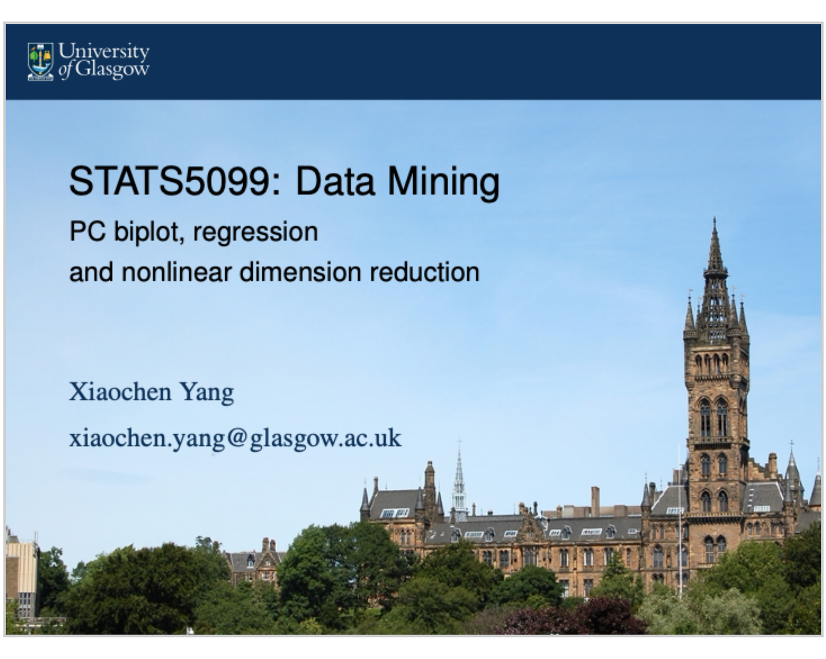
silde2

Thursday, May 20, 2021 4:16 PM

TutorialSlide2

University of Glasgow



STATS5099: Data Mining

PC biplot, regression and nonlinear dimension reduction

Xiaochen Yang
xiaochen.yang@glasgow.ac.uk

This week's content

- PCA biplots { scores, loadings
- multidimensional scaling
- principal component regression

Covariance or correlation-based PCA?

Prefer covariance-based PCA / original variables when:
- variables are comparable, e.g. gene expression from the same platform with similar range and scale
- differences in variation are of interest or meaningful themselves, e.g. rates in USArrests, stock data (high beta stocks will have higher loadings but they probably should)

Prefer correlation-based PCA / standardised variables when:
- difference in the variable variation caused by measurements on different scales, e.g. the trunk diameter in cm, biomass of leaves in kg, the number of branches, overall height in meters
- using correlation will throw out a tremendous amount of information → often need more correlation-PCs than covariance-PCs to retain an equivalent proportion of total variation

The "best" method to use is based on a subjective choice, careful thought and some experience.

https://stats.stackexchange.com/questions/53/pca-on-correlation-or-covariance

Multidimensional scaling (MDS)

- (often) a nonlinear dimension reduction method
- input: dissimilarity matrix $\delta_{ij}, i, j = 1, \ldots, n$ do not need original data
- output: **configuration points** $z_1, \ldots, z_n$
  configuration distances d_{ij}
- objective: preserve pairwise distances δ_{ij} as well as possible

$$\min_{z_1,\ldots,z_n} \sum_{i<j} L(\underbrace{\delta_{ij}}_{\text{dissimilarity}}, \overbrace{d_{ij}}^{\text{output-configuration}})$$

The loss function is termed **stress** or **strain**.

Different types of MDS

- distance scaling *vs* inner product scaling
  - distance: given dissimilarities, fit by $d_{ij} = \|z_i - z_j\|$ configuration distance
  - inner product: given similarities, fit by $\langle z_i, z_j \rangle$
- metric scaling *vs* nonmetric scaling
  - metric: δ_{ij} are interval or ratio, e.g. distance matrix (eurodist), inter-correlations among a set of variables (crimes)
  - nonmetric: δ_{ij} are ordinal; ranks matter more than actual values (wish) input: ranking

	distance scaling	inner product scaling
metric scaling	normalised stress Sammon mapping	classical scaling
nonmetric scaling	Stress-1	–

Implementing MDS and R commands

1. compute the dissimilarity matrix
   - convert similarity to dissimilarity: `sim2diss(s, method)`
   - convert data matrix to dissimilarity: `dist`
2. apply the appropriate MDS method
   - classical MDS: `cmdscale`
   - metric MDS: `mds(delta, type=c("ratio","interval")`
   - Sammon mapping: `Sammon`
   - non-metric MDS: `mds(delta, type="ordinal")` only rank
3. choose the number of dimensions: scree plot (stress *vs* no. of dim) ⇒ elbow
   choose between different metric MDS methods: Shepard diagram (fitted distance *vs* dissimilarity)
4. interpretation: configuration points z_i, z_j

