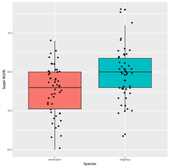


The figure illustrates the distribution of sepal widths from two species (‘versicolor’ and ‘virginica’), from the ‘Iris’ dataset in R.

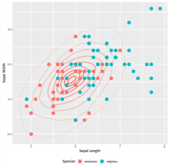
Wilcoxon rank sum test

Given two random samples $X_1, \dots, X_m$ and $Y_1, \dots, Y_n$, construct

$$W = \sum_j r(Y_j, \mathscr{D}_F \cup \mathscr{D}_G), \qquad r(Y, \mathscr{D}) = \sum_{Z \in \mathscr{D}} \mathbb{1}(Z \le Y).$$

This is *distribution free* under the null hypothesis that both samples have the same underlying distribution.

The two-sided Wilcoxon rank sum test gives a *p*-value of 0.005, hence we reject the null hypothesis that the true location shift is zero.



The red lines are spatial depth contours, drawn with reference to the ‘versicolor’ data.

A *depth function* quantifies how central a point $\boldsymbol{x}$ is with respect to a distribution F.

Points which are *more central* are said to be *deeper*.

This framework allows many rank based nonparametric techniques to be translated to a broader class of data, e.g. multivariate and functional data.

- Depth induces a *center-outwards* ordering of points.
- Contrast with the notion of rank which gives a *lowest-highest* ordering in the univariate setting.

- Hypothesis tests – Two sample quality index
- Exploratory data analysis – D-D plots

1. Inference procedures
   - Hypothesis tests
   - Rank tests
   - Multivariate quantiles
   - Confidence regions
2. Exploratory data analysis
3. Classification and clustering
4. Outlier detection

Depth contours

The *region of depth d* is defined by

$$\mathcal{R}(d, F) = \{\boldsymbol{x} \in \mathbb{R}^p \mid D(\boldsymbol{x}, F) \geq d\}.$$

The boundary $\partial\mathcal{R}(d, F)$ is called the *contour of depth d*.

Define
$$R(\boldsymbol{x}, F) = P(D(Y, F) \geq D(\boldsymbol{x}, F) \mid Y \sim F).$$

Then, as long as $D(\cdot, F)$ is continuous, the probability integral transform gives
$$R(X, F) \sim \mathsf{Uniform}[0, 1].$$

Liu, R.Y., Parelius, J.M, & Singh, K. (1999) Multivariate analysis by data depth: descriptive statistics, graphics and inference

- Depth contours are analogous to univariate quantiles.
- Sample points ordered with respect to their corresponding $R(X_j, \hat{F}_n)$ are analogous to order statistics.
- The distribution-free nature of $R(X, F)$ is analogous to how $F(X) \sim \mathsf{Uniform}[0, 1]$.

Why not use likelihood contours?

The 'Curse of Dimensionality'.

Additionally, consider a uniform distribution, say on a unit ball. This has non-trivial depth contours, but no proper density contours.

Depth-Depth plots

Let F, G be two distributions on $\mathbb{R}^p$, and let D be a depth function. We construct the D-D plot

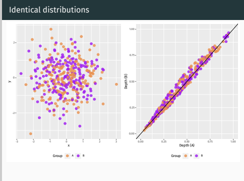
$$DD(F, G) = \{(D(\mathbf{x}, F),\ D(\mathbf{x}, G)) : \mathbf{x} \in \mathbb{R}^p\}.$$

Given data $\mathscr{D}_F, \mathscr{D}_G$, we may instead look at the D-D plot

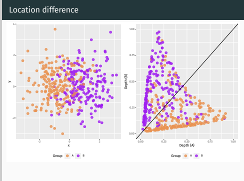
$$DD(\hat{F}_m, \hat{G}_n) = \left\{\left(D(\mathbf{x}, \hat{F}_m),\ D(\mathbf{x}, \hat{G}_n)\right) : \mathbf{x} \in \mathscr{D}_F \cup \mathscr{D}_G\right\}.$$

Liu, R.Y., Parelius, J.M, & Singh, K. (1999) Multivariate analysis by data depth: descriptive statistics, graphics and inference

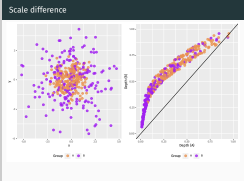
- The area of $DD(F, G)$ can serve as an affine-invariant measure of the discrepancy between F and G.
- The D-D plot gives an ℓ-variate visualization of ℓ groups of data regardless of what the original data looks like (multivariate, functional).



