COMPARISON OF DISTANCE-BASED CLASSIFIERS FOR ELLIPTICALLY CONTOURED DISTRIBUTIONS

Submitted by
Mahbuba Haque

A thesis submitted to the Department of Statistics in partial fulfillment of the requirements for a two-year Master degree in Statistics in the Faculty of Social Sciences

Supervisor
Rauf Ahmad

Spring, 2017
A simulation study is carried out to compare three distance-based classifiers for their misclassification and asymptotic distributions when the data follow certain elliptically contoured distributions. The data are generated from multivariate normal, multivariate $t$ and multivariate normal mixture distributions with varying covariance structures, sample sizes and dimension sizes. In many of the simulated cases, the dimensions of the data are much larger than the sample size. The simulations show that for small dimension sizes, the centroid classifier generally performs better. The nearest neighbour classifier shows superior performance compared to the other classifiers when the covariance structure is of compound symmetry form. All three classifiers showed to have asymptotic normal distribution, regardless of the underlying distribution of the data.
1 Introduction

In multivariate analyses, many of the techniques used for modeling data rely on the assumption that the data follow a multivariate normal distribution. When the assumption of normality does not hold, it can be of interest to relax the normality assumption. A competing family of distributions is the elliptically contoured distributions, including the multivariate normal. Two other, most frequently used non-normal distributions in the family, are the multivariate $t$ distribution and the multivariate normal mixture distribution. The multivariate $t$ distribution can offer flexibility when modeling data following a longer and fatter tailed distribution, relative to the multivariate normal, which is more realistic in many cases.

This thesis deals with distance-based classifiers when the data come from elliptically contoured distributions. In classification, the goal is to classify an observation into one of several prespecified classes (Anderson, 2003). Classifiers are evaluated based on their misclassification rates and convergence to asymptotic normality under a variety of parameters, i.e. sample size $n$, dimension size $p$ and covariance structure $\Sigma$.

Section 2 begins with definitions of elliptically contoured distributions and the distance-based classifiers evaluated in this study. The simulation design is described in Sec. 3. Results and discussion of the simulations are presented in Sec. 4 and a conclusion of the results is reserved for Sec. 5.
2 Method

2.1 Elliptically Contoured Distributions

A $p$-dimensional random vector $\mathbf{X} = (X_1, ..., X_p)'$ is said to follow a general elliptically contoured distribution (ECD) if it has a density of the form

$$f_{\mathbf{X}}(\mathbf{x}) = c_m|\Sigma|^{-1/2}h((\mathbf{x} - \mu)'\Sigma^{-1}(\mathbf{x} - \mu))$$  

for some function $h \in \mathbb{R}$, where $\mu \in \mathbb{R}^p$ is the mean vector and $\Sigma \in \mathbb{R}^{p \times p}$ is the covariance matrix which is to be assumed positive definite ($\Sigma > 0$).

From the kernel of the pdf in Eqn. (1), it follows that, the ECDs can be characterized by their constant elliptical contours (Muirhead, 1982). The EC class consists of many useful distributions, including multivariate normal (MVN). Two other commonly used distributions from this class are the multivariate $t$ (MVT) and the multivariate normal mixture (MVNM).

A random vector $\mathbf{X}$ follows a MVN, MVT or MVNM distribution if the pdf in (1) reduces to the following forms, respectively,

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}}e^{-(\mathbf{x} - \mu)'\Sigma^{-1}(\mathbf{x} - \mu)/2}$$  

$$= \frac{\Gamma((\nu + p)/2)}{\Gamma(\nu/2)(\nu\pi)^{p/2}|\Sigma|^{1/2}} \left[ 1 + \frac{1}{\nu}(\mathbf{x} - \mu)'\Sigma^{-1}(\mathbf{x} - \mu) \right]^{-(\nu+p)/2}$$  

$$= (1 - \epsilon)\mathcal{N}(\mu_1, \Sigma_1) + \epsilon\mathcal{N}(\mu_2, \Sigma_2)$$

where $-\infty < x_i < \infty, \nu$ is the degrees of freedom and $\epsilon$ is the level of contamination with restriction $0 \leq \epsilon \leq 1$.

