Hertzsprung–Russell Data

Our first example is an astronomy dataset discussed by Rousseeuw and Leroy (1987). The data were taken from the Hertzsprung–Russell diagram of the star cluster CYG OB1 and consist of measurements for 47 stars in the direction of Cygnus. Two variables reported are the logarithm of the surface temperature and the logarithm of the light intensity. A scatterplot of the data is given in Figure 3(a). This is an example in which MVE is known to work well. In this case, NNVE and NN-BR both yield practically the same answer as MVE, but all are quite different from the standard nonrobust method. The entropy plot, Figure 3(b), as described in Section 2.3, suggests that the best choice of K is around 8 or 9, but the estimated covariances for K from 8 to 15 are almost identical. The estimated covariances using $K = 8$ are given in Table 1.

4.2 BR's Simulated Example

This simulated example is an extension of BR's simulation study. The goal is to show that our approach has kept the strength of BR's original method when there is a high proportion of noise points, and corrects for the underestimation of variances when there is no noise. In the first part of the simulation study, our setup is the same as that of BR, section 4. The data are bivariate normal with mean 0 and covariance matrix diag(4, 25). The outliers have the same distribution but are multiplied by 10. Each dataset contains 500 observations.

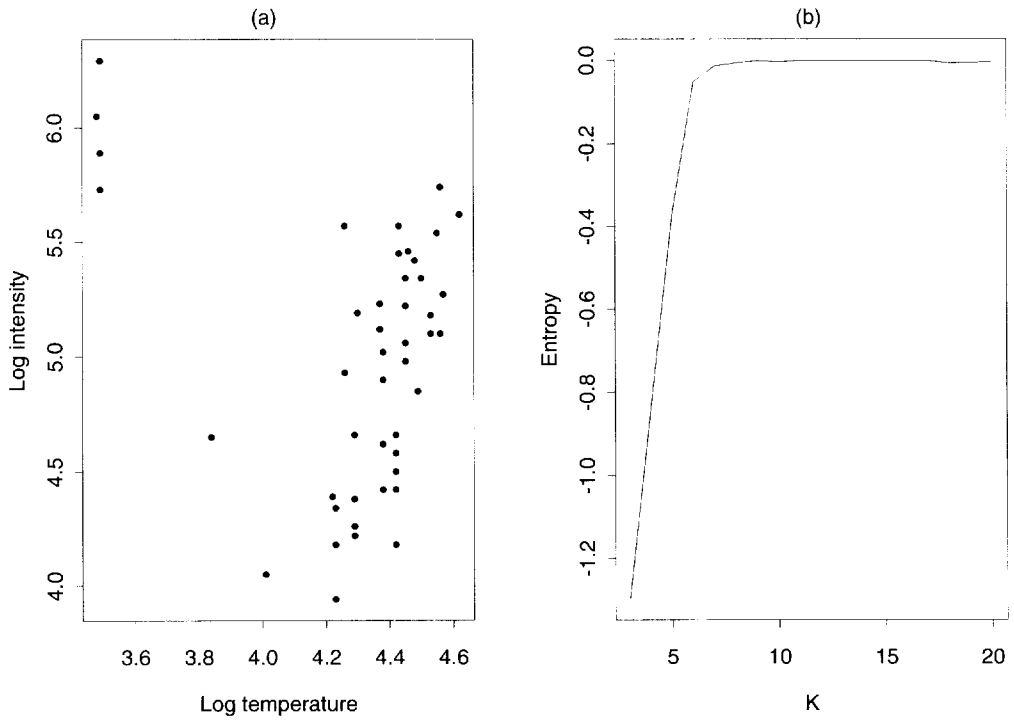

Figure 3. Plots for Hertzsprung–Russell Data. (a) Scatterplot of the original data; (b) entropy plot for choosing a good value of K.

Table 1. Covariance Estimates for the Star Data

Standard		NNVE		MVE		NN-BR	
.0846	−.0350	.0115	.0343	.0112	.0376	.0115	.0343
−.0350	.3263	.0343	.2390	.0376	.2350	.0343	.2390

Five cases were considered, in which the proportion of outliers is 0, 5%, 33%, 50%, and 67%.

The Monte Carlo means, standard errors, mean squared errors (MSEs); and relative MSE based on 500 simulations are displayed in Table 2. The relative MSE (%MSE) was calculated by dividing the MSE by that of NNVE. As expected, the standard nonrobust method breaks down immediately, giving very poor estimates even with 5% outliers. NNVE outperforms MVE in every case when there are outliers, and performs dramatically better when there are 50% or 67% outliers. When there are 50% outliers, the MSE of the variance estimates is at least 100 times greater for MVE than for NNVE. (The MSE for the standard method is a further three orders of magnitude greater than for MVE.) When there are 67% outliers, the superiority of NNVE over MVE is even more marked. NNVE and MVE perform about equally well when there are no outliers.

NN-BR underestimates the true variances when there are no outliers, and NNVE largely corrects this bias, reducing it by about two-thirds. With 5%–50% outliers, NN-BR performs well, but NNVE outperforms it. With 67% outliers, NN-BR outperforms NNVE.

In the second part of this simulation study, we investigated how the different methods performed in the presence of slight skewness of the signal distribution. To do this, we used the skewed bivariate normal distribution of Azzalini and Valle (1996) with skewness parameters $\delta_1 = .9$ and $\delta_2 = -.9$. The $(1, 1)$, $(2, 2)$, and $(1, 2)$ entries of the true covariance matrix are 4, 25, and -5. The δ parameters regulate the degree of skewness. Otherwise, the setup for this scenario is identical to the normal case in the first part of the simulation study. Figure 4 is a scatterplot of a simulated dataset with no outliers and density contours superimposed. The graph indicates that the distribution is a slightly skewed modification of a bivariate normal distribution. The results of the second part of the simulation are given in Table 3.

As in the normal case, the standard estimation method broke down immediately. An interesting phenomenon is that for no outliers or for 5% outliers, the MVE variance estimates are more downwardly biased than in the normal case. This suggests that MVE may be somewhat sensitive to mild violations of the normality assumption, which is not totally surprising given that fitting ellipses is at the heart of the MVE algorithm. On the other hand, the performances of NNVE and NN-BR are similar to those in the normal case, with NNVE again clearly outperforming MVE.

4.3 Australian Athletes Data

The next example is based on data from $n = 202$ athletes collected at the Australian Institute of Sport (Cook and Weisberg 1994). The variables considered here are percentage of body fat (BFAT), body mass index (BMI), red cell count (RCC), lean body mass (LBM), and plasma ferritin concentration (FERR). Figure 5 displays the pairwise scatterplots of these five variables.

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

	Standard			NNVE			MVE			NN-BR		
	4.0	*25.0*	*0*	*4.0*	*25.0*	*0*	*4.0*	*25.0*	*0*	*4.0*	*25.0*	*0*
0% outliers												
Mean	4.00	24.97	−.02	3.52	21.32	−.02	3.51	21.93	.01	2.37	14.62	−.03
SE	.25	1.55	.46	.27	1.75	.45	.32	2.13	.62	.27	1.84	.44
MSE	.06	2.41	.21	.30	16.62	.20	.34	13.99	.39	2.73	111.05	.20
%MSE	.20	.15	1.05	1.00	1.00	1.00	1.13	.84	1.95	9.10	6.68	1.00
5% outliers												
Mean	23.97	146.11	.18	3.76	22.80	−.01	3.61	22.53	0	3.64	22.00	−.02
SE	5.79	32.84	9.68	.28	1.80	.47	.32	2.03	.57	.28	1.79	.47
MSE	432.19	1.57E4	93.52	.13	8.07	.22	.26	10.19	.33	.21	12.22	.22
%MSE	3,324.54	1,945.48	425.09	1.00	1.00	1.00	2.00	1.26	1.50	1.62	1.51	1.00
33% outliers												
Mean	135.24	844.93	.71	3.86	23.08	−.01	4.28	26.52	.02	3.62	21.64	−.01
SE	14.13	95.04	26.17	.35	2.17	.53	.40	2.47	.65	.33	2.12	.51
MSE	1.74E4	6.81E5	684.00	.14	8.37	.28	.24	8.41	.42	.25	15.80	.26
%MSE	1.24E5	8.14E4	2,442.86	1.00	1.00	1.00	1.71	1.00	1.50	1.79	1.89	.93
50% outliers												
Mean	201.85	1261.30	2.02	4.06	23.72	−.01	8.51	53.39	.09	3.72	21.79	.02
SE	17.97	109.72	32.88	.48	2.63	.71	1.80	11.07	2.93	.42	2.37	.67
MSE	3.95E4	1.54E6	1,082.78	.23	8.53	.50	23.55	928.27	8.58	.25	15.9	.45
%MSE	1.72E5	1.81E5.27	2,165.56	1.00	1.00	1.00	102.39	108.82	17.16	1.09	1.86	.90
67% outliers												
Mean	269.84	1,688.67	−1.20	4.76	26.44	.11	129.35	797.39	.92	4.17	23.61	.05
SE	19.15	126.24	35.58	7.38	40.39	2.23	35.53	219.6	79.32	4.58	21.34	1.49
MSE	7.10E4	2.78E6	1,264.77	54.87	1,630.42	4.99	1.70E4	6.45E5	6,279.85	20.99	456.25	2.21
%MSE	1,293.97	1,705.08	253.46	1.00	1.00	1.00	309.82	395.60	1,258.49	.38	.28	.44

NOTE: Each dataset has 500 observations. Each relative MSE was were calculated by dividing the MSE by that of NNVE. The reported results are based on 500 simulations.

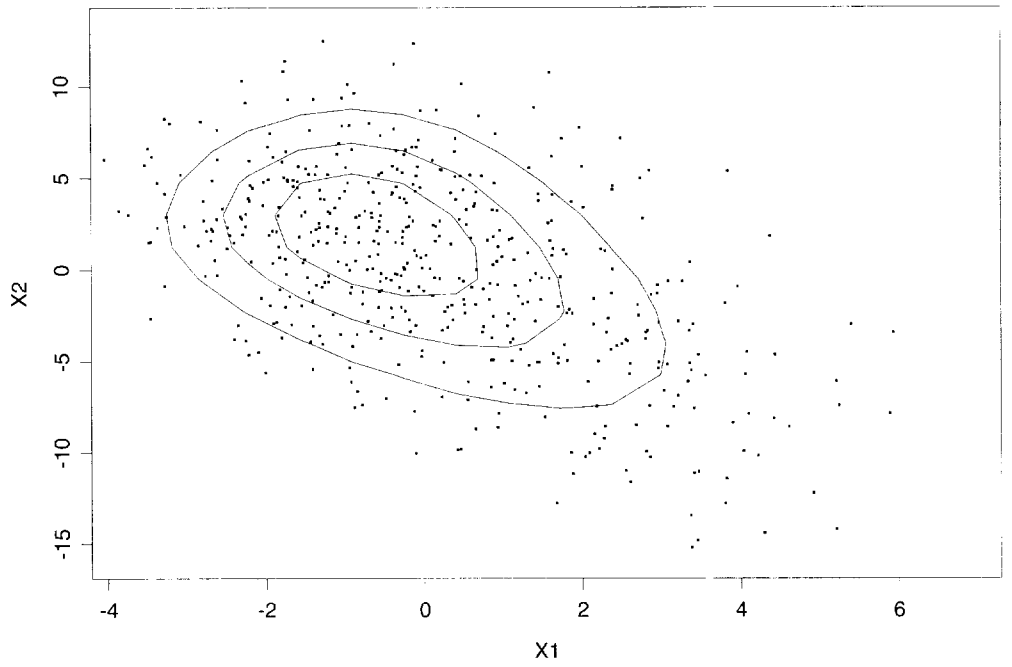

Figure 4. One Dataset From the Skew Normal Distribution Used in the Simulation, With No Outliers. The solids lines are density contours.

