Intro ML
ML. 2. Regression models

DataLab CSIC

Contents from

- Bishop Ch 3
- ISLR Ch 6
- ESL Chs 3 (+18)
- CASI ch 16
- Gelman, Hill, Vehtari

Objectives and schedule

An introduction to ‘modern’ topics in linear regression

- Conceptual discussion of linear regression
  - Subset selection
  - Shrinkage/regularisation
  - Dimension reduction
- Bayesian linear regression
- Large p
- Limitations

Useful in other contexts

- Nonlinear regression (eg neural nets)
- Classification (eg through glms)
- Variable and model selection….

+ Case study + Examples in lab

Linear regression

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p$$

A bit more…Linear basis function models

$$y = \beta_0 + \beta_1 \phi_1(x) + \dots + \beta_M \phi_M(x)$$

Polynomial regression, incl. interactions
Dummy variables
Other 'usual' transforms of variables like log,…
Gaussian basis functions
Sigmoidal basis functions
Wavelets
…..

Linear regression model

Least squares

$$\min \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_{1i} - \cdots - \beta_p x_{pi} \right)^2$$

Estimation

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

QR...

Linear regression model

Uncorrelated observations with constant variance

$$\mathrm{Var}(\hat{\beta}) = (X^T X)^{-1} \sigma^2$$

$$\hat{\sigma}^2 = \frac{1}{n-p-1} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Linear regression model

Normal errors

$$y = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p + \varepsilon \qquad \varepsilon \sim N(0, \sigma^2)$$

$$\hat{\beta} \sim N(\beta, (X^T X)^{-1} \sigma^2) \qquad \hat{\beta}, \hat{\sigma}^2 \ \text{INDEP}$$

$$(n-p-1)\,\hat{\sigma}^2 \sim \sigma^2 \chi^2_{n-p-1}$$

Linear regression model

- Non-linearity of response-predictor relationship
- Correlation of error terms
- Non-constant variance of error terms
- Outliers
- High leverage points
- Collinearity

Back to the Bias-variance tradeoff

$$Y = f(X) + \varepsilon \qquad E(\varepsilon) = 0 \qquad Var(\varepsilon) = \sigma^2$$

$$EPE = E\left(Y - \hat{f}(x)\right)^2$$

$$EPE = E\left(\hat{f}(x) - f(x)\right)^2 \qquad BIAS^2$$

$$+$$

$$E\left(\hat{f}(x) - E(\hat{f}(x))\right)^2 \qquad VAR$$

$$+$$

$$\sigma^2 \qquad NOISE$$

Error due to using much simpler model

How approximation changes if different training set used

Generally, more flexible method: variance increases, bias decreases

Prediction Accuracy of LR

If ‘true’ relationship response-predictors approx. linear, least squares have low bias

If n>>p, also low variance. Perform well on test

If not, there could be lot of variability. Overfitting >Poor predictions

If p>n, no longer unique LS: variance superbig, little use

By constraining coefficients, reduce variance with small impact on bias

Interpretability of LR

Frequently, several variables in LR not associated with response

Including irrelevant vars leads to unnecessary model complexity

By excluding them, improve interpretability

Feature selection, Variable selection

Three strategies

- Subset selection. Identify subset, then fit LR
- Shrinkage (regularisation). Fit with all predictors, but coeffs shrunken towards zero.
- Dimension reduction. Projecting p predictors into a space of lower dimension. Then fit LR.

Subset selection

Subset selection

Choose a subset of predictors (Discard the rest)

Fit LR with the subset

Subset selection: Best subset regression

1. Start with $m = 0$ and the null model $\hat{\eta}_0(x) = \hat{\beta}_0$, estimated by the mean of the y_i.
2. At step $m = 1$, pick the single variable j that fits the response best, in terms of the loss L evaluated on the training data, in a univariate regression $\hat{\eta}_1(x) = \hat{\beta}_0 + x_j'\hat{\beta}_j$. Set $\mathcal{A}_1 = \{j\}$.
3. For each subset size $m \in \{2, 3, \ldots, M\}$ (with $M \leq \min(n-1, p)$) identify the best subset $\mathcal{A}_m$ of size m when fitting a linear model $\hat{\eta}_m(x) = \hat{\beta}_0 + x_{\mathcal{A}_m}'\hat{\beta}_{\mathcal{A}_m}$ with m of the p variables, in terms of the loss L.
4. Use some external data or other means to select the "best" amongst these M models.

https://arxiv.org/pdf/1507.03133.pdf

Subset selection: Forward stepwise regression

1 Start with $m = 0$ and the null model $\hat{\eta}_0(x) = \hat{\beta}_0$, estimated by the mean of the y_i.

2 At step $m = 1$, pick the single variable j that fits the response best, in terms of the loss L evaluated on the training data, in a univariate regression $\hat{\eta}_1(x) = \hat{\beta}_0 + x'_j \hat{\beta}_j$. Set $\mathcal{A}_1 = \{j\}$.

