Dimension reduction and Principal components regression

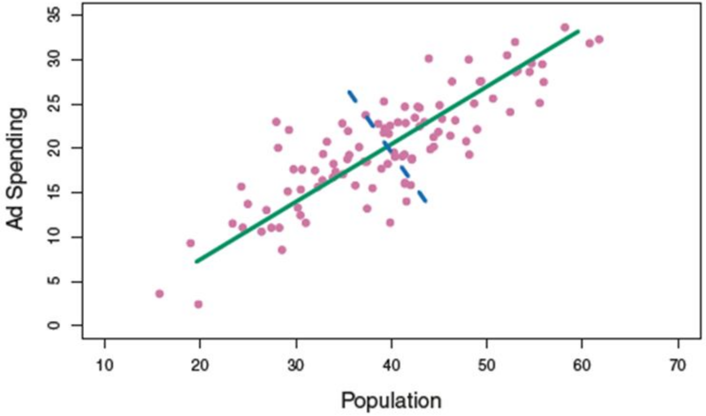
Outline

- Dimension reduction methods
- Principal components analysis
- Principal components regression
- Considerations in high dimensions

Dimensions reduction methods

… make new variables by combining existing variables.

$$Z_m = \sum_{j=1}^{p} \phi_{jm} X_j$$

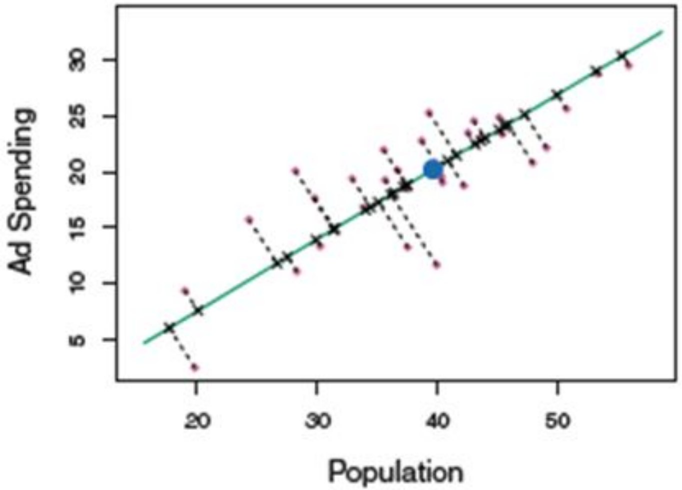


Income	Age	Gender	Utility	Childcare	Groceries	Leisure	Health
14.891	34	Male	0.54	0.62	0.57	0.18	0.61
106.025	82	Female	1.45	0	0.95	0.58	0.15
104.593	71	Male	1.95	0	0.64	0.28	0.42
148.924	36	Female	0.7	0.5	0.74	0.1	0.28
55.882	68	Male	1.32	0	0.56	0.43	0.15
80.18	77	Male	1.53	0	0.66	0.13	0.73
20.996	37	Female	1.56	0.51			
71.408	87	Male	0.56	0			
15.125	66	Female	1.15	0			
71.061	41	Female	0.76	0.89			

Income	Age	Gender	Indispensable	Dispensable
14.891	34	Male	1.73	0.79
106.025	82	Female	2.4	0.73
104.593	71	Male	2.59	0.7
148.924	36	Female	1.94	0.38
55.882	68	Male	1.88	0.58
80.18	77	Male	2.19	0.86
20.996	37	Female	2.65	0.9
71.408	87	Male	1.32	1.21
15.125	66	Female	1.71	0.92
71.061	41	Female	2.43	1.01

Principal components analysis (PCA)

- A technique to reduce dimension of an n x p data matrix
- Determining new variables by linearly combining the existing variables
- There will be at most p principal components, but …
- The *1st principal component* contains *most variability (information)* of the original data. Its direction represents the *line closest to the original data*.
- All principal components are uncorrelated => their directions are perpendicular to each other.
- Relying only on the predictors => an unsupervised analysis method



Steps to determine the principal components

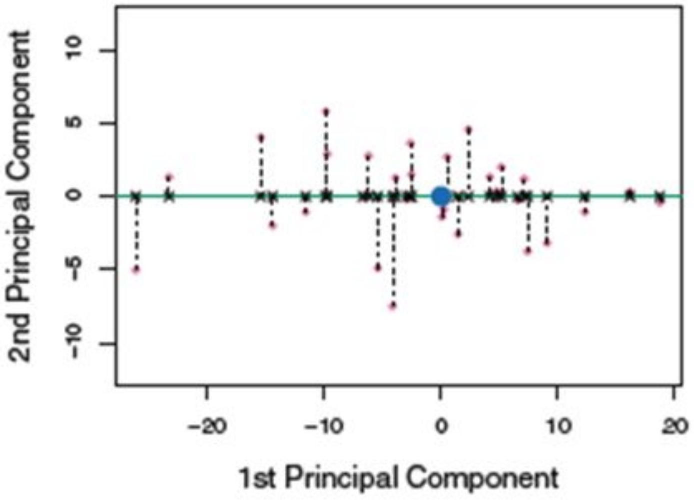
Step 1. Standardise the variables (or high variance variable would dominate the principal components)

Step 2. Compute covariance matrix

Step 3. Calculate the principal components

- Calculate eigenvectors and eigenvalues of the covariance matrix
- Eigenvector of the largest eigenvalue is the 1st PC, and so on

Step 4. Project the original variables onto the direction of the (selected) principal components



Detail tutorial available at https://online.stat.psu.edu/stat505/lesson/11/11.1

Principal components regression (PCR)

PCR is OLS regression of the response on the 1st M principal components of the original predictors.

