ETC3250/5250: Dimension reduction

Semester 1, 2020

Professor Di Cook

Econometrics and Business Statistics
Monash University

Week 4 (b)

PCA vs LDA

Discriminant space: is the low-dimensional space where the class means are the furthest apart relative to the common variance-covariance.

The discriminant space is provided by the eigenvectors after making an eigen-decomposition of $\Sigma^{-1}\Sigma_B$, where

$$\Sigma_B = \frac{1}{K}\sum_{i=1}^{K}(\mu_i - \mu)(\mu_i - \mu)' \quad \text{and} \quad \Sigma = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{n_k}\sum_{i=1}^{n_k}(x_i - \mu_k)(x_i - \mu_k)'$$

Mahalanobis distance

For two p-dimensional vectors, Euclidean distance is

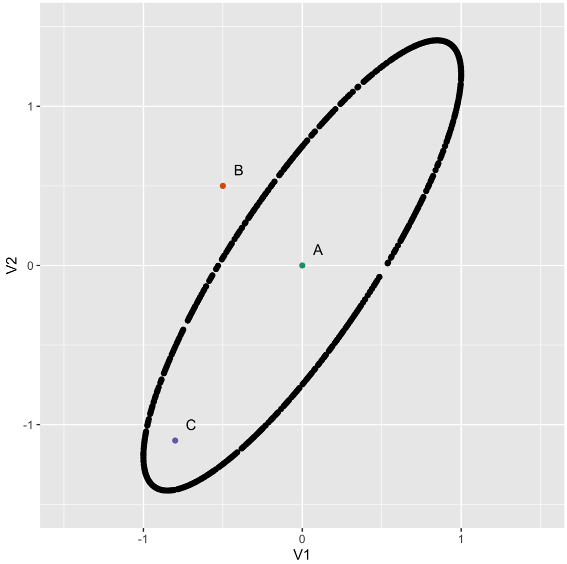
$$d(x,y) = \sqrt{(x-y)'(x-y)}$$

and Mahalanobs distance is

$$d(x,y) = \sqrt{(x-y)'\Sigma^{-1}(x-y)}$$

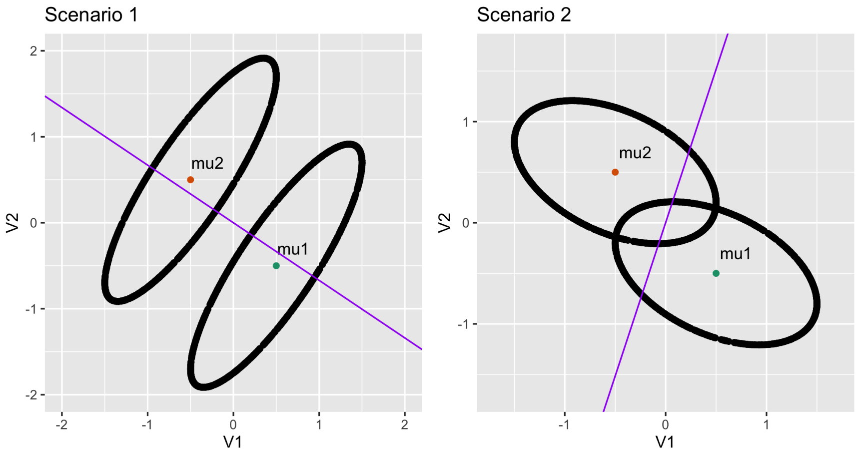
Which points are closest according to Euclidean distance? Which points are closest relative to the variance-covariance?

00:30



Discriminant space

Both means the same. Two different variance-covariance matrices.
Discriminant space depends on the variance-covariance matrix.



