Mahalanobis' Distance : A Brief History and Some Observations

Probal Chaudhuri

Indian Statistical Institute, Kolkata

(Based on Joint Work with Subhajit Dutta of IIT-Kanpur and Anil K. Ghosh of ISI-Kolkata)

Conference in Honour of Rajeeva L. Karandikar
March 30, 2022

Brief History

- Fisher, R. A. (1936), 'The use of multiple measurements in taxonomic problems', *Ann. Eugenics*, **7**, 179–188.
- Fisher was interested in the taxonomic classification of different species of Iris. He had measurements on the lengths and the widths of the sepals and the petals of the flowers of three different species, namely, *Iris setosa, Iris virginica* and *Iris versicolor*.

Brief History (Contd.)

- Mahalanobis, P. C. (1936), 'On the generalized distance in statistics', *Proc. Nat. Acad. Sci., India*, **12**, 49–55.
- Mahalanobis met Nelson Annandale at the 1920 Nagpur session of the Indian Science Congress. Annandale asked Mahalanobis to analyze anthropometric measurements of Anglo-Indians in Calcutta. This eventually led to the development of Mahalanobis' distance.

Brief History (Contd.)

- Mahalanobis' distance of an observation $\mathbf{x}$ from a population with mean μ and dispersion $\mathbf{\Sigma}$:

$$D^2(\mathbf{x}, \mu, \mathbf{\Sigma}) = (\mathbf{x} - \mu)'\mathbf{\Sigma}^{-1}(\mathbf{x} - \mu).$$

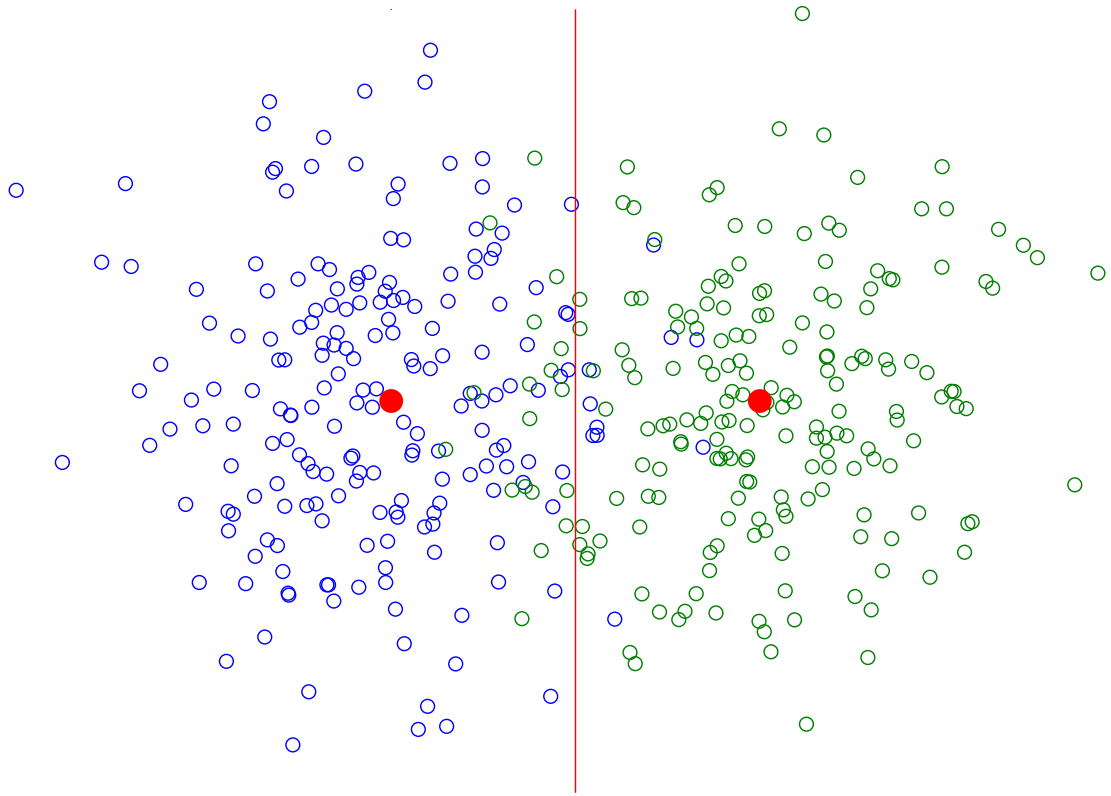
- Fisher's linear discriminant function for two populations with means μ_1 and μ_2 and a common dispersion $\mathbf{\Sigma}$:

$$(\mathbf{x} - (\mu_1 + \mu_2)/2)'\mathbf{\Sigma}^{-1}(\mu_1 - \mu_2).$$

Brief History (Contd.)

