International Journal of Statistical Sciences *ISSN 1683-5603*
Vol. 16, 2018, pp 69-86

Minimum β-Divergence Estimators of Multivariate Location and Scatter Parameters: Some Properties and Applications

Md. Nurul Haque Mollah and S. K. Bhattacharjee

Department of Statistics
University of Rajshahi, Rajshahi, Bangladesh

Abstract

The minimum β-divergence estimators of multivariate Gaussian location and scatter parameters are highly robust against outliers. Since estimating the location and scatter parameters are the cornerstone of many multivariate statistical methods, the minimum β-divergence estimators of those parameters are the important building block when developing robust multivariate techniques including robust principal component analysis, factor analysis, canonical correlation analysis, independent component analysis, multiple regression analysis, cluster analysis and discriminant analysis. It also serves as a convenient tool for detection of multivariate outliers. The minimum β-divergence estimators of multivariate Gaussian location and scatter parameters are reviewed, along with its main properties such as affine equivariance, breakdown value, and influence function. We discuss its computation and some applications in applied and methodological multivariate statistics.

Keywords: Multivariate Analysis, Multivariate Normal Distribution, Minimum β-Divergence Estimators, Orthogonal Affine Equivariance and Robustness.

1. Introduction

The parameter estimation of location vector and scatter matrix is the cornerstone of multivariate data analysis, as it provides necessary inputs in the subsequent inferential statistical methods [Anderson, 2003; Johnson and Wichern, 2007)]. The sample mean and the sample covariance matrix are the most common estimators of multivariate location vector and scatter matrix, respectively. In the multivariate location and scatter setting, the data are stored in an $n \times p$ data matrix $X_n = (x_1, \ldots, x_n)^T$ with $x_i = (x_{i1}, \ldots, x_{ip})^T$ the i-th vector observation. Here n stands for the number of objects and p for the number of variables and the superscript T for the transpose. Then the estimators of location vector μ and the scatter matrix V are as follows

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{1}$$

$$\hat{V} = \frac{1}{n}\sum_{i=1}^{n} (x_i - \hat{\mu})^T (x_i - \hat{\mu}), \tag{2}$$

which are also known as sample mean vector $\hat{\mu}$ and the sample covariance matrix $\hat{V}$, respectively. They are the optimal estimators in a multivariate Gaussian context, which can maximizes likelihood function as well as minimizes Kullback Leibler (KL) divergence. Both estimators are affine equivariant. However, they are very much sensitive to outlying observations. Any real dataset may often contaminated by outlying observations due to several steps involves in the data generating processes. So the field of robustness becomes popular among statisticians and researchers. Several authors have proposed various robust estimators of location and scatter to overcome the problem of outlying observations in multivariate analysis [Rousseeuw, 1985; Hampel, 1986; Croux, 1999; Maronna and Zamar, 2002]. In this paper, we introduce the minimum β-divergence estimators as an alternative high-breakdown robust and equivariant estimators of multivariate location and scatter parameters [Mollah et al. 2007, 2008a, 2010a]. We also review some of its other interesting properties and applications in different areas of multivariate analysis.

2. Description of the Minimum β-Divergence Estimators of Multivariate Location and Scatter Parameters

We assume that the observations are sampled from multivariate Gaussian distribution with location parameter μ and scatter parameter V, where μ is a vector with p components and V is a positive definite $p \times p$ matrix. The multivariate Gaussian distribution is elliptically symmetric and unimodal, which is defined in the form

$$f(x|\mu, V) = k.exp\left\{-\frac{1}{2}d^2(x|\mu, V)\right\}, \tag{3}$$

where, $k = |2\pi V|^{-1/2}$ and

$$d(x|\mu, V) = \sqrt{(x-\mu)/V^{-1}(x-\mu)} \tag{4}$$

which is known as the Mahalanobis distance between a data vector x and its mean vector μ.

Then the β-divergence between the true density $g(x)$ and model density $f(x|\theta = \{\mu, V\})$ is defined as

$$\mathfrak{D}_\beta\big(g(x), f(x|\theta)\big) = \int \left[\frac{1}{\beta}\{g^\beta(x) - f^\beta(x|\theta)\}g(x) - \frac{1}{\beta+1}\{g^{\beta+1}(x) - f^{\beta+1}(x|\theta)\}\right] dx \quad (5)$$

for $\beta > 0$, which is modified version of the density power divergence [Basu et al. 1998, Minami and Eguchi, 2002] . When $\beta \to 0$, the β-divergence reduces to KL-divergence, that is

$$\lim_{\beta\to 0} \mathfrak{D}_\beta\big(g(x), f(x|\theta)\big) = \int g(x) log \frac{g(x)}{f(x|\theta)} dx = \mathfrak{D}_{KL}\big(g(x), f(x|\theta)\big) \quad (6)$$

Both divergences measure the discrepancy between two densities and satisfy the inequalities

$$\mathfrak{D}_{KL}\big(g(x), f(x|\theta)\big) \geq 0 \text{ and } \mathfrak{D}_\beta\big(g(x), f(x|\theta)\big) \geq 0,$$

[Minami and Eguchi, 2002; Mollah et al. 2006]. The equality holds if and only if $g(x) = f(x|\theta)$ for all x and $\theta = \{\mu, V\}$ for both cases. Therefore, minimizers of both divergences would be the optimal solution for $\theta = \{\mu, V\}$ in absence of outlying observation. It should be noted here that the minimizer of KL-divergence is equivalent to the maximizer of likelihood function (LF). The minimum β-divergence estimator for $\theta = \{\mu, V\}$ is defined as

$$\hat{\theta}_\beta = arg\min_\theta \widehat{\mathfrak{D}}_\beta(g(x),\ f(x|\theta)) = arg\max_\theta L_\beta(\theta|X) \quad (7)$$

where

$$L_\beta(\theta|X) = \frac{1}{\beta}\left[\frac{1}{n l_\beta(\theta)}\sum_{i=1}^n f^\beta(x|\theta) - 1\right] \quad (8)$$

which is known as β-likelihood function with $l_\beta(\theta) = (1+\beta)^{-p/2}|2\pi V|^{-\frac{\beta^2}{2(1+\beta)}}$.