These distributions will be denoted as $\mathbf{X} \sim \mathcal{N}_p(\mu, \Sigma)$, $\mathbf{X} \sim t_\nu(\mu, \Sigma)$ and $\mathbf{X} \sim MVNM(\mu_i, \Sigma_i, \epsilon)$ (Muirhead, 1982).
2.2 Distance-Based Classifiers

Let \( x_{ki} \) be \( p \)-vectors drawn from \( \pi_i, \ i = 1, 2 \) \( k = 1, \ldots, n_i \) and \( n_i \) be the sample size of each class. Let \( x_0 \) be a new observation that is to be classified to one of these two classes using this rule. A distance-based classifier is the one that discriminates and classifies based on the distance between the observation to be classified and the data from the populations being considered (McLachlan, 2012).

The classical likelihood-based linear discriminant function for two normal populations, assuming equal priors and costs, is

\[
T = (\bar{x}_1 - \bar{x}_2)' \Sigma^{-1} x_0 - \frac{1}{2} (\bar{x}_1' \Sigma^{-1} \bar{x}_1 - \bar{x}_2' \Sigma^{-1} \bar{x}_2)
\]  

(5)

where \( \bar{x}_i = n_i^{-1} \sum_{k=1}^{n_i} x_{ki}, \ i = 1, 2, \) \( \Sigma \) is a pooled estimator of common \( \Sigma \) and \( \hat{\Sigma}_i = \sum_{k=1}^{n_i} (x_{ki} - \bar{x}_i)(x_{ki} - \bar{x}_i)' / (n_i - 1) \) is a pooled estimator of \( \Sigma \) and \( d_i = a_i (x_{ki} - \bar{x}_i)(x_{ki} - \bar{x}_i)' \) is the quadratic form of \( x_{ki} \) from \( x_{li} \) and \( a_i = \frac{1}{n_i} (n_i - 1) \).

This classifier is not applicable when data are high-dimensional relative to the sample size, since \( \hat{\Sigma} \) is singular in that case. One possible modification is to remove \( \hat{\Sigma} \) from the classifier. This results into a biased term consisting of the trace function of the unknown covariance matrix. Chan and Hall (2009) suggested an adjustment for the bias term and called the resulting classifier scale-adjusted classifier. This adjustment is composed of

\[
d_i = a_i \sum_{k=1}^{n_i} \sum_{l=1}^{n_i} (x_{ki} - x_{li})'(x_{ki} - x_{li})
\]  

(6)

where \( d_i \) denotes the average variability of \( x_{ki}, \ i = 1, 2, \) \( k = 1, \ldots, n_i, \) \( (x_{ki} - x_{li})'(x_{ki} - x_{li}) \) is the quadratic form of \( x_{ki} \) from \( x_{li} \) and \( a_i = \frac{1}{n_i(n_i - 1)} \).

We chose three distance-based classifiers for evaluation, specifically the average distance classifier, nearest neighbour classifier and the centroid method defined as

\[
T_{\text{avg}} = \frac{1}{n_1} \sum_{k=1}^{n_1} q_1 - \frac{1}{n_2} \sum_{k=1}^{n_2} q_2 - B
\]  

(7)

\[
T_{\text{nn}} = \min_{k=1,\ldots,n_1} q_1 - \min_{k=1,\ldots,n_2} q_2 - B
\]  

(8)

\[
T_{\text{cen}} = r_1 - r_2 - B
\]  

