MedBioInfo - Machine Learning

# Statistical principles for supervised learning: Overfitting, regularization and all that - Part 1

(Generalised) linear models and regularisation

Manuela Zucknick

Department of Biostatistics, University of Oslo

manuela.zucknick@medisin.uio.no

May 03, 2022

Some of the figures in this presentation are taken from "Elements of Statistical Learning" (Springer, 2009) and "An Introduction to Statistical Learning, with applications in R" (Springer, 2013) with permission from the authors.

# Schedule for today

| | |
|---|---|
| **09:00 - 11:30** | **Lecture (with breaks): Statistical principles for supervised learning.**<br>*Some topics: variance vs bias, avoiding overfitting, regularisation, penalised regression & Bayesian alternatives, model selection, assessment & validation, cross-validation and bootstrapping* |
| **11:30 - 12:30** | **Lunch** |
| **12:30 - 14:30** | **Computer lab (with R).**<br>*Some starting points for an integrated analysis of miRNA, mRNA and protein data with (penalised) regression. Paper: Aure et al, Integrated analysis reveals microRNA networks coordinately expressed with key proteins in breast cancer, Genome Medicine, 2015.* |
| **14:45 - 16:00** | **Lecture: Examples for extensions - Statistical machine learning for cancer drug screens**<br>*Some topics: multi-view multi-task learning via multivariate penalised regression, tailor the penalty terms to reflect known structure in the data* |

# Some topics for this morning

## Part 1

- Supervised learning
- (Generalised) linear models
- Regularisation
- Penalised regression & Bayesian alternatives

## Part 2

- Overfitting
- Variance vs bias
- Model selection, assessment & validation
- Prediction performance
- Resampling: Cross-validation & bootstrap

# Further reading

**James G, Witten D, Hastie T and Tibshirani R (2021)**, An Introduction to Statistical Learning with Applications in R, Springer, 2nd edition. `https://www.statlearning.com`

**Hastie T, Tibshirani R, Friedman J (2009)**, The Elements of Statistical Learning, Springer, 2nd edition. `https://hastie.su.domains/ElemStatLearn/`

**Holmes S, Huber W (2019)**, Modern Statistics for Modern Biology, Cambridge University Press. `https://www.huber.embl.de/msmb/`

(some chapters on supervised/ unsupervised machine learning)

# Part 1: Outline

1. Introduction
   - Example: Integrative omics for personalized cancer therapy
   - Supervised learning (with generalised linear models)

2. Regularisation: penalties and hierarchical priors
   - Penalised regression
   - Bayesian variable selection
   - Support vector machines as a penalisation method

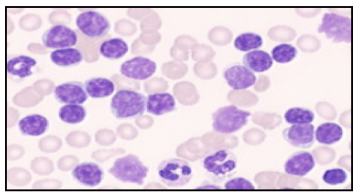
# Introductory example:

# Integrative omics for personalized cancer therapy

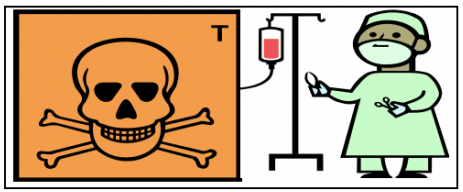
# Personalized cancer therapy

...aims to find the best therapy for each patient based on data about the patient and tumor (e.g. genomic data).