From these plots, we see that data points corresponding to certain athletes might follow patterns different from the majority. The standard sample covariance estimate is given in Table 4, and the estimates using MVE and NNVE are shown in Table 5.

Except for the variable FERR, for which the variance is about 13% larger when estimated by MVE than by NNVE, the NNVE and MVE estimates agree well for the rest of the entries in the covariance matrix, being within about 10% or less of each other. It is interesting to observe what happens when we consider only subsets of the variables. When only the first four variables are considered, the results are similar to those for all five variables. Table 6 shows estimates based on the first three variables and the first two variables.

Table 3. Monte Carlo Averages, Standard Errors, MSEs, and Relative MSEs of Estimated Covariance for Various Levels of Contamination in the Bivariate Skew Normal Scenario

	Standard			*NNVE*			*MVE*			*NN-BR*		
	4.0	*25.0*	*−5.0*	*4.0*	*25.0*	*−5.0*	*4.0*	*25.0*	*−5.0*	4.0	25.0	−5.0
0% outliers												
Mean	3.99	24.94	−4.96	3.35	20.47	−3.87	3.08	19.34	−3.27	2.27	13.95	−2.55
SE	.26	1.77	.55	.26	1.87	.53	.31	2.15	.75	.28	1.75	.50
MSE	.07	3.13	.31	.49	23.97	1.53	.94	36.69	3.51	3.06	125.06	6.19
%MSE	.14	.13	.20	1.00	1.00	1.00	1.92	1.53	2.29	6.24	5.22	4.05
5% outliers												
Mean	24.03	149.04	−29.97	3.63	22.17	−4.27	3.18	19.91	−3.43	3.51	21.38	−4.09
SE	6.53	36.38	12.86	.30	1.76	.59	.33	2.02	.77	.29	1.74	.58
MSE	443.55	1.67E4	789.00	.23	11.11	.87	.79	30.01	3.03	.33	16.14	1.14
%MSE	1,928.48	1,503.15	906.90	1.00	1.00	1.00	3.43	2.70	3.48	1.43	1.45	1.31
33% outliers												
Mean	134.70	838.63	−167.51	3.67	22.11	−4.20	4.07	25.20	−4.79	3.45	20.75	−3.91
SE	15.90	98.68	31.47	.35	2.31	.69	.42	2.81	.89	.35	2.26	.69
MSE	1.73E4	6.72E5	2.74E4	.23	13.67	1.10	.18	7.94	.82	.43	23.14	1.64
%MSE	7.52E4	4.92E4	2.49E4	1.00	1.00	1.00	.78	.58	.75	1.87	1.69	1.49
50% outliers												
Mean	202.21	1,260.15	−252.33	3.81	22.88	−4.35	8.72	53.78	−10.54	3.48	20.90	−3.92
SE	18.53	119.92	38.36	.51	3.04	.99	1.93	11.69	3.61	.40	2.67	.81
MSE	3.96E4	1.54E6	6.26E4	.30	13.72	1.39	25.98	964.43	43.84	.43	23.95	1.80
%MSE	1.32E5	1.12E5	4.49E4	1.00	1.00	1.00	86.60	70.29	31.54	1.43	1.75	1.29
67% outliers												
Mean	268.86	1,685.03	−336.69	6.67	40.31	−7.09	124.63	765.56	−160.05	5.28	31.95	−5.84
SE	21.97	137.5	45.17	21.72	130.42	21.59	28.78	181.81	57.72	12.74	76.91	14.37
MSE	7.06E4	2.77E6	1.12E5	478.01	1.72E4	469.81	1.54E4	5.81E5	2.74E4	163.74	5,951.96	206.86
%MSE	147.70	161.05	238.39	1.00	1.00	1.00	32.22	33.78	58.32	.34	.35	.44

NOTE: The set up is the same as that in Table 2.

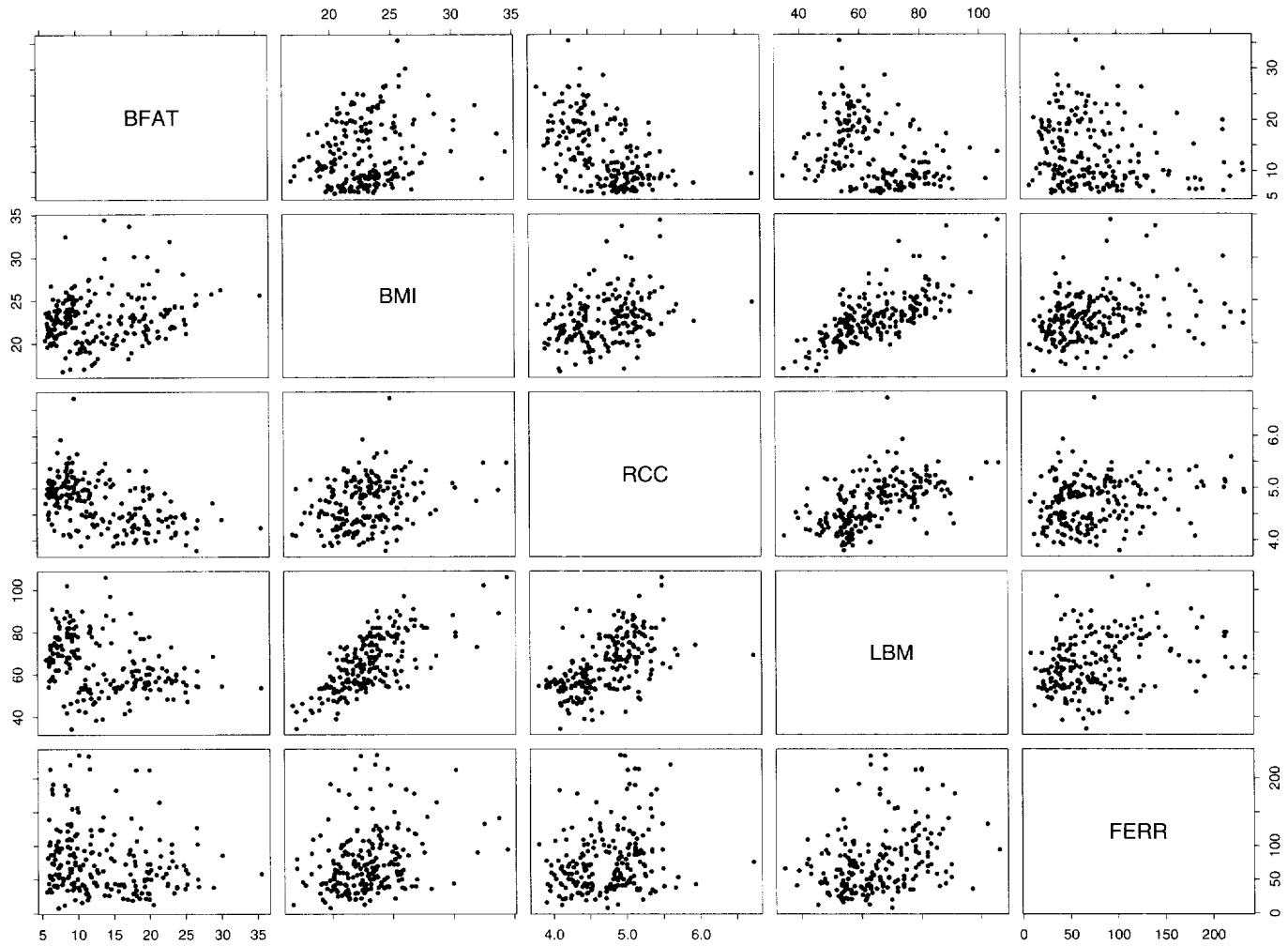

Figure 5. Pairwise Scatterplots for the Australian Athletes Data. The five variables are percentage of body fat (BFAT), body mass index (BMI), red cell count (RCC), lean body mass (LBM), and plasma ferritin concentration (FERR).

When we consider the first three variables (BFAT, BMI, and RCC), the NNVE and MVE covariance estimates are similar to each other and to the top 3×3 submatrices in Table 5.

When we consider just the first two variables (BFAT and BMI), however, we see a dramatic difference; the MVE estimate of the variance of BFAT is about half the NNVE estimate, and less than half as large as the estimates from all the methods using more variables. This behavior of MVE seems strange and may be due to the lack of elliptical symmetry nature of the bivariate distribution of BFAT and BMI (see Fig. 5).

Also, when considering just BFAT and BMI, the MVE estimate of the covariance between BFAT (percentage of body fat) and BMI (body mass index) is negative, whereas the NNVE estimate of the same covariance is positive. Which is correct? When all of the variables are considered, all of the estimates of this covariance are positive. Also, the nature of the two variables suggests a positive relationship. These considerations suggest that the sign of the MVE estimate is incorrect in this instance.

Table 4. Australian Athletes Data: Standard Covariance Estimation

			Standard		
BFAT	38.314	3.325	−1.399	−29.274	−53.920
BMI	3.325	8.202	.393	26.721	41.160
RCC	−1.399	.393	.210	3.298	5.457
LBM	−29.275	26.721	3.298	170.830	197.170
FERR	−53.920	41.160	5.457	197.170	2,256.368

In general, we do not expect the estimated covariances to stay the same when new variables are added. This is because, as is well known, some outliers in a high-dimensional setup cannot be detected when only a subset of the variables is considered. The phenomenon that we have observed here is the opposite, though: The results suggest that there are some points that MVE identifies as outliers when only BFAT and BMI are considered, but that no longer look like outliers to MVE when RCC is also taken into account. Based on the (BFAT, BMI) scatterplot and our simulation results in the skew-normal case, we believe that this underestimation on the part of MVE may well be due to the lack of elliptical symmetry of the distributions. Note that this problem does not arise with NNVE.

4.4 Example of Results With Irregular Dimensional Discrepancy

To investigate the phenomenon observed in Section 4.3, we conducted the following simulation study. Signal points were simulated from a mixture of two normal distributions to create

Table 5. Australian Athletes Data: MVE and NNVE Robust Covariance Estimates Based on All Five Variables

	MVE					NNVE				
BFAT	36.86	2.45	−1.37	−29.65	−34.29	37.25	2.19	−1.43	−31.00	−42.29
BMI	2.46	5.56	.27	19.79	18.89	2.19	5.02	.22	18.25	18.56
RCC	−1.37	.27	.18	3.01	2.48	−1.43	.22	.18	2.86	2.98
LBM	−29.65	19.79	3.01	147.75	124.24	−31.00	18.25	2.86	144.64	128.37
FERR	−34.29	18.89	2.48	124.24	1,079.12	−42.29	18.56	2.98	128.37	1,218.86

a four-dimensional dataset to cover a slightly "bent" region. Ten points from two clusters were then added as outliers. The exact simulation setup is given in the Appendix. The practice of using normal mixtures to approximate a wide variety of distributions has been discussed by Marron and Wand (1992) and Roeder and Wasserman (1997).

We applied both MVE and NNVE with $K = 8$ to the full datasets and obtained very similar four-dimensional covariance matrices using both methods. However, when we concentrate on two of the four variables, $v1$ and $v2$, a similar phenomenon to that observed in the previous section occurred. Sample averages of the estimated variances and correlations of the first two variables based on 500 simulations are displayed in Table 7. The rows from top to bottom correspond to the true values, the estimates based on all four variables and the estimates based on only the first two variables.