(9)

where \( q_i = (x_0 - x_{ki})'(x_0 - x_{ki}) \), \( k = 1, \ldots, n_i, \ i = 1, 2, \) \( r_i = (x_0 - \bar{x}_i)'(x_0 - \bar{x}_i) \) and \( B = \sum_{i=1}^{2} d_i / n_i \).
The average distance classifier ($T_{\text{avg}}$) calculates the average squared distance between $x_0$ and the observations from $\pi_1$ and compares it to the average squared distance between $x_0$ and the observations from $\pi_2$. $x_0$ is allocated to the population that is closest to $x_0$. The nearest neighbour classifier ($T_{\text{nn}}$) compares the squared distance from $x_0$ to the observation closest to $x_0$ from $\pi_1$ with the squared distance from $x_0$ to the observation closest to $x_0$ from $\pi_2$. The centroid method ($T_{\text{cen}}$) compares the squared distance from $x_0$ to the average of the observations from $\pi_1$ with the same distance but with the observations now coming from $\pi_2$.

Let $T$ be any of the three classifiers $T_{\text{avg}}$, $T_{\text{nn}}$ or $T_{\text{cen}}$. The classification rule for all three classifiers can then be stated as follows.

Rule: Classify $x_0$ into $\pi_2$ if $T > 0$, and into $\pi_1$ otherwise.
3 Simulation Design

To evaluate the performance of the three classifiers in Eqn. (7) - (9), a simulation study is performed with data generated from the three elliptical distributions presented in Sec. 2.1. The aim of the simulation study is to evaluate the misclassification rates and the asymptotic distribution. The results presented in Sec. 4 are averages over 1000 simulation runs.

The sample sizes $n = \{10, 20, 50, 100\}$ are used to see how the classifiers perform for different sample sizes. To evaluate the impact dimension size has on performance, data of dimension sizes $p = \{5, 10, 20, 50, 100, 300, 500, 1000\}$ are generated. Four different covariance structures are evaluated in this study. The first covariance structure represents the most basic structure where the covariance matrix has variance one and no correlation between the observations. For the second covariance structure, the population variances are allowed to differ. The compound symmetry (CS) covariance structure allows for correlation between the observations but restricts the variances and covariances to be constant. The structure is defined as $\text{Cov}_i(X_k, X_l) = \sigma_i + \sigma^2 I(k = l)$. A slightly more flexible structure is the first order autoregressive covariance (AR(1)) structure, which has homogeneous variances and where the correlations decrease as the time lag increases. The AR(1) structure is defined as $\text{Cov}_i(X_k, X_l) = \sigma^2 \rho_i |k - l|$.

For the simulations, a 2-fold cross-validation is implemented where the data set is split into two parts, a training sample and a test sample. The training sample is used to train the classifier, which is used to classify each observation in the test sample. The roles of training sample and test sample are then reversed and the same procedure is repeated.

The value of a classifier is usually based on its misclassification rate and asymptotic distribution. As the true misclassification rate can not be computed, since the parameters are unknown, a simple formula is commonly preferred. It is called apparent error rate (APER) and is computed as seen in table 1. Table 1 shows how observations are divided into correctly classified and misclassified groups. Observations from $\pi_1$ wrongly classified as belonging to $\pi_2$ are grouped into $m_1$, and wrongly classified observations from $\pi_2$ into $\pi_1$ are denoted $m_2$. 

5
Table 1: Computation of APER

<table>
<thead>
<tr>
<th>Predicted membership</th>
<th>$\pi_1$</th>
<th>$\pi_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actual membership</td>
<td>$\pi_1$</td>
<td>$n_1 - m_1$</td>
</tr>
<tr>
<td></td>
<td>$\pi_2$</td>
<td>$m_2$</td>
</tr>
</tbody>
</table>

where $m_i, i = 1, 2$, denotes the number of misclassified observations of $n_i$.