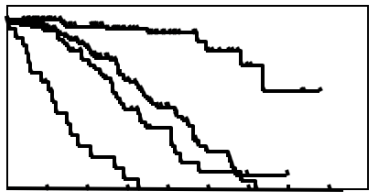
one diagnosis



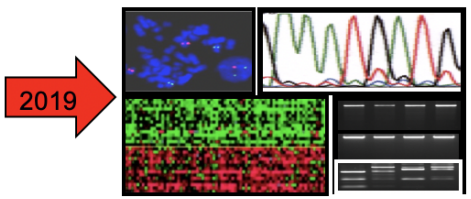
uniform therapy



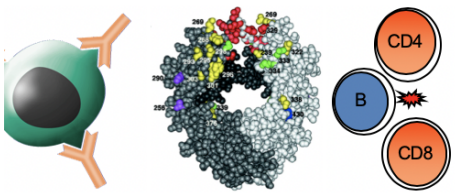

variable results



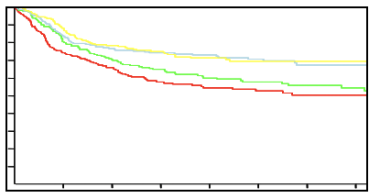
molecular subtypes



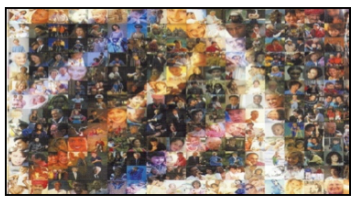

targeted therapy



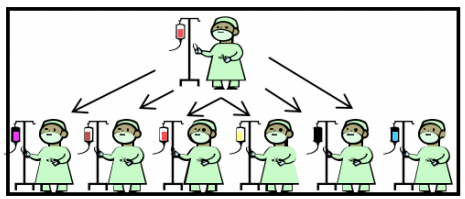

better results



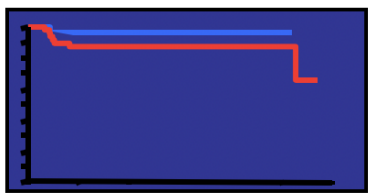
individual profile



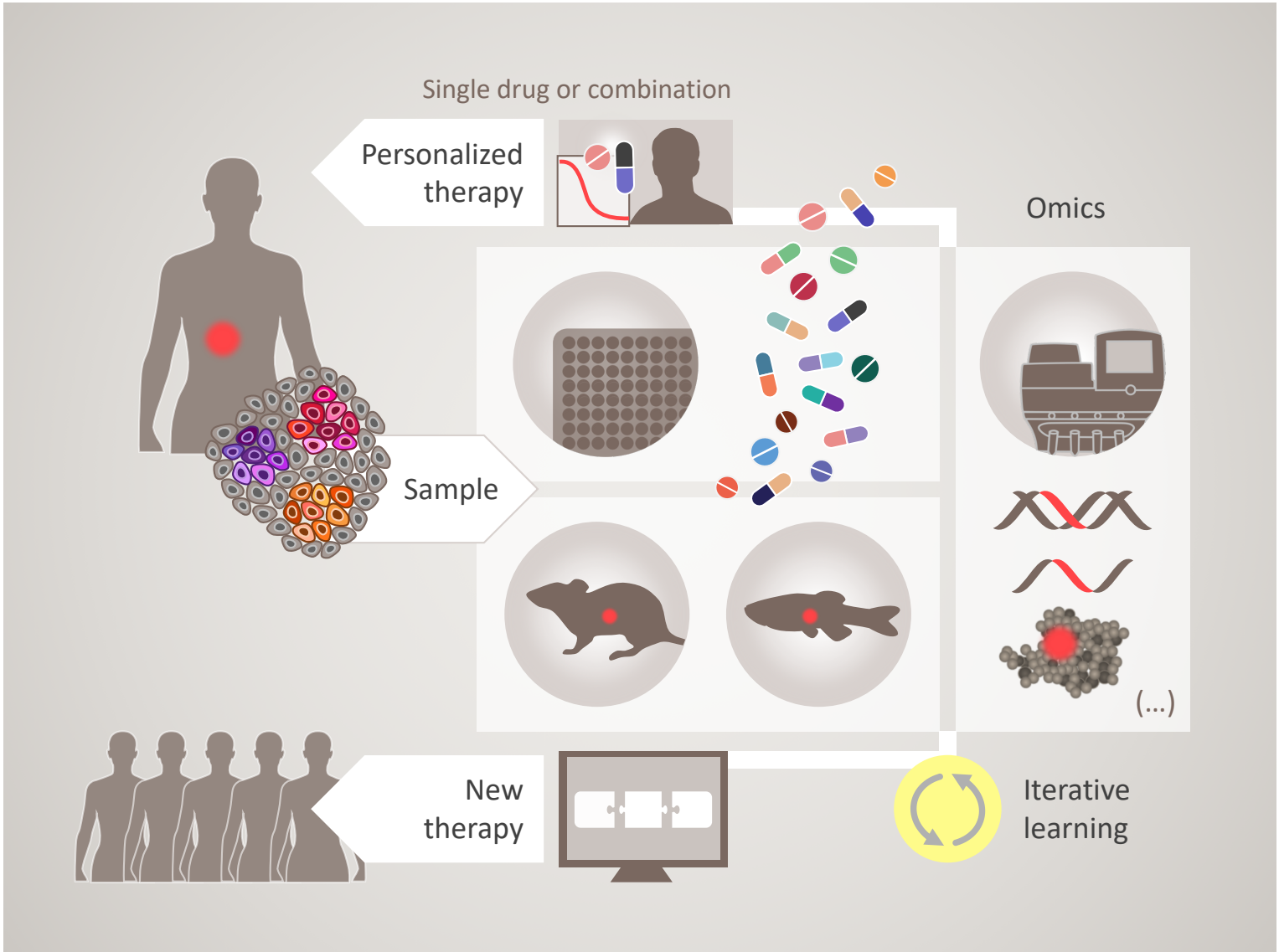
personal therapy



optimal results



slide by Stephan Pfister



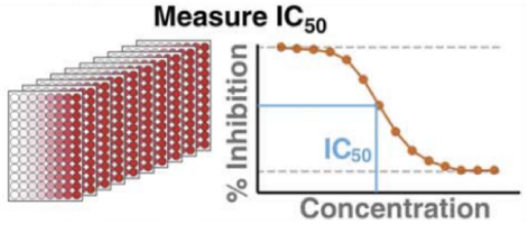

slide by Kjetil Taskén

# Predict sensitivity to multiple drugs $\boldsymbol{Y}$ from multi-omics $\boldsymbol{X}$

$$\boldsymbol{Y} = \boldsymbol{XB} + \epsilon$$

- **Multivariate Y:**
  Drug dose response

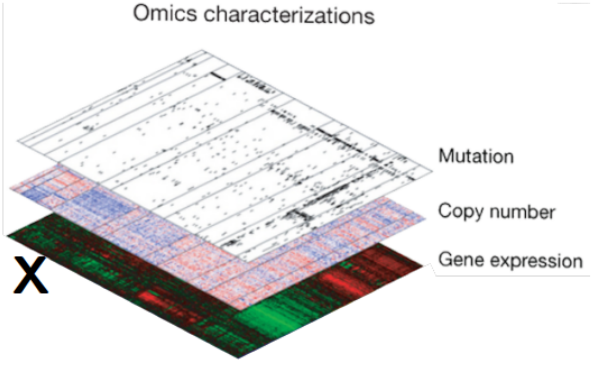
$$n \text{ cell lines} \overbrace{\begin{bmatrix} | & & | \\ \mathbf{y}_{\bullet 1} & \dots & \mathbf{y}_{\bullet m} \\ | & & | \end{bmatrix}}^{\textit{drug sensitivity}} = \mathbf{Y}$$




Source: Yang, et al. 2017

- **Heterogeneous X:**
  Integrative omics

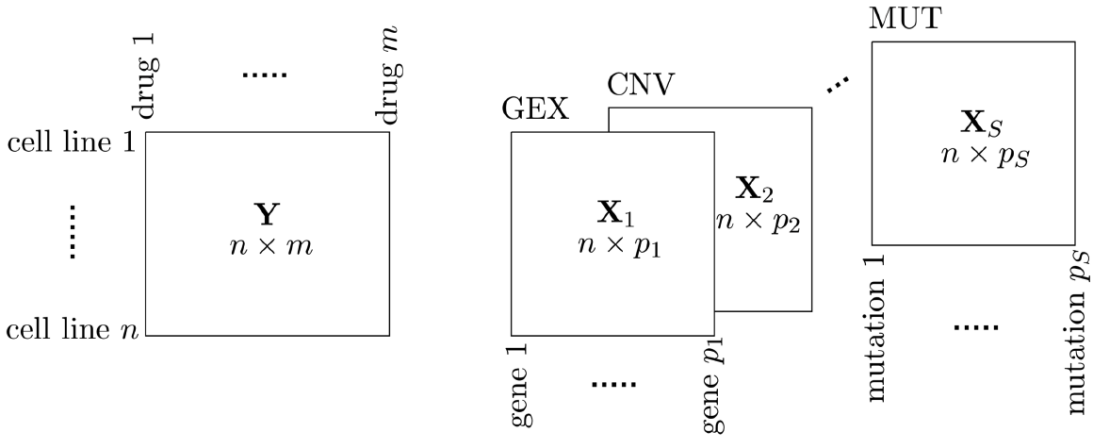
$$n \text{ cell lines} \overbrace{\begin{bmatrix} \mathbf{X}_1 & \vdots & \mathbf{X}_2 & \vdots & \mathbf{X}_3 \end{bmatrix}}^{\textit{gene expression}\ \ \textit{copy number}\ \ \textit{mutation}} = \mathbf{X}$$



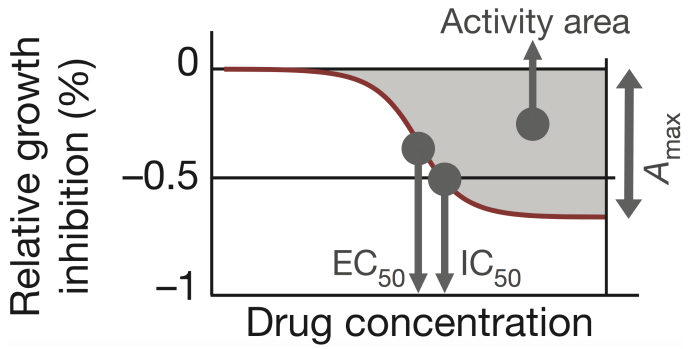

Source: TCGA, 2013

# Challenges and opportunities (1)

- Small sample size
- Several types of input data **X**:
  E.g., gene expression, copy number, mutation
- Multivariate response **Y**



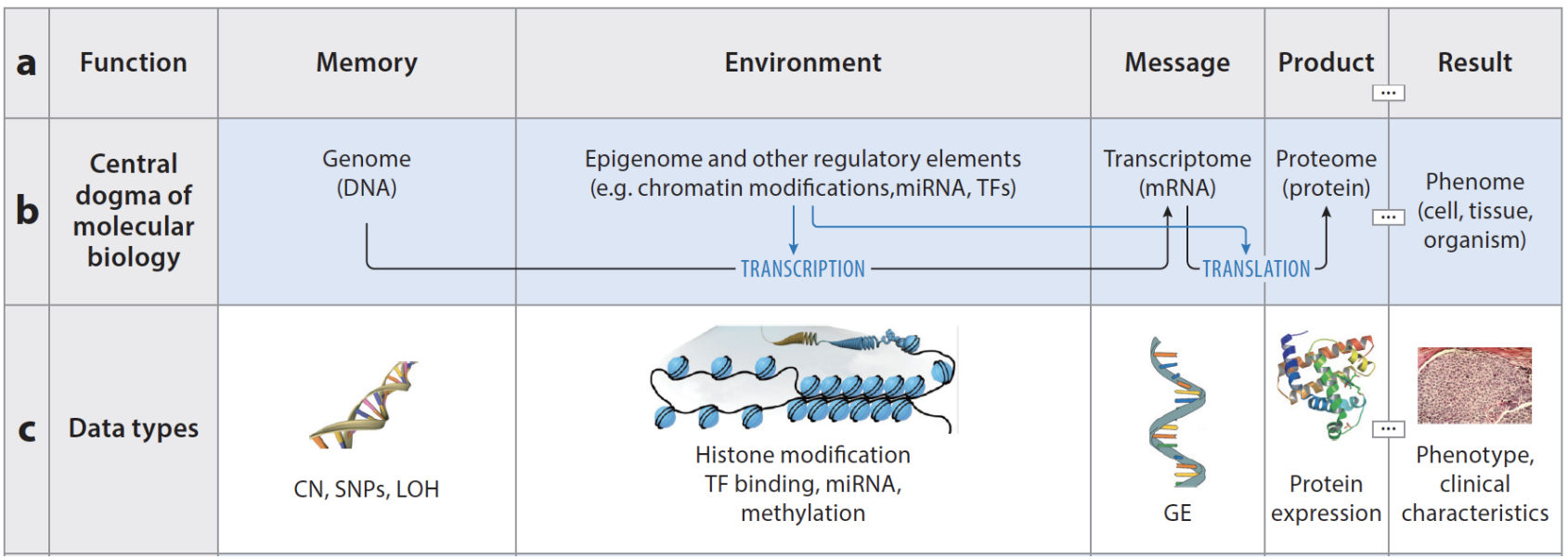

- Unclear how to define **Y**



# Challenges and opportunities (2)

The data are highly structured:

1. In **Y**: relationships between drugs, e.g. due to similar chemical drug composition, same target genes/pathways
2. In **X**: relationships between molecular data sources




Ickstadt et al. (2018)

# Supervised learning

# (with generalised linear models)

# "Observation = Signal plus Noise"

Assume the output is a (deterministic) function of $x$ plus (random) noise that has zero mean. So the expectation is the function. For instance, function $f$ below:

$$y = f_\beta(x) + \epsilon$$

Let's assume for now that $f$ is linear. We could assume that the noise term ("error term") follows a Gaussian distribution with some unknown variance.