- Fisher's linear discriminant function corresponds to the separating hyperplane that separates the points, which are closer in Mahalanobis' distance to one population from the points, which are closer to the other population.
- Fisher's choice of his linear discriminant function was motivated by the fact that this linear function maximizes Fisher's two-sample *t*-statistic computed from linear functions of the data.

Brief History (Contd.)



Brief History (Contd.)

- Rao, C. R. (1948), ‘The utilization of multiple measurements in problems of biological classification’, *JRSS–B*, **10**, 159–203.

Rao considered the problem of classifying human skulls recovered in archaeological excavation into Iron Age or Bronze Age

- The linear discriminant function has Bayes risk optimality for Gaussian class distributions, which differ in their locations but have the same dispersion.
- In fact, the Bayes risk optimality holds for elliptically symmetric and unimodal class distributions, which differ only in their locations.

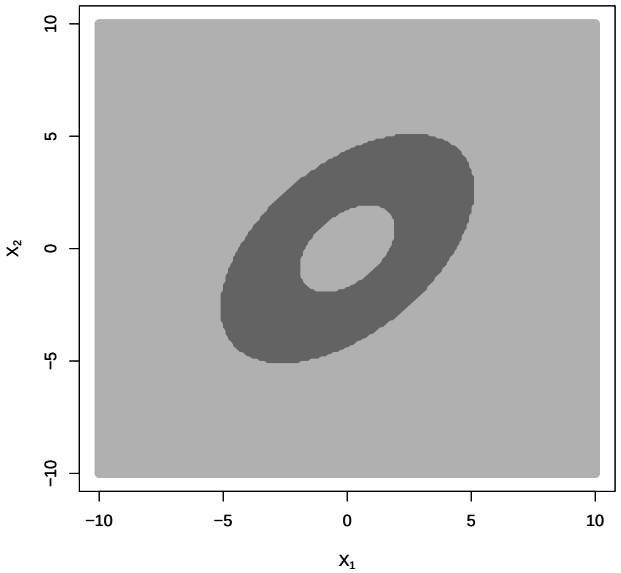
Examples

- Example **(a)** : Class 1 : Mixture of $N_d(\mathbf{0}, \mathbf{\Sigma})$ and $N_d(\mathbf{0}, 10\mathbf{\Sigma})$; and Class 2 : $N_d(\mathbf{0}, 5\mathbf{\Sigma})$.
  $\mathbf{\Sigma} = [0.5\mathbf{1}'\mathbf{1} + 0.5\mathbf{I_d}]$, N_d is the d-variate normal distribution.

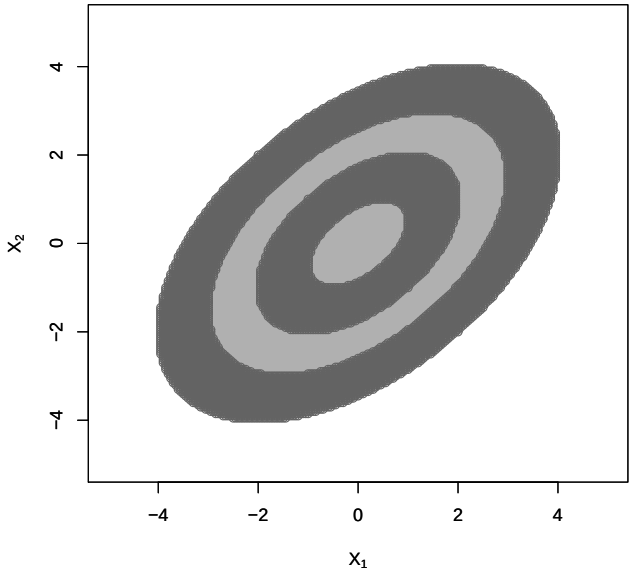
- Example **(b)** : Class 1 : Mixture of $U_d(\mathbf{0}, \mathbf{\Sigma}, 0, 1)$ and $U_d(\mathbf{0}, \mathbf{\Sigma}, 2, 3)$; and Class 2 : $U_d(\mathbf{0}, \mathbf{\Sigma}, 1, 2)$ and $U_d(\mathbf{0}, \mathbf{\Sigma}, 3, 4)$. $U_d(\boldsymbol{\mu}, \mathbf{\Sigma}, r_1, r_2)$ denotes the uniform distribution over the region $\{\mathbf{x} \in \mathbb{R}^d : r_1 < \|\mathbf{\Sigma}^{-1/2}(\mathbf{x} - \boldsymbol{\mu})\| < r_2\}$.