The β-Likelihood reduces to average of log-likelihood function when $\beta \rightarrow 0$. That is,

$$\lim_{\beta \to 0} L_\beta(\theta|X) = L_0(\theta|X) = \frac{1}{n}\sum_i^n \log f(x_i|\theta) \tag{9}$$

The minimum β-divergence estimators $\hat{\theta}_\beta = \{\hat{\mu}_\beta, \hat{V}_\beta\}$ of $\theta = \{\mu, V\}$ are obtained iteratively as follows:

$$\hat{\mu}_\beta^{(r+1)} = \frac{\sum_{j=1}^n W_\beta(x_j|\hat{\mu}_\beta^{(r)}, \hat{V}_\beta^{(r)}) x_j}{\sum_{j=1}^n W_\beta(x_j|\hat{\mu}_\beta^{(r)}, \hat{V}_\beta^{(r)})} \tag{10}$$

$$\hat{V}_\beta^{(r+1)} = (1+\beta)\frac{\sum_{j=1}^n W_\beta(x_{jk}|\hat{\mu}_\beta^{(r)}, \hat{V}_\beta^{(r)})(x_j - \hat{\mu}_\beta^{(r)})\left(x_j - \hat{\mu}_\beta^{(r)}\right)'}{\sum_{j=1}^n W_\beta(x_j|\hat{\mu}_\beta^{(r)}, \hat{V}_\beta^{(r)})} \tag{11}$$

and

$$W_\beta(x|\hat{\mu}_\beta^{(r)}, \hat{V}_\beta^{(r)}) = exp\left\{-\frac{\beta}{2} d^2\left(x|\hat{\mu}_\beta^{(r)}, \hat{V}_\beta^{(r)}\right)\right\} \tag{12}$$

The formulation of equations (10-12) is described in Mollah et al. (2007, 2010a). The function in equation (12) is called the β-weight function, which plays the key role for robust estimation of the parameters. If β tends to 0, then the equations (10) and (11) are reduced to the classical non-iterative estimates of mean and covariance matrix as given in equations (1) and (2) respectively. The robustness performance of the minimum β-divergence estimator $\hat{\theta}_\beta = \{\hat{\mu}_\beta, \hat{V}_\beta\}$ of $\theta = \{\mu, V\}$ depends on the value of the tuning parameter β and initialization of the parameters.

2.1. β-Selection using *k*-Fold Cross Validation

To select the appropriate β by *k*-fold cross validation (CV), the tuning parameter β is fixed to β_0. The steps for selecting the appropriate β by *k*-fold cross validation is given below:

Step-1: Dataset $S = \{x_i; i = 1,2, \dots, n\}$ randomly split into *k* subsets $S_1, S_2, \dots, S_k$ where $S_j = \{x_{(t)}|x_{(t)} \in S, t = 1,2, \dots, n_j\}$ and $\sum_{j=1}^k n_j = n$

Step-2: Let S_j^c be the complement set of S_j, $j=1,2,\dots,k$.

Step-3: Estimate $\hat{\mu}_\beta$ and $\hat{V}_\beta$ iteratively by equations (7-8) based on dataset S_j^c

Step-4: Compute $\mathrm{CV}_j(\beta)$ using the dataset S_j, for j=1, 2,.., k

$\mathrm{CV}_j(\beta) = L_{\beta_0}(\hat{\mu}_\beta, \hat{V}_\beta \mid S_j)$, where

$$L_{\beta_0}(\hat{\mu}_\beta, \hat{V}_\beta \mid S_j) = \frac{1}{\beta_0}\left[1 - \frac{1}{n_j}\left|\hat{V}_\beta\right|^{-\frac{\beta_0}{2(1+\beta_0)}} \sum_{x_j \in S_j} W_{\beta_0}(x_j \mid \hat{\mu}_\beta, \hat{V}_\beta\right]$$

Step-5: Compute $\hat{\beta} = \underset{\beta}{argmin}\,\mathrm{CV}(\beta)$

where $\mathrm{CV}(\beta) = \frac{1}{n}\sum_{j=1}^{k} CV_j(\beta)$

More discussion about β selection also can be found in Mollah et al. (2007, 2010a).

2.2. Influence Function

The influence function for the estimator T at x under the distribution F is defined as

$$IF(x; T, F) = \lim_{t \to 0} \frac{T[(1-t)F + t\Delta_x] - T(F)}{t}$$

where Δ_x is the probability measure that puts mass 1 at the point x. If the gross error sensitivity (GES), that is, $\lim_x |IF(x; T, F)|$ is finite, then the estimator T is said to be B-robust under the distribution F (c.f. chapter 5 of Hampel et al. (1986)).

The robustness of the minimum β-divergence estimators were investigated by the influence function (Mollah et el. 2007). The influence function for the location estimator $\hat{\mu}_\beta = \mu_\beta(X)$ at x under the distribution F is given by

$$IF\left(x; \mu_\beta(X), F\right) = -\mu + \frac{W_\beta(x|\mu,V)x}{E_X\{W_\beta(X|\mu,V)\}} \tag{13}$$

The influence function for the scatter estimator $\hat{V}_\beta = V_\beta(X)$ at x under the distribution F is given by

$$IF\left(x; V_\beta(X), F\right) = -V + \frac{W_\beta(x|\mu,V)(x-\mu)(x-\mu)'}{E_X\{W_\beta(X|\mu,V)\}} \tag{14}$$

Obviously, the gross error sensitivity (GES), that is, $lim_x \text{IF}(x; T, F)$ is finite for both location and scatter estimators, since if the components of x become larger, then the corresponding weight $W_\beta(x|\mu, V)$ becomes smaller for both IF. Thus both estimators are known as B-robust under the distribution F. More discussion for both IF can be found in Mollah et al. (2007).