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon, \text{ where } \epsilon \sim N(0, \sigma^2)$$

After estimating the model parameters as $\hat{\beta}_0, ..., \hat{\beta}_p$, we can predict a new outcome as

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p$$

.

# Supervised learning

## Supervised learning

refers to the task of inferring a functional relationship between **input data matrix X** (e.g. gene expression array measurements) and **output data vector** $Y$ (= response/ outcome).

The input data are used for predicting the outcome.

$$Y = f_\beta(\mathbf{X}) + \epsilon,$$

where $\epsilon$ captures measurement errors and other discrepancies, e.g. by $\epsilon \sim N(0, \sigma^2 I_n)$.

In classical statistics, this task is usually performed by (generalised) linear regression models.

# Linear models

## Linear model

refers to the regression coefficients $\beta$ being in an additive relation (linear combination). The input data $X$ can be transformed in some non-linear way:

$$\begin{aligned} Y &= \beta_0 + \beta_1 h_1(X_1) + \beta_2 h_2(X_2) + \ldots + \beta_p h_p(X_p) + \epsilon \\ &= \beta_0 + \beta_1 X_1^* + \beta_2 X_2^* + \ldots + \beta_p X_p^* + \epsilon \\ &= \mathbf{X}^* \beta + \epsilon \\ \\ E(Y) &= \mathbf{X}^* \beta \end{aligned}$$

## Linear predictor:

We call the linear combination $\mathbf{X}^*\beta$ - or $\mathbf{X}\beta$ for untransformed input data - the linear predictor.

# Recap: Ordinary least squares regression

## Estimation of the model parameters (regression coefficients):

$$\begin{aligned} \text{RSS} &= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_p x_{ip})^2 \end{aligned}$$

- We estimate the regression coefficient estimates $\hat{\beta}_i$ by minimising the sum of squared residuals (RSS) $\rightarrow$ method of least squares

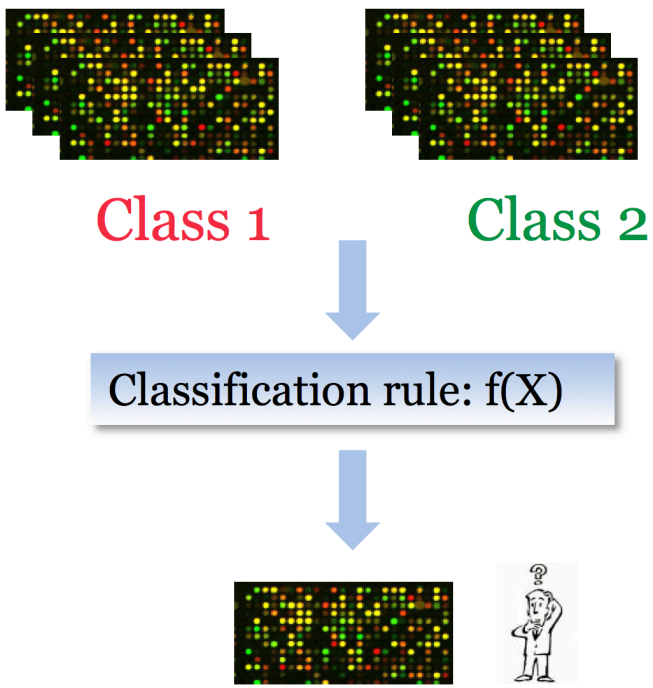
- RSS = negative log-likelihood function $\ell(\beta)$. $\rightarrow$
- Minimise RSS $\equiv$ maximise (log-)likelihood $\rightarrow$
- Least-squares is here equivalent to Maximum-likelihood.

# Regression versus classification

## Classification

If the outcome is binary/ categorical, the prediction task is called classification. In that case, a prediction model is called classifier.

- Data:
  Samples + pre-defined classes
- Goal:
  1. find a classification rule (e.g. gene signature)
  2. predict class of a new sample



# Regression versus classification

- We tend to refer to problems with a quantitative response as regression problems,
- while those involving a qualitative response are often referred to as classification problems.

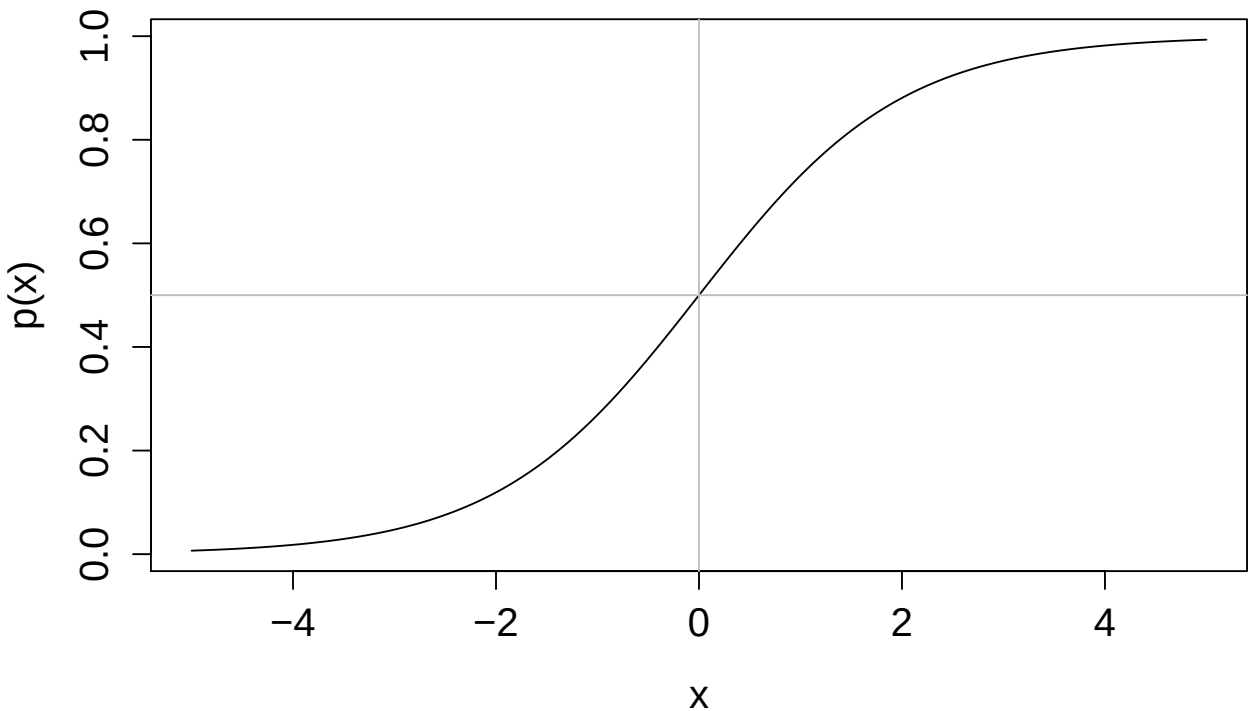
- But logistic regression is often used for classification...
  - If we use the class probabilities: regression
  - If we only use the predicted class labels: classification

- For classification tasks: assume in the following, that we want to predict binary class labels coded as $y_i \in \{0, 1\}$.