- Classes have same location **0** but different scatters and shapes.

Bayes Class Boundaries



(a) Example **(a)**



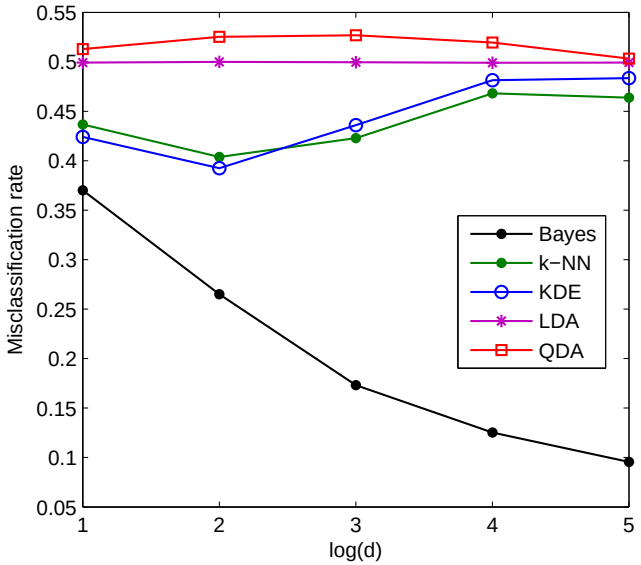
(b) Example **(b)**

Figure: Bayes class boundaries in $\mathbb{R}^2$.

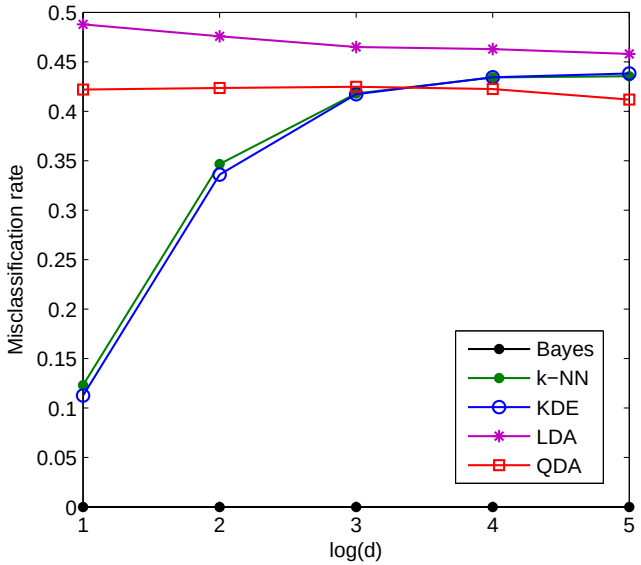
Bayes class boundaries (Contd.)

- Class distributions involve elliptically symmetric distributions.
- They have same location (i.e., **0**), and they differ in their scatters as well as shapes.
- No linear or quadratic classifier will work here as they will fail to capture the Bayes class boundaries!!

Performance of some standard classifiers



(a) Example **(a)**



(b) Example **(b)**

Figure: Misclassification rates of LDA, QDA, two nonparametric classifiers and the Bayes classifier for $d = 2, 5, 10, 20, 50$ and 100.