Projection pursuit (PP) generalises PCA

PCA:

$$\underset{\phi_{11},\ldots,\phi_{p1}}{\text{maximize}} \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{p}\phi_{j1}x_{ij}\right)^2 \text{ subject to } \sum_{j=1}^{p}\phi_{j1}^2 = 1$$

PP:

$$\underset{\phi_{11},\ldots,\phi_{p1}}{\text{maximize}}\, f\left(\sum_{j=1}^{p}\phi_{j1}x_{ij}\right) \text{ subject to } \sum_{j=1}^{p}\phi_{j1}^2 = 1$$

MDS

Multidimensional scaling (MDS) finds a low-dimensional layout of points that minimises the difference between distances computed in the *p*-dimensional space, and those computed in the low-dimensional space.

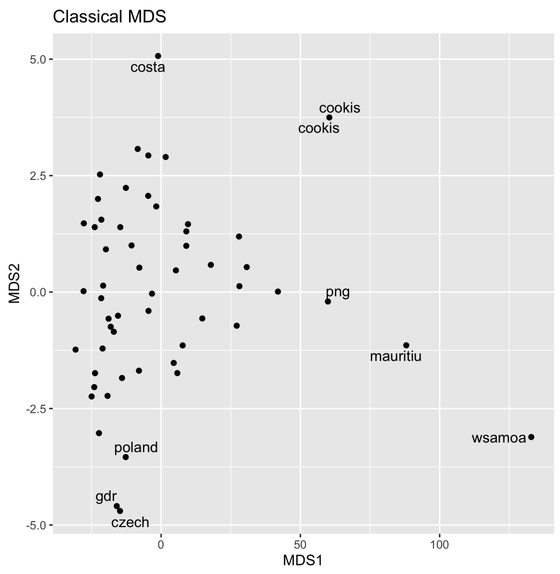
$$\mathrm{Stress}_D(x_1, \ldots, x_N) = \left(\sum_{i,j=1; i \neq j}^{N} (d_{ij} - d_k(i,j))^2 \right)^{1/2}$$

where D is an $N \times N$ matrix of distances (d_{ij}) between all pairs of points, and $d_k(i, j)$ is the distance between the points in the low-dimensional space.

MDS

- Classical MDS is the same as PCA
- Metric MDS incorporates power transformations on the distances, d_{ij}^r.
- Non-metric MDS incorporates a monotonic transformation of the distances, e.g. rank

```
track <- read_csv("data/womens_track.csv'
track_mds <- cmdscale(dist(track[,1:7]))
  as_tibble() %>%
  mutate(country = track$country)
```



Challenge

For each of these distance matrices, find a layout in 1 or 2D that accurately reflects the full distances.

```
## # A tibble: 3 x 4
##   name      A     B     C
##   <chr> <dbl> <dbl> <dbl>
## 1 A       0.1   3.2   3.9
## 2 B       3.2  -0.1   5.1
## 3 C       3.9   5.1   0
```

```
## # A tibble: 4 x 5
##   name      A     B     C     D
##   <chr> <dbl> <dbl> <dbl> <dbl>
## 1 A       0.1   0.9   2.1   3
## 2 B       0.9   0     1.1   1.9
## 3 C       2.1   1.1   0.1   1.1
## 4 D       3     1.9   1.1  -0.1
```

00:30

Non-linear dimension reduction

📊 T-distributed Stochastic Neighbor Embedding (t-SNE): similar to MDS, except emphasis is placed on grouping observations into clusters. Observations within a cluster are placed close in the low-dimensional representation, but clusters themselves are placed far apart.

Non-linear dimension reduction

📊 Local linear embedding (LLE): Finds nearest neighbours of points, defines interpoint distances relative to neighbours, and preserves these proximities in the low-dimensional mapping. Optimisation is used to solve an eigen-decomposition of the knn distance construction.

Non-linear dimension reduction

Self-organising maps (SOM): First clusters the observations into $k \times k$ groups. Uses the mean of each group laid out in a constrained 2D grid to create a 2D projection.

👩‍💻 Made by a human with a computer

Slides at https://iml.numbat.space.

Code and data at https://github.com/numbats/iml.

Created using R Markdown with flair by **xaringan**, and **kunoichi** (female ninja) style.