Again on bias - variance trade - off

- decompose squared error in bias^2 and variance

$f(x)$ true value

$\hat{f}(x)$ our estimator $\rightarrow \hat{y}$

$$E\left[\left(\hat{f}(x) - f(x)\right)^2\right]$$

$$= E\left[\left(\hat{f}(x) - \underline{E[\hat{f}(x)]} + \underline{E[\hat{f}(x)]} - f(x)\right)^2\right]$$

$$= E\left[\left(\hat{f}(x) - E[\hat{f}(x)]\right)^2\right] + E\left[\left(E[\hat{f}(x)] - f(x)\right)^2\right] +$$

$$+ 2E\left[\left(\hat{f}(x) - E[\hat{f}(x)]\right)\left(E[\hat{f}(x)] - f(x)\right)\right]$$

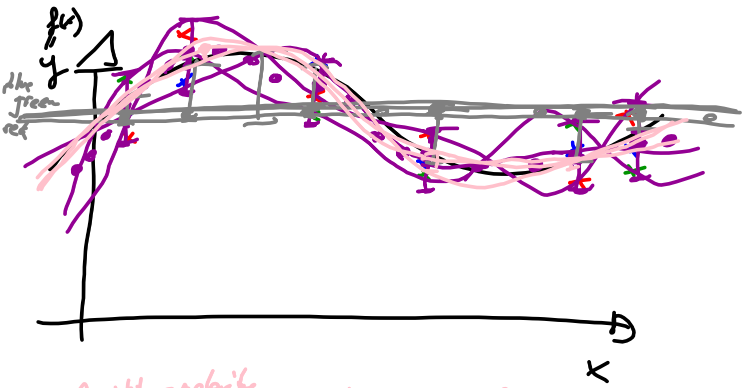
$$= Var\left(\hat{f}(x)\right) + \left(E[\hat{f}(x)] - f(x)\right)^2 + 2\left(E[\hat{f}(x)] - f(x)\right)E\left[\hat{f}(x) - E[\hat{f}(x)]\right]$$

$$\overset{!}{=} Var\left(\hat{f}(x)\right) + bias^2 + 0$$

Empirical visualization

$$Var(\hat{f}(x)) = E\left[\left(\hat{f}(x) - E[\hat{f}(x)]\right)^2\right]$$

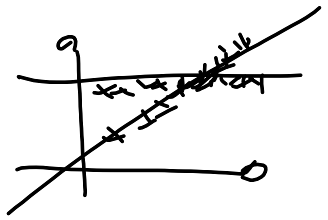
$$bias^2 = \left(E[\hat{f}(x)] - f(x)\right)^2$$

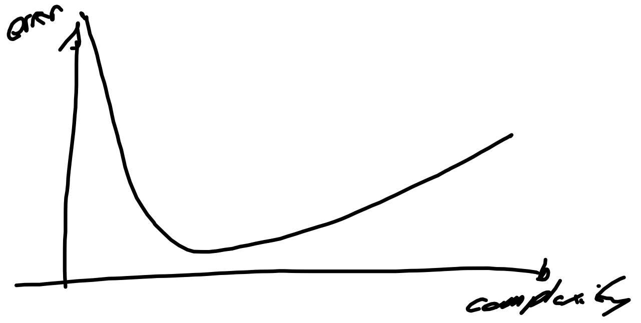


$\hat{f}(x)$ complex bias very small large variance

$\hat{f}(x)$ simple small variance large bias

$\hat{f}$ right complexity best trade-off between bias and variance





Methods for model selection

GOAL : find the "best" model among all possible models explaining/predicting y given X

Usually, we have a vector of response $y = (y_1, \dots, y_n)^T$

and a design matrix
matrix of covariates

$$X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix} = (\underset{\to}{1}, x_1, \dots, x_p)$$

$n \times (p+1)$ matrix → n observations, → p variables (+1 for the intercept)

intercept: not strictly necessery, but most often used in practice

if we use $y - \bar{y}$ as response → intercept $= 0$

Only considering linear effects, with p variables there are 2^p possible models to relate X and y

complexity $\downarrow$

$$\underline{y = \beta_0 + \varepsilon} \quad \leftarrow \text{null model (only intercept)}$$

$$\vdots$$

$$\underline{y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p + \varepsilon} \quad \leftarrow \text{full model (all variables included)}$$

- find the best model: 1- Best Subset Selection
  - fit all the models
  - find the model that minimizes an information criterion
  - it is very computationally heavy ← we need to fit 2^p models
- alternatively: simplified procedure (mostly used in practice)
  - backward elimination
  - forward selection
  - stepwise selection
  - stepback selection

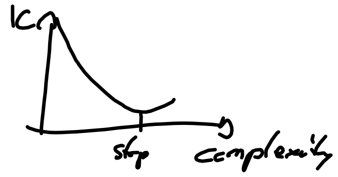
Backward elimination

→ start with the full model

→ remove one by one the least important variables (those which increase the least the residual variance)

→ proceed until the removal of the least important variable increase the chosen information criterion

$$\underline{IC} = \underset{\uparrow}{\text{deviance}} + \underset{\downarrow}{\text{penalty}}$$



Forward selection

→ start from the null model

→ add each time the variable that reduces the residual variance the most (most useful varariable to explain the variability of y)

→ stop when adding a new variable results in an increase in the IC

$$IC = \underset{\downarrow}{\text{deviance}} + \underset{\uparrow}{\text{penalty}}$$

<u>Stepwise selection</u>: like forward selection, but at each step we allow the removal of a variable previously added.