# Logistic regression

- A logit transformation will transform probabilities $p(x_i)$ to a continuous scale.
- Can then go back to fitting a linear model:

$$g(p(x_i)) = \log \frac{p(x_i)}{1 - p(x_i)} = \beta_0 + \beta_1 x_i$$



# Examples of (generalised) linear regression

$$\begin{aligned} E(Y) &= g^{-1}(\mathbf{X}\beta) \\ Var(Y) &= g^{-1}(\mathbf{X}\beta) \end{aligned}$$

**Different outcome types modeled with link functions $g(.)$:**

- Continuous outcome $\rightarrow$ *(multiple) linear regression*
  - expression of a particular gene/protein/miRNA, biomarkers
- Binary/categorical (groups) $\rightarrow$ *logistic regression*
  - tumour vs. normal tissue, therapy responders vs. non-responders, mutation status
- Counts $\rightarrow$ e.g. *Poisson* or *Negative binomial regression*
  - gene expression for a single gene as measured by RNA-seq
- Survival (time to an event) $\rightarrow$ *Cox PH regression*
  - overall survival, progression-free survival

# Key points

- Example objective:
  Identify omics features of tumour samples that predict response to specific cancer therapies of individual patients
- Machine learning goals:
  Predictive modelling by supervised learning with some structure learning.
- "Short, fat" data sets:
  small sample size $n$ (10s to 100s) compared to large number of features $p$ (100s to 10,000s): $p >> n$
- Key:
  Use known structure in the data to restrict the model space.

# "Classical" regularisation: penalties and hierarchical priors

# Regularisation

Key to dealing with the $p >> n$ problem: Restrict the parameter space (i.e. add constraints to $\beta$) $\rightarrow$ Regularisation!

It is often reasonable to assume that only a small number of all genes are linked to the response: Good models are often sparse.

**Advantages of regularisation**

- It can make the prediction method more accurate by reducing the effect of noisy genes.
- It can do automatic variable selection.
  In particular, if a gene effect estimate is shrunk to zero, then it is eliminated from the prediction rule.

# Penalised (generalised) linear regression

# Penalised regression

- Standard regression cannot deal with $p >> n$:
  - The maximum-likelihood estimate $\hat{\beta} = \arg\max_\beta \ell(\beta)$ does not exist ($\ell$ = log-likelihood).

- **Solution:**
  Penalise the likelihood function by subtracting a penalty term and maximise penalised log-likelihood instead:

$$\hat{\beta} = \arg\max_\beta(\ell(\beta) - \lambda||\beta||)$$

  - $\lambda$ is a penalty parameter,
  - $||\beta||$ represents the size of the regression coefficient vector,
  - The larger $\lambda$ is chosen, the more the algorithm is encouraged to find a solution where $||\beta||$ is small $\rightarrow$ shrinkage.

# Penalised regression

- Examples for penalty terms:
  - Ridge regression (Hoerl and Kennard 1970):
    $\lambda||\beta|| := \lambda \sum_{g=1}^{p} \beta_g^2$ $\rightarrow$ **$L_2$ penalty**
  - Lasso regression (Tibshirani 1996):
    $\lambda||\beta|| := \lambda \sum_{g=1}^{p} |\beta_g|$ $\rightarrow$ **$L_1$ penalty**
  - Elastic net (Zou and Hastie 2005):
    Combination of both ridge and lasso penalty:
    $\lambda_1 \sum_{g=1}^{p} |\beta_g| + \lambda_2 \sum_{g=1}^{p} \beta_g^2$
- Advantage of lasso and elastic net:
  Both will produce a sparse solution, where only a few genes have estimate $\hat{\beta}_g \neq 0$.

# Example 1

LETTER

doi:10.1038/nature11003

## The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity

Jordi Barretina[1,2,3]†*, Giordano Caponigro[4]*, Nicolas Stransky[1]*, Kavitha Venkatesan[4]*, Adam A. Margolin[1]†*, Sungjoon Kim[5], Christopher J. Wilson[4], Joseph Lehár[4], Gregory V. Kryukov[1], Dmitriy Sonkin[4], Anupama Reddy[4], Manway Liu[4], Lauren Murray[1], Michael F. Berger[1]†, John E. Monahan[4], Paula Morais[1], Jodi Meltzer[4], Adam Korejwa[1], Judit Jané-Valbuena[1,2], Felipa A. Mapa[4], Joseph Thibault[5], Eva Bric-Furlong[4], Pichai Raman[4], Aaron Shipway[5], Ingo H. Engels[5], Jill Cheng[6], Guoying K. Yu[6], Jianjun Yu[6], Peter Aspesi Jr[4], Melanie de Silva[4], Kalpana Jagtap[4], Michael D. Jones[4], Li Wang[4], Charles Hatton[3], Emanuele Palescandolo[3], Supriya Gupta[1], Scott Mahan[1], Carrie Sougnez[1], Robert C. Onofrio[1], Ted Liefeld[1], Laura MacConaill[3], Wendy Winckler[1], Michael Reich[1], Nanxin Li[5], Jill P. Mesirov[1], Stacey B. Gabriel[1], Gad Getz[1], Kristin Ardlie[1], Vivien Chan[6], Vic E. Myer[4], Barbara L. Weber[4], Jeff Porter[4], Markus Warmuth[4], Peter Finan[4], Jennifer L. Harris[5], Matthew Meyerson[1,2,3], Todd R. Golub[1,3,7,8], Michael P. Morrissey[4]*, William R. Sellers[4]*, Robert Schlegel[4]* & Levi A. Garraway[1,2,3]*

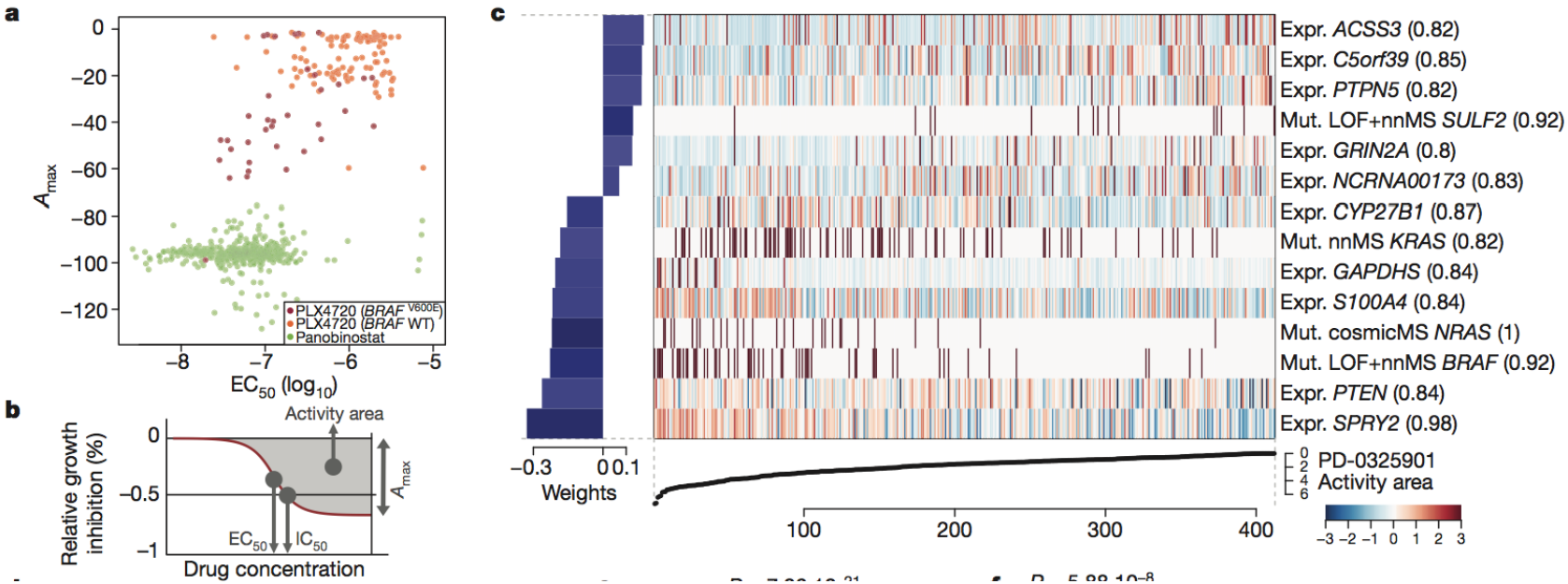



Figure 2. Predictive modelling of pharmacological sensitivity using CCLE genomic data. a, b, Drug responses for panobinostat (green) and PLX4720 (orange/purple) represented by the high-concentration effect level (Amax) and transitional concentration (EC50) for a sigmoidal fit to the response curve (b). c, Elastic net regression modelling of genomic features that predict sensitivity to PD-0325901. [...]