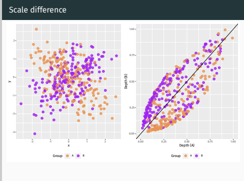
Both groups from standard normal distributions.



Means shifted to $(-1, 0)^\top$ and $(1, 0)^\top$.



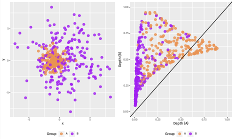
Covariances $\mathbb{I}_2$ and $4\mathbb{I}_2$.



Covariances

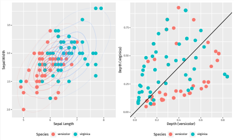
$$\begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}.$$

Location and scale difference

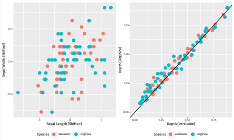


- Means $(-1, 0)^\top$ and $(1, 0)^\top$.
- Covariances $\mathbb{I}_2$ and $9\mathbb{I}_2$.

2023-12-11



D-D plot indicative of location difference.



Data has been shifted so that the locations coincide.

Maximum depth classifiers

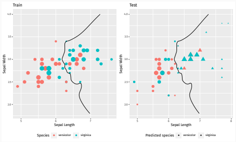
Given a point $\boldsymbol{x} \in \mathbb{R}^p$, assign it to the class with respect to which it has maximum depth. In other words, choose

$$\hat{k}(\boldsymbol{x}) = \arg\max_j D(\boldsymbol{x}, \hat{F}_j).$$

Under certain conditions, this asymptotically performs on par with the Bayes classifier.

Ghosh, A.K., & Chaudhuri, P. (2005) On maximum depth and related classifiers

Maximum depth classification corresponds to using the $x = y$ line to separate points in the D-D plot.



- This figure illustrates maximum depth classification on the same multivariate data shown earlier, using spatial depth.
- The depth contours are learned from training data.
- The black curve denotes the learned decision boundary.
- Classification accuracies hover around 70%.

Relative data depth

The *relative data depth*

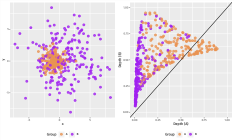
$$\mathrm{ReD}(\boldsymbol{x}) = D(\boldsymbol{x}, \hat{F}_{\hat{k}(\boldsymbol{x})}) - \max_{j \neq \hat{k}(\boldsymbol{x})} D(\boldsymbol{x}, \hat{F}_j)$$

gives a measure of confidence in the classification of $\boldsymbol{x}$.

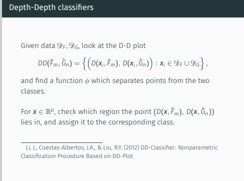
Jörnsten, R. (2004) Clustering and classification based on the L_1 data depth

- This can be used to identify and remove 'noisy' examples from the training set.
- This can also be used as a measure of dissimilarity in clustering, with an objective function

$$\frac{1}{N} \sum_{k} \sum_{\boldsymbol{x}_i \in C(k)} \mathrm{ReD}(\boldsymbol{x}_i).$$



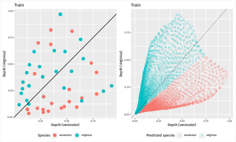
This illustrates that the maximum depth classifier may not always be appropriate.



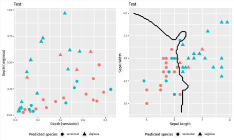
- The D-D plot converts the ℓ-class classification problem to one in a ℓ-variate setting, regardless of what the original data looks like (multivariate, functional).
- The separating function ϕ is approximated by searching in a class of functions Γ, for instance, the family of increasing functions, or the family of polynomials.
- The two class DD classifier is easily extended to ℓ groups, in the form of the DD^G classifier. The data transformed via

$$\mathbf{x} \mapsto (D(\mathbf{x}, \hat{F}_1), \ldots, D(\mathbf{x}, \hat{F}_\ell))$$

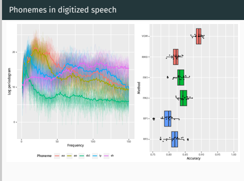
  can be classified using any existing multivariate classifier (LDA, kNN, GLM, etc).