PCR assumes that *‘the directions in which X1, ..., Xp show the most variation are the directions that are associated with Y’*

(There’s no guarantee that the above assumption is correct in all cases.)

$$y_i = \theta_0 + \sum_{m=1}^{M} \theta_m z_{im} + \epsilon_i, \quad i = 1, \ldots, n,$$

where

$$Z_m = \sum_{j=1}^{p} \phi_{jm} X_j$$

Principal components regression (PCR)

PCR is a dimension reduction method (if $M<p$) but not a feature selection method

Interpreting PCR may be challenging, especially when needs to be related to the original predictors

The number of principal components used can be decided by cross-validation

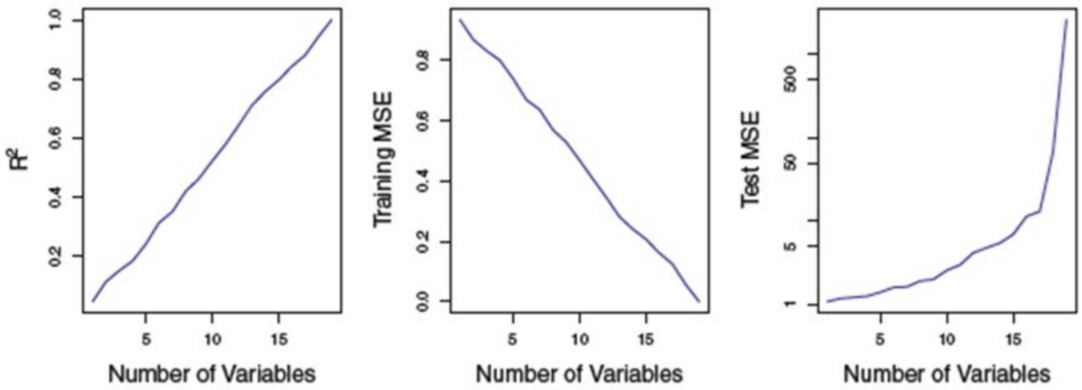
Using a trained PCR model for prediction (e.g. of a test set)

- Standardise predictors in the test set
- Project the standardised predictors onto the axis of the principal components
- Use the projected predictors for input into the PCR model

Considerations in high dimensions

What goes wrong in high dimensions

- Most traditional learning methods weren't designed for high dimensions (and when n is not much larger than p)
- Below are results from an example of regressing 20 response values against 1 to 20 predictors, all of which *unrelated to the response.*

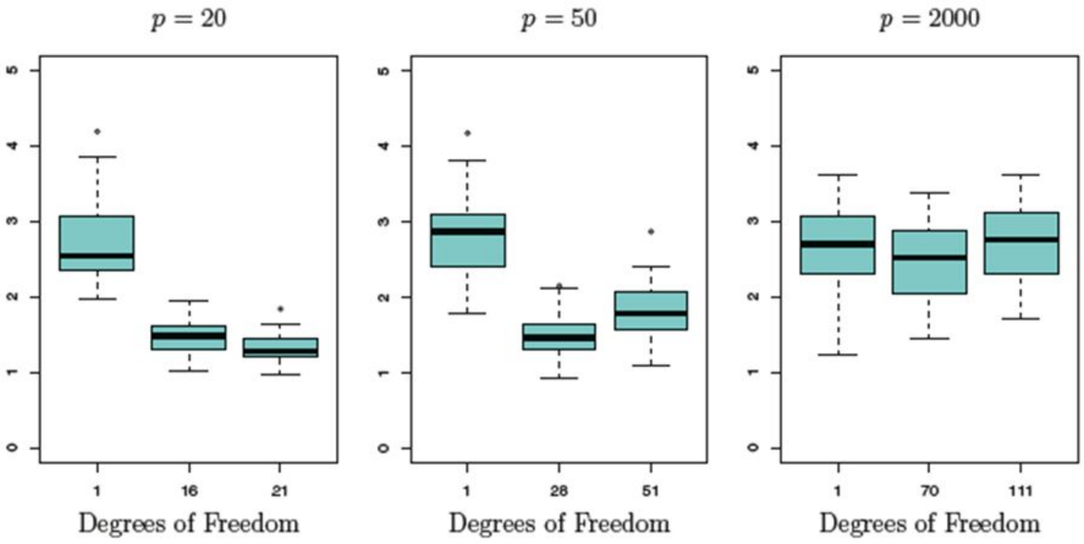


Considerations in high dimensions

Subset selection methods, regularisation and PCR are useful for regression in high dimensions => *avoiding overfitting* by using *less flexible approach* than OLS

The example on the right (from a lasso) say 3 things

- Lasso helps to reduce dimension
- Correct penalty needed for good predictive
- The curse of dimensionality



Considerations in high dimensions

More features would do more harm than good

- Deteriorating the (prediction) quality of the fitted model
- More time needed to do feature selection
- Costly data collection and preparation
- Risks of not having the features available in real-life applications

=> choosing variables relevant to the response is critical and must involve subject matter experts in the model building process.

Considerations in high dimensions

Interpreting results in high dimensions

- High dimensions present a very good chance of multicollinearity => not sure which variables are predictive of the response
- Large regression coefficients may be assigned to *variables that are correlated to the variables that are truly predictive* of the response
- Be very cautious to conclude a subset of predictors is better than others in predicting the response.
- More reasonable to say the selected subset of predictors forms *one of many possible models* to predict the response.
- *Never* use metrics on the training data (p-value, R2) to report the quality of fit - *always* use independent test sets where possible.