- Outcome $Y$: Activity area (from dose response curve)
- Model: Elastic net

# Example 2


ARTICLE

Received 18 Aug 2014 | Accepted 18 Nov 2014 | Published 9 Jan 2015

DOI: 10.1038/ncomms6901 OPEN

## Combining gene mutation with gene expression data improves outcome prediction in myelodysplastic syndromes

Moritz Gerstung[1,*], Andrea Pellagatti[2,*], Luca Malcovati[3,4], Aristoteles Giagounidis[5], Matteo G. Della Porta[3,6], Martin Jädersten[7], Hamid Dolatshad[2], Amit Verma[8], Nicholas C.P. Cross[9], Paresh Vyas[10], Sally Killick[11], Eva Hellström-Lindberg[7], Mario Cazzola[3,4], Elli Papaemmanuil[1], Peter J. Campbell[1] & Jacqueline Boultwood[2]

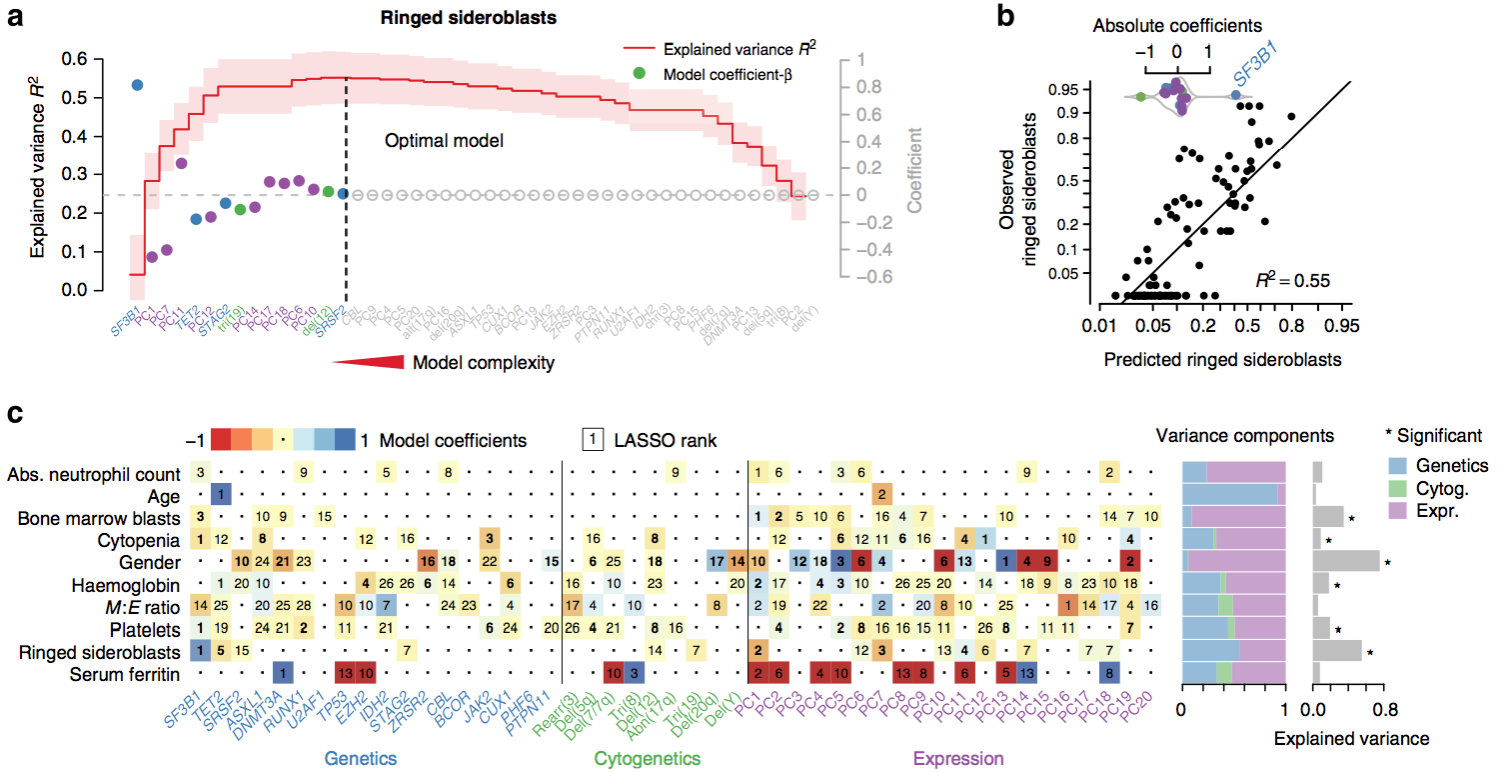



**Figure 3 | Prediction of blood and bone marrow counts.** (**a**) Variance explained by selected driver genes, cytogenetic lesions and first 20 transcriptome principal components (red line ±1 s.d.; fivefold cross validation) ordered by their occurrence in a LASSO penalized model. The optimal model maximizes the explained variance $R^2$. The right axis indicates the effect of each standardized covariate in the optimal model. (**b**) Scatter plot of predicted and observed amounts of ringed sideroblasts on a double logit axis. The inset shows the model coefficients indicating the magnitude of each fold change of driver alterations or a unit fold change in the expression components. (**c**) Heatmap of optimal model coefficients for eight blood and bone marrow counts plus gender and age. LASSO-selected coefficients are coloured. The numbers on each tile denote the order in which variables are included indicating their relative importance. Bold fonts are used for highly significant coefficients in which the explained variance is one s.d. below the maximum. The right bar plot shows the estimated distribution of variance explained by genetic, cytogenetic and transcriptomic variables. Stars (*) denote models where $R^2$ is greater than zero by a margin of more than one s.d.

- Outcome $Y$: Blood and bone marrow counts, age, gender, ...
- Model: Elastic net

# Prostate cancer example (Hastie et al, 2009):

- Predict the level of (log) prostate specific antigen (PSA) by a number of clinical measures, in 97 men who were about to receive a radical prostatectomy.

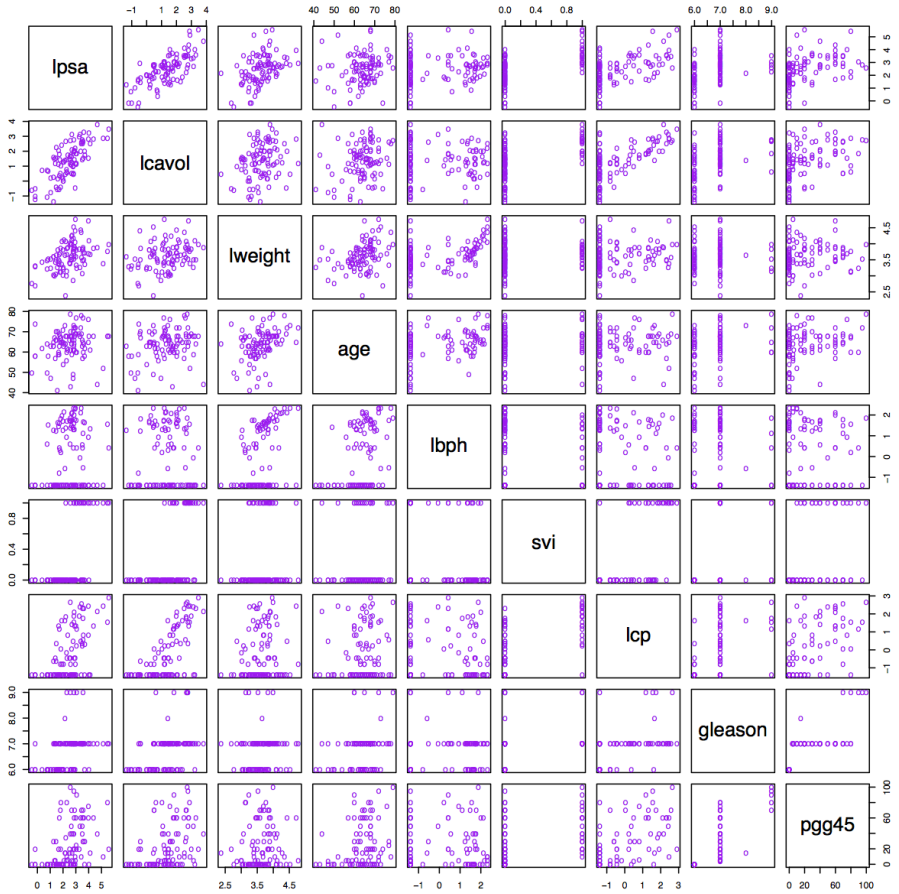



**FIGURE 1.1.** *Scatterplot matrix of the prostate cancer data. The first row shows the response against each of the predictors in turn. Two of the predictors,* `svi` *and* `gleason`*, are categorical.*