For MVE, the sample average of the estimated var($v1$) using $v1$ and $v2$ only is about 55% of the estimated var($v1$) using all four variables. The results obtained by NNVE are much more consistent across different dimensions. The correlation for these two variables is very close to 0 and is so estimated by both methods using all variables. However, when using two variables only, MVE estimated a negative correlation of $-.224$. A scatterplot of the first and the second variables from one randomly chosen dataset is shown in Figure 6. It can be seen from the plot that one group of outliers becomes undetectable when we consider only the first two variables. Nonetheless, including them barely changes the two-dimensional estimated variances. Based on this plot, we believe that most people would be likely to conclude that there are only four outliers and that the rest of the data are signal points. However, MVE produced much smaller variance estimates using $v1$ and $v2$ only. Equivalent findings were also observed when the parameter of interest is the determinant of the covariance matrix. The estimates obtained by MVE using $v1$ and $v2$ are less than 50% as large on average as those obtained by MVE using all four variables. NNVE does not have this problem.

4.5 Linear Minefield

Our last example is the simulated linear minefield considered by BR and by Dasgupta and Raftery (1998). A partially mined area is imaged by an aircraft and the resulting image is processed to identify possible mines; many of these are not in fact mines, but are actually clutter. The result is a set of points in which one region (the minefield) has a higher density than the rest (clutter). The goal of the analysis is to estimate characteristics of the minefield; here we focus on the covariance matrix (which in turn summarizes information about its orientation, its area, and aspects of its shape).

The data were simulated based on specifications given by Muise and Smith (1992) to reflect typical datasets of this kind. Here we use the same dataset as used by BR; a scatterplot of the data is given in Figure 7(a). The proportion of noise in this example is about 88%, which is extremely high. The diagonal elements of the sample covariance of the signal points, after being multiplied by 100, are .05 and 6.26.

MVE overestimated the variances, and obtained estimates of 8.04 and 7.90. Thus MVE overestimated the variance of the first variable by a factor of more than 100. The entropy plot for NNVE (not shown) suggested the best K to be around 35 to 40. Using NNVE with $K = 35$, we obtained estimated variances equal to 1.50 and 6.22. Counting the points with $\hat{h}_i > .5$ as signal, the estimated proportion of noise is about 75%. Even though the result of NNVE has been a great improvement over that of MVE, with so many outliers we found that some outliers were misclassified simply by chance.

As pointed out in Remark 3 in Section 2.3, in a situation with such a high proportion of outliers, a careful diagnosis of whether there is only one main data cloud is needed. We suggest a diagnostic tool that consists of applying the MCLUST procedure in S-PLUS using the "EI" option with the number of normal components chosen based on the Bayes information criterion (BIC) (Fraley and Raftery 1999). Here the "EI" option implies that the data are assumed to be a mixture of normals with a spherical shape and sharing an equal variance. To be precise, we assume that the covariance matrix, Σ_k, for

Table 6. Australian Athletes Data; MVE and NNVE Covariance Estimates Based on Three Variables (BFAT, BMI, and RCC) and Based on Two Variables (BFAT and BMI)

	MVE			NNVE			MVE		NNVE	
BFAT	31.46	.49	−1.37	32.80	.50	−1.46	17.44	−.21	33.04	.58
BMI	.49	4.84	.27	.50	4.24	.28	−.21	5.54	.58	4.16
RCC	−1.37	.27	.19	−1.46	.28	.18				

Table 7. MVE and NNVE Estimated Covariances

	MVE			NNVE		
	var(v1)	var(v2)	cor(v1, v2)	var(v1)	var(v2)	cor(v1, v2)
True	4.968	1.699	.091	4.968	1.699	.091
All variables	4.180	1.375	−.052	4.588	1.249	.035
v1, v2	2.333	1.240	−.224	4.109	1.057	.008

each normal component k is $\sigma^2 I$. Using NNVE with $K = 35$, the scatterplots for points with $\hat{h}_i > .5$ are given in Figure 7(b). Each circle in this plot corresponds to the 95% quantile contour of a normal component obtained using MCLUST/EI. One consequence of the spherical and equal variance constraints imposed by "EI" is that the data are covered by small circles (or balls in higher dimensions). That is, MCLUST produces more small normal components to best approximate the likelihood of the data. Two circles/balls do not overlap if the distance between the two centers is larger than $\sigma\sqrt{\chi^2_{d,q}}$, with d being the dimension of the data and q being the quantile level (e.g., .95).

Two issues are worth noting:

1. The concept of covering a compact set of any shape with small open balls is established elementary calculus. That is, we can use small overlapping balls to locate the main data cloud regardless of its shape.

2. Any points within a circle that does not overlap with the circles that cover the main data cloud provide the location of outliers that cannot be removed by one application of NNVE. Note that the large D_k's of these data imply that they are from tight clusters containing more than K points. In practice, without prior information that there is only one main data cloud, these points could very likely be signal points and should be carefully examined before their further removal is considered. The outcomes of MCLUST/EI provide identification of these data points. Furthermore, instead of considering a large number of data points, application of MCLUST allows us to simply concentrate on a much smaller number of normal components.

The following procedure for checking whether there is more than one separate data cloud in the data after cleaning is computationally straightforward. Within this diagnostic procedure, each mean of a normal mixture component is referred to as a "point." In the minefield data, after one application of NNVE, BIC suggests that there are 11 such "points" to be considered.

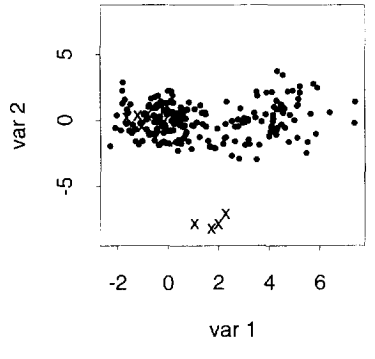

Figure 6. Scatterplot of v1 Versus v2 for a Dataset in Section 4.4. Outliers are indicated by "x".

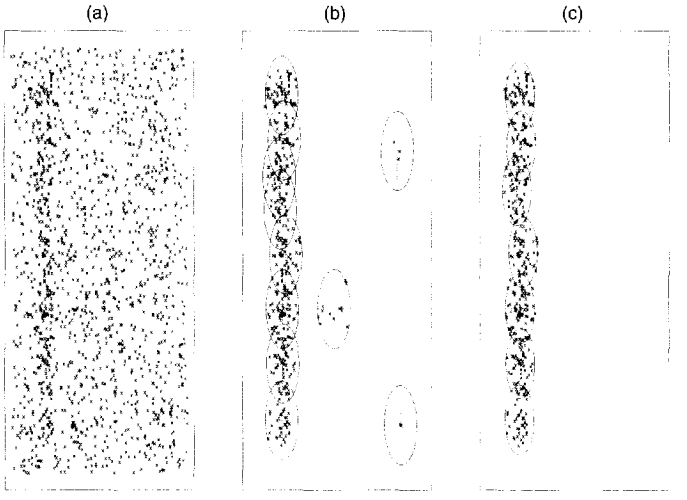

Figure 7. The Dasgupta and Raftery Linear Minefield Example. (a) The original dataset. (b) The result of one application of NNVE. (c) The result of two applications of NNVE. The values of K used for the first and second application of NNVE were 25 and 15.

Step 1. Start with any point and locate all other points within a $2 \times \sigma\sqrt{\chi^2_{d,q}}$ radius; index these as elements of group 1.

Step 2. For all points within this group, repeat step 1 until no other points can be included in this group.

Step 3. Choose a new point that has not been classified and repeat steps 1 and 2 with a new group ID.

Step 4. Stop the procedure when all points are classified. The group that contains the largest number of data points is considered the main data cloud.

For the mine data, we know that there is only one minefield. The separate circles in Figure 7(b) thus suggest a potential violation of assumption (A2). That is, based on the prior information that the signal points are from one main data cloud, some data points are "far" away from the center of the main signal points, μ_1, but still have small Kth NN distances. Note that this diagnostic technique requires no prior knowledge of the location of μ_1 or of its estimation during the procedure. Even though the illustration was given for a two-dimensional dataset, the procedure obviously is applicable to data of higher dimensions. As discussed in Remark 3 of Section 2.3, there are two potential alternatives when there is only one main data cloud. Estimating the sample covariance of points in the "main" data cloud provides estimates of .15 and 6.22. Reapplying NNVE to the data after the first cleaning as if the remaining points are regular data, we obtain estimates of .11 and 6.18. A scatterplot of points with $\hat{h}_i > .5$ after the second cleaning and the diagnostic circles are shown in Figure 7(c). We can see that after the second application of NNVE, all of the outliers were successfully removed. The diagnostic plot also indicates that just one main data cloud remains. The results for the second application are much closer to the true values of .05 and 6.26 than any of the other methods considered. This simple technique also provides a convenient tool for better understanding the signal structure after cleaning out the noise points.

5. DISCUSSION

BR introduced the NNC method for removing clutter in spatial point patterns. This can potentially be used also to remove or downweight outliers from other datasets and so provide robust estimation methods. In practice, however, the NNC procedure often has a downward bias when the signal points are from a unimodal distribution and there are no outliers. Here we have modified this method so that it no longer suffers from this problem, while retaining (and indeed, often enhancing) its good performance in the presence of outliers. The modification is based on the idea of artificially introducing some outliers into the data; this reduces the bias when there are no outliers, but hardly changes the results when there are outliers.

We have shown a consistency result for the resulting NNVE method, as well as the fact that under certain assumptions, each data point has bounded influence on the estimates and it remains bounded regardless of the proportion of outliers. Simulation studies show better performance than the popular MVE method when the proportion of outliers is large (especially $\geq 50\%$), and when the signal distribution is not elliptically symmetry.

When the proportion of outliers is massive ($\geq 50\%$), it seems advantageous to apply NNVE twice rather than once. We have found that the good performance of the method does not depend crucially on the precise choice of the values of K used in the two applications of NNVE. Moreover, when the proportion of outliers is very high, the original NNC method of BR (NN-BR) actually performs better than NNVE. This holds out the possibility of an adaptive procedure: If the proportion of outliers is high enough, then revert to the NN-BR method; otherwise, use NNVE. How and whether to do this and, if so, how to do it, remain open questions.

As pointed out by a referee, even though NNVE is scale equivariant, it is not affine equivariant. However, the concept of affine equivariance seems to be more meaningful when the underlying distribution of the signal points is elliptically symmetric than in other scenarios. The examples in Sections 4.2, 4.3, and 4.4 suggest that NNVE performs particularly well when the signal points do not follow an elliptically symmetric distribution. Because in practice one usually does not know either the distribution of the signal points or the proportion of outliers, NNVE provides a useful option, particularly during the data exploration stage.

In the context of data mining, there is considerable interest in estimating characteristics of massive datasets with plenty of noise. MVE and similar methods have difficulties with such problems. NNVE, however, has the advantage of reducing the calculation of the robustness weights to a one-dimensional problem, as well as the ability to clean out a high proportion of noise points, and thus may be more readily applicable than some other methods to robust covariance estimation in very large datasets.

APPENDIX: TECHNICAL DETAILS

Proof of Corollary 1

Denote the observed data by W, and let $h_i(\theta^*) = \Pr(Z_i = 1|W_i)$. It is easy to see that $n^{-1}\sum h_i(\theta^*)$ converges to $\mathrm{E}[\mathrm{E}\{I_{(Z=1)}|W\}] = \Pr(Z = 1)$. Let $f_W(w)$ be the density of the observed data. Then

$$\mu_1^* = \int x \frac{\Pr(Z=1|w)}{\Pr(Z=1)} f_W(w)\,dw$$
$$= \int x f(w|Z=1)\,dw = \mathrm{E}(X|Z=1) = \mu_1.$$

Similar derivations show that $\Sigma_1^* = \mathrm{var}(X|Z=1) = \Sigma_1$.