Elliptic Distributions

- Suppose that the class densities are elliptically symmetric

$$f_i(\mathbf{x}) = |\mathbf{\Sigma}_i|^{-1/2} g_i(\|\mathbf{\Sigma}_i^{-1/2}(\mathbf{x} - \boldsymbol{\mu}_i)\|)$$

$$= \psi_i(\mathrm{MD}(\mathbf{x}, \boldsymbol{\mu}_i, \mathbf{\Sigma}_i)) \text{ for all } i = 1, 2.$$

- The class posterior probabilities are

$$p(1|\mathbf{x}) = \pi_1 f_1(\mathbf{x}) / (\pi_1 f_1(\mathbf{x}) + \pi_2 f_2(\mathbf{x}))$$

and

$$p(2|\mathbf{x}) = 1 - p(1|\mathbf{x}).$$

- It is easy to see that

$$\log\{p(1|\mathbf{x})/p(2|\mathbf{x})\} = \log(\pi_1 f_1(\mathbf{x})/\pi_2 f_2(\mathbf{x})) =$$

$$\log(\pi_1/\pi_2) + \log \psi_1(\mathrm{MD}(\mathbf{x}, \boldsymbol{\mu}_1, \mathbf{\Sigma}_1)) - \log \psi_2(\mathrm{MD}(\mathbf{x}, \boldsymbol{\mu}_2, \mathbf{\Sigma}_2)).$$

Elliptic Distributions (Contd.)

- The posteriors turn out to be of the form

$$p(1|\mathbf{x}) = p(1|\mathbf{z}(\mathbf{x}))$$

$$= \frac{\exp(\log \psi_1(z_1(\mathbf{x})) - \log \psi_2(z_2(\mathbf{x}))}{[1 + \exp(\log \psi_1(z_1(\mathbf{x})) - \log \psi_2(z_2(\mathbf{x}))]},$$

$$p(2|\mathbf{x}) = p(2|\mathbf{z}(\mathbf{x}))$$

$$= \frac{1}{[1 + \exp(\log \psi_1(z_1(\mathbf{x})) - \log \psi_2(z_2(\mathbf{x}))]},$$

where $\mathbf{z}(\mathbf{x}) = (z_1(\mathbf{x}), z_2(\mathbf{x})) = (\mathrm{MD}(\mathbf{x}, \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1), \mathrm{MD}(\mathbf{x}, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2))$.

- The posterior probabilities satisfy a generalized additive model (Hastie and Tibshirani, 1990).

Gaussian Distributions

- If we have two normal populations $\mathrm{N}_d(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$ and $\mathrm{N}_d(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$, we get linear logistic regression model for the posterior probabilities. This is related to Fisher's linear discriminant analysis.

- If we have two normal populations $\mathrm{N}_d(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and $\mathrm{N}_d(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$, we get quadratic logistic regression model for the posterior probabilities. This is related to quadratic discriminant analysis.

Nonparametric Multinomial Additive Logistic Regression Model

- For any $1 \leq i \leq (J-1)$, it is easy to see that

$$\log\{p(i|\mathbf{x})/p(J|\mathbf{x})\} =$$

$$\log(\pi_i/\pi_J) + \log\psi_i(\mathrm{MD}(\mathbf{x}, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)) - \log\psi_J(\mathrm{MD}(\mathbf{x}, \boldsymbol{\mu}_J, \boldsymbol{\Sigma}_J)),$$

where $p(i|\mathbf{x})$ is the posterior probability of the i-th class.

- For any $1 \leq i \leq (J-1)$, the posteriors are of the form

$$p(i|\mathbf{x}) = p(i|\mathbf{z}(\mathbf{x})) = \frac{\exp(\Phi_i(\mathbf{z}(\mathbf{x})))}{[1 + \sum_{k=1}^{(J-1)} \exp(\Phi_k(\mathbf{z}(\mathbf{x})))]},$$

$$p(J|\mathbf{x}) = p(J|\mathbf{z}(\mathbf{x})) = \frac{1}{[1 + \sum_{k=1}^{(J-1)} \exp(\Phi_k(\mathbf{z}(\mathbf{x})))]},$$

where $\mathbf{z}(\mathbf{x}) = (\mathrm{MD}(\mathbf{x}, \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1), \cdots, \mathrm{MD}(\mathbf{x}, \boldsymbol{\mu}_J, \boldsymbol{\Sigma}_J))$.

Nonparametric Multinomial Additive Logistic Regression Model (Contd.)