2.3. Parameters Initialization and Breakdown Points of the Estimates

The robustness of the minimum β-divergence estimator $\hat{\theta}_\beta = \{\hat{\mu}_\beta, \hat{V}_\beta\}$ for the Gaussian parameter $\theta = \{\mu, V\}$ is measured by means of finite-sample replacement breakdown point suggested by Donoho and Huber (1983). The breakdown point of an estimator measures the smallest fraction m/n of outlying observations that carry the estimates beyond all bounds (Lopuhaa and Rousseeuw, 1991; Hubert and Debruyne, 2010). Denote $X_{n,m}$ as the data matrix obtained by replacing m data vectors $x_{j+1}, \ldots, x_{j+m}$ of X_n by outlying observations satisfying the Tukey-Huber contamination model (THCM; Agosinelli et al. 2015).The breakdown point for location estimator $\hat{\mu}_\beta = \mu_\beta(X_n)$ is defined as

$$\varepsilon^*(\hat{\mu}_\beta; X_n) = \min_{1\le m\le n}\left\{\frac{m}{n}: \sup_m \left\|\mu_\beta(X_n) - \mu_\beta\left(X_{n,m}\right)\right\| = \infty\right\}, \tag{15}$$

where the supremum (sup) is taken over all possible m outlying observations in $X_{n,m}$. The breakdown point for the scatter estimator $\hat{V}_\beta = V_\beta(X_n)$ is defined as:

$$\varepsilon^*(\hat{V}_\beta; X_n) = \min_{1\le m\le n}\left\{\frac{m}{n}: \sup_m \varphi\left(V_\beta(X_n), V_\beta\left(X_{n,m}\right)\right) = \infty\right\}, \tag{16}$$

where $\varphi(A, B)=\max\{|\boldsymbol{\lambda}_1(A)- \boldsymbol{\lambda}_1(B)|, |\boldsymbol{\lambda}_p(A)^{-1}- \boldsymbol{\lambda}_p(B)^{-1}|\}$, with $\boldsymbol{\lambda}_1(A)\ge \ldots.\ge\boldsymbol{\lambda}_p(A)$ being the ordered eigen values of the matrix A. However, the breakdown points in equations (15-16) depend on the value of the tuning parameter β and the initialization of the parameters in the iterative equations (10-12). During the first iteration ($r=0$), the mean vector $\hat{\mu}_\beta^{(r)}$ in the β-weight function (eq.12) is initialized by the coordinate-wise sample median vector (x_{md}), since mean vector and the coordinate-wise median vector are same for multivariate Gaussian distribution and the coordinate-wise median vector is highly robust estimator of location parameter against outliers with breakdown point $[(n+1)/2]/n \approx 0.5$ (Lopuhaa and Rousseeuw, 1991). The covariance matrix $\hat{V}_\beta^{(r)}$ in the β-weight function (eq.12) is initialized

by the identity matrix *I*. Then the first iterative solution $\hat{\theta}_\beta^{(1)} = (\hat{\mu}_\beta^{(1)}, \hat{V}_\beta^{(1)})$ moves from $\hat{\theta}_\beta^{(0)} = (\hat{\mu}_\beta^{(0)}, \hat{V}_\beta^{(0)})$ towards the optimal solution of $\theta = \{\mu, V\}$ in presence of outlying observations also. To confirm it, let *m* data vectors $x_{j+1}, \ldots, x_{j+m}$ in the data matrix X_n are contaminated by the extreme outliers. With these outlying data vectors, the Mahalanobis distance produces $d^2\left(x_{j+k} | \hat{\mu}_\beta^{(r)} = x_{md}, \hat{V}_\beta^{(r)} = I\right) \to \infty$ and the corresponding β-weight function produces $W_\beta(x_{i+k} | \hat{\mu}_\beta^{(r)} = x_{md}, \hat{V}_\beta^{(r)} = I) \to 0,\ (k = 1,2, \ldots, m)$ by equation (12). On the other hand, with the usual data vectors, the Mahalanobis distance produces $d^2\left(x_i | \hat{\mu}_\beta^{(r)} = x_{md}, \hat{V}_\beta^{(r)} = I\right) \to 0$ and the corresponding β-weight function produces $W_\beta(x_i | \hat{\mu}_\beta^{(r)} = x_{md}, \hat{V}_\beta^{(r)} = I) \to 1, (i = 1, \ldots, j, j + m + 1, \ldots, n)$ by equation (12). Thus, outlying observations cannot influence the estimates at all computed by equations (10-11) during first iteration.