For asymptotic distribution, we evaluate the distributional convergence of a classifier to normality, i.e

$$\frac{T - E(T)}{\sqrt{Var(T)}} \overset{D}{\to} N(0, 1) \text{ as } n, p \to \infty$$

(10)

where $T$ is any of the three classifiers in Eqns. (7) - (9), $E(T)$ denotes the expected value of the classifier, $Var(T)$ denotes the variance and $N(0, 1)$ denotes the standard normal distribution. The objective is to evaluate each $T$ for its convergence in Eqn. (10) for $n, p \to \infty$. 
4 Results and Discussion

4.1 Misclassification Rates of the Classifiers

Figure 1 shows the APER of the three classifiers when data are generated from $\mathcal{N}_p(\mu_i, \Sigma_i)$ where $\mu_1 = 0$ and $\mu_2$ has half of the elements set to 0 and the other half set to 1. $\Sigma_i$ has AR(1) structure, with parameters set to $\sigma^2 = 1$, $\rho_1 = 0.3$ and $\rho_2 = 0.7$. The second row of figure 1 shows the APER of the three classifiers but with data generated from $t_\nu(\mu_i, \Sigma_i)$, with $\nu = 10$ and $\mu_i, \Sigma_i$ set exactly as for MVN on the row above. On the first row, the nearest neighbour classifier and the centroid classifier show consistent rate of misclassification regardless of sample size ($n = \{10, 20, 50, 100\}$). The APER converges to 0 for the centroid classifier at approximately $p = 300$, while it converges to 0 for the nearest neighbour classifier at $p = 500$. The average distance classifier converges to the centroid classifier as the sample size increases and completely overlaps when $n = 100$. When data are generated from MVT, the centroid classifier clearly outperforms both the average distance classifier and the nearest neighbour classifier for small sample sizes. The average distance and nearest neighbour classifiers do not perform much better as the dimension increases when the sample size is small. As can be seen from the eight graphs in figure 1, the centroid classifier performs very similar regardless of sample size or whether the data come from MVN or MVT.

Data are generated from $\mathcal{N}_p(\mu_i, \Sigma_i)$ and $t_\nu(\mu_i, \Sigma_i)$ where the mean vectors are set as for the populations in figure 1. The covariances matrices are of compound symmetry structure, with parameters set to $\sigma^2 = 1$, $\sigma_1 = 0.3$ and $\sigma_2 = 0.7$. Figure 2 shows, on the first row, the APER of the three classifiers where the data come from MVN, and on the second row the APER is shown for the MVT generated data. For both rows, the centroid and nearest neighbour classifiers perform quite similar for all sample sizes. The average distance classifier shows major improvement when the sample size increases and has an APER close to zero for $n > 50$, high dimensions and when data are from MVN.
Figure 1: The first (second) row shows APER of the three classifiers with data from MVN (MVT) with AR(1) covariance structure where $\sigma^2 = 1, \rho_1 = 0.3$ for $\pi_1, \sigma^2 = 1, \rho_2 = 0.7$ for $\pi_2$ and $n = 10, 20, 50, 100$ (L to R each row).
Figure 2: The first (second) row shows APER of the three classifiers with data from MVN (MVT) with CS covariance structure where $\sigma^2 = 1$, $\sigma_1 = 0.3$ for $\pi_1$, $\sigma^2 = 1$, $\sigma_2 = 0.7$ for $\pi_2$ and $n = 10, 20, 50, 100$ (L to R each row).
Figure 3 shows APER of classifiers when data are generated from a distribution with identity matrix as its covariance structure and another distribution which has an AR(1) covariance structure. The parameters of the AR(1) covariance structure are $\sigma^2 = 1$ and $\rho = 0.7$. The results are very similar to the results in figure 1, where both distributions have AR(1) covariance structure. The differences in APER between MVN and MVT in figure 3 are almost completely eliminated when $n > 50$. Figure 4 shows misclassification rates for data generated from a distribution with CS covariance structure and a distribution with identity covariance matrix. The parameters of the CS covariance structure are set to $\sigma^2 = 1$ and $\sigma_2 = 0.7$. Compared to figure 2, the nearest neighbour and centroid classifiers perform better while the average distance classifier still shows relatively poor performance.