<u>Stepback selection</u>: like backward elimination, but each step we allow the re-adding of a variable previously eliminated.

Advantages of backward elimination/stepback selection
- we start with a legitimate model
- better handling of correlation between variables

Advantages of forward selection / stepwise selection
- works also for $p > n$
- in general, much better for large p

# observations : n
# covariates : p

Model selection techniques:
- work in general situations (linear models, GLM, GAM, Cox model, ...)
- easy to implement, used a lot in practice
- various issues as:
  - underestimation of the s.e.
  - re-use of data
  - multiple testing

$$\beta_B X_B + \beta_C X_C + \beta_D X_D + \beta_E X_E$$

categorical variables

$$X \begin{cases} \underline{A} \\ B \\ C \\ D \\ E \end{cases}$$

$x_1 = A$
$x_2 = C$
$x_3 = D$
$\vdots$

intercept

$$\begin{array}{ccccc} \boxed{X_A} & \boxed{X_B} & X_C & X_D & X_E \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \end{array}$$

$$\text{deviance} + 2p$$

using AIC is approximately like testing at level $\alpha = 0.157$

Principal Component Analysis/Regression

- we want to go towards a simpler (smaller p) model
  → variable selection: remove useless variables;
- alternatively:
  - construct new variables as linear combinations of the original
  - try to keep as much of the original information in as less new variables as possible

Consider a matrix X of data, suppose with mean 0 and variance Σ

↳ if mean $\neq 0$, we can always center the variables $x_i = x_i - \bar{x}_i$

e.g. $X \sim N_p(\underline{0}, \Sigma)$

We want new variables z on the form αX

$$z = \alpha X \Rightarrow Var(z) = \alpha^T \Sigma \alpha$$

The goal is to find α which makes $Var(z)$ the largest among all normalized linear combinations of the columns of X

$$\max_{\alpha} Var(z) = \max_{\alpha} \alpha^T \Sigma \alpha$$

$$\text{subject to } \|\alpha\| = 1$$

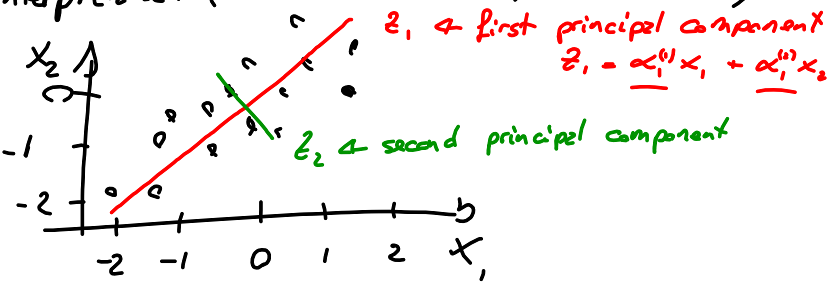
Once we did this (we obtained the α, let us say α_1) we look for another linear combination, orthogonal to the first one

$$\max_{\alpha} Var(z) = \max_{\alpha} \alpha^T \Sigma \alpha$$

$$\text{subject to } \|\alpha\| = 1 \quad \text{and} \quad \alpha^T \alpha_1 = 0$$

And so on...

Graphical interpretation (in 2 dimensions, $X = (x_1, x_2)$)



Matematically, this can be done by using the spectral decomposition of the matrix Σ

$$\Sigma = \underset{\text{eigenvectors}}{(\alpha_1 \;\dots\; \alpha_p)} \underset{\text{eigenvalues}}{\begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_p \end{pmatrix}} \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_p \end{pmatrix} \qquad \lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p$$

$z_1 = \alpha_1 X$

↳ first vector of principal loadings

↳ first principal component

λ_k is a measure of the original variance explained by the first principal component

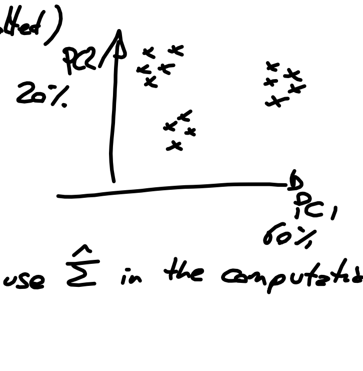
- each k-th principal component explains $\dfrac{\lambda_k}{\sum_{k=1}^{p} \lambda_k}$ of the original variance
- the first K principal components explain $\sum_{k=1}^{K} \lambda_k / \sum_{k=1}^{p} \lambda_k$ " " "

useful to select how many of the principal components to use in a model
new variables

e.g.: 90% of the original variance → $\underset{J}{4}$ principal components

PCR $y = \beta_0 + \beta_1 \underline{z_1} + \dots + \beta_J \underline{z_J} + \varepsilon$

- very useful in case of large p (only k<p components are used)
  ⇒ it can be used when p>n
- removing less useful principal components (those that do not explain anything of the original variance) reduce the model complexity
- used to solve collinearity problems (PC are orthogonal)
- used for graphical purposes (usually the first two principal components are plotted)



NOTES
- obviously, Σ is unknown, so we need to use $\hat{\Sigma}$ in the computations;
- there are some drawbacks:
  - reduced interpretability
  - reduced portability wrt variable selection

$\underset{n\times p}{X} \xrightarrow{PC} z$ $\qquad \hat{y}_{OLS} = X\hat{\beta}$ $\qquad \hat{y}_{PC} = z\hat{\gamma}$

$\hat{y}_{OLS} = \hat{y}_{PC}$ when we use all the principal components in our PCR (J=p)