During the second iteration (*r=1*), the mean vector $\hat{\mu}_\beta^{(r)}$ in the β-weight function (eq.12) is replaced by the updated mean vector $\hat{\mu}_\beta^{(1)}$ which is more close to the optimal solution of the mean vector. Then the covariance matrix $\hat{V}_\beta^{(r)}$is replaced by the updated covariance matrix $\hat{V}_\beta^{(1)}$. Then the second iterative solution $\hat{\theta}_\beta^{(2)} = (\hat{\mu}_\beta^{(2)}, \hat{V}_\beta^{(2)})$ moves from $\hat{\theta}_\beta^{(1)} = (\hat{\mu}_\beta^{(1)}, \hat{V}_\beta^{(1)})$ towards the optimal solution of $\theta = \{\mu, V\}$ more accurately. This is because, with the previous *m* outlying data vectors, the Mahalanobis distance again produces $d^2\left(x_{j+k} | \hat{\mu}_\beta^{(r)} = \hat{\mu}_\beta^{(1)}, \hat{V}_\beta^{(r)} = \hat{V}_\beta^{(1)}\right) \to \infty$ and the corresponding β-weight function produces $W_\beta(x_{i+k} | \hat{\mu}_\beta^{(r)} = \hat{\mu}_\beta^{(1)}, \hat{V}_\beta^{(r)} = \hat{V}_\beta^{(1)}) \to 0,\ (k = 1,2, \ldots, m)$ by equation (12) more accurately. On the other hand, with the usual data vectors, the Mahalanobis distance produces $d^2\left(x_i | \hat{\mu}_\beta^{(r)} = \hat{\mu}_\beta^{(1)}, \hat{V}_\beta^{(r)} = \hat{V}_\beta^{(1)}\right) \to 0$ and the corresponding β-weight function produces $W_\beta(x_i | \hat{\mu}_\beta^{(r)} = \hat{\mu}_\beta^{(1)}, \hat{V}_\beta^{(r)} = \hat{V}_\beta^{(1)}) \to 1, (i = 1, \ldots, j, j + m + 1, \ldots, n)$ by equation (12) more accurately. It should be noted here that if $n<p$ or $|\hat{V}_\beta^{(r)}| = 0$, then $\hat{\Lambda}_\beta^{(r)} = \text{diag}\left(\hat{\sigma}_{11,\beta}^{(r)}, \hat{\sigma}_{22,\beta}^{(r)}, \ldots, \hat{\sigma}_{pp,\beta}^{(r)}\right)$, the diagonal matrix, can be used instead of the covariance matrix $\hat{V}_\beta^{(r)}$ in the β-weight function (12) to calculate the

weight for each data vector, where $\hat{\sigma}_{ii,\beta}^{(r)}$ is the *i*th diagonal element of $\hat{V}_{\beta}^{(r)}$ (for *r>0*). Thus, outlying observations cannot influence the estimates obtained by equations (10-11) at all during second iteration also. Similarly, in each iteration, the β-weight function with an appropriate β produces larger weights with usual (uncontaminated) data vectors and smaller weight for the outlying (contaminated) data vectors which leads the convergence of the iterative equations (10-11) to the optimal solution of $\theta = \{\mu, V\}$ accurately when upto *m=n/2*=50% of data vectors in $X_{n,m}$ are contaminated by outliers. If more than *m=n/2* of data vectors are outlying in $X_{n,m}$, then coordinate-wise median vector fails to initialize the mean vector $\hat{\mu}_{\beta}^{(r)}$ to the good part of the dataset. Then the iterative equations (10-11) may also fail to converge in the optimal solution of $\theta = \{\mu, V\}$. Thus the minimum β-divergence estimators $\hat{\theta}_{\beta} = \{\hat{\mu}_{\beta,}\hat{V}_{\beta}\}$ with an appropriate β are claimed as highly robust estimators against outliers with breakdown point $[(n+1)/2]/n \approx 0.5$ if mean vector is initialized by the coordinate-wise median vector in equation (12). However, the minimum β-divergence estimators $\hat{\theta}_{\beta} = \{\hat{\mu}_{\beta,}\hat{V}_{\beta}\}$ with an appropriate β can also produce reasonable estimates when more than *m=n/2* of data vectors are outlying in $X_{n,m}$ if the mean vector $\hat{\mu}_{\beta}^{(r)}$ is initialized by a data vector belonging to the good part in $X_{n,m}$ (Mollah et al., 2010a). Thus the breakdown point $[(n+1)/2]/n \approx 0.5$ of $\hat{\theta}_{\beta} = \{\hat{\mu}_{\beta,}\hat{V}_{\beta}\}$ can be increased based on the initialization of the parameters which may be the open challenge to the researcher.

2.4 Equivariance Property of the Estimators

The minimum β-divergence estimators $\hat{\theta}_{\beta} = \{\hat{\mu}_{\beta} = \mu_{\beta}(X),\ \hat{V}_{\beta} = V_{\beta}(X)\}$ for the Gaussian parameters $\theta = \{\mu, V\}$ satisfy the affine equivariance properties as follows

$$\mu_{\beta}(AX + b) = \sum_{i=1}^{n} \varphi_i (Ax_i + b) = A \sum_{i=1}^{n} \varphi_i x_i + b = A\mu_{\beta}(X) + b \qquad (17)$$

and

$$V_{\beta}(AX + b) = \sum_{i=1}^{n} \varphi_i \big[Ax_i + b - \mu_{\beta}(Ax_i + b)\big][Ax_i + b - \mu_{\beta}(Ax_i + b)]^{/} = AV_{\beta}(X)A^{/} \qquad (18)$$

where A be a $p\times p$ non-singular/orthogonal matrix and b be a non-zero p-vector and

$$\varphi_i = W_\beta\left(x_i|\mu_\beta(X), V_\beta(X)\right) / \sum_{j=1}^{n} W_\beta\left(x_j|\mu_\beta(X), V_\beta(X)\right)$$

with $W_\beta(x|\mu_\beta(X), V_\beta(X)) = exp\left\{-\frac{\beta}{2} d^2\left(x|\,\mu_\beta(X), V_\beta(X)\right)\right\}$