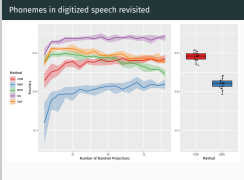
- The figure on the left shows the D-D plot for the training data.
- The figure on the right shows the locations of points (originally taken from a grid in the real data space) in the D-D plot. They are coloured according to the class predicted by the DD classifier, using polynomial boundaries.
- In this instance, the classification rule agrees closely with the maximum depth classifier rule. This is illustrated by the decision boundary in the D-D plot almost coinciding with the diagonal.



- The figure on the left shows the predictions for the testing data on the D-D plot.
- The figure on the right shows the predictions for the testing data in the original space.
- The black curve denotes the decision boundary.
- Classification accuracies hover aruond 70%.



- This figure illustrated periodograms obtained from digitized speech.
- Different groups correspond to the pronunciation of different phonemes.
- The thicker lines denote the median curves from the corresponding group.
- This data is available as ‘phoneme data’ from the `fds` package in R.



- The random functions $\phi_1, \ldots, \phi_\ell$ have been generated by a Gaussian process with an exponential covariance kernel.
- The last three methods employ the maximum depth classifier (with the corresponding depths), applied on the transformed data
$$X \mapsto \left(\langle X, \phi_1 \rangle, \ldots, \langle X, \phi_\ell \rangle\right).$$
- The degradation in performance of the Mahalanobis classifier is likely due to the worsening estimate of the covariance matrix as the number of projections (hence the dimension) ℓ increases.

Do depth functions completely characterize probability distributions?

Sometimes!

This has implications in the consistency of depth based tests and classifiers, where all information about the given data/distribution is obtained via depth.

Halfspace depth revisited

The halfspace depth characterizes discrete probability distributions, i.e. if $D_H(\cdot, P) = D_H(\cdot, Q)$ and one of P, Q is discrete, then $P = Q$.

The halfspace depth also characterizes elliptic probability distributions.

Cuesta-Albertos, J.A., & Nieto-Reyes, A. (2008) The Tukey and the random Tukey depths characterize discrete distributions
Kong, L., & Zuo, Y. (2010) Smooth depth contours characterize the underlying distribution

The halfspace depth characterizes distributions P in $\mathbb{R}^p$ with contiguous support such that the depth contours for $0 < p < 1/2$ are *smooth* and the *maximal mass of P at a hyperplane*

$$\Delta(P) = \sup P(v^\top X = c) = 0.$$

A counterexample

Next, if $\psi_Z(\boldsymbol{t}) = \psi(\|\boldsymbol{t}\|_\alpha)$, then $\boldsymbol{v}^\top \boldsymbol{Z} \overset{d}{=} \|\boldsymbol{v}\|_\alpha Z_1$. Such distributions are called *α-symmetric*.

Using this, it can be shown that

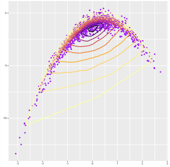
$$D_H(\boldsymbol{x}, P) = D_H(\boldsymbol{x}, Q) = F(-\|\boldsymbol{x}\|_\infty),$$

where F is the cdf of X_1.

- Observe that

$$\begin{aligned} D_H(\boldsymbol{x}, F_Z) &= \inf_{\boldsymbol{v} \neq 0} P(\boldsymbol{v}^\top Z \leq \boldsymbol{v}^\top \boldsymbol{x}) \\ &= \inf_{\boldsymbol{v} \neq 0} P\left(Z_1 \leq \frac{\boldsymbol{v}^\top \boldsymbol{x}}{\|\boldsymbol{v}\|_\alpha}\right) \\ &= P\left(Z_1 \leq \inf_{\|\boldsymbol{v}\|_\alpha = 1} \boldsymbol{v}^\top \boldsymbol{x}\right). \end{aligned}$$

- The infimum $-\|\boldsymbol{x}\|_\infty$ is achieved when $\boldsymbol{v} = \boldsymbol{e}_j$.
- This is easy to see when $\alpha = 1$ (optimization over a convex hull). When $0 < \alpha \leq 1$, use $\|\boldsymbol{v}\|_\alpha \geq \|\boldsymbol{v}\|_1$.



Halfspace depth contours of data drawn from a 'banana' shaped distribution, generated by first drawing

$$X \sim \mathcal{N}(0, \Sigma), \qquad \Sigma = \begin{bmatrix} 1 & 0.9 \\ 0.9 & 1 \end{bmatrix},$$

then setting

$$Y = \begin{bmatrix} aX_1 \\ X_2/a + b((aX_1)^2 + a^2) \end{bmatrix}, \qquad a = 1, \quad b = 1.$$