3 For each subset size $m \in \{2, 3, \ldots, M\}$ (with $M \leq \min(n-1, p)$) identify the variable k that when augmented with $\mathcal{A}_{m-1}$ to form $\mathcal{A}_m$, leads to the model $\hat{\eta}_m(x) = \hat{\beta}_0 + x'_{\mathcal{A}_m} \hat{\beta}_{\mathcal{A}_m}$ that performs best in terms of the loss L.

4 Use some external data or other means to select the "best" amongst these M models.

Subset selection: Indirect estimation of test error

From training MSE=RSS/n, with d variables. Optimistic….

Mallows C_p

$$C_p = \frac{1}{n}\left(RSS + 2d\hat{\sigma}^2\right)$$

Akaike Info Crit

$$AIC = \frac{1}{n\hat{\sigma}^2}\left(RSS + 2d\hat{\sigma}^2\right)$$

Bayesian Info Crit

$$BIC = \frac{1}{n}\left(RSS + \log(n)\, d\hat{\sigma}^2\right)$$

Adjusted R^2

$$Adj.R^2 = 1 - \frac{RSS/(n-d-1)}{TSS/(n-1)}$$

Subset selection: Indirect estimation of test error

Use validation or cross-validation

(see Lab)

Final comments

Usable with other models

Variable/feature selection

Shrinkage/Regularisation

Regularisation

- Fit a model with all p predictors trying to constrain or regularize coeficients, aka shrink coeficientes towards 0
- Limit model complexity by adding a regularisation term
- Introduce sparsity

- Can significantly reduce variance in exchange of a small bias increase

Ridge regression

$$\min \sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda \sum_{i=1}^{p} \beta_i^2 \longrightarrow \hat{\beta}^R_\lambda$$

- First term. Minimize RSS. Fit data well
- Second term. Min. shrinkage penalty. Small when coeffs (1,…,p) close to 0
- λ tuning parameter. 0 vs infty
- Chosen via cross-validation. Values in grid, choose that with smaller CV error
- As parameter increases, flexibility decreases, variance decreases, bias increases
- But tends to preserve all coefficients… (interpretability if p large)

Lasso (least absolute shrinkage and selection operator)

$$\min \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \longrightarrow \hat{\beta}^L_\lambda$$

- First term. Minimize RSS. Fit data well
- Second term. Min. shrinkage penalty. Small when coeffs (1,…,p) close to 0
- λ tuning parameter. 0 vs infty
- Chosen via cross-validation as before
- As parameter increases, flexibility decreases, variance decreases, bias increases
- Forces some coeffs to 0 if parameter big, variable selection, sparse models