The equations (17-18) satisfy the orthogonally affine equivariance property from the fact that the Mahalanobis distance $d(x|\hat{\mu}_\beta^{(r)}, \hat{V}_\beta^{(r)}) = \sqrt{\left(x-\hat{\mu}_\beta^{(r)}\right)' \hat{V}_\beta^{(r)^{-1}} \left(x-\hat{\mu}_\beta^{(r)}\right)}$ in the β-weight function (12) is orthogonally affine invariant, since $\hat{\mu}_\beta^{(r)}$ is initialized by the coordinate-wise sample median vector (x_{md}) and $\hat{V}_\beta^{(r)}$ is initialized by the identity matrix I at $r=0$. It should be noted here again that the coordinate-wise sample median vector is equivalent to the sample mean vector in the case of multivariate normal distribution. Also the equations (17-18) satisfy the affine equivariance property from the fact that the Mahalanobis distance $d(x|\hat{\mu}_\beta^{(r)}, \hat{V}_\beta^{(r)}) = \sqrt{\left(x-\hat{\mu}_\beta^{(r)}\right)' \hat{V}_\beta^{(r)^{-1}} \left(x-\hat{\mu}_\beta^{(r)}\right)}$ in the β-weight function (eq. 12) is affine invariant if $\hat{\mu}_\beta^{(r)}$ and $\hat{V}_\beta^{(r)}$ are initialized by any of the affine equivariance estimators of μ and V at $r=0$. Obviously, the reweighted estimators (17-18) satisfy both the affine and orthogonally-affine equivariance property in each of the iterations. However, the minimum β-divergence estimators $\hat{\theta}_\beta = \left\{\hat{\mu}_\beta = \mu_\beta(X),\ \hat{V}_\beta = V_\beta(X)\right\}$ satisfying the affine equivariance property can achieve the breakdown point $[(n-p+1)/2]/n$ <0.5 for p>1. On the other hand, it can achieve the breakdown point $[(n+1)/2]/n \approx 0.5$ satisfying the affine equivariance property orthogonally.

3. Applications

3.1 Multivariate Outlier Detection using β-Weight Function

A data vector x in a dataset is said to be outlying if at least one component of $x = \{x_1, x_2, \ldots\ldots, x_p\}$ is contaminated by outlier. To derive a criterion whether the

data vector x is contaminated or not, the β-weight function (12) is rewrite as follows

$$W_\beta(x|\hat{\mu}_\beta, \hat{V}_\beta) \exp\left\{-\frac{\beta}{2} d^2(x|\hat{\mu}_\beta, \hat{V}_\beta)\right\}, \tag{19}$$

where $(\hat{\mu}_\beta, \hat{V}_\beta)$ are the minimum β-divergence estimators of (μ, V) obtained by iterative equations (10-12). The values of this weight function lie between 0 and 1as discussed previously. This weight function produces larger weight if x is a usual data vector and smaller weight if x is an unusual data vector. Therefore, the β-weight function (eq.19) is used to detect outlier as follows:

$$W_\beta(x|\hat{\mu}_\beta, \hat{V}_\beta) = \begin{cases} > \partial, & \text{if } x \text{ is usual data vector} \\ \leq \partial, & \text{if } x \text{ is outlying data vector} \end{cases} \tag{20}$$

The threshold value ∂ can be determined by the quantile values of $W_\beta(x|\hat{\mu}_\beta, \hat{V}_\beta)$ for j = 1,2,……,n with probability

$$\Pr\{W_\beta(x|\hat{\mu}_\beta, \hat{V}_\beta) \leq \partial\} \leq p, \tag{21}$$

where p is the probability for selecting the cut-off value ∂ as a quantile value based on the empirical distribution of $W_\beta(x|\hat{\mu}_\beta, \hat{V}_\beta)$. The value of p should less than 0.1 to fix the cut-off value ∂ for detection of outlying data vector using equation (20). This idea was first introduced in Mollah et al. (2012).

The criteria whether an unlabeled data vector x is contaminated by outlier or not, is defined as follows:

$$w_\beta(x) = \sum_{k=1}^{K} W_{k,\beta}(x|\hat{\mu}_{k,\beta}, \hat{V}_{k,\beta}) = \begin{cases} \geq \partial, & \text{if } x \text{ is not outlying} \\ < \partial, & \text{if } x \text{ is outlying} \end{cases} \tag{22}$$

where, $\partial = \sum_{k=1}^{K} \partial_k$, here ∂_k is the cut-off value for outlier detection in the kth population obtained by equations (19-20) and $(\hat{\mu}_{k,\beta}, \hat{V}_{k,\beta})$ are the estimators of (μ, V) for kth population.

3.2. Clustering and Classification

Clustering is an unsupervised learning which plays the key role in the field of data mining. Basically, there are three types of clustering approaches known as partitioned based, model based and hierarchical clustering (HC). The later HC approach seems to be more useful than the former partitioned and model based

approaches, since HC does not require to knowing the number of clusters unlike the former two approaches. It becomes popular for high-throughput high-dimensional gene expression data analysis from the research work of Eisen et al. (1998). The HC approaches are formulated based on the distance matrix or dissimilarity matrix using single, complete or average linkages. The dissimilarity matrix D is defined based on the correlation matrix R. However, the correlation matrix R as well as the distance matrix or dissimilarity matrix D are sensitive to outlying observations, which leads the misleading clustering results by HC. To overcome this problem, Mollah et al. (2009) proposed β-HC by robustifying R based on the minimum β-divergence estimator $\hat{V}_\beta$ of the covariance matrix V as follows:

Let $\hat{V}_\beta = \left[\hat{\sigma}_{ij}\right]_{p\times p}$, which implies $\hat{R}_\beta = \left[\hat{r}_{ij}\right]_{p\times p}$, the minimum β-divergence estimator of the correlation matrix R, where $\hat{r}_{ij} = \hat{\sigma}_{ij}/\sqrt{\hat{\sigma}_{ii}\hat{\sigma}_{jj}}$. Then the β-dissimilarity matrix is defined as $\widehat{D}_\beta = \left[\hat{d}_{ij}\right]_{p\times p}$, where $\hat{d}_{ij} = 1 - \hat{r}_{ij} \geq 0$. Then the β-dissimilarity matrix $\widehat{D}_\beta$ is used instead of traditional dissimilarity matrix D for formulating HC algorithms from the robustness viewpoints. More discussion about β-HC and its application for gene expression data analysis can be found in Mollah et al. (2009). Badsha (2010) and Badsha et al. (2013) extended β-HC to β-CHC for complementary hierarchical clustering (CHC; Nowak and Tibshirani, 2008) from the robustness viewpoints. Kabir (2018) and Kabir and Mollah (2018) also proposed the robustification of the model based clustering using the minimum β-divergence estimators of the mean vectors μ's and the covariance matrices V's obtained by the EM algorithm.