- We replace the original feature variables by the Mahalanobis' distances from different classes.

$$\mathbf{x} \rightarrow \mathbf{z}(\mathbf{x}) = (\mathrm{MD}(\mathbf{x}, \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1), \cdots, \mathrm{MD}(\mathbf{x}, \boldsymbol{\mu}_J, \boldsymbol{\Sigma}_J)).$$

- One can use the backfitting algorithm (Hastie and Tibshirani, 1990) to estimate the posterior probabilities from the training data.

More general Class Distributions

- Non-elliptic class distributions.
- Multi-modal class distributions.
- Mixture models for class distributions.

Finite Mixture of Elliptically Symmetric Densities

- Assume

$$f_i(\mathbf{x}) = \sum_{k=1}^{R_i} \theta_{ik} |\mathbf{\Sigma}_{ik}|^{-1/2} g_{ik}(\|\mathbf{\Sigma}_{ik}^{-1/2}(\mathbf{x} - \boldsymbol{\mu}_{ik})\|),$$

  where θ_{ik}s are positive satisfying $\sum_{k=1}^{R_i} \theta_{ik} = 1$ for all $1 \leq i \leq J$.
- The posterior probability for the i-th class is

$$p(i|\mathbf{x}) = \sum_{r=1}^{R_i} p(c_{ir}|\mathbf{x}) \text{ for all } 1 \leq i \leq J,$$

  where c_{ir} denotes the r-th sub-class in the i-th class.
- The posterior probability $p(c_{ir}|\mathbf{x})$ satisfies a multinomial additive logistic regression model because the distribution of the sub-population c_{ir} is elliptically symmetric.

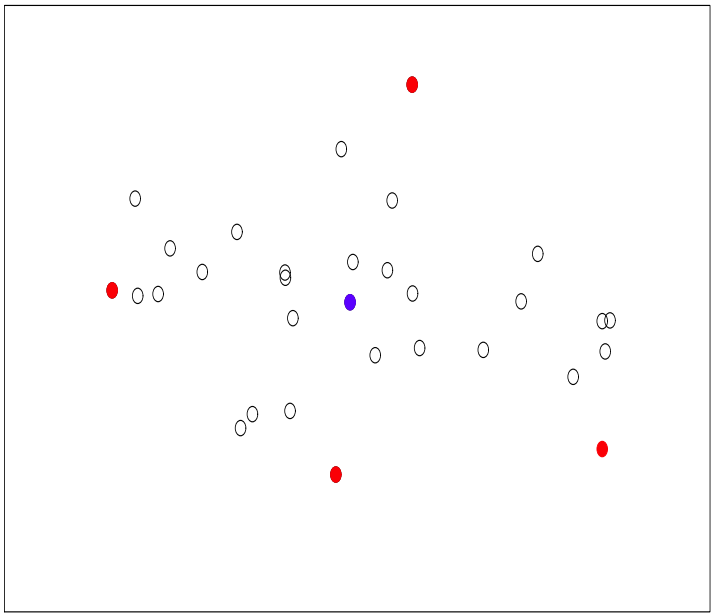
The Missing Data Problem

- In the training data, we have the class labels, but the sub-class labels are not available.
- If we had known the sub-class labels, we could once again use the backfitting algorithm to estimate the sub-class posteriors.
- Sub-class labels can be treated as missing observations. We can use an EM-type algorithm.

The Algorithm

- Initial E-step : Sub-class labels are estimated by appropriate cluster analysis of the training data in each class.
- Initial and later M-steps : Once the sub-class labels are obtained, sub-class posteriors can be estimated by fitting a nonparametric multinomial additive logistic regression model using the backfitting algorithm.
- Later E-steps : The sub-class labels are estimated by sub-class posterior probabilities.
- Iterations are carried out until posterior estimates stabilize.
- An observation is classified into the class having the largest posterior probability.