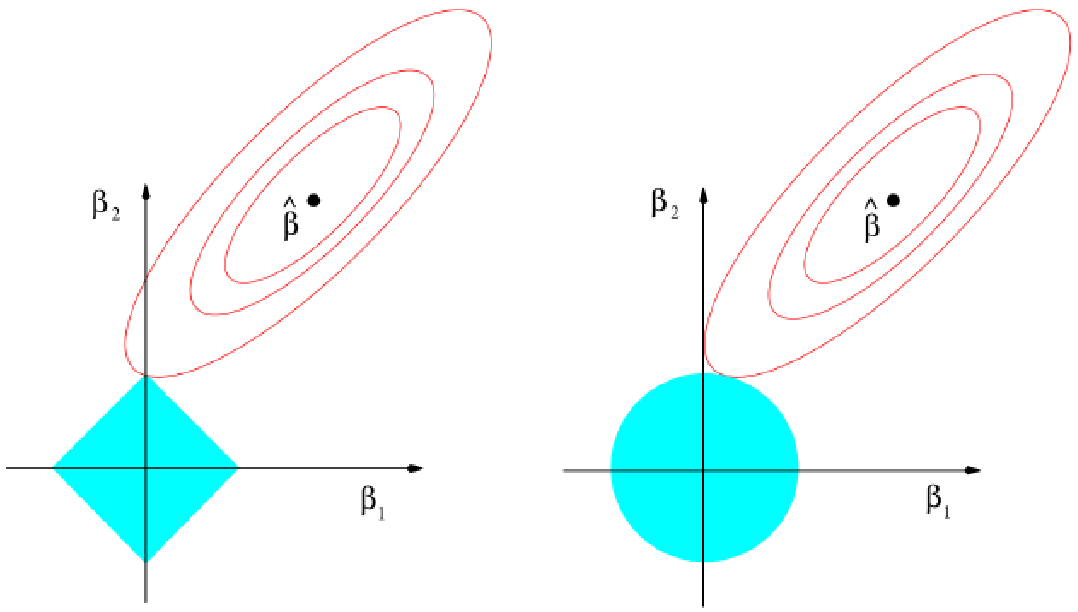
Lasso and ridge regression

$$\min\left(\sum \left(y_i-\beta_0-\sum_j \beta_j x_{ij}\right)^2\right) \quad \text{s.t.} \quad \sum_{j=1}^{p} |\beta_j| \le s$$

$$\min\left(\sum \left(y_i-\beta_0-\sum_j \beta_j x_{ij}\right)^2\right) \quad \text{s.t} \quad \sum_{j=1}^{p} \beta_j^2 \le s$$

$$\min\left(\sum \left(y_i-\beta_0-\sum_j \beta_j x_{ij}\right)^2\right) \quad \text{s.t} \quad \sum_{j=1}^{p} I(\beta_j \neq 0) \le s$$

Lasso and ridge regression



No free lunches….

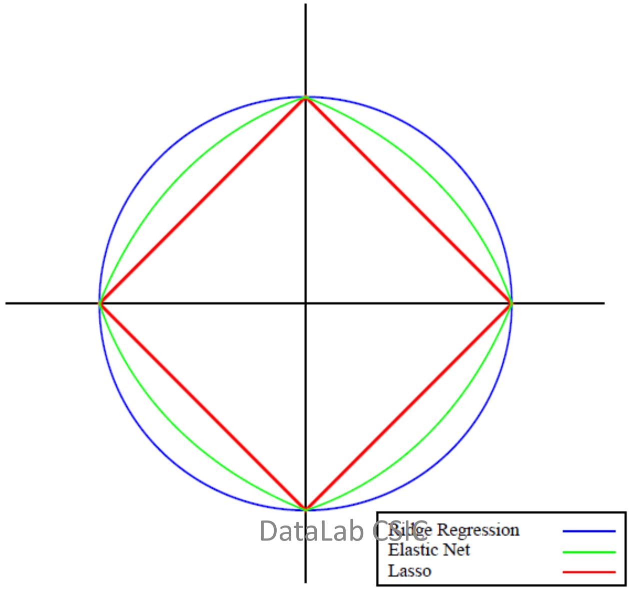
If small number of predictors dominates, Lasso

else ridge regression

Elastic Net

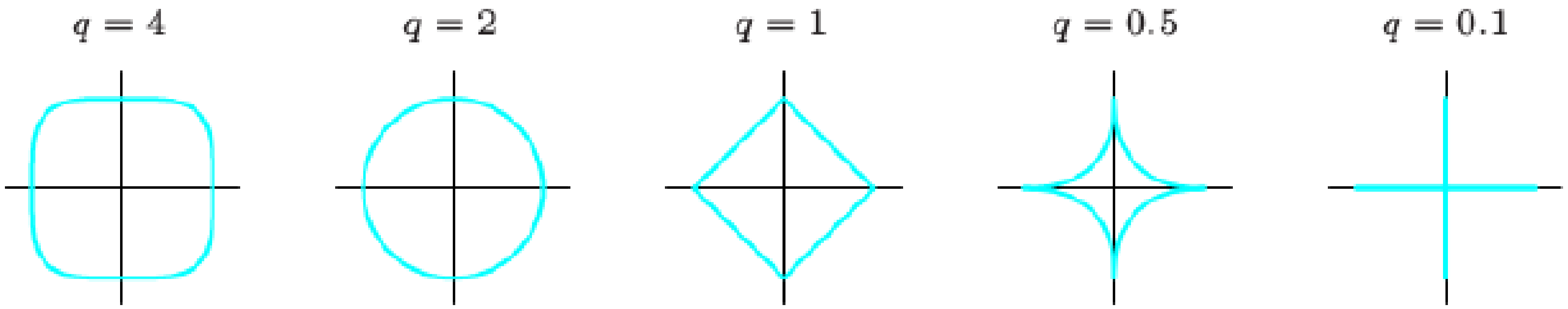
$$\min \left(\sum y_i - \beta_0 - \sum_i \beta_i x_{ij} \right)^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=1}^{p} \beta_j^2$$

$$\min \left(\sum y_i - \beta_0 - \sum_i \beta_i x_{ij} \right)^2 + \lambda \left(\alpha \sum_{i=1}^{p} |\beta_i| + (1-\alpha) \sum_{i=1}^{p} \beta_i^2 \right)$$