Proof of Lemma 1

It is obvious that $D^\ell\{h^{-1}(D,\theta)+1\}^{-1}$ is larger than $D^\ell h^{-1}(D,\theta)$ for positive D. Part (a) is a direct result of the fact that the former function is concave with a unique bounded maximum for $D > 0$. Part (b) follows L'Hôpital's rule. To prove part (c), first let $\ell^* = (\ell+1)/p - 1$. By changing variables in the integration, it is easy to see that there exist positive M_1 and M_2 such that the integral of interest is smaller than $\int M_1 y^{\ell^*}\exp(-M_2 y)\,dy$, which for certain $M_3 > 0$ and a positive integer k is less than

$$\int_0^1 M_1 y^{\ell^*}\,dy + \int_1^\infty M_3 y^k \exp(-M_2 y)\,dy. \tag{9}$$

Because $\ell^* > -1$, it is straightforward to show that (9) is bounded.

Simulation Setup of Section 4.4

Out of 200 signal points, 65% were generated from a MVN distribution with mean 0 and variance M_1, whereas the rest were from a MVN with mean $(4, 0, 4.5, 2)'$ and variance M_2. The upper triangular elements of M_1, from left to right, are 1.107, −.169, .094, .067, 1.268, −.150, −.106, 1.083, .059, and 1.042. Those for M_2 are 1.750, 1.061, .612, .433, 2.5, .866, .612, 1.5, .534, and 1.25. Clusters of 6 and 4 noise points were generated from MVN distributions with mean and variance μ_3, M_3 and μ_4, M_4, where $\mu_3 = (-1, 0, 5, 2.5)'$ and $\mu_4 = (1.5, -8, 2, 1.5)'$. The upper triangular elements of M_3 are .3, .035, .023, −.057, .25, −.0299, .019, .282, .007, and .368. M_4 is a diagonal matrix with all variances equal to .2.

[Received January 2000. Revised October 2001.]

REFERENCES

Azzalini, A., and Valle, A. D. (1996), "The Multivariate Skew-Normal Distribution," *Biometrika*, 83, 715–726.

Bryant, P., and Williamson, J. A. (1978), "Asymptotic Behaviour of Classification Maximum Likelihood Estimates," *Biometrika*, 65, 273–282.

Byers, S. D., and Raftery, A. E. (1998), "Nearest-Neighbor Clutter Removal for Estimating Features in Spatial Point Processes," *Journal of the American Statistical Association*, 93, 577–584.

Campbell, N. A. (1980), "Robust Procedures in Multivariate Analysis I: Robust Covariance Estimation," *Applied Statistics*, 29, 231–237.

Carroll, R. J., Ruppert, D., and Stefanski, L. A. (1995), *Measurement Error in Nonlinear Models*, London: Chapman and Hall.

Cook, R. D., and Weisberg, S. (1994), *An Introduction to Regression Graphics*, New York: Wiley.

Dasgupta, A., and Raftery, A. E. (1998), "Detecting Features in Spatial Point Processes With Clutter via Model-Based Clustering," *Journal of the American Statistical Association*, 93, 294–302.

Day, N. E. (1969), "Estimating the Components of a Mixture of Two Normal Distributions," *Biometrika*, 56, 463–474.

Feng, Z., and McCulloch, C. E. (1992), "Statistical Inference Using Maximum Likelihood Estimation and the Generalized Likelihood Ratio When the True Parameter is on the Boundary of the Parameter Space," *Statistics & Probability Letters*, 13, 325–332.

Fraley, C., and Raftery, A. E. (1999), "MCLUST: Software for Model-Based Clustering and Discriminant Analysis," *Journal of Classification*, 16, 297–306.

Huber, P. J. (1981), *Robust Statistics*, New York: Wiley.

Lopuhaä, H. P. (1989), "On the Relation Between *S*-Estimators and *M*-Estimators of Multivariate Location and Covariance," *The Annals of Statistics*, 17, 1662–1683.

Maronna, R. A. (1976), "Robust M-Estimators of Multivariate Location and Scatter," *The Annals of Statistics*, 4, 51–67.
Maronna, R. A., and Yohai, V. J. (1995), "The Behavior of the Stahel–Donoho Robust Multivariate Estimator," *Journal of the American Statistical Association*, 90, 330–341.
Marron, J. S., and Wand, M. P. (1992), "Exact Mean Integrated Squared Error," *The Annals of Statistics*, 20, 712–736.
Muise, R., and Smith, C. (1992), "Nonparametric Minefield Detection and Localization," Technical Report CSS-TM-591-91, Naval Surface Warfare Center, Coastal Systems Station.
Poston, W. L., Wegman, E. J., Priebe, C. E., and Solka, J. L. (1997), "A Deterministic Method for Robust Estimation of Multivariate Location and Shape," *Journal of Computational and Graphical Statistics*, 6, 300–313.
Roeder, K., and Wasserman, L. (1997), "Practical Bayesian Density Estimation Using Mixtures of Normals," *Journal of the American Statistical Association*, 92, 894–902.
Rousseeuw, P. J., and Leroy, A. M. (1987), *Robust Regression and Outlier Detection*, New York: Wiley.
Ruppert, D. (1992), "Computing S Estimators for Regression and Multivariate Location/Dispersion," *Journal of Computational and Graphical Statistics*, 1, 253–270.
Tyler, D. E. (1994), "Finite Sample Breakdown Points of Projection Based Multivariate Location and Scatter Statistics," *The Annals of Statistics*, 22, 1024–1044.
Woodruff, D. L., and Rocke, D. M. (1994), "Computable Robust Estimation of Multivariate Location and Shape in High Dimension Using Compound Estimators," *Journal of the American Statistical Association*, 89, 888–889.

Comment

Christophe CROUX and Stefan VAN AELST

First, we would like to congratulate the authors with their article, proposing an entirely new way of robust covariance matrix estimation using nearest neighbors. This intuitively appealing approach allows reliable covariance matrix estimation in the presence of high amounts of scattered noise. The initial step will downweight isolated and very small clusters of outliers, so what remains is the main group of the data and groups of clustered outliers. After this cleaning step, the data can be further analyzed. Clusters can then be detected using a clustering algorithm such as the MCLUST procedure of Fraley and Raftery (1999). Both steps together thus lead to a robust clustering method. To construct a robust covariance estimator, the authors add artificial outlying data points to make sure that the dataset always contains more than one group.

We believe that the proposed method is very useful for robust cluster analysis and is ideally suited for such applications as the detection of minefields as outlined by Dasgupta and Raftery (1998) and Byers and Raftery (1998). Robust cluster analysis is an important new area of research that turns out to be quite difficult. We refer here to recent work of Rocke and Woodruff (2000) that gives another way of doing robust cluster analysis. However, we have some concerns regarding the resulting estimator of robust covariance. In the first part of this discussion we focus on what is commonly expected from a robust covariance matrix estimator. In the second part we comment on some computational aspects of the NNVE and provide some further simulation results.

THE NEAREST-NEIGHBOR VARIANCE ESTIMATOR AS A ROBUST COVARIANCE ESTIMATOR

The population covariance matrix is a key quantity in multivariate statistics. If we are able to estimate it robustly, then we can use this estimate for robust principal component analysis, robust correlation analysis, robust factor analysis, and other applications. Suppose that we have p-variate observations $X_1, \ldots, X_n$ independently and identically generated from a certain distribution H. Then we denote the population covariance matrix as $\Sigma(H) = \mathrm{cov}_H(X)$. The NNVE $\widehat{\Sigma}_{1,NN}$ is now nothing else but a weighted sample covariance matrix, where the weights depend in a quite complicated way on the Kth NN Euclidean distances $D_{K1}, \ldots, D_{Kn}$.

Is the NNVE Estimating the Population Covariance Matrix?

The aim is that the NNVE estimates the population covariance matrix of the signal distribution. So if no outliers are present, then $\widehat{\Sigma}_{1,NN}$ should estimate $\Sigma(H)$. The authors show that $\widehat{\Sigma}_{1,NN}$ is a consistent estimator for its population counterpart, which we can denote by $\Sigma_1^*(H)$. Now in general, $\Sigma(H)$ will be different from $\Sigma_1^*(H)$. The authors suggest, and the simulations confirm, that in many applications $\Sigma_1^*(H)$ will be close to $\Sigma(H)$. However, from a mathematical standpoint, we cannot say that the NNVE consistently estimates the covariance matrix. Indeed, even for normal distribution there will be a (slight) difference between $\Sigma(H)$ and $\Sigma_1^*(H)$. But we do consistently estimate another population quantity.

This is reminiscent of many other robust estimators, including the minimum volume ellipsoid estimator (MVEE). The MVEE will, under some weak regularity conditions, consistently estimate its population counterpart which, we denote by MVE(H). The quantity MVE(H) will in general be different from $\Sigma(H)$, because it describes another feature of a multivariate distribution. But suppose now that H is an elliptically symmetric distribution, meaning that its density can be written in the form $g((x-\mu)'\Sigma^{-1}(x-\mu))$ for a certain real-valued function g. Then it is known that the MVE(H) equals

Christophe Croux is Professor, Department of Applied Economics, K. U. Leuven, Leuven, B-3000 Leuven, Belgium (E-mail: *christophe.croux@econ.kuleuven.ac.be*). Stefan Van Aelst is Professor, Department of Applied Mathematics and Computing, Ghent University, B-9000 Gent, Belgium (E-mail: *Stefan.VanAelst@rug.ac.be*).

Journal of the American Statistical Association
December 2002, Vol. 97, No. 460, Theory and Methods
DOI 10.1198/016214502388618799

(after multiplication by a consistency factor not depending on μ and Σ) the population covariance matrix, whenever the latter exists. This means that in elliptical models, the MVEE (like almost all affine-equivariant covariance matrix estimators) is a consistent estimator of $\Sigma(H)$. Particularly when sampling from a normal $N(\mu, \Sigma)$, the MVEE (and almost all other affine equivariant covariance determinant matrix estimators) will consistently estimate the parameter Σ, but this does not hold for the NNVE.

When sampling from nonelliptical distributions, like the skewed normal distributions of the authors' Table 3, the MVEE turns out to be a very bad estimator for $\Sigma(H)$. This is not surprising, because MVEE estimates an intrinsically different population quantity. In the sequel of this discussion we will therefore compare the NNVE with another weighted covariance matrix estimator, the reweighted minimum covariance determinant matrix estimator (RMCD). The latter estimator is affine equivariant and asymptotically normal, has a positive breakdown point, is implemented in some of the major statistical software packages, and is fast to compute (Rousseeuw and Van Driessen 1999). The weights are computed as follows:

$$w_i = \begin{cases} 1 & \text{if } (x_i - t_n)' C_n^{-1} (x_i - t_n) \le \chi^2_{p,.975} \\ 0 & \text{otherwise,} \end{cases}$$

where (t_n, C_n) are the initial MCD estimates. For defining this MCD estimator, consider all the subsets of size h $(h \le n)$ from the sample and keep that subset whose covariance matrix has the smallest determinant. Then the location and scatter MCD estimates are given by the average and covariance matrix computed over this optimal subset. Typically, the size of the subset equals $h = [n(1-\alpha)]$, with $\alpha = .5$ or $\alpha = .25$.

Do We Need Affine Equivariance?

Because the NNVE works with Euclidean distances and coordinatewise standardization of the data in the first step of the procedure, we do not have affine equivariance. We say that a covariance matrix estimator $\widehat{\Sigma}$ is affine equivariant if

$$\widehat{\Sigma}(AX_1 + b, \ldots, AX_n + b) = A\Sigma(X_1, \ldots, X_n)A'$$

for any regular matrix A and vector b. We believe that affine equivariance is an important property, and not only when working with elliptical model distributions. It also has some advantages at the finite-sample level. If we work with an affine-equivariant covariance matrix estimator $\widehat{\Sigma}$, then Mahalanobis distances based on $\widehat{\Sigma}$ are invariant to linear transformations of the data, the principal components remain the same under an orthogonal transformation of the data, factor loadings are equivariant under linear transformations of the data, and so on.