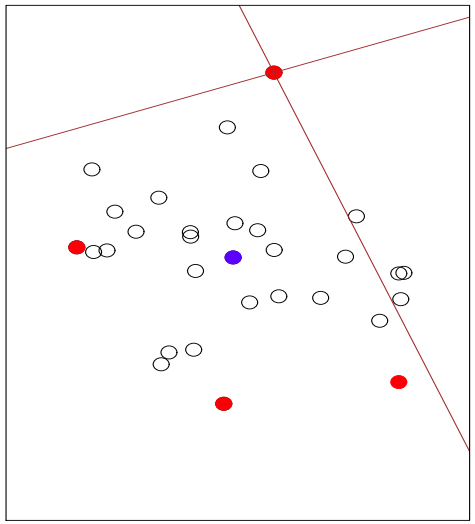
Data Depth

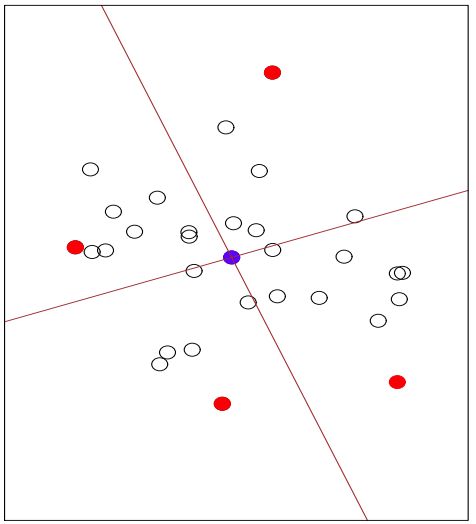


Data Depth (Contd.)

- Data depth of a point measures the relative position of that point in a given data cloud.
- It gives a center-outward ordering of the points relative to the data cloud.
- How do we measure the depth of a point?
- One approach is to consider lines through the point (hyper-planes for $d \geq 3$) and look at the proportion of data points lying on the two sides (half-spaces) of the line.

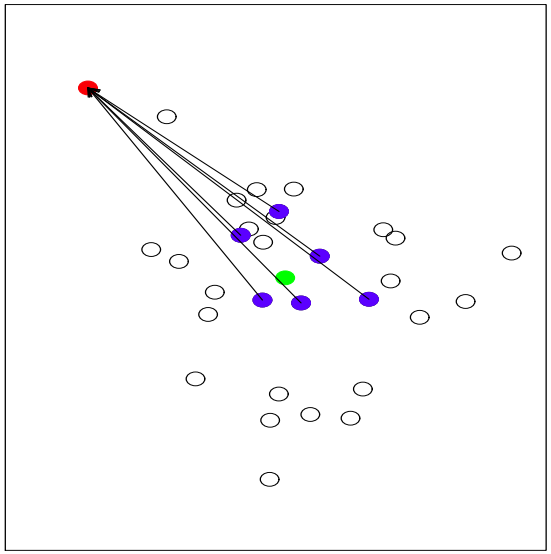
Half-Space Depth

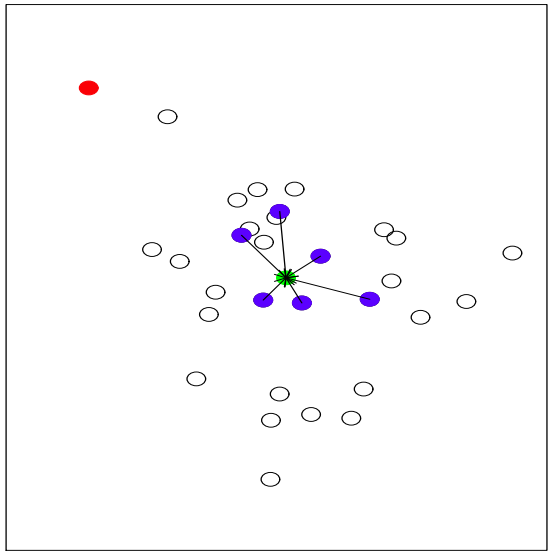




- We can take the minimum of the proportion of data points in any half-space of the line through a point after considering all the possible lines through it. This is called the half-space or Tukey depth. (Tukey, 1975).
- All the data points lie on one half-space of one line through the red point, while any line through the blue point has almost equal proportion of data points in both the half-spaces.