The two populations where data are sampled from have parameters set to $\mu_1 = 0$, $\Sigma_1 = 2I$ and $\mu_2 = 1$, $\Sigma_2 = 4I$. In figure 5 the centroid classifier has low APER for data generated from both MVN and MVT. The other two classifiers show extremely poor performance regardless of the distribution of the data. For the MVT data, on the second row of figure 5, the nearest neighbour classifier gets better as sample size increases.
Figure 3: The first (second) row shows APER of the three classifiers with data from MVN (MVT) where $\Sigma_1 = I$ for $\pi_1$, AR(1) covariance structure for $\pi_2$ with parameters $\sigma^2 = 1, \rho = 0.7$ and $n = 10, 20, 50, 100$ (L to R each row).
Figure 4: The first (second) row shows APER of the three classifiers with data from MVN (MVT) where $\Sigma_1 = I$ for $\pi_1$, CS covariance structure for $\pi_2$ with parameters $\sigma^2 = 1, \sigma^2 = 0.7$ and $n = 10, 20, 50, 100$ (L to R each row).
Figure 5: The first (second) row shows APER of the three classifiers with data from MVN (MVT) where $\Sigma_1 = 2I$ for $\pi_1$, $\Sigma_1 = 4I$ for $\pi_2$ and $n = 10, 20, 50, 100$ (L to R each row).
Data are generated from two multivariate normal mixture distributions, specified as in Eqn. 4. The first distribution has mean vectors set to $\mu_1 = \mu_2 = 0$ and AR(1) covariance structure with parameters $\sigma^2 = 1, \rho_1 = 0.3$ and $\rho_2 = 0.7$. The second distribution has mean vectors set to $\mu_1 = \mu_2 = 1$ and CS covariance structure with $\sigma^2 = 1, \sigma_1 = 0.3$ and $\sigma_2 = 0.7$. The level of contamination, $\epsilon$, is set to 0.1 for both distributions. Figure 6 shows that the APER do not change as sample size increases for data generated from these two specified MVNM distributions. All three classifiers show misclassification rates between approximately 10\%–15\%, and while the centroid and average distance classifier stabilizes at $p = 20$, the nearest neighbour classifier shows improvement as the dimension increases.

Figure 6: The figure shows APER of the three classifiers with data from MVNM with AR(1) covariance structure for $\pi_1$ with parameters $\sigma^2 = 1, \rho_1 = 0.3, \rho_2 = 0.7$, CS covariance structure for $\pi_2$ with parameters $\sigma^2 = 1, \sigma_1 = 0.3, \sigma_2 = 0.7$ and $n = 10, 20, 50, 100$ (top L to bottom R).
Figure 7 shows misclassification rates for multivariate normal mixture generated data where one population has a CS covariance structure with $\sigma^2 = 1, \sigma_1 = 0.3$ and $\sigma_2 = 0.7$ and the other has diagonal covariance structure with $\Sigma_1 = 1I$ and $\Sigma_2 = 2I$. The level of contaminatio, $\epsilon$ is set to 0.1 for both distributions. Just as for the case in figure 6, the classifiers reach their minimum misclassification rate at small sample sizes and low dimensions.

Figure 7: The figure shows APER of the three classifiers with data from MVNM with CS covariance structure for $\pi_1$ with parameters $\sigma^2 = 1, \sigma_1 = 0.3, \sigma_2 = 0.7, \Sigma_1 = 1I, \Sigma_2 = 2I$ for $\pi_2$ and $n = 10, 20, 50, 100$ (top L to bottom R).
4.2 Asymptotic Distribution of the Classifiers

The second property investigated is the asymptotic distribution of the classifiers. In figures 8 - 10 are the asymptotic distribution of the three classifiers for AR(1) covariance structure. The standard normal density curve is added in each graph for comparison. Both MVN and MVT generated data show very similar asymptotic distribution for both $p = 100$ and $p = 500$. Figures 11-16 in the Appendix show the asymptotic results for the classifiers under CS and diagonal covariance structures.