Another inconvenience of the lack of affine equivariance is that the statistical precision of the procedure may depend on the value of the population covariance matrix. The simulation study in Section 4.2 uses a bivariate normal with covariance matrix diag(4, 25) as signal distribution. The relative mean squared errors will not remain the same for other signal distributions, and one needs to perform simulations for several choices of the signal covariance matrix to get a more complete picture. This would not have been necessary when working with an affine-equivariant procedure.

If one is willing to give up the property of affine equivariance, then it becomes fairly easy to construct new robust estimators of the covariance matrix. For example, one could compute the Euclidean distance between every observation and the coordinatewise median of the sample. If such a distance is large, then the observation will receive a lower weight in the calculation of a weighted covariance matrix estimator. Such an estimator will be robust, simple, and fast to compute. Another idea, in the same spirit as the NNVE, is to compute first the Kth nearest Euclidean distances D_{Ki} for $i = 1, \ldots, n$. Let D_- be the smallest and D_+ the largest distance. Then define weights as

$$w_i = \begin{cases} 1 & \text{if } D_{Ki} \le D_- + (D_+ - D_-)/3 \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

and denote the corresponding weighted sample mean and covariance matrix by $\widehat{\Sigma}_{0,NN}$. This estimator is explicitly defined and poses no computational difficulties. In fact, the weights in (1) are used by the authors as starting values for the EM algorithm for computing $\widehat{\Sigma}_{1,NN}$, and one might wonder whether it is not worth studying $\widehat{\Sigma}_{0,NN}$ as an estimator in its own right, describing the shape of the core of the data cloud.

The NNVE Does Not Have the Exact-Fit Property

A covariance matrix estimator that has the exact-fit property can easily detect lower-dimensional data structures. In analogy with the regression context (Rousseeuw and Leroy 1987, p. 120), we say that the data are lying in an exact-fit position if they are concentrated on the same hyperplane. We say that $\widehat{\Sigma}$ has the exact-fit property if it has no maximal rank as soon as more as half of the data are lying in an exact-fit position. The smallest eigenvalue of a covariance matrix estimator with the exact-fit property equals 0 as soon as a majority of the data lies in a lower-dimensional subspace. Robust estimators like MVEE or RMCD do have this property, but NNVE does not. As a counterexample, consider the data represented in Figure 1. All observations except two are lying perfectly on a straight line. The NNVE is giving a correlation coefficient of .87 and is not detecting that the data, aside from the two outliers, are intrinsically one-dimensional. The reason is that the Kth NN Euclidean distances of these two observations are not large. We call these point correlation outliers: measured in Euclidean distance, they are not far away from the main data cloud, but measured in the statistical or Mahalanobis distance, they are far away.

We Do Not Know the Limit Distribution of the NNVE

The limit distribution of the RMCD was shown to be normal (Butler, Davies, and Jhun 1993; Lopuhäa 1999) and its asymptotic variance was obtained by Croux and Haesbroeck (1999). It is then possible to obtain limit results for multivariate analysis procedures based on robust covariance matrix estimators, as was done in the context of principal components (Croux and Haesbroeck 2000), multivariate regression (Rousseeuw, Van Aelst, and Van Driessen 2000),

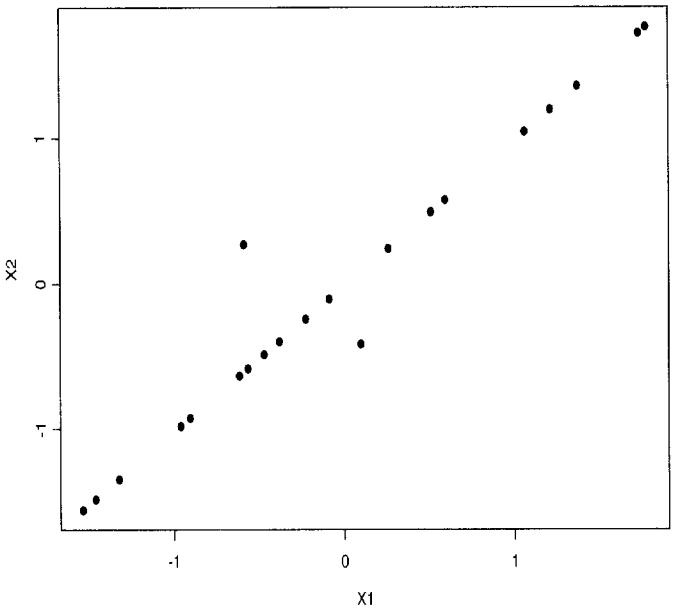

Figure 1. Example of a Bivariate Data Cloud Lying in an Exact-Fit Position.

canonical correlations (Croux and Dehon 2002) and factor analysis (Pison, Rousseeuw, Filzmoser, and Croux 2002). Not knowing the limit distribution of NNVE makes it hard to obtain some theoretical backbone for further application of the NNVE to classical multivariate analysis problems.

To conclude this section, we do agree with Wang and Raftery that NNVE provides a useful option during the data exploration stage. It can highlight features in the data that estimators like MVEE and RMCD could miss. But it has several conceptual weaknesses: no consistency at elliptical distributions, no affine equivariance, no limit distribution available, and an inability to detect correlation outliers.

COMPUTATIONAL ASPECTS AND A SIMULATION EXPERIMENT

The NNVE has been implemented by the authors in S-PLUS code, which they kindly provided to us. The weak point in the algorithm of the NNVE is the convergence of the EM algorithm. Numerical experiments using the default value for $K = 12$ showed that nonconvergence occasionally occurs, mainly in situations with no or little outliers. But numerical problems arose almost systematically for large samples in high dimensions. It seems that it is difficult to go beyond 1,000 observations in 15 dimensions. The computation time of the NNVE is at least comparable to that of the RMCD (using the S-PLUS implementation). The NNVE is therefore more suitable for analyzing small and moderate-size datasets in an interactive way than for serving as a datamining tool.

We found it useful to plot the weights that the observations receive in the NNVE procedure, allowing us to discriminate the signal and the noise points. Figure 2 plots the weights for a dataset generated as in Section 4.2 of the article, in the case where no outliers are present. In (a) the weights of the observations are plotted versus their Mahalanobis distances. It is striking that the weights are not decreasing in the Mahalanobis distance, and that already from the 3rd quartile of the chi-squared distribution with 2 degrees of freedom signal points are getting downweighted. Figure 2(b) plots the weights versus the Kth NN distances, together with an interpolating weight function. Now the weight function is decreasing in D_{Ki}, and descending at a fast rate from the part where the weights equal 1 (signal points) to the part where the weights equal 0 (noise points).

Wang and Raftery only compare the NNVE with the MVEE estimator. Many other robust covariance estimators do exist, however. Maronna and Yohai (1998) have provided an overview. In the last part of this discussion we repeat the simulation study of Section 2 of Wang and Raftery, but now

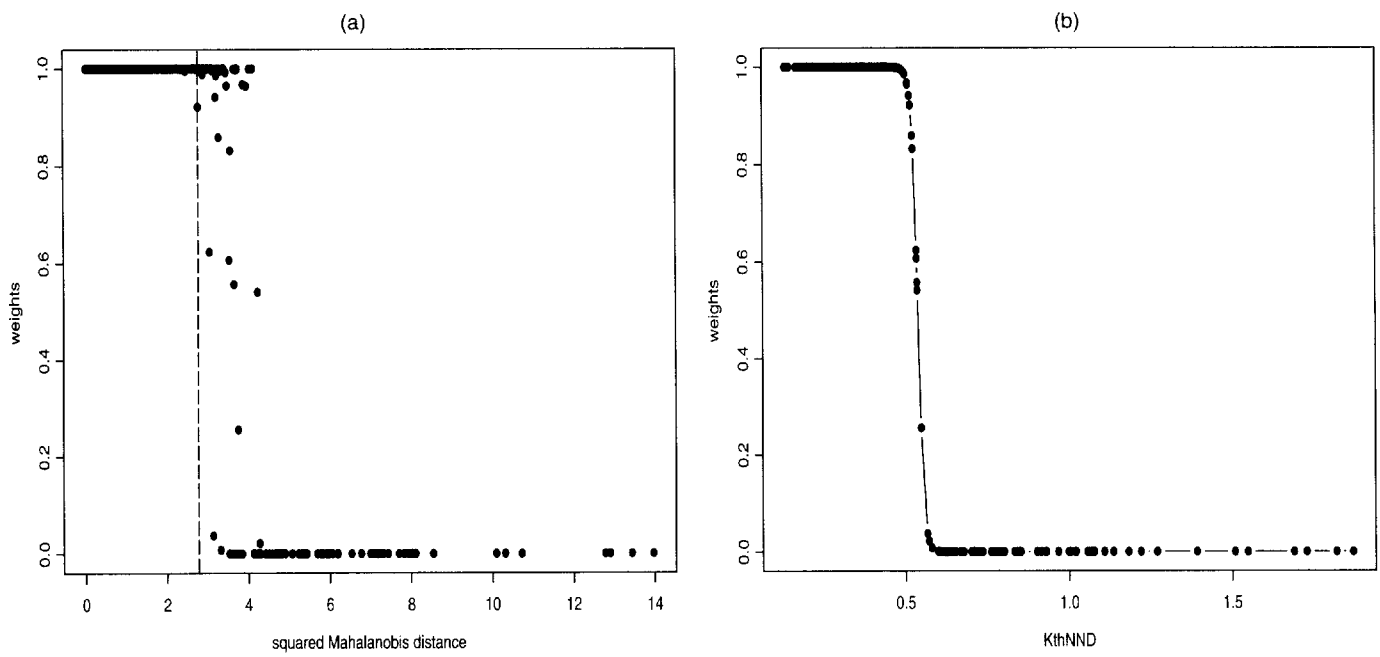

Figure 2. Weights w_i Versus (a) the Squared Mahalanobis Distances, Where the Vertical Line is Indicating the 75% Quantile of a χ^2_2 Distribution, and (b) the Kth NN Distances With the Interpolating Weight Function.

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

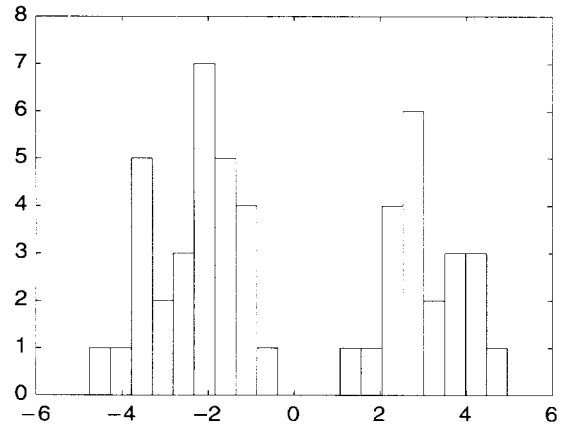

Figure 1. Histogram of First Canonical Variate for 10-Dimensional Simulated Normal Dataset of Size $n = 50$.

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.

Comment

Ehsan S. Soofi and Ali Dadpay

Robustness as a notion has been a topic of discussion among statisticians, econometricians, and engineers in the contexts of various problems. Robust estimation and regression and Bayesian robustness are now parts of the common statistics vocabulary (see, e.g., Portnoy and He 2000). Application and development of robust methods in econometrics are growing rapidly (see, e.g., Maddala and Rao 1997). Robust control theory for engineers is dated back to 1960s (see, e.g., Dorato 1987). Wang and Raftery (WR) break some new ground in robust covariance estimation and improve substantially over the existing methods for certain data conditions.