# Penalised regression

## Linear regression models to explain log(PSA):

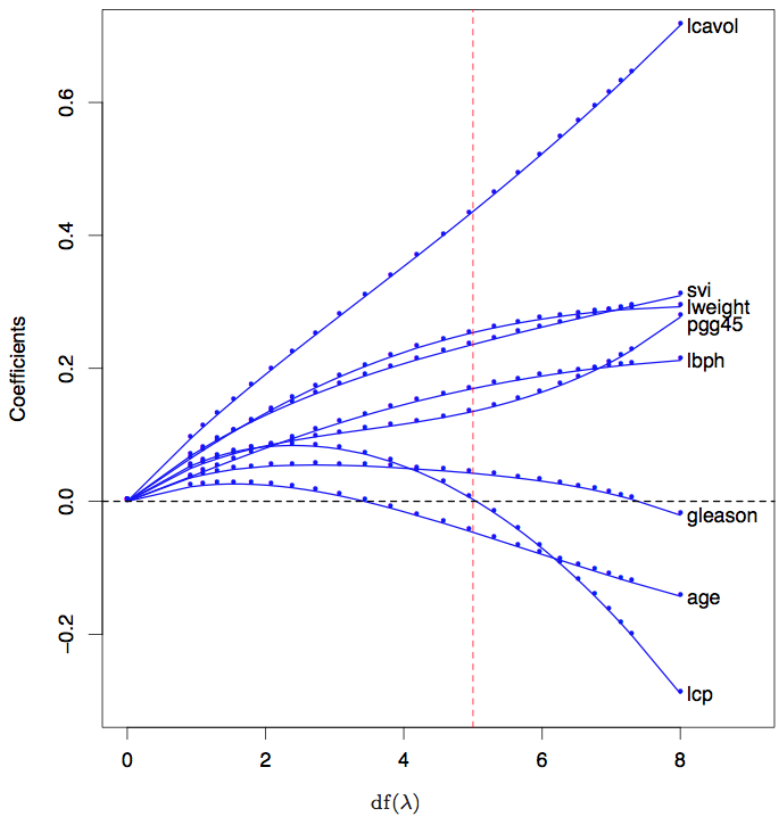
| Term | LS | Best Subset | Ridge | Lasso |
|---|---|---|---|---|
| Intercept | 2.465 | 2.477 | 2.452 | 2.468 |
| lcavol | 0.680 | 0.740 | 0.420 | 0.533 |
| lweight | 0.263 | 0.316 | 0.238 | 0.169 |
| age | −0.141 | | −0.046 | |
| lbph | 0.210 | | 0.162 | 0.002 |
| svi | 0.305 | | 0.227 | 0.094 |
| lcp | −0.288 | | 0.000 | |
| gleason | −0.021 | | 0.040 | |
| pgg45 | 0.267 | | 0.133 | |
| Test Error | 0.521 | 0.492 | 0.492 | 0.479 |
| Std Error | 0.179 | 0.143 | 0.165 | 0.164 |

Hastie et al. (2009), Table 3.3

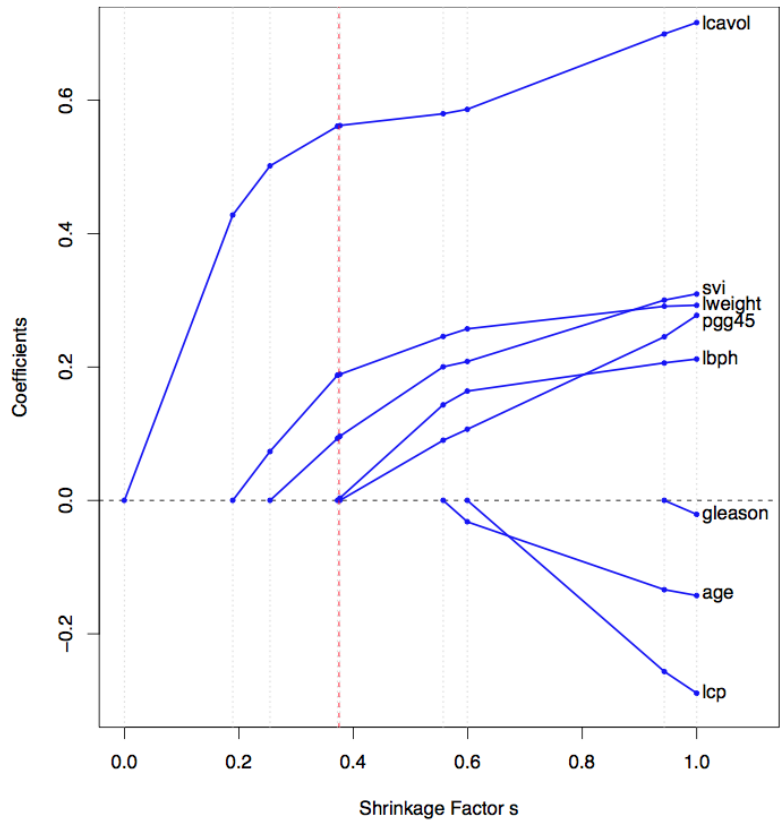
# Penalised regression

Examples for coefficient paths relative to penality $\lambda$:

## Ridge regresson




## Lasso regression




Hastie et al. (2009), Figures 3.8 and 3.10

# Penalised regression

- Ridge regression $L_2$: shrinks all coefficients to small, but non-zero values.
- Lasso regression $L_1$: shrinks some coefficients to exactly zero.
- Elastic net: mixture of the two: does shrink some coeffients to exactly zero. Keeps more variables if there is correlation.

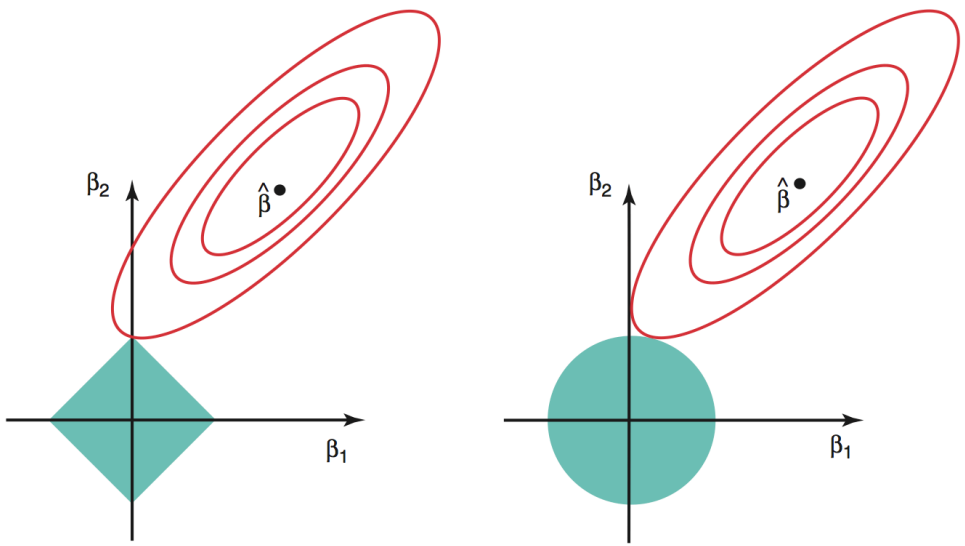
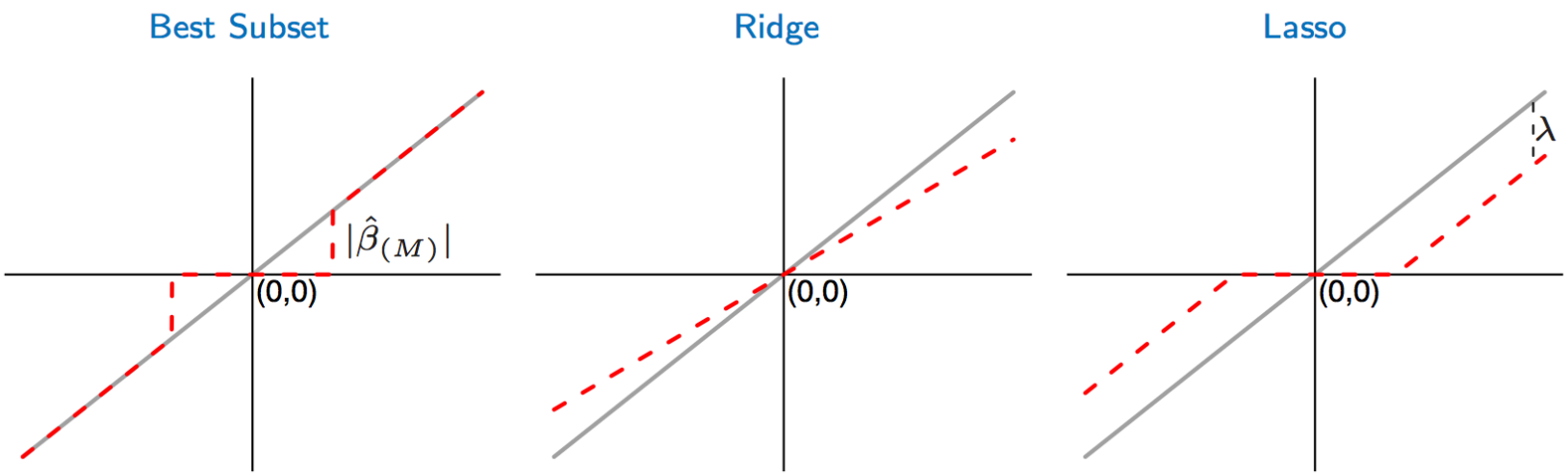



