Conjecture on Minimum Covariance Determinant (MCD) Location and Scatter

Maman A. Djauhari

Department of Mathematics
Faculty of Sciences
Universiti Teknologi Malaysia

maman@utm.my

Abstract

We review Rousseeuw and van Driessen's basic theorem of Fast MCD algorithm. Later on, we present a conjecture that if minimum covariance determinant in Fast MCD is replaced by minimum vector variance (MVV), then the results are not changed. The importance of this conjecture lies in the computational efficiency of the algorithm where a multilinear form is replaced by a quadratic form.

Introduction

On December 2005, in his public lecture at the University Malaya, Kuala Lumpur, Calyampudhi R. Rao pointed out that "The current statistical methodology based on probabilistic models applied on small data sets appears to be inadequate to meet the needs of the society in terms of quick processing of data and making the information available for practical purposes." This statement reflects a very fundamental problem in modern society, namely, quick data processing. It is in the spirit to respond to this problem in multivariate scheme that this paper is presented.

The most important step in multivariate analysis is to produce highly robust estimates of location and scatter. A well accepted and widely used methodology that has received considerable attention in literature is the so-called fast minimum covariance determinant (Fast MCD) introduced by Rousseeuw and van Driessen [1]. It is affine-equivariant and has high breakdown point and bounded influence function. The original version of Fast MCD appeared more than two decades ago when Rousseeuw in 1985 [2] introduced the methodology called minimum covariance determinant (MCD). In-depth discussions on MCD can be found, for example, in [1], [3], [4], [5], [6], [7], [8] and [9]. However, the computational efficiency of MCD was unsatisfactory. It needs fourteen years before Rousseeuw and van Driessen [1] introduce Fast MCD; a fast version of MCD.

Fast MCD becomes more and more popular after the work of Hubert *et al.* in [9] who improve its performance in order to give a closer solution to the minimum global, and Hubert *et al.* in [10] who comprehensively show its role in robust multivariate methods. However, it is not without limitation. Its computational efficiency is very challenging for improvement. The use of Mahalanobis distance and covariance determinant in Fast MCD are not apt when the data sets are of high dimension because it becomes computationally inefficient. On the other hand, computational efficiency is as important as effectiveness (see [11]).
This paper presents a conjecture that if the objective function in Fast MCD algorithm is substituted by minimum vector variance, then the MCD location and scatter are not changed. This conjecture is justified by numerous simulation experiments and is developed based on the fact that vector variance can also be used as a measure of multivariate dispersion (see [12]). Furthermore, the used of minimum vector variance to replace minimum covariance determinant as the objective function in Fast MCD algorithm is advantageous (see [13]).

The rest of the paper is organized as follows. In Section 2, Rousseeuw and van Driessen's basic theorem of Fast MCD algorithm will be reviewed and discussed. Later on, in Section 3, we recall vector variance as a complementary measure of dispersion to covariance determinant. The use of both measures will provide more information about the covariance structure than if we use a single measure. Section 4 presents the conjecture that the MCD location and scatter will not change if minimum covariance determinant in Fast MCD algorithm is substituted by minimum vector variance.

Rousseeuw and van Driessen's Theorem

We begin with the following theorem, which is the basic theorem mentioned about Fast MCD, introduced by Rousseeuw and van Driessen. See Theorem 1 in [1].

Theorem 1. *Let $X_1, X_2, \ldots, X_n$ be a set of i.i.d. random vectors of p dimension where the second moment exist. Let H_1 be a subset of $\mathbf{X} = \{X_1, X_2, \ldots, X_n\}$ of h elements,*

$$\bar{X}_{H_1} = \frac{1}{h}\sum_{X_i \in H_1} X_i \quad \text{and}$$

$$S_1 = \frac{1}{h}\sum_{X_i \in H_1}\left(X_i - \bar{X}_{H_1}\right)\left(X_i - \bar{X}_{H_1}\right)^t.$$

If $|S_1| \neq 0$, let $d_i^2 = \left(X_i - \bar{X}_{H_1}\right)^t S_1^{-1}\left(X_i - \bar{X}_{H_1}\right)$ for all $i = 1, 2, \ldots, n$. Let also $H_2 = \{X_{\sigma(1)}, X_{\sigma(2)}, \ldots, X_{\sigma(h)}\}$, where σ is a permutation on the index set, such that $d^2_{\sigma(1)} \leq d^2_{\sigma(2)} \leq \ldots \leq d^2_{\sigma(n)}$. If $\bar{X}_{H_2}$ and S_2 are mean vector and covariance matrix associated to H_2, respectively, then, $|S_2| \leq |S_1|$ with equality if and only if $\bar{X}_{H_1} = \bar{X}_{H_2}$ and $S_2 = S_1$.