The notion of noise also appears in various contexts in many fields. Traditionally, noise is viewed as distortion integrated in the magnitude of an observed value of a variable and is modeled as an unobservable and uncontrollable error component of the observation. The objective is to clean the observation and extract the signal part. WR define noise as irrelevant data points. In this formulation some points in the dataset are considered as signal and some other points are considered as noise (outliers), which are judged to be irrelevant. WR's objective is to clean the dataset by separating the signal points from the noise points and then estimate the covariance of the signal distribution. The NNVE uses the NNC procedure developed by Byers and Raftery (1998) to clean the data and then estimate the covariance matrix of the signal distribution.

We comment on two issues. First, we briefly comment on WR's elaboration of an "entropy-type" plot proposed by Byers and Raftery (1998) for selecting K for the NNC procedure. We then focus on some general issues regarding covariance estimation, including the dimension of the data, criteria for comparing and evaluating covariance estimators, and the impact of various covariance estimates on some multivariate techniques. We construct an example to illustrate these issues. This example shows that comparison of the multivariate quantities can provide insight about the geometry of the data trimming in the higher dimension as well.

1. ENTROPY PLOT

Byers and Raftery (1998) proposed an entropy-type plot for selection of K in the NN classification. In Remark 2, WR elaborate on the underpinning logic of the entropy-type graph. Here the negative entropy of the signal-outlier classification is defined as $\mathcal{E}(K) = \sum h_i \log(h_i)$, where $h_i = h_i(d, \theta)$, $\theta = (\tau, \lambda_1, \lambda_2)$ is shown in eq. (2) of WR. However, $\mathcal{E}(K)$ is not negative entropy of the distribution of the signal-outlier classification, because $\boldsymbol{h} = (h_1, \ldots, h_n)$ is not a single probability vector, $\sum_{i=1}^{n} h_i \neq 1$. Each $h_i = (h_i, 1 - h_i)$ is a probability vector, and $h_i = h(Z_i)$, $i = 1, \ldots, n$ gives the probability distributions of the signal-outlier classification indicators Z_i, $i = 1, \ldots, n$.

The entropy of Z_i is $H(h_i) = -h_i \log(h_i) - (1 - h_i)\log(1 - h_i)$, which quantifies uncertainty about the signal-noise classification of the ith datum. The overall uncertainty about the signal-noise classification is given by the joint entropy

$$\begin{aligned} H(\boldsymbol{h}) &= H(Z_1, \ldots, Z_n) \\ &= -\sum H(h_i) \\ &= -\sum [h_i \log(h_i) + (1 - h_i)\log(1 - h_i)]. \end{aligned}$$

The summation is implied by the independence of Z_i's. We note that if $h_i \approx .5$ for all $i = 1, \ldots, n$, then $2\mathcal{E}(K) \approx -H(\boldsymbol{h})$. Otherwise, $-2\mathcal{E}(K) - H(\boldsymbol{h}) \neq 0$, and the difference can be substantial when many h_i's are far from .5.

A useful information index for this problem is the normalized information index $I(\boldsymbol{h}) = 1 - H(\boldsymbol{h})/n \log 2$. This measure provides information about the signal-noise classification in terms of the fraction of the uncertainty reduction from the uniform prior, that is, a 50%–50% chance. For a general prior $\boldsymbol{p} = (p, 1 - p)$ for Z_i, one may use the entropy difference index $I_p(\boldsymbol{p}) = 1 - H(\boldsymbol{h})/nH(\boldsymbol{p})$. The entropy difference index $I_p(\boldsymbol{p})$ may be positive when the data are informative or negative when the data provide a surprise (Lindley 1956). Alternatively, one may use an information index based on the relative entropy $K(\boldsymbol{h} : \boldsymbol{p}) = \sum \{h_i \log(h_i/p) + (1 - h_i)\log[(1 - h_i)/(1 - p)]\}$, which is nonnegative (see Soofi 1994 for details).

2. COMPARISON OF COVARIANCE MATRICES

WR compare the NNVE procedure with the standard estimator and the MVE in two simulation experiments. The comparison is elementwise and is limited to the MSE criteria. Comparison of covariance matrices based on bias (mean difference) and MSE of individual elements is too simplistic.

In some applications, the objective may be estimation of the covariance matrix. However, often the objective of an analysis is not covariance estimation per se, but covariance is estimated for the purpose of discovering patterns of relationships among a set of variables. Covariance matrix estimation is just a first step in the analysis. Regression and many multivariate statistical techniques require estimation of the covariance matrix as a first step. The items of interest are various functions of the covariance matrix, which in general are nonlinear functions of the elements of the matrix. Therefore, examining performance of covariance estimators according to criteria that are

Ehsan S. Soofi is Professor of Business Statistics, School of Business Administration, University of Wisconsin-Milwaukee, Milwaukee, WI 53201 (E-mail: *esoofi@uwm.edu*), and Research Associate, Center for Research on International Economics. Ali Dadpay is a doctoral student, Department of Economics, University of Wisconsin-Milwaukee, Milwaukee, WI 53201 (E-mail: *adadpay@uwm.edu*).

Journal of the American Statistical Association
December 2002, Vol. 97, No. 460, Theory and Methods
DOI 10.1198/016214502388618816

Table 1. Two Estimates of the Variable Correlations and the Correlations Between Least Squares Coefficients

	Estimate A					Estimate B				
Correlations of variables										
	x_1	x_2	x_3	x_4	x_5	x_1	x_2	x_3	x_4	x_5
x_1	1.00					1.00				
x_2	.19	1.00				.16	1.00			
x_3	−.49	.30	1.00			−.55	.23	1.00		
x_4	−.36	.71	.55	1.00		−.42	.68	.56	1.00	
x_5	.18	.30	.25	.32	1.00	−.20	.24	.20	.31	1.00
Correlations of least squares coefficients of predictors $x_1, \ldots, x_5$										
	b_1	b_2	b_3	b_4	b_5	b_1	b_2	b_3	b_4	b_5
b_1	1.00					1.00				
b_2	−.69	1.00				−.65	1.00			
b_3	.45	−.20	1.00			.35	−.06	1.00		
b_4	.66	−.82	−.03	1.00		.58	−.79	−.21	1.00	
b_5	−.43	.22	−.29	−.34	1.00	.14	−.13	.00	−.03	1.00

additive in discrepancies between the elements and comparing results of various covariance estimators elementwise may not suffice. Loss functions and error indices that include inverse and determinant of the matrix are commonly used in covariance modeling. Such criteria are also needed for evaluating the overall performance of a covariance estimation procedure.

The simulation results reported by WR are convincing on the better elementwise MSE performance of NNVE compared with MVE (for the admissible <50% outliers cases) for two-dimensional data where the covariance matrix of the outliers is a 10-fold multiple of the covariance matrix of the mass of data. This represents a very simple scenario. Detecting outliers and clusters in two dimensions based on scatterplots is not difficult; sophisticated methodologies are useful when simple devices fail to do the job.

3. ILLUSTRATIVE EXAMPLE

WR compare the results of three estimation methods by estimating the covariance matrices in a few examples. The Australian Athletes data is the only example that includes more than two variables. We compared some regression and principal components quantities of the correlation matrices of the three estimates, standard, MVE, and NNVE. We found that for this dataset, the three methods produced very similar multivariate results.

For the purpose of illustration, we continued as follows. We reversed the sign of the standard estimate of the covariance between FERR and BFAT and computed the regression and principal components quantities. We found some interesting results that illustrate the general issues pointed out earlier. We note that FERR and BFAT have the weakest correlation in the set (.18), and thus a change of sign for such a weak correlation by a robust estimation should not be too surprising. In this example we report on the comparison of some regression and principal components quantities computed from the modified standard correlation matrix and the correlation matrix of the NNVE. Therefore, in the scenario of this example, we have two hypothetical estimates, "estimate A" and "estimate B," of a covariance matrix $\Sigma = [\sigma_{ij}]$ of a five-dimensional variable $\boldsymbol{x} = (x_1, x_2, x_3, x_4, x_5)$.

The correlation matrices of the hypothetical estimates A and B are shown in top part of Table 1. Except for the sign of $r(x_1, x_5)$, the correlations under "estimate A" are the same as the standard correlations of the real data. All the correlations under "estimate B" are the same as the NNVE of the real data. We note that, other than $r(x_1, x_5)$, the correlations differ slightly in the two sets.

As a first step, we compare the overall discrepancy between the two estimates of Σ using two well-known measures,

$$RMSR(\Sigma_B, \Sigma_A) = \sqrt{\frac{2}{p(p+1)} \sum_{j=1}^{p} \sum_{i \le j} (\sigma_{B_{ij}} - \sigma_{Aij})^2}$$

$$K(\Sigma_B : \Sigma_A) = \frac{1}{2}\left[Tr\left(\Sigma_B \Sigma_A^{-1}\right) - \log\left|\Sigma_B \Sigma_A^{-1}\right| - p\right],$$

Table 2. Coefficients and R^2 of Regression of Each Variable on the Other Four Variables

	Estimate A					Estimate B				
	y_1	y_2	y_3	y_4	y_5	y_1	y_2	y_3	y_4	y_5
Predictor										
x_1		.60	−.57	−.53	.64		.56	−.43	−.46	−.18
x_2	.80		.29	.75	−.37	.79		.08	.71	.20
x_3	−.35	.13		.02	.33	−.29	.04		.14	−.05
x_4	−.83	.90	.05		.63	−.74	.87	.33		.13
x_5	.29	−.13	.25	.18		−.09	.08	−.00	.02	
R^2	.68	.76	.46	.80	.29	.61	.71	.44	.76	.11

Table 3. Principal Component Results of Two Estimates of a Correlation Matrix

	Estimate A							Estimate B						
	Correlation with PCs					Cumulative		Correlation with PCs					Cumulative	
	v_1	v_2	v_3	v_4	v_5	R_1^2	R_2^2	v_1	v_2	v_3	v_4	v_5	R_1^2	R_2^2
x_1	−.37	.88	−.06	.26	.14	.14	.91	−.59	.73	−.01	.31	.15	.35	.88
x_2	.72	.50	−.44	.06	−.18	.52	.77	.59	.76	−.17	−.03	−.21	.35	.93
x_3	.77	−.37	.22	.47	.03	.59	.73	.78	−.33	−.22	.49	.00	.61	.72
x_4	.92	−.00	−.23	−.25	.20	.85	.85	.90	.20	−.18	−.24	.23	.81	.86
x_5	.48	.52	.69	−.17	−.05	.23	.50	.50	.11	.85	.07	.01	.25	.27
λ_i	2.33	1.42	.77	.38	.10	2.33	3.75	2.38	1.27	.84	.39	.12	2.38	3.65

where $Tr(\cdot)$ and $|\cdot|$ denote the trace and determinant and $p = 5$ is the dimension of the matrix. The root mean square residual (RMSR) corresponds to the quadratic loss and is used as the unweighted least squares fit function in covariance modeling. The second measure is the Kullback–Liebler information between two multivariate normal distributions with equal means. This measure is commonly used as the maximum likelihood fit function in covariance modeling. It also has been used as the entropy loss for covariance estimation by Haff (1980) and others. The first measure provides an elementwise comparison, but the second measure includes inverse and determinant, which are important functions of the covariance matrix used in regression and multivariate techniques.

For the correlation matrices shown in Table 1, $RMSR = .105$ and $K = .471$. To appreciate the magnitudes of these indices, compare them with the discrepancy between correlation matrix of NNVE and the original standard correlation matrix, $RMSR = .037$ and $K = .049$. Note that with modification of only one element, the Kullback–Liebler measure shows nearly a 10-fold increase, much higher than the $2\frac{1}{2}$-fold RMSR increase. This comparison suggests that a small change in the elements of a covariance matrix may have more serious multivariate consequences than a measure of elementwise differences can capture.