**FIGURE 6.7.** *Contours of the error and constraint functions for the lasso* (left) *and ridge regression* (right). *The solid blue areas are the constraint regions,* $|\beta_1| + |\beta_2| \leq s$ *and* $\beta_1^2 + \beta_2^2 \leq s$, *while the red ellipses are the contours of the RSS.*

James et al. (2013)

# Penalised regression

| Estimator | Formula |
|---|---|
| Best subset (size $M$) | $\hat{\beta}_j \cdot I[\text{rank}(\lvert\hat{\beta}_j\rvert \leq M)$ |
| Ridge | $\hat{\beta}_j/(1+\lambda)$ |
| Lasso | $\text{sign}(\hat{\beta}_j)(\lvert\hat{\beta}_j\rvert - \lambda)_+$ |




Hastie et al. (2009), Table 3.4

# Penalty terms: Desirable properties (e.g. Fan & Li 2001)

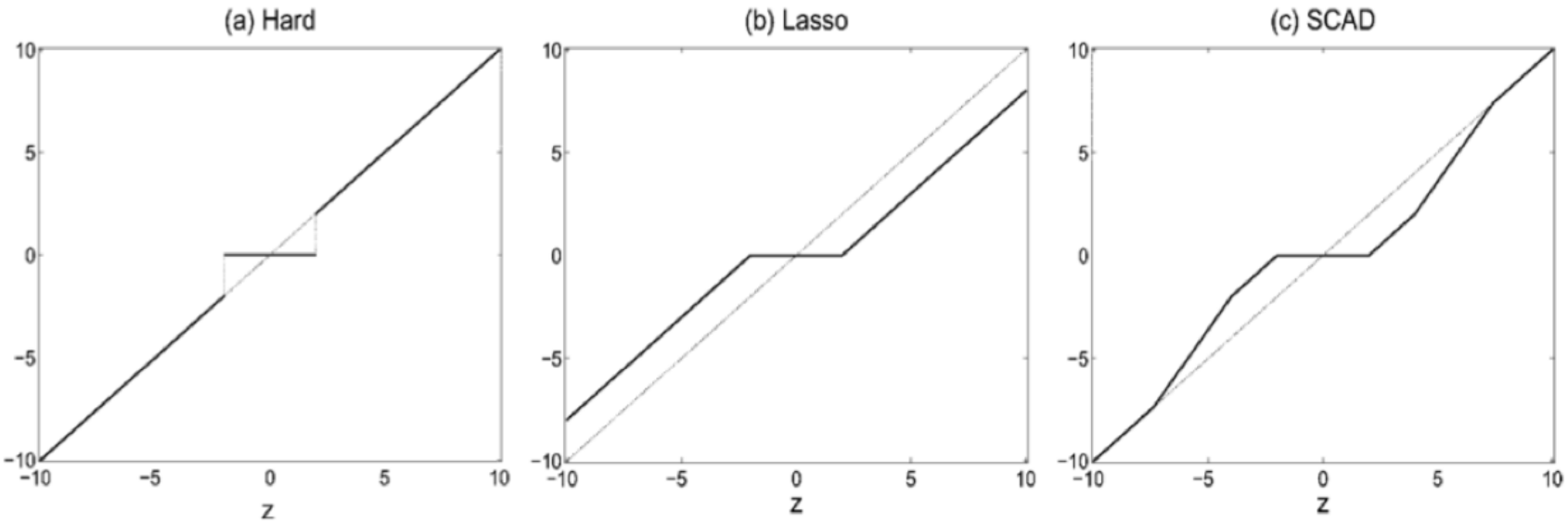
A good penalty function should result in an estimator with the following properties:

- Unbiasedness: the estimator is nearly unbiased when the true unknown parameter is large - to avoid unnecessary modelling bias.
- Sparsity: the estimator is a thresholding rule, which automatically sets small estimated coefficients to zero - to reduce model complexity.
- Continuity: the estimator is continuous - to avoid instability in model prediction.

- Ridge, lasso and elastic net estimators are continuous.
- Lasso and elastic net estimators are sparse.
- All three estimators are biased, even for large parameters.

# Penalty terms: Introducing unbiasedness

- Adaptive lasso (Zou 2006):
  $p_\lambda(|\beta|) = \lambda \sum_{j=1}^{p} w_j |\beta_j|$, e.g. with weights $w_j = 1/|\beta_j^{OLS}|$
- SCAD (Smoothly Clipped Absolute Deviation, Fan & Li 2001):
  $p_\lambda(0) = 0$ and $p_\lambda(|\beta|)$ for $|\beta| > 0$, such that
  $p'_\lambda(|\beta|) = \lambda I(|\beta| \leq \lambda) + \frac{(a\lambda - |\beta|)_+}{(a-1)} I(|\beta| > \lambda)$ with $a > 2$




Fan and Li (2001), Figure 2

# Model selection consistency and oracle properties

## Model selection consistency

A model is model selection consistent, if the true model is found asymptotically: $\lim_{n\to\infty} Pr(\hat{M}_n = M) = 1$,
where $M$ is the set of variables in the true model of size $d$
$M = \{j : \beta_j \neq 0\}$.

## Oracle property

An estimator has the oracle property, if one could not improve on the asymptotic results given by the estimator, even if one would know the true model $M$ in advance.

- The oracle property implies model selection consistency and
- that parameter estimates of the true model are asymptotically unbiased.

# Model selection consistency and oracle properties

- Model selection consistency: fulfilled by elastic net, lasso, adaptive lasso, SCAD
- Oracle property: Adaptive lasso and SCAD $\rightarrow$ but only if the true model size $d << n$

**Problems with adaptive lasso and SCAD:**

- Performance depends on the unknown "truth": they will only perform well, if number of true predictors $d$ is much smaller than sample size $n$.
- Based on iterative algorithms, starting solutions (with $p^* < n$).
- Performance depends on good choice for starting solution.

$\rightarrow$ Do not use, if no prior knowledge about expected size of true model is available.

# Model selection consistency and oracle properties

- Model selection consistency: fulfilled by elastic net, lasso, adaptive lasso, SCAD
- Oracle property: Adaptive lasso and SCAD $\rightarrow$ but only if the true model size $d << n$

**Problems with adaptive lasso and SCAD:**

- Performance depends on the unknown "truth": they will only perform well, if number of true predictors $d$ is much smaller than sample size $n$.
- Based on iterative algorithms, starting solutions (with $p^* < n$).
- Performance depends on good choice for starting solution.

$\rightarrow$ Do not use, if no prior knowledge about expected size of true model is available.