And other generalisations…

$$\min \left(\sum y_i - \beta_0 - \sum_j \beta_j x_{ij} \right)^2 + \lambda \sum_{j}^{p} |\beta_j|^q$$



Final comments

Usable with many models

In particular, with neural nets

Very important: Bayesian interpretation later on!!!!

Dimension reduction

Dimension reduction

Transform predictors to a smaller number of variables

Fit a linear regression based on transformed variables

$$(X_1, \dots, X_p) \xrightarrow[M<p]{} (Z_1, \dots, Z_M) \qquad Z_m = \sum_{j=1}^{p} \phi_{jm} X_j$$

$$y_i = \theta_0 + \sum_{m=1}^{M} \theta_m z_{im} + \varepsilon_i \qquad i = 1, \dots, n$$

$$\sum_{1}^{M} \theta_m z_{im} = \sum_{m=1}^{M} \theta_m \sum_{j=1}^{p} \phi_{jm} x_{ij} = \sum_{j=1}^{p} \sum_{m=1}^{M} \theta_m \phi_{jm} x_{ij} = \sum_{j=1}^{p} \beta_j x_{ij}$$

$$\beta_j = \sum_{m=1}^{M} \theta_m \phi_{jm}$$

Dimension reduction: Principal component regression

Construct first M principal components
Fit through least squares with PCs as predictors

If first M components explain variability nicely and M<<p, prevents overfitting
Not a feature selection method!!! Interpretability lags…
M by cross validation
First standardise variables (as in PCA)

Dimention reduction: Partial least squares

Standardise predictors

First PLS direction. Regress Y on each predictor to obtain coefficient for Z_1

Regress each variable on Z_1 to obtain residuals

Second PLS direction. Regress Y on each residual to obtain coefficient for Z_2

….

Not a feature selection method!!! Interpretability lags…

M by cross validation

Bayesian Linear Regression

Bayesian linear regression

Model

$$y_i = \beta_0 + \beta_1 x_{1i} + \dots + \beta_p x_{pi} + \varepsilon_i \quad i = 1,\dots,n$$

$$\varepsilon_i \sim N(0,\sigma^2)$$

$$y \mid \beta, \sigma, X \sim N(X\beta, \sigma^2 I)$$

Prior   Non-informative prior

$$p(\beta,\sigma^2) = p(\beta \mid \sigma^2)\, p(\sigma^2) \propto 1 \times \sigma^{-2} = \sigma^{-2}$$

Other priors mentioned later
(for econs g-priors)

Bayesian linear regression

Posterior

$$p(\beta,\sigma^2|y) = p(\beta|\sigma^2,y)\, p(\sigma^2|y)$$

$$\beta|\sigma^2,y \sim N(\hat{\beta}, V_\beta \sigma^2) \qquad V_\beta = (X^T X)^{-1}$$

$$\hat{\beta} = V_\beta X^T y$$

$$\sigma^2|y \sim \frac{p(\beta,\sigma^2|y)}{p(\beta|\sigma^2,y)} \sim \text{Inv-}\chi^2(n-p, s^2) \qquad s^2 = \frac{1}{(n-p)} (y-X\hat{\beta})^T(y-X\hat{\beta})$$

Simulating

FOR I = 1 TO N

SAMPLE $\sigma^2 \sim \sigma^2|y$

SAMPLE $\beta \sim \beta|\sigma^2,y$

Bayesian linear regression

Predictive

$y \quad (\tilde{X}, \tilde{y}) \quad f(\tilde{y} \mid \tilde{X}, y) = \iint f(\tilde{y} \mid \beta, \sigma, \tilde{X})\, f(\beta, \sigma \mid y)\, d\sigma\, d\beta$

$$\tilde{y} \mid \tilde{X}, \sigma, y \sim N\left(\tilde{X}\hat{\beta},\ (I + \tilde{X} V_\beta \tilde{X})\,\sigma^2\right)$$

$$\tilde{y} \mid \tilde{X}, y \sim t_{n-p}\left(\tilde{X}\hat{\beta},\ s^2 (I + \tilde{X} V_\beta \tilde{X})\right)$$