The bottom part of Table 1 compares the correlations of the least squares estimates when $x_1, \ldots, x_5$ are the standardized explanatory variables in a problem. In such problems, the variance of least squares is $\text{var}(\boldsymbol{b}) = R_X^{-1}\sigma^2$ and the correlation matrix is given by $R_b = DR_X^{-1}D$, where $D = diag[1/d_1, \ldots, 1/d_p]$, d_i^2 being the diagonal elements of R_X^{-1} generally referred to as variance inflation factors. Table 1 shows the two correlation matrices R_b. We note that the discrepancies between the elements of the two sets of correlations for the least squares estimates are more pronounced than the correlation matrices of the variables. The RMSR for comparing the elements of the two R_b's is .216, about twice of the RMSR found for the case of R_X.

Table 2 shows the least squares coefficients and R^2 of regression of each variable on the other four variables. For each regression, the column shows the dependent variable $x_i = y_i$ and the rows show the predictor variables x_j, $j \neq i$, $j = 1, \ldots, 5$. The coefficients are computed by $\boldsymbol{b}_i = R_{(-i)}^{-1}\boldsymbol{r}_i$, where $R_{(-i)}$ is the correlation matrix of the variables excluding $x_i = y_i$ and $\boldsymbol{r}_i$ is the vector containing the correlations between $x_i = y_i$ and the other four variables. Note that not only do the two covariances lead to different regression estimates for y_1 and y_5 whose correlation was modified, but also the results are quite different for other regressions. For example, in the regression of y_3, the ratio of the coefficients of x_2 and x_4 changes from about 6/1 to less than 1/4. The regression of y_5 is the most revealing about the data trimming. The reduction of R^2 from .29 to .11 reveals that the cluster of data points trimmed by estimate B is close to the hyperplane of the regression of x_5 on the other four variables given by estimate A.

Table 3 compares the results pertaining to the principal components analysis based on the two correlation matrices. The table shows that results are more discrepant here than when the two correlation matrices were compared. In the case of the first component, we see that $r(x_1, v_1)$ is much weaker and $r(x_2, v_1)$ stronger for estimate A than for estimate B. The other four components are substantially different for the two correlation matrices, both in terms of the strength of the correlation with the variables and the directions of the principle components. The cumulative R_i^2 for the individual variables are also different for the first two components. The first two components have variances (eigenvalues λ_i) >1, which are usually retained in the principle components analysis.

This hypothetical example illustrated some important points. Evaluation and comparison of covariance estimates based on criteria that involve important functions of a matrix, such as the inverse, can reveal useful information about their use in multivariate analysis. However, the elementwise MSE criteria falls short in this respect. Examining impacts of covariance estimation on some subsequent multivariate analyses is needed for appreciating the scope of application of the estimation procedure. The rich geometry of multivariate statistics can be revealing about the geometry of data trimming in higher dimensions. A multivariate investigation is needed to determine the scope of applicability of NNVE.

ADDITIONAL REFERENCES

Dorato, P. (1987), *Robust Control*, New York: IEEE Press.

Haff, L. R. (1980), "Empirical Bayes Estimation of the Multivariate Normal Covariance Matrix," *The Annals of Statistics*, 8, 586–597.

Lindley, D. V. (1956), "On a Measure of Information Provided by an Experiment," *The Annals of Mathematical Statistics*, 27, 986–1005.

Maddala, G. S., and Rao, C. R. (1997), *Robust Inference*, New York: Elsevier.

Portnoy, S., and He, X. (2000), "A Robust Journey in the New Millennium," *Journal of the American Statistical Association*, 95, 1331–1335.

Soofi, E. S. (1994), "Capturing the Intangible Concept of Information," *Journal of the American Statistical Association*, 89, 1243–1254.

Comment

Douglas G. SIMPSON

The authors present an intriguing new approach to robust covariance estimation. By introducing Kth NN weighting of the elements of the covariance matrix estimate, they build local weighting into estimation process. Intuitively and mathematically they show that outliers are down weighted in the procedure because of the large distances to their "neighbors." The simulation studies and examples demonstrate a level of robustness to clusters of outliers that compares well with the minimum volume ellipsoid estimator, a well-known affine-equivariant estimator with a high breakdown point.

Although the proposed estimator is not affine equivariant, it is scale and orthogonal equivariant. This type of equivariance is useful for principal component analysis, especially in highly multivariate settings where the dimension of the data vectors exceeds the sample size. (For one such example see Locantore et al. 1999 and discussion, in which a "spherical principal component analysis" is used for robust functional data analysis.) In principal component analysis, the bias in the estimate of the magnitude of the covariance is of less importance than capturing the directions of greatest variation. In such applications one may avoid the need for the data augmentation proposed here to reduce bias in the magnitude of the covariance estimate when all the data are "good." A simplified version of the NN analysis may prove useful for highly multivariate problems.

The authors have done a thorough job of identifying practical issues that arise in implementing the NN idea and in developing workable solutions. I suspect that there is substantial scope for further work on generalizing and refining the approach. For example, the derivation of (1) and (2) depends on a mixture point process model that appears to be inessential to the approach; rather, it is the mathematical properties of the resulting Kth NN weights in (2) and (6) that make the procedure work. By focusing on properties of the estimator, it may be possible to develop a broad class of weighting functions, h, that provide good properties without reference to a spatial mixture model. The data augmentation involves some ad hoc choices and statistical complexity (in terms of analyzing the properties of the procedure), which might be sidestepped by introducing a theoretical adjustment factor based on a target model such as the multivariate normal distribution. An example is the MAD estimator, median $|\mathrm{Xi} - \mathrm{median}(\mathrm{Xi})|/\{2\Phi(1) - 1\}$, where Φ denotes the standard normal distribution function and where the scale factor is used to make this consistent as an estimator of the standard deviation if the data come from a normal distribution.

I congratulate the authors on their novel approach to covariance estimation, and I look forward to seeing the new possibilities that this line of research opens.

ADDITIONAL REFERENCE

Locantore, N., Marron, J. S., Simpson, D. G., Tripoli, N., Zhang, J. T., and Cohen, K. L. (1999), "Robust Principal Component Analysis for Functional Data" (with discussion), *Test*, 8, 1–73.

Douglas G. Simpson is Professor, Department of Statistics, University of Illinois at Urbana-Champaign, Champaign, IL 61820 (E-mail: *dgs@uiuc.edu*).

Journal of the American Statistical Association
December 2002, Vol. 97, No. 460, Theory and Methods
DOI 10.1198/016214502388618825

Rejoinder

Naisyin WANG and Adrian E. RAFTERY

We are very grateful to the seven discussants for their thoughtful comments, which raise important issues and suggest many possible advances and areas for further research. We particularly endorse Simpson's comment that there is substantial scope for further work generalizing and refining the approach in our article. We hope that the article and the discussions will stimulate this research. We now address some of the issues raised.

1. COMPARISON ESTIMATOR

Croux and Van Aelst reiterate our conclusion that MVE is a poor estimator for nonelliptical distributions, and compare NNVE instead with the RMCD estimator. Their simulation study (Table 1) is very striking. As they point out, NNVE outperforms RMCD in almost all situations, by a factor of up to 8 in MSE.

The one exception is the case with 67% outliers, in which the RMCD with $\alpha = .75$ performs better than NNVE. Even there, it is worth noting that this is at the cost of quite poor performance when there are few outliers. The delicate task is to achieve good performance both when there are many outliers and when there are few, and this is what NNVE tries to achieve. The standard choice of α for RMCD is .5, and we suspect that with this choice the performance with 67% outliers would be quite poor (as is the case for MVE). However, it seems likely that when there are fewer than 50% outliers, RMCD with $\alpha = .5$ would be more efficient than RMCD with $\alpha = .75$. It seems doubtful that there is any value of α for which RMCD would work well both when there are few outliers and when there are many outliers. How to select α for RMCD in practice, when the proportion of outliers is unknown, is an issue yet to be addressed.

MCD is available in S-PLUS, but the currently available software that we were able to find does not allow an α larger than .5. In practice, users of NNVE would select K using the data (e.g., via the entropy-based plot we discussed in our article) and would also check for small outlying clusters; see Section 6 of this rejoinder. Given the time constraints for writing this rejoinder, we were not able to conduct a comparison between RMCD and a version of NNVE that does incorporate these elements, such as the one implemented in Section 6. However, the improvements observed in the simulation study reported in Section 6 lead us to suspect that NNVE could uniformly outperform RMCD with $\alpha = .75$.

2. AFFINE EQUIVARIANCE

NNVE is not affine equivariant. Croux and Van Aelst ask whether we need affine equivariance, and they answer in the affirmative. They point out some advantages of affine equivariance. Simpson points out that NNVE is scale and orthogonal equivariant, and so, for example, it is invariant under scale changes in the variables.

Affine equivariance has costs as well as benefits. For example, requiring affine equivariance excludes methods that shrink toward independence, such as Bayesian estimation with a prior centered on independence and (for regression) ridge regression. This seems unfortunate, because there is considerable evidence that shrinkage of this kind improves performance, at least in regression (e.g., Dempster, Schatzoff, and Wermuth 1977).

And, of course, as Croux and Van Aelst point out, requiring affine equivariance makes it harder to find good robust covariance estimators. They suggest several promising non–affine-equivariant estimators that could be alternatives to or refinements of NNVE, and we look forward to further investigation of these.

Affine equivariance does seem desirable when the variables are measurements of the same quantity and are on the same scale, and when linear combinations are actually used. This can arise in, for example, psychometrics when the variables may be scores on similar tests or questions and composite scores are formed by taking sums. But in many applications, it seems of doubtful relevance. For example, in the two real datasets in our article, the variables are measurements of quite different things, and it would not seem to make much sense to take linear combinations of them. In the Australian Athletes data, for example, two of the variables are BFAT and RCC, and taking a linear combination of these (or of any other variables in this dataset) would seem to make little scientific sense. Thus affine equivariance seems of doubtful relevance in this application, especially when weighed against the costs that it carries.

Having said all this, it is possible to modify NNVE slightly so as to achieve affine equivariance, as follows. First compute an affine-equivariant covariance estimator, such as MVE (even though this by itself often would not give good results when there is a high proportion of outliers). Next, sphere the data using the resulting covariance estimate. Then compute NNVE, and finally transform the NNVE covariance estimate back to the original scale. The resulting estimator would be affine equivariant and might inherit the good properties of NNVE. However, whether it would perform as well as the original NNVE, particularly in nonelliptical situations, requires further investigation.

3. THE EXACT-FIT PROPERTY

Croux and Van Aelst point out that NNVE does not have the exact-fit property. One consequence of the exact-fit property is that the smallest eigenvalue of a covariance matrix equals 0 as soon as a majority of the data lie in a lower-dimensional subspace. Croux and Van Aelst illustrate this with their Figure 1,

Journal of the American Statistical Association
December 2002, Vol. 97, No. 460, Theory and Methods
DOI 10.1198/016214502388618834

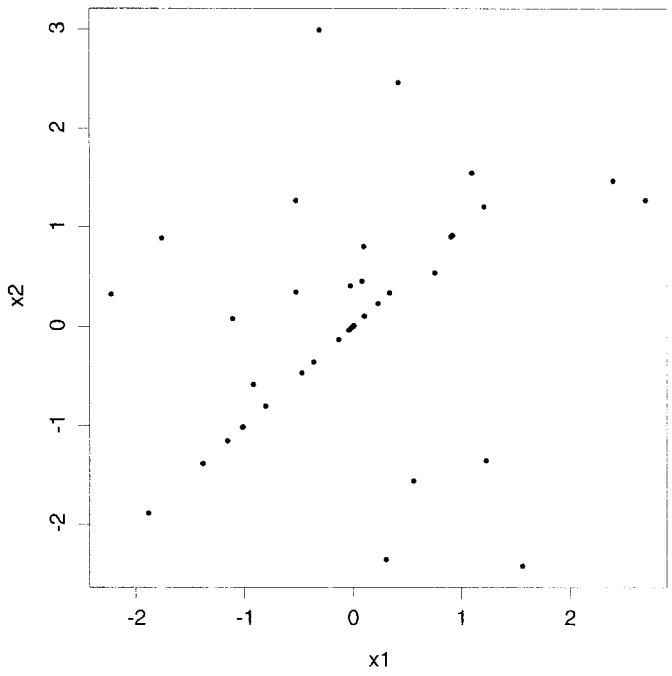

Figure 1. Another Example of a Bivariate Data Cloud Lying in an Exact-Fit Position.

suggesting that a good robust estimator of the correlation in this plot should be exactly equal to 1, whereas the NNVE estimator is .87.