This theorem states that, if H_2 is more concentrated than H_1, then the covariance matrix S_2 of all vectors belonging to H_2 has a lower determinant than that of the covariance matrix S_1 of all vectors in H_1. This is the necessary condition for H_2 to be more concentrated than H_1 used by Rousseeuw and van Driessen in [1] in constructing their algorithm. There are three other important results given by Rousseeuw and van Driessen that we want to discuss. First, if the procedure in that theorem is repeated several times, the results are convergent. More precisely, there exist an integer m such that $|S_m| = |S_{m-1}|$, i.e., H_m is as concentrated as H_{m-1}. The second important result is a consequence of the first. Let $\lambda_{k1} \geq \lambda_{k2} \geq \ldots \geq \lambda_{kp} > 0$ be the ordered eigenvalues of S_k ; $k = 1, 2$. If H_2 is more concentrated than H_1, then $\lambda_{21}\ \lambda_{22} \ldots\ \lambda_{2p} < \lambda_{11}\ \lambda_{12} \ldots \lambda_{1p}$. It is this implication that will be discussed in the rest of the paper.

Third, the MCD subset H of $\mathbf{X}$ is separated from $\mathbf{X}\backslash H$ by an ellipsoid. This result is very important in the context of the present paper because it implies that, if H_2 is more concentrated than H_1, the smallest ellipsoid that covers H_2 has smaller volume than that of the smallest ellipsoid that covers H_1. There exists then an affine transformation such that the transformed former ellipsoid is contained entirely in the latter. This viewpoint will lead us to another notion of "more concentrated data subset" which will be defined in the next section.

Proposed Measure of Data Concentration

Let H_1 and H_2 be the two subsets of $\mathbf{X}$ defined in Theorem 1. Thus, H_2 is more concentrated than H_1. As the smallest ellipsoid that covers H_2 has smaller volume than that of the smallest ellipsoid that covers H_1, after an appropriate affine transformation, the transformed former ellipsoid is contained entirely in the transformed latter ellipsoid. Hence, due to the fact that covariance matrix is affine-equivariant, the above notion of "more concentrated data subset" has the following implication. If H_2 is more concentrated than H_1, then $\lambda_{2i} < \lambda_{1i}$ for all $i = 1, 2, \ldots, p$. See [2], [5] and [7] for further discussion on the property that MCD scatter is affine-equivariant. Consequently, because the covariance structure is involved and all eigenvalues are assumed to be positive, by using the notion of *vec* operator, see [14] and [15], we get the following theorem.

Theorem 2. *Let us denote* $vec(A)$ *the vector obtained from a matrix A by stacking its columns one underneath the other. If* H_2 *and* H_1 *are the two subsets of* $\mathbf{X}$ *defined in Theorem 1, then* $\|vec(S_2)\|^2 < \|vec(S_1)\|^2$.

This theorem is straight forward. Let S be a covariance matrix of size $(p \times p)$. Then, the squared Frobenious norm $\|vec(S)\|^2$ of S is equal to the trace of S^2. Consequently, if $\lambda_{2i} < \lambda_{1i}$ for all $i = 1, 2, \ldots, p$, then $\sum_{i=1}^{p} \lambda_{2i}^2 < \sum_{i=1}^{p} \lambda_{1i}^2$. Hence, $\|vec(S_2)\|^2 < \|vec(S_1)\|^2$. Theorem 2 gives us another necessary condition for H_2 to be more concentrated than H_1. If H_2 is more concentrated than H_1 in the sense of Theorem 1, then $\|vec(S_2)\|^2 < \|vec(S_1)\|^2$. In statistical literature, $\|vec(S)\|^2$ is called vector variance (see [16]). More information on the role of the squared Frobenious norm of S in statistics, can be found in [12] for the distributional properties of vector variance, [13] for its role in robust estimation of location and scatter, [16] for the original ideas of vector covariance and vector variance, [17] for the distributional properties of vector covariance, and [18] for its application in multivariate process variability monitoring.

In the next section we present a conjecture on MCD location and scatter if we use minimum vector variance as the objective function in Fast MCD algorithm.

Conjecture

According to Theorem 2, like covariance determinant, vector variance can also be used as multivariate dispersion measure. Both are complementary to each other. The use of them simultaneously will provide more information than if we use a single measure. More over, if instead of minimum covariance determinant we use minimum vector variance as the objective function in Fast MCD algorithm, the formulation of Rousseeuw and van Driessen's theorem mentioned in the Section 2 is slightly modified as follows.

Theorem 3. *Let* $X_1, X_2, \ldots, X_n$ *be a set of i.i.d. random vectors of p dimension where the second moment exist. Let* H_1 *be a subset of* $\mathbf{X} = \{X_1, X_2, \ldots, X_n\}$ *of h elements,*

$$\bar{X}_{H_1} = \frac{1}{h} \sum_{X_i \in H_1} X_i \quad \text{and}$$

$$S_1 = \frac{1}{h} \sum_{X_i \in H_1} \left(X_i - \bar{X}_{H_1}\right)\left(X_i - \bar{X}_{H_1}\right)^t.$$

If $|S_1| \neq 0$, *let* $d_i^2 = \left(X_i - \bar{X}_{H_1}\right)^t S_1^{-1} \left(X_i - \bar{X}_{H_1}\right)$ *for all i = 1, 2, ... , n. Let also* $H_2 = \{X_{\acute{o}(1)}, X_{\acute{o}(2)}, \ldots, X_{\sigma(h)}\}$,

where σ *is a permutation on the index set, such that* $d^2_{\sigma(1)} \leq d^2_{\sigma(2)} \leq \ldots \leq d^2_{\sigma(n)}$. *If* $\bar{X}_{H_2}$ *and* S_2 *are mean vector and covariance matrix associated to* H_2 *, respectively, then,* $\|vec(S_2)\|^2 \leq \|vec(S_1)\|^2$ *with equality if and only if* $\bar{X}_{H_1} = \bar{X}_{H_2}$ *and* $S_2 = S_1$.