Simulating

FOR I = 1 TO N

SAMPLE $\sigma^2 \sim \sigma^2 \mid y$

SAMPLE $\beta \sim \beta \mid \sigma^2, y$

SAMPLE $\tilde{y} \sim \tilde{y} \mid \beta, \sigma^2, \tilde{X}$

Ridge regression as MAP estimation

Normal prior

$$p(y|x,\beta) \qquad \beta_i \sim N(0,\sigma_0^2)$$

Posterior

$$p(\beta|y) \propto \left[\prod_{i=1}^{n} p(y_i|x_i,\beta)\right]\left[\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^p \exp\left(-\frac{1}{2\sigma^2}\sum\beta_i^2\right)\right]$$

$$\log p(\beta|y) \propto \left(\sum_{i=1}^{n} \log p(y_i|x_i,\beta)\right) - \frac{1}{2\sigma^2}\sum\beta_i^2 + \cancel{\text{const}}$$

$$-\log p(\beta|y) \propto \underbrace{-\left(\sum_{i=1}^{n} \log p(y_i|x_i,\beta)\right)} + \underbrace{\frac{1}{2\sigma^2}\left(\sum\beta_i^2\right)}$$

Role of prior varinace

Lasso as MAP estimation

Normal prior

Posterior

$$p(y \mid x, \beta) \qquad \beta_i \sim \mathrm{Lap}(\tau) \quad \frac{1}{2\tau} \exp\left(-\frac{|\beta|}{\tau}\right)$$

$$p(\beta \mid y) \propto \left[\prod_{i=1}^{n} p(y_i \mid x_i, \beta)\right] \left[\left(\frac{1}{2\tau}\right)^{p} \exp\left(-\frac{|\beta_i|}{\tau}\right)\right]$$

$$\log p(\beta \mid y) \propto \left(\sum_{i=1}^{n} \log p(y_i \mid x_i, \beta)\right) - \frac{1}{\tau} \Sigma |\beta_i| + \text{const}$$

$$-\log p(\beta \mid y) \propto \underbrace{-\left(\sum_{i=1}^{n} \log p(y_i \mid x_i, \beta)\right)} + \underbrace{\frac{1}{\tau} \Sigma |\beta_i|}$$

MAP estimation with flat prior

Normal prior

$$p(y|x,\beta) \qquad p(\beta_i) \propto 1$$

Posterior

$$p(\beta|y) \propto \left[\prod_{i=1}^{n} p(y_i|x_i,\beta)\right] \cdot 1$$

$$\log p(\beta|y) \propto \sum_{i=1}^{n} \log p(y_i|x_i,\beta) \rightarrow \text{MLE}$$

Roy Lichtenstein (1923-1997) ???

Wow!!!

Bayes prevents from overfitting!!!!!