On the other hand, classification is a supervised learning which plays the key role in the field of machine learning for class prediction or pattern recognition. In the literature, there are several approaches addressed for classifications (Anderson 2003; Johnson and Wichern 2007), where Gaussian Bayes classifier is one of the most popular candidate. However, most of the existing classifiers including Gaussian Bayes classifiers are very much sensitive to outliers. So, they can produce misleading prediction results in presence of outliers. To overcome this problem, Matiur (2012) and Matiur and Mollah (2018) proposed the robustification of Gaussian Bayes Classifier based on the minimum β-divergence estimators of the mean vectors μ's and the covariance matrices V's. The

classification region R_j for classifying the test vector x to the jth population by the Gaussian Bayes β-classifier is defined as

$$R_j: \widehat{U}_{ij,\beta}(x) > \log\frac{[q_i C(j|i)]}{[q_j C(i|j)]}, i = 1,2,\dots,m\ (i \neq j) \tag{23}$$

where q_i's are the mixing proportions, C($j|i$) is the cost of misclassifying an observation from jth population as from ith population and

$$\widehat{U}_{ij,\beta}(x) = \frac{1}{2}\log\frac{\left|\hat{V}_{i,\beta_i}\right|}{\left|\hat{V}_{j,\beta_j}\right|} + \frac{1}{2}\left(x-\hat{\mu}_{i,\beta_i}\right)'\hat{V}_{i,\beta_i}^{-1}\left(x-\hat{\mu}_{i,\beta_i}\right) - \frac{1}{2}\left(x-\hat{\mu}_{j,\beta_j}\right)'\hat{V}_{j,\beta_j}^{-1}\left(x-\hat{\mu}_{j,\beta_j}\right) \tag{24}$$

which is known as the Gaussian Bayes β-classifier. It is non-linear. It reduces to the conventional non-linear Gaussian Bayes classifier for $\beta=0$. If we assume homogeneous populations (i.e. $V_1=V_2=\dots=V_m$), the non-linear β-classifier reduces to the linear classifier as follows

$$\widehat{U}_{ij,\beta}(x) = x'\hat{V}_{\beta}^{-1}\left(\hat{\mu}_{j,\beta_j}-\hat{\mu}_{i,\beta_i}\right) - \frac{1}{2}\left(\hat{\mu}_{j,\beta_j}+\hat{\mu}_{i,\beta_i}\right)'\hat{V}_{\beta}^{-1}\left(\hat{\mu}_{j,\beta_j}-\hat{\mu}_{i,\beta_i}\right) \tag{25}$$

which is also known as β-LDA. It reduces to the Gaussian Bayes LDA for $\beta=0$. Here $(\hat{\mu}_{i,\beta_i}, \hat{V}_{i,\beta_i})$ are minimum β-divergence estimates of (μ_i, V_i) computed by equations (10-12) based on the training samples from the ith multivariate Gaussian population and $\hat{V}_{\beta} = \frac{1}{n}\sum_{i=1}^{m} n_i \hat{V}_{i,\beta_i}$, the pooled variance. More discussion about robustification of Bayes classifiers and their applications can be found in Matiur (2012), Ahmed et al. (2017) and, Matiur and Mollah (2018).

3.3 Dimension Reduction

In statistics, machine learning, and information theory, dimensionality reduction or dimension reduction is the process of reducing the number of random variables under consideration by obtaining a set of principal variables. It is essential to produce the inputs of some statistical approaches those are suffering from the high-dimensionality of the dataset. There are two types of dimension reduction techniques (i) feature selection and (ii) feature projection. Feature selection approaches try to find a subset of the original variables. There are three feature

selection strategies (i) filtering strategy (ii) the wrapper strategy (e.g. search guided by accuracy), and (iii) the embedded strategy (features are selected to add or be removed while building the model based on the prediction errors). Feature projection transforms the data in the high-dimensional space to a space of fewer dimensions. Principal component analysis (PCA), factor analysis (FA) and canonical correlation analysis (CCA) are considered as the most popular feature projection approaches for dimension reduction. The estimation of the mean vector μ and covariance matrix V plays the key role in each of PCA, FA and CCA. However, traditional sample mean vector and covariance matrix (1-2) are sensitive to outlying observations though they are affine equivariant. There are some popular affine equivariant robust estimators like MCD and MVE for (μ, V), but their robustness performance gradually decreases if the number of variables p increases in the dataset. Mollah et al. (2010b) proposed robust PCA based on the minimum β-divergence estimators $(\hat{\mu}_{\beta,}\hat{V}_{\beta})$ of (μ, V) computed by equations (10-12) which is more robust than the other existing robust estimators in the literature as discussed previously. In general, the β-PCA aims to extract the most informative q-dimensional output vector $y_j = (y_{j1,}y_{j2,} \dots, y_{jq})^{/}$ from the input vector $x_j = (x_{j1,}x_{j2,} \dots, x_{jp,})^{/}$ of dimension $p \geq q$ whose components are assumed to be linearly correlated to each other. This is achieved by learning the $p \times q$ orthogonal matrix $\hat{\Gamma}_{\beta} = [\,\hat{\gamma}_{1,}\,\hat{\gamma}_{2}, \dots .\hat{\gamma}_{q}]$ which relates x_j to y_j by