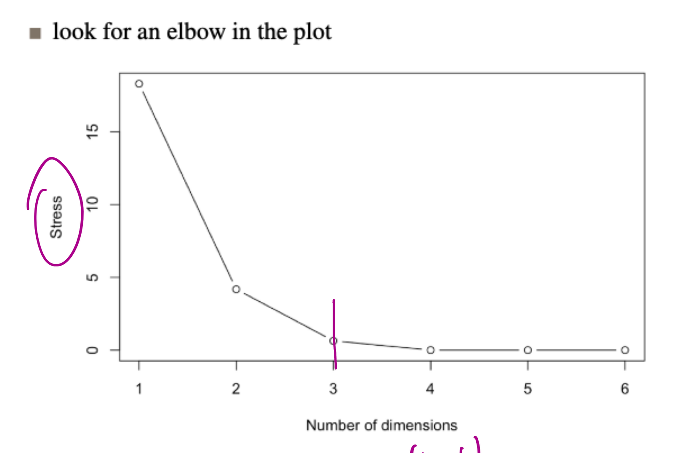
sim22diss

- `sim22diss(s, method=c("corr",z))`

Method (Argument)	Conversion Formula
Correlation ("corr")	$\delta_{ij} = \sqrt{1 - r_{ij}}$
Reverse ("reverse")	$\delta_{ij} = \min(s_{ij}) + \max(s_{ij}) - s_{ij}$
Reciprocal ("reciprocal")	$\delta_{ij} = 1/s_{ij}$
Membership ("membership")	$\delta_{ij} = 1 - s_{ij}$
Rank orders ("ranks")	$\delta_{ij} = \text{rank}(-s_{ij})$
Exponential ("exp")	$\delta_{ij} = -\log(s_{ij}/\max(s_{ij}))$
Gaussian ("Gaussian")	$\delta_{ij} = \sqrt{-\log(s_{ij}/\max(s_{ij}))}$
Transition frequencies ("transition")	$\delta_{ij} = 1/\sqrt{f_{ij}}$
Co-occurrences ("cooccurrence")	$\delta_{ij} = \left(1 + \frac{f_{ij}\sum_{i,j} f_{ij}}{\sum_i f_{ij} \sum_j f_{ij}}\right)^{-1}$
Gravity ("gravity")	$\delta_{ij} = \sqrt{\frac{\sum_i f_{ij} \sum_j f_{ij}}{f_{ij} \sum_{i,j} f_{ij}}}$
Confusion proportions ("confusion")	$\delta_{ij} = 1 - p_{ij}$
Probabilities ("probability")	$\delta_{ij} = 1/\sqrt{\arcsin(p_{ij})}$
Integer value z	$\delta_{ij} = z - s_{ij}$

scree plot

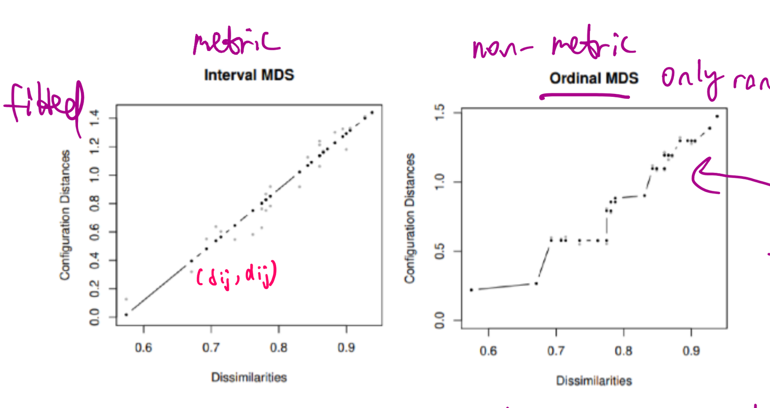
- look for an elbow in the plot



- reduce the feature dimension to 2 (n−1)

Shepard diagram

how well is the conf. points?



fitted similar to diss.

along close to the line

Q&As

MDS for dimension reduction
{ PCA linear
MDS non-linear

Tutorial question 1

This confusion matrix shows the frequency with which a stimulus letters shown for some milliseconds only is confused with another letter:

Letter	B	C	D	F	G
B	–				
C	3	–			
D	7	5	–		
F	3	5	7	–	
G	7	12	2	2	–

Task: visualise the letters as points in 1/2/3D and interpret the plot
1. Is this a dissimilarity matrix?
2. Which MDS method is more appropriate?
3. How many dimensions would you keep?
4. Do you see any clusters?
5. Investigate the sensitivity to different parameters.
6. Investigate the sensitivity to initial configurations.