Figure 8: Asymptotic distribution of the average distance classifier for MVN (graph 1 and 2) and MVT (graph 3 and 4) for $p = 100, 500, 100, 500$ (L to R) with AR(1) covariance structure with parameters $\sigma^2 = 1, \rho_1 = 0.3$ and $\rho_2 = 0.7$. 


Figure 9: Asymptotic distribution of the nearest neighbour classifier for MVN (graph 1 and 2) and MVT (graph 3 and 4) for $p = 100, 500, 100, 500$ (L to R) with AR(1) covariance structure with parameters $\sigma^2 = 1, \rho_1 = 0.3$ and $\rho_2 = 0.7$.

Figure 10: Asymptotic distribution of the centroid classifier for MVN (graph 1 and 2) and MVT (graph 3 and 4) for $p = 100, 500, 100, 500$ (L to R) with AR(1) covariance structure with parameters $\sigma^2 = 1, \rho_1 = 0.3$ and $\rho_2 = 0.7$. 
5 Conclusion

The simulation study showed that the centroid classifier generally had a lower misclassification rate than the other two classifiers. For a given covariance structure, the APER of the centroid classifier stayed the same regardless of sample size and distribution of the data. However, for data with CS covariance structure, the nearest neighbour classifier greatly outperformed the other two classifiers for larger sample sizes. The ability of the classifiers to correctly classify observations did not differ much between data from MVN and MVT. Only minor differences were detected, and that was for small sample sizes \((n = 10)\). For MVNM generated data, the APER reached its minimum value for dimension sizes as small as \(p = 20\) and it also did not change as sample sizes increased. The asymptotic distribution of all classifiers, for all covariance structures investigated, showed to be approximately normal.
References


Figure 11: Asymptotic distribution of the average distance classifier for MVN (graph 1 and 2) and MVT (graph 3 and 4) for $p = 100, 500, 100, 500$ (L to R) with CS covariance structure with parameters $\sigma^2 = 1, \sigma_1 = 0.3$ and $\sigma_2 = 0.7$.

Figure 12: Asymptotic distribution of the nearest neighbour classifier for MVN (graph 1 and 2) and MVT (graph 3 and 4) for $p = 100, 500, 100, 500$ (L to R) with CS covariance structure with parameters $\sigma^2 = 1, \sigma_1 = 0.3$ and $\sigma_2 = 0.7$. 
Figure 13: Asymptotic distribution of the centroid classifier for MVN (graph 1 and 2) and MVT (graph 3 and 4) for $p = 100, 500, 100, 500$ (L to R) with CS covariance structure with parameters $\sigma^2 = 1, \sigma_1 = 0.3$ and $\sigma_2 = 0.7$.

Figure 14: Asymptotic distribution of the average distance classifier for MVN (graph 1 and 2) and MVT (graph 3 and 4) for $p = 100, 500, 100, 500$ (L to R) where $\Sigma_1 = 2I$ for $\pi_1$ and $\Sigma_1 = 4I$ for $\pi_2$. 
Figure 15: Asymptotic distribution of the nearest neighbour classifier for MVN (graph 1 and 2) and MVT (graph 3 and 4) for $p = 100, 500, 100, 500$ (L to R) where $\Sigma_1 = 2I$ for $\pi_1$ and $\Sigma_1 = 4I$ for $\pi_2$.

Figure 16: Asymptotic distribution of the centroid classifier for MVN (graph 1 and 2) and MVT (graph 3 and 4) for $p = 100, 500, 100, 500$ (L to R) where $\Sigma_1 = 2I$ for $\pi_1$ and $\Sigma_1 = 4I$ for $\pi_2$. 

22