$$y_j = \hat{\Gamma}_{\beta}^{/}(x_j - \hat{\mu}_{\beta}) \tag{26}$$

such that components of y_j are mutually uncorrelated satisfying the variance inequality property of principal components (Higuchi and Eguchi, 2004; Mollah et al. 2010a). The orthogonal matrix $\hat{\Gamma}_{\beta}$ is determined by $\hat{\Gamma}_{\beta} = \text{eigen}\,(\hat{V}_{\beta})$ such that $\hat{\Gamma}_{\beta}^{/}\hat{V}_{\beta}\hat{\Gamma}_{\beta} = \text{diag}(\hat{\lambda}_1, \hat{\lambda}_2, \dots \hat{\lambda}_q)$ satisfying the inequality $\hat{\lambda}_1 > \hat{\lambda}_2 > \cdots > \hat{\lambda}_q$, where $\hat{\lambda}_i$ is the variance of the *i*th principal component (PC). Mollah et al. (2010b) extended β-PCA for exploring the local PCA structures. Similarly, Ahsan (2012) and Ahsan et al. (2012) robustify factor analyzers (FA) and, Singha (2013) and Singha et al. (2014) robustify canonical correlation analyzers (CCA) based on the minimum β-divergence estimators $(\hat{\mu}_{\beta,}\hat{V}_{\beta})$ of (μ, V) computed by equations (10-12) for high-dimensional molecular OMICS data analysis from the robustness viewpoints

3.4 Blind Source Separation

Preprocessing of data is necessary in some adaptive independent component analysis (ICA) algorithms for Blind Source Separation (BSS), because it reduces the complexity of the ICA problems [Hyv¨arinen et al., 2001; Cichocki and Amari, 2002]. For example, robust FastICA fixed-point algorithm [Hyv¨arinen et al., 2001] is a popular algorithm for BSS, however, it produces misleading results in presence of outliers due to the utilization of non-robust prewritten dataset. To overcome this problem, Mollah et al. (2007) robustify the prewhitening procedure based on the minimum β-divergence estimators $\left(\hat{\mu}_{\beta,}\hat{V}_{\beta}\right)$ of (μ, V) computed by equations (10-12) as follows:

Let us consider the linear ICA model for an observable random vector x of dimension p as

$$x = As \tag{27}$$

where $A \in Rm \times m$ and s is an unobservable source vector whose components are assumed to be independent and non-Gaussian. A random vector z is said to be whiten or sphere if $E(Z)=0$ and $E(ZZ^{/})=I_p$ (identity matrix). In the β-prewhitening procedure, the prewhitten data vector z_j is obtained from x_j by the following equation

$$z_j = \hat{\mathrm{V}}_{\beta}^{-1/2}\left(x_j - \hat{\mu}_{\beta}\right) \tag{26}$$

More discussion about β-prewhitening and its application to BSS can be found in [Mollah et al. 2007, 2009, 2010c].

4. Conclusion

In this paper we have illustrated that the minimum β-divergence estimators of multivariate Gaussian location and scatter parameters are highly robust against outliers. We have discussed how the minimum β-divergence estimators of those parameters playing key role when developing robust multivariate techniques including robust principal component analysis, factor analysis, canonical correlation analysis, independent component analysis, multiple regression analysis, cluster analysis and discriminant analysis. It also serves as a convenient tool for detection of multivariate outliers. The minimum β-divergence estimators of multivariate Gaussian location and scatter parameters are reviewed, along with

its main properties such as affine equivariance, breakdown value, and influence function. We discuss its computation and some applications in applied and methodological multivariate statistics. Finally we have provided a detailed reference list with applications and generalizations of the minimum β-divergence estimators in the theoretical and applied research.

References

[1] Anderson, T. W. (2003). An Introduction to Multivariate Statistical Analysis. Wiley Interscience.

[2] Agostinelli, C., Leung, A., Yohai, V., and Zamar, R. (2015). Robust estimation of multivariate location and scatter in the presence of cell-wise and case-wise contamination, TEST. 24, 441-461.

[3] Ahsan, M. A. (2012). Robustification of factor Analyzers and Its Application for Gene Expression Data Analysis. Unpublished M.Sc. Thesis, Dept. of Statistics, University of Rajshahi, Bangladesh.

[4] Ahsan, M. A., Rahaman, M. M., Monir, M. M., Hossain, M. R. and Mollah, M. N. H. (2012). Robustification of Factor Analyzers and Its Application for Microarray Gene Expression Data Analysis. Proceedings of the International Conference on Bioinformatics, Health, Agriculture and Environment - 2012, University of Rajshahi, Bangladesh, ISBN-978-984-33-5876-9.

[5] Ahmed, M. S., Shahjaman, M., Rana, M. M. and Mollah, M. N. H. (2017). Robustification of Naïve Bayes Classifier and Its Application for Microarray Gene Expression Data Analysis. BioMed Research International. Volume 2017, Article ID 3020627, 17 pages, https://doi.org/10.1155/2017/3020627.

[6] Basu, A., Harris, I. R., Hjort, N. L. and Jones, M. C. (1998). Robust and efficient estimation by minimizing a density power divergence. Biometrika, 85, 549-559.

[7] Badsha, M. B. (2010). Robustification of Complementary Hierarchical Clustering for Gene Expression Data Analysis. Unpublished M.Sc. Thesis, Dept. of Statistics, University of Rajshahi, Bangladesh.

[8] Badsha, M. B., Jahan, N., Kurata, H. and Mollah, M. N. H. (2013). Robust Complementary Hierarchical Clustering for Gene Expression Data Analysis by β-divergence. Journal of Bioscience and Bioengineering (JBB), Vol-116 (3), pp. 397-407.