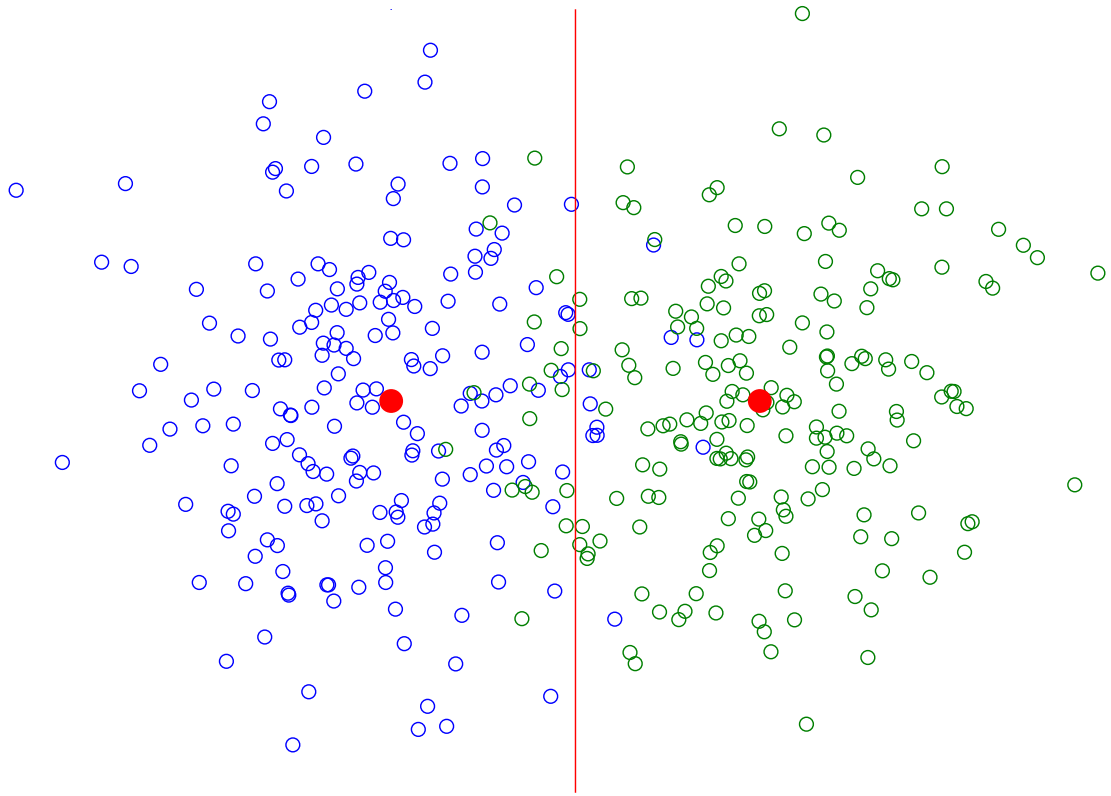
Spatial Depth





- If a point is “central” then the resultant of all the unit vectors from the sample points towards it will have norm close to zero. On the other hand, if a point is “very outlying”, this resultant will have a norm close to 1.

Data Depth and Classification



Data Depth and Classification (Contd.)

- For an elliptically symmetric and strongly unimodal distribution, the depth of a point relative to that distribution is a decreasing function of the Mahalanobis distance of that point from the centre of that distribution.

Classification in Infinite Dimensional Space

- Rao, C. R. and Varadarajan, V. S. (1963), 'Discrimination of Gaussian Processes', *Sankhyā*, **25**, 159–203.

Rao and Varadarajan considered the discrimination problem of Gaussian measures in Hilbert spaces. They considered Hellinger distance between two probability distributions

Classification in Infinite Dimensional Space (Contd.)

- Two Gaussian probabilities with positive definite covariances in finite dimensional spaces are mutually absolutely continuous.
- In infinite dimensional spaces, two Gaussian probabilities with positive definite covariance operators are either orthogonal or mutually absolutely continuous.
- Orthogonality implies perfect separation between populations. In a sense, the classification problem becomes “degenerate” and “trivial”.

Classification in Infinite Dimensional Space (Contd.)

- Two Gaussian probabilities with a common positive definite covariance $\mathbf{\Sigma}$ opertator are mutually absolutely continuous if and only if the difference between their means $(\mu_1 - \mu_2)$ lies in the range space of the square-root $\mathbf{\Sigma}^{1/2}$ of their covariance operator.

- Then Mahalanobis distance between the two distributions
  $D^2 = (\mu_1 - \mu_2)'\mathbf{\Sigma}^{-1}(\mu_1 - \mu_2) = \|\mathbf{\Sigma}^{-1/2}(\mu_1 - \mu_2)\|^2$ is a well defined monotonic function of the Hellinger distance.

- The classification problem is “non-degenerate” and “non-trivial”. The likelihood ratio is “well-behaved” and yields the Bayes classifier.