Regularisation equivalent to (MAP with) sparsity inducing priors

(MAP) with flat prior equivalent to least squares

Further thoughts on large scale problems

Low vs high dimension problems

Low dimensional problems: n>>p

High dimensional problems: p>n

- Bias-variance tradeoff
- Danger of overfitting

High dimension problems

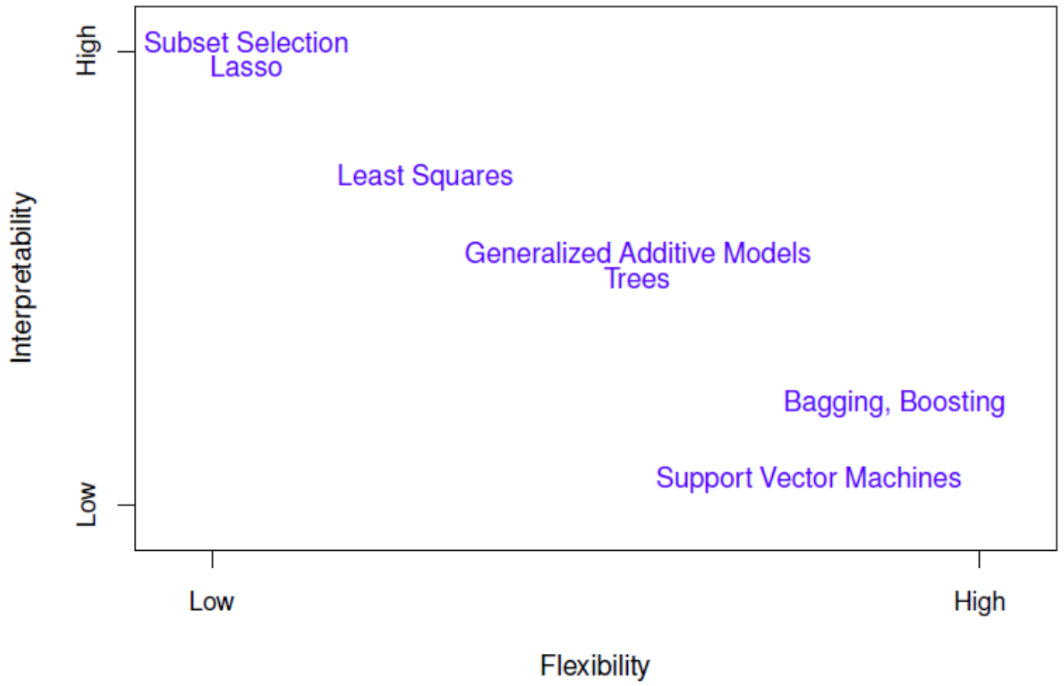
If p>n

Least squares should not be performed. Perfect fit with zero residuals, overfit -→ Typically terrible fit in an independet testing set. Too flexible model

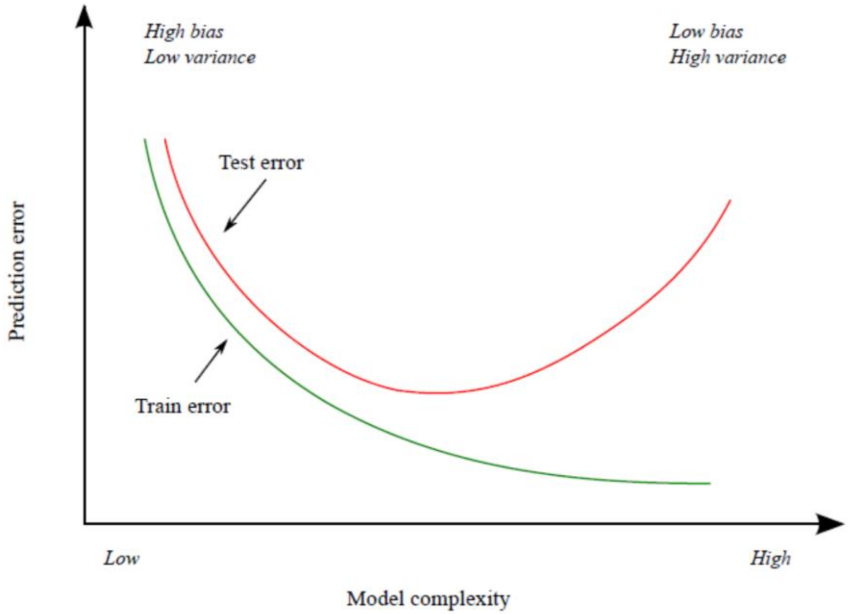
Cp, AIC, BIC not appropriate (estimate for variance is zero….)

Forward stepwise, ridge, lasso, PCR may still be relevant. Avoid overfitting by using a less flexible approach than least squares

From lecture 1-2



A missing dimension



High dimension problems

Adding additional signal features truly associated with response will improve model, reduce test set error

Adding noise features not truly associated with response will deteriorate model, increase test set error

Multicollinearity exacerbates

High dimension problems

Care with RSS, p-values, R^2 etc…

Report on independent test set or cross-validate…

High dimension problems

Or do Bayes

Roy Lichtenstein (1923-1997) ???

When n is very big!!!

e.g. biglm

Linear regression on datasets larger than memory available

https://cran.r-project.org/web/packages/biglm/biglm.pdf

The limits of linear (basis function) models

The limits of linear models

Useful properties

Closed form solutions to least squares

Very tractable Bayesian treatment

Model arbitrary nonlinearities, with choices of basis functions

But important limitations

Number of functions needs to grow rapidly as p grows

Two properties to be exploited

Data actually tend to live in space of smaller dimension

Target variables my depend only significantly on a few directions in data manifold

More

Further read

CASI. Ch16

Gelman, Hill, Vehtari

See lab