[9] Cichocki, A. and Amari, S. (2002). Adaptive Blind Signal and Image Processing, Wiley, New York.

[10] Croux, C. and Haesbroeck, G. (1999). Influence function and efficiency of the Minimum Co-variance Determinant scatter matrix estimator. Journal of Multivariate Analysis, 71:161-190.

[11] Donoho, D. L. and Huber, P. J. (1983). The notion of breakdown point. In a Festschrift for Ericion.

[12] Eisen, M. B., Spellman, P. T., Brown, P. O. and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns, Proc. Natl. Acad. Sci. USA, 95, 14863-14868.

[13] Hampel, F.R. Ronchetti, E.M. Rousseeuw, P.J. and Stahel, W.A. (1986). Robust Statistics: The Approach Based on Influence Functions. Wiley, New York.

[14] Higuchi, I. and Eguchi, S. (2004). Robust principal component analysis with adaptive selection for tuning parameters. Journal of Machine Learning Research, 5, 453–471.

[15] Hyv¨arinen, A., Karhunen, J. and Oja, E. (2001). Independent Component Analysis, Wiley, New York, 2001.

[16] Johnson, R. A. and Wichern, D. W. (2007). Applied multivariate statistical analysis. Sixth edition, Prentice-Hall.

[17] Hubert, M. and Debruyne, M. (2010). Minimum Covariance Determinant. Advanced Review. Vol.2, John Wiley & Sons. Inc.

[18] Kabir, M. H. (2018). Development of Statistical Algorithm for Data Mining in Bioinformatics. Unpublished PhD Thesis, Dept. of Statistics, University of Rajshahi, Bangladesh.

[19] Kabir, M. H. and Mollah, M. N. H. (2018). A Semi-Supervised Robust Model based Clustering and Its Application for Gene Expression Data Analysis. International Conference on New Paradigms in Statistics for Scientific and Industrial Research, January 4-6, 2018. Kolkata, India.

[20] Lopuha, H.P. and Rousseeuw, P.J. (1991). Breakdown points of an equivariant estimators of multivariate location and covariance matrices. The Annals of Statistics, 19:229-248.

[21] Mollah, M. N. H., Minami, M. and Eguchi, S. (2006). Exploring Latent Structure of Mixture ICA Models by the Minimum β-Divergence Method. Neural Computation, Vol.18, pp.166-190.

[22] Mollah, M. N. H., Minami, M. and Eguchi, S. (2007). Robust Prewhitening for ICA by Minimizing β-Divergence and its Application to Fast ICA. Neural Processing Letters, Vol. 25, pp. 91-110.

[23] Mollah, M. M. H., Hossain, M. G. and Mollah, M. N. H. (2008a). Robust Estimation for Multivariate Normal Distribution. Journal of Applied Statistical Science (JASS), Vol. 16, pp. 377-386.

[24] Mollah, M. N. H., Mari, P., Komori, O. and Eguchi, S. (2009). Robust Hierarchical Clustering for Gene Expression Data Analysis. Communications of SIWN, Vol. 6, pp. 118-122.

[25] Mollah, M. N. H., Sultana, N., Minami, M. and Eguchi, S. (2010a). Robust Extraction of Local Structures by the Minimum β-Divergence method., Neural Network, Vol. 23, pp. 226-238.

[26] Mollah, M. M. H., Hossain, M. G. and Mollah, M. N. H. (2010b). Robust Principal Component Analysis Based on Robust Estimation of Multivariate Normal Distribution. International Journal of Statistical Science (IJSS), Vol.10, pp. 19-35.

[27] Mollah, M. N. H. (2010c). Robust Image Processing by β-Prewhitening Based FastICA algorithm. Int. Journal of Tomography and Statistics (IJTS), Vol. 13, pp. 126-136.

[28] Mollah, M. H., Mollah, M. N. H. and Kishino, H. (2012). β-Empirical Bayes inference and model diagnosis of microarray data . BMC Bioinformatics, 13:135.

[29] Maronna, R.A. and Zamar, R.H. (2002). Robust estimates of location and dispersion for high dimensional data sets. Technometrics, 44:307-317.

[30] Nowak, G. and Tibshirani, R. (2008). Complementary hierarchical clustering, Biostatistics, 9, 467-483.

[31] Rahman, M. M. (2012). Robustification of Bayes Classifier and Its Application for Gene Expression Data Analysis. Unpublished M.Sc. Thesis, Dept. of Statistics, University of Rajshahi, Bangladesh.

[32] Rahman, M. M. and Mollah, M. H. (2018). Robustification of Gaussian Bayes Classifier by the Minimum β-Divergence Method. Accepted for publication in the Journal of Classification. Springer.

[33] Rousseeuw, P.J. (1985). Multivariate estimation with high breakdown point. In W. Grossmann, G. Pug, I. Vincze, and W. Wertz, editors, Mathematical Statistics and Applications, Vol. B, pages 283-297, Dordrecht, 1985. Reidel Publishing Company.

[34] Rousseeuw, P.J. and Van Driessen, K. (1999). A fast algorithm for the Minimum Covariance Determinant estimator. Technometrics, 41:212-223.

[35] Singha, A. C. (2013). Statistical Phylogenetic Modeling and Its Application for DNA and Protein Sequence Analysis. Unpublished M.Sc Thesis, Dept. of Statistics, University of Rajshahi, Bangladesh.

[36] Singha, A. C., Ahmed, M. S., Rana, M. M., Ahsan, M. A. and Mollah, M.N.H. (2014). Robust Phylogenetic Canonical Correlation Analysis. International Conference on Applied Statistics (ICAS), 26-28, Dec., 2014, ISRT, University of Dhaka, Bangladesh.