It is not clear to us that the exact-fit property is desirable when there is no prior knowledge that the "clean" data lie in a lower-dimensional subspace. Consider Figure 1. The data there lie in an exact fit position, as do the data in Croux and Van Aelst's Figure 1, and for any covariance estimator with the exact fit property, the estimated correlation will be exactly 1. However, it is far from clear that this is a good thing. For example, this would imply that one could predict X_2 exactly from X_1 in future data, which seems unlikely. The standard correlation is .26. A good summary of the data would include the fact that many of the data do lie on a line, but would also mention that many do not.

The NNVE estimate of the correlation is .88 (using $K = 5$, which is suggested by the entropy-based plot). Outliers far away from the main data feature were downweighted, and so the NNVE estimate of correlation is much higher than the standard estimate. However, the NNVE estimate of correlation does reflect the fact that the relationship between X_1 and X_2 is not known exactly, and this seems desirable, at least in the absence of external information to the contrary.

4. CONSISTENCY

Croux and Van Aelst mention that NNVE lacks consistency at elliptical distributions, which MVE and most other affine-equivariant estimators possess. What was not made clear in their discussion is that for estimators such as MVE to be consistent, the distribution family of the data (e.g., multivariate normality) must be known. Different elliptical distributions of the data can have different characteristic functions after being sphered (Cambanis, Huang, and Simons 1981). If the distribution family is known, then a consistency factor for MVE can then be calculated to ensure consistency for that family (e.g., Rousseeuw and Leroy 1987, p. 260).

On the other hand, NNVE requires a separate set of assumptions to achieve consistency. These come from a spatial mixture model; see Corollary 1 in Section 3 of our article. Obviously, these assumptions are different from the traditional requirements. However, Simpson points out that this could be overcome by introducing a theoretical adjustment factor based on a target model such as the multivariate normal distribution. This seems like a worthwhile refinement of NNVE.

Nonetheless, it is not clear that this is of major practical importance. MVE and other affine-equivariant estimators lack consistency at nonelliptical distributions. So consistency at a particular family of elliptical distributions seems relevant only when we are sure that the underlying distribution really belongs to that family, which will rarely be the case. Indeed, part of the rationale for robustness in the first place is to achieve good performance when the true underlying distribution is not known precisely. Also, in our simulations, NNVE was no more biased than MVE for the multivariate normal distribution (the 0% panel of Table 2), and was substantially less biased for the nonelliptical distribution we considered (the 0% panel of Table 3).

5. COMPUTATIONAL ASPECTS

Croux and Van Aelst point out that the S-PLUS code that we provided often has numerical problems for large samples in high dimensions. This is not too surprising, because this is research code, and S-PLUS tends to be slow. Production code for this method would likely be written in a faster, lower-level language, such as C or Fortran. We would expect a dramatic improvement in computational efficiency if this were done.

Also, there has been considerable work recently on efficient computational methods for implementing the EM algorithm and similar approaches in very large datasets, mainly in the data-mining community (see, e.g., Bradley, Fayyad, and Reina 1998, 1999). We would expect that the use of ideas such as these will lead to much more efficient implementations of NNVE in very large datasets.

6. COMPARISON WITH MIXTURE MODELING

McLachlan and Hamaty carry out a very interesting comparison between NNVE and mixture modeling with normal and t components. For the Hertzspring–Russell data, with 6 outliers out of 47 data points, NNVE and mixture modeling give similar results. For the Australian Athletes data, the mixture analysis uncovers structure in the data that goes beyond the mere estimator of an overall covariance matrix. We agree that such analyses are very useful; covariance estimation could be viewed as a preliminary step before such more sophisticated analyses.

Our experience, however, is that when there are many outliers, direct mixture modeling often breaks down, and a preliminary application of NN cleaning can be very useful, even when mixture modeling is the ultimate goal. (For an example of this, see sec. 5.3 of Fraley and Raftery 2002.) We suspect that a direct application of McLachlan and Hamaty's form of mixture modeling would also run into difficulties in our minefield example.

Table 1. Monte Carlo Averages, Standard Errors, MSEs and Relative MSEs of Estimated Covariance for 67% Outlier Contamination in the Bivariate Normal Scenario

	NNBR			*NNVE-Naive*			*NNVE-Select K*			*NNVE-K/MCLUST*		
Mean	4.06	25.71	−.04	4.33	27.40	−.06	3.91	24.42	.01	3.82	23.92	.03
SE	3.70	25.24	1.78	4.88	31.56	2.30	.80	3.84	1.08	.66	4.02	.87
MSE	13.69	636.25	3.16	23.88	999.60	5.30	.65	15.02	1.16	.46	17.31	.75
%MSE	.02	.02	.28	.01	.01	.17	.52	.79	.76	.74	.68	1.17

NOTE: Each dataset has 500 observations. Each relative MSE were calculated by dividing the MSE of normal mixture, as given in Table 2 of McLachlan and Hamaty, by that of NNVE. The reported results are based on 500 simulations.

McLachlan and Hamaty present a simulation study comparing NNVE with mixture modeling when the correct mixture model is assumed known and is used. Of course, one would expect the mixture model to outperform any other method in this situation, but the study is still of interest nonetheless. NNVE performed well, perhaps surprisingly so, when there are 33% and 50% outliers.

However, when there were 67% outliers, NNVE had extremely low relative efficiency in McLachlan and Hamaty's study, on the order of 1%. The large MSEs of NNVE in this case were caused mainly by one or two of the 500 simulated datasets whose results dominated. However, this is still a cause of concern, and we looked into it in more detail.

In practice, there are two things that NNVE users would do that are not reflected in our simulation study. One of these is to use the entropy-based plot to choose K; the other is to check at the end whether there are small clusters of outliers left, using MCLUST or other methods, as in our minefield application, for example. To incorporate these aspects in the automatic procedure in a simple way, we implemented a very simple selection of K using the entropy-based criterion of Byers and Raftery (1998), hereafter referred to as BR. We selected K among $K = 8$, 12, and 16. The rule that we used was to use the smallest K unless any of the K's above it has an absolute value of $\mathcal{E}(K)$ 70% or less of the current absolute value of $\mathcal{E}(K)$. Of our 500 simulated datasets, 284 used $K = 8$, 92 used $K = 12$, and 124 used $K = 16$. In practice, users would perhaps use the entropy-based K plot to perform a more sophisticated selection than the simple automatic selection that we have implemented here.

The estimator with this added K selection feature was denoted by NNVE-Select K. The estimator with K selected and with the second MCLUST-based phase of outlier removal was denoted by NNVE-K/MCLUST. The simulation results using the same scenario as in the 67% panel of Table 2 in our article are given in Table 1.

After incorporating these refinements, that represent more closely what users would actually do, we find that the efficiency of NNVE relative to the normal mixture model has greatly improved from what McLachlan and Hamaty obtained. It is at least 68%, compared with the 1% for the naive NNVE as reported by McLachlan and Hamaty.

7. EVALUATION CRITERIA

Soofi and Dadpay point out that the entropy-like measure, $\mathcal{E}(K)$, that we adopted from BR as the basis of a way of choosing K, is not, strictly speaking, itself an entropy. They are right about this. In fact, $\mathcal{E}(K)$ measures a sum of partial negative entropies and is simply an "entropy-based" quantity. However, we have found $\mathcal{E}(K)$ to be useful for choosing K. Soofi and Dadpay suggest using the joint entropy, $H(\mathbf{h})$, or, equivalently in term of choosing K, the normalized information $I(\mathbf{h})$ instead. These may also work well, and this could be a topic of further research.

Soofi and Dadpay also point out that elementwise comparison of estimated and true covariance matrices is only one way of assessing the quality of covariance matrix estimators, and that other error measures may be useful in the context of particular tasks in multivariate analysis. We agree fully, and look forward to the development and reporting of such error measures in future research.

8. OTHER ISSUES

McLachlan and Hamaty suggest carrying out a bootstrap test of the distribution of NN distances to see if it has two components, and proceeding with NNC only if the test suggests that it does. This is a potentially interesting way of reducing or eliminating the bias in the original NNC/NN-BR method of Byers and Raftery (1998) when there are no outliers. It could be viewed as an alternative to our approach of introducing artificial outlying data. It would be of interest to compare these two approaches in a systematic way.

Croux and Van Aelst point out that the limit distribution of NNVE is not yet available. This is true, but not too surprising since NNVE has just been introduced. Limit theory for other robust estimators has tended to lag the introduction of the estimator itself, as Croux and Van Aelst point out, for example, for the RMCD. This is one of the issues that should be addressed in the future research that Simpson calls for.

In Figure 2 of Croux and Van Aelst, they note that when there are no outliers, a good proportion of signal points could still be downweighted. In our experience, when there are no outliers, the original NN-BR has a misclassification rate of classifying signal as noise ranging from 5% to 25%. This rate is reduced in NNVE, as shown by the numerical comparison between NN-BR and NNVE in the 0% panel of Table 2. Further, from the same panel we see that the downward bias of NNVE is no more severe than that of MVE, a consistent estimator under this scenario. We suspect that this slight downward bias of NNVE is not unique among robust covariance estimators, at least in finite samples, as a result of guarding against outliers.

In the same figure, Croux and Van Aelst also note that the weights obtained by NNVE are not decreasing in the Mahalanobis distance. This observation is somewhat expected. The

weights of NNVE are obtained based on local features (the Kth NN distances) rather than on the global distances from the data points to the center of the data. This "local" nature of NNVE allows it to be less sensitive to deviations from elliptical symmetry than methods such as MVE.

On the other hand, even though the NNVE weights are based on local quantities completely different in concept from global distances such as the Mahalanobis distance, Figure 2 of Croux and Van Aelst nicely shows that the (local) NNVE weights nevertheless capture the major features represented by such global distances. So far, few robust methods based on local properties have been proposed. Our good results with NNVE suggest that local robust methods in general may be promising, and we hope that there will be more robustness research from a local perspective in the future.

The S-PLUS code for implementing NNVE is available at the first author's website, *http://stat.tamu.edu/~nwang*.

ADDITIONAL REFERENCES

Bradley, P. S., Fayyad, U. M., and Reina, C. A. (1998), "Scaling EM (Expectation-Maximization) Clustering to Large Databases," Technical Report 98-35, Microsoft Research.

——— (1999), "Mixture Model Estimation Using EM Over Large Databases," in *Proceedings of the Statistical Computing Section, American Statistical Association*, pp. 10–16.

Cambanis, S., Huang, S., and Simons, G. (1981), "On the Theory of Elliptically Contoured Distributions," *Journal of Multivariate Analysis*, 81, 368–385.

Dempster, A. P., Schatzoff, M., and Wermuth, N. (1977), "A Simulation Study of Alternatives to Ordinary Least Squares," *Journal of American Statistical Association*, 72, 77–91.

Fraley, C., and Raftery, A. E. (2002), "Model-Based Clustering, Discriminant Analysis, and Density Estimation," *Journal of the American Statistical Association*, 97, 611–631.