Numerous simulation experiments with several values of sample size n and dimension p show that:

1. If the procedure in this theorem is repeated, then there exist an integer m such that $\|vec(S_m)\|^2 = \|vec(S_{m-1})\|^2$;
2. The MCD location and scatter are not changed.

In these experiments we use p = 2, 5, 10, 20, 30, 40, 50, and 100, $n = 10p$ with the number of replications equal 100. Thus, there are 800 simulation experiments and their results lead us to the following conjecture.

Conjecture. Let the procedures in Theorem 1 and Theorem 3 be repeated such that they reach their convergence. Let also, (T_{MCD}, C_{MCD}) and (T_{MVV}, C_{MVV}) be the pairs of location and scatter issued from Theorem 1 and Theorem 3, respectively. Then,

$$(T_{MCD}, C_{MCD}) = (T_{MVV}, C_{MVV}).$$

Additional Remarks

Robust Malahanobis distance issued from the algorithm based on Theorem 1 and that issued from the algorithm based on Theorem 3 are equal. More over, the use of Theorem 3 is computationally more efficient than Theorem 1. The former only needs a quadratic form while the latter a multilinear form.

References

[1] Rousseeuw, P. J. and van Driessen, K. 1999. A Fast Algorithm for the Minimum Covariance Determinant Estimator. *Technometrics,* **41** (3), 212–223.

[2] Rousseeuw, P. J. 1985. *Multivariate Estimation with High Breakdown Point*. In: Grossman, B. W., Pflug, G., Vincze, I., Wertz, W., eds. Mathematical Statistics and Applications. D. Reidel Publishing Company, 283–297.

[3] Rousseeuw, P. J., and Leroy, A. M. 1987. *Robust Regression and Outlier Detection*. John Wiley & Sons, New York.

[4] Rousseeuw, P. J. and van Zomeren, B. C. 1990. Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association,* **85** (411), 633–639.

[5] Lopuhaa, H. P. and Rousseeuw, P. J. 1991. Breakdown points of affine equivariance estimators of multivariate location and covariance matrices. *Annales of Statististics,* **19** (1), 229–248.

[6] Hadi, A.S. 1992. Identifying multivariate outlier in multivariate data. *Journal of Royal Statistical Society B,* **53**, 761–771.

[7] Croux, C. and Haesbroeck, G. 1999. Influence Function and Efficiency of The Minimum Covariance Determinant Scatter matrix Estimator. *Journal of Multivariate Analysis,* **71**, 161–190.

[8] Werner, M. 2003. Identification of Multivariate Outliers in Large Data Sets. Ph.D. Thesis, University of Colorado at Denver.

[9] Hubert, M., Rousseeuw, P.J. and van Aelst, S. 2005. *Multivariate Outlier Detection and Robustness*. Handbook of Statistics, **24**, Elsevier B.V., 263–302.

[10] Hubert, M., Rousseeuw, P.J. and van Aelst, S. (2008). *High-Breakdown Robust Multivariate Methods*. Statistical Science, **21** (1), 92–119.

[11] Angiulli, F. and Pizzuti, C. 2005. Outlier mining and large high-dimensional data sets. *IEEE Transaction Knowledge Data Engineering,* **17** (2), 203–215.

[12] Djauhari, M.A. 2007. A Measure of Data Concentration. *Journal of Applied Probability and Statistics*, **2** (2), 139–157.

[13] Herwindiati, D.E., Djauhari, M.A. and Mashuri, M. 2007. Robust Multivariate Outlier Labeling. *Communication in Statistics: Simulation and Computation*, **36** (6), 1287–1294.

[14] Muirhead, R.J. 1982. *Aspect of Multivariate Statistical Theory*. John Wiley & Sons, Inc., New York.

[15] Schott, J.R. 1997. *Matrix Analysis for Statistics*. John Wiley & Sons, Inc., New York.

[16] Escoufier, Y. 1973. Le Traitement des variables vectorielles. *Biometrics,* **29**, 751–760.

[17] Cleroux, R. 1987. Multivariate Association and Inference Problems in Data Analysis. *Proceedings of the Fifth International Symposium on Data Analysis and Informatics*, **1**, Versailles, France.

[18] Djauhari, M.A., Mashuri, M., and Herwindiati, D.E. 2008. Monitoring Multivariate Process Variability. *Communication in Statistics – Theory and Methods*, **37** (11), 1742–1754.