Principal component regression

Limitations of linear regression
- multicollinearity (features strongly correlated)
- high-dimensional problem, i.e. when $p > n$ features > observations

→ $\mathbf{X}^\top\mathbf{X}$ is near-singular or singular (recall $\hat{\beta}_{\text{OLS}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$)

Principal component regression: apply linear regression on PC scores in higher dim case
- when keeping all PCs, can get back the original regression coefficients from PC regression coefficients and loadings
- when keeping the first q PCs, reduce variance but increase bias MSE ↓

feature

obs. $X = \begin{pmatrix} 1 & 5 \\ 2 & 10 \\ 3 & 15 \end{pmatrix}$

X^TX is not invertible

Bias-variance trade-off

Assume $Y = f(\boldsymbol{x}) + \epsilon$ where $\mathbb{E}(\epsilon) = 0$ and $\text{Var}(\epsilon) = \sigma^2$. The prediction error of a regression fit $\hat{f}(\boldsymbol{x})$ at an input point $\boldsymbol{x} = \boldsymbol{x}_0$:

prediction error

$$\begin{aligned} \text{PE} &= \mathbb{E}[(Y - \hat{f}(\boldsymbol{x}_0))^2 | \boldsymbol{x} = \boldsymbol{x}_0] \\ &= \sigma^2 + \mathbb{E}[(f(\boldsymbol{x}_0) - \hat{f}(\boldsymbol{x}_0))^2] \\ &= \text{Irreducible error} + \text{MSE} \\ &= \sigma^2 + (f(\boldsymbol{x}_0) - \mathbb{E}[\hat{f}(\boldsymbol{x}_0)])^2 + \mathbb{E}[(\hat{f}(\boldsymbol{x}_0) - \mathbb{E}[\hat{f}(\boldsymbol{x}_0)])^2] \\ &= \text{Irreducible error} + \text{Bias}^2 + \text{Variance} \end{aligned}$$

linear reg → PCA , var ↓ bias ↑

Bias-variance trade-off

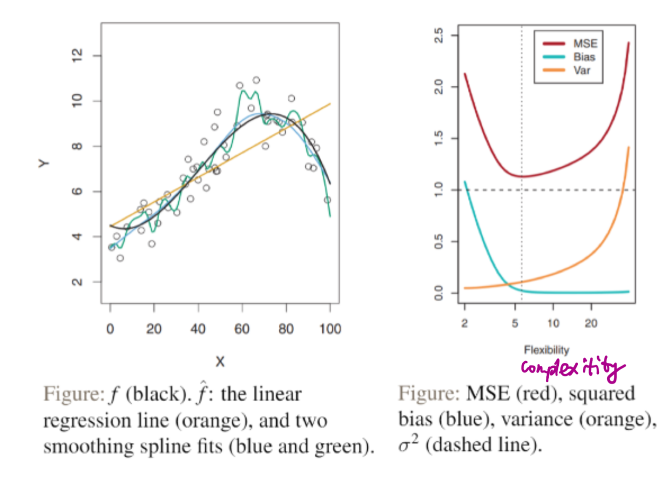


Figure: f (black). $\hat{f}$: the linear regression line (orange), and two smoothing spline fits (blue and green).

Figure: MSE (red), squared bias (blue), variance (orange), σ^2 (dashed line).

Bias-variance trade-off

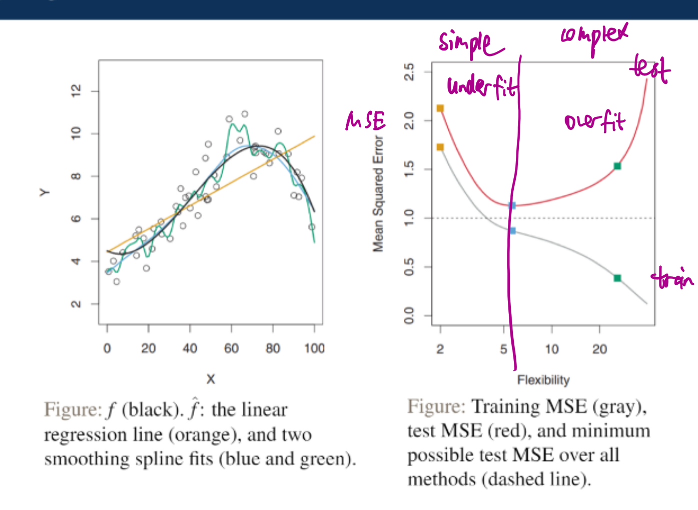


Figure: f (black). $\hat{f}$: the linear regression line (orange), and two smoothing spline fits (blue and green).

Figure: Training MSE (gray), test MSE (red), and minimum possible test MSE over all methods (dashed line).

Before leaving the tutorial ...

Check if you feel confident about the learning outcomes:
- Interpreting a biplot.
- Understanding different types of MDS.
- Performing MDS in R and interpreting its output.
- Understanding the bias-variance tradeoff.
- Understanding principal component regression and its advantages (over linear regression).

Quick homework: compare PCA and MDS – what are the differences, pros and cons?