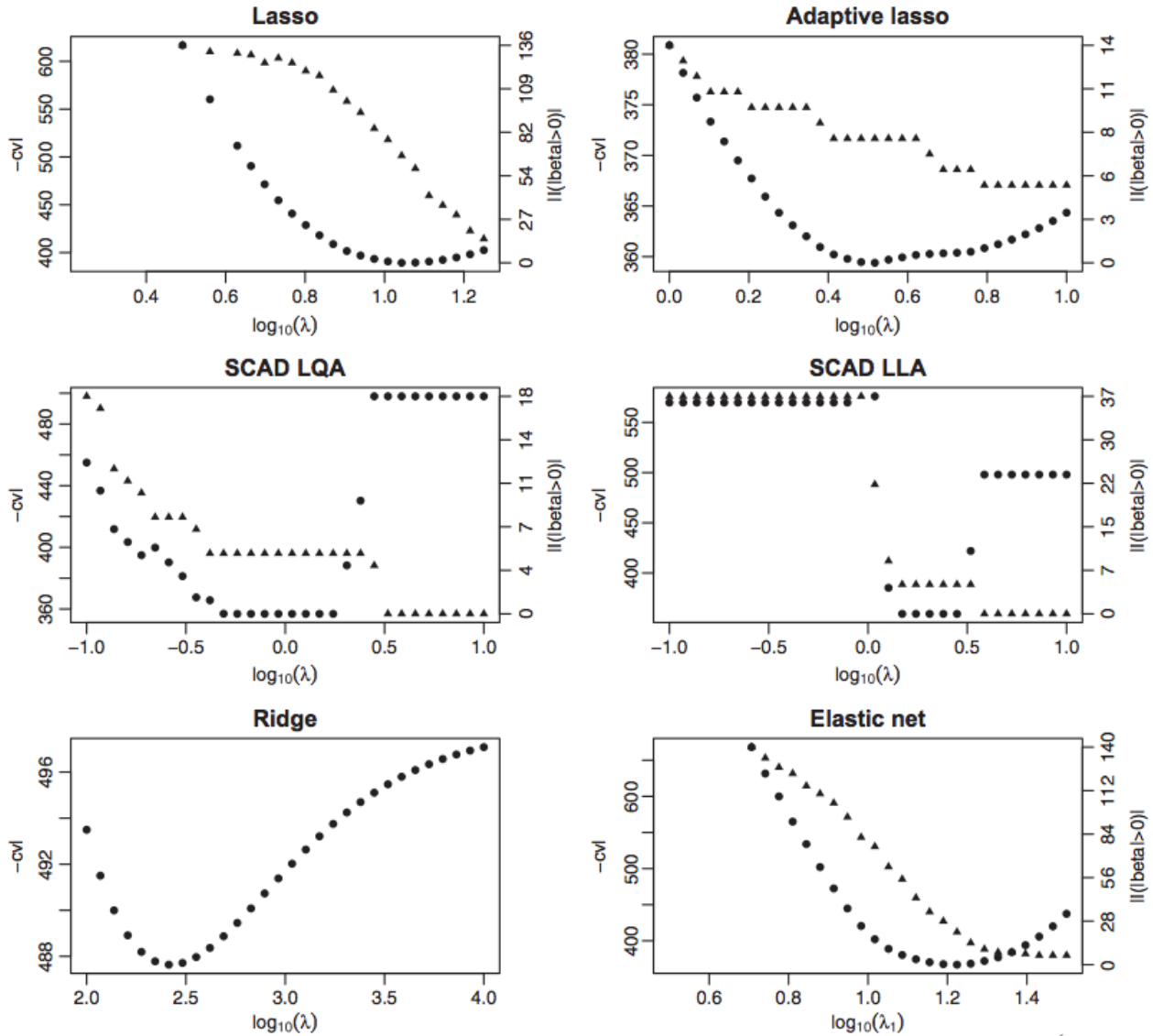
# Model selection: Competing goals

**What makes a good prediction model in molecular biology?**

- Make good predictions
- Learn about the biology: Which genes play an important role

- Prediction accuracy: low generalisation error
- Interpretability
  - Parsimony: small number of genes that can be followed up in biological experiments
  - Interpretable model: explicit modelling of relationship between genes and response (no "black box")
  - Stability: little variation in resulting profile when training data are varied (resampled)

# Comparison of penalty functions




Benner et al. (2010), Figure 1

# Comparison of penalty functions

## Simulation 1: Very sparse scenario

**Table 2** High-dimensional example: Results of simulations with $d = 5$ response-related variables and independence.[a)]

| | $MSE(\hat{\beta}_{true})$ | FP | FN | IBS | $R^2_{IBS}$ |
|---|---|---|---|---|---|
| SCAD-LQA | 0.02 | 0 | 0 | 0.073 | 0.692 |
| SCAD-LLA | 0.03 | 0 | 0 | 0.074 | 0.684 |
| Adaptive lasso | 0.05 | 1 | 0 | 0.075 | 0.684 |
| Lasso | 0.20 | 68 | 0 | 0.088 | 0.620 |
| Elastic net | 0.04 | 21 | 0 | 0.078 | 0.665 |
| Ridge | 2.05 | – | – | 0.217 | 0.064 |
| Oracle Cox | 0.02 | – | – | 0.072 | 0.693 |

a) The median values across 100 simulation runs are shown.

Benner et al. (2010), Table 2

# Comparison of penalty functions

## Simulation 2: Moderately sparse scenario

**Table 4** High-dimensional example: Results of simulations with with $d = 30$ response-related variables and independence.[a)]

| | $MSE(\hat{\beta}_{true})$ | FP | FN | IBS | $R^2_{IBS}$ |
|---|---|---|---|---|---|
| SCAD-LQA | 2.21 | 0 | 29 | 0.238 | 0.007 |
| SCAD-LLA | 2.25 | 0 | 30 | 0.241 | 0.000 |
| Adaptive lasso | 2.04 | 4 | 20 | 0.206 | 0.147 |
| Lasso | 1.86 | 57 | 5 | 0.166 | 0.320 |
| Elastic net | 1.86 | 56 | 5 | 0.166 | 0.321 |
| Ridge | 2.14 | – | – | 0.224 | 0.084 |
| Oracle Cox | 0.13 | – | – | 0.036 | 0.852 |

a) The median values across 100 simulation runs are shown.

Benner et al. (2010), Table 4

# Key points

- "Classical regularisation": add penalty term to loss function (constrained optimisation)
- Goal: Choose penalty term that fits with known data structure and modeling goals
- Competing modeling goals:
  Prediction accuracy vs interpretability (parsimony, stability)

- Sparsity: if only a minority of features expected to be relevant
- Unbiasedness: if unbiased estimation of effects is important (not relevant if prediction is only goal)
- Model selection consistency: if we care about finding the "right" features (e.g. to learn about biological processes)

# Different penalties for different types of data

Assume two data matrices $\mathbf{X}$ and $\mathbf{Z}$:

$$Y = \mathbf{X}\beta + \mathbf{Z}\gamma + \epsilon$$

- Mandatory covariates: Do not penalise the parameters $\gamma$:

$$\ell_{\text{pen}}(\beta, \gamma) = \ell(\beta, \gamma) - \lambda||\beta||$$

e.g. with R packages `glmnet` or `penalized`

- Several types of molecular data sets:
Allow different penalties for $\beta$ and $\gamma$:

$$\ell_{\text{pen}}(\beta, \gamma) = \ell(\beta, \gamma) - \lambda_\beta||\beta|| - \lambda_\gamma||\gamma||$$

e.g. with R packages `GRridge` (Van de Wiel *et al.*, 2016)
`http://www.few.vu.nl/~mavdwiel/grridge.html`)

# Different penalties for different types of data

Assume two data matrices **X** and **Z**:

$$Y = \mathbf{X}\beta + \mathbf{Z}\gamma + \epsilon$$

- Several types of molecular data sets:
- Alternative: **Combine all data and use one penalty**, after scaling all features to unit variance to ensure that the data sources are treated equally.
- Example: Elastic Net models in Barretina et al. (2012)

# Bayesian Variable Selection

(see O'Hara and Sillanpää, 2009, for a comprehensive review)

# Bayesian interpretation of penalised regression

- Ridge regression as a penalised log-likelihood problem ...

$$\hat{\beta} = \arg\max_{\beta}(\ell - \lambda \sum_{i=1}^{p} \beta_i^2)$$

- ... is equivalent to *maximum a posteriori* solution of Bayesian linear regression with Gaussian prior

$$p(\beta|\tau) = N(0, \tau I_p), \text{ where } \tau = 1/(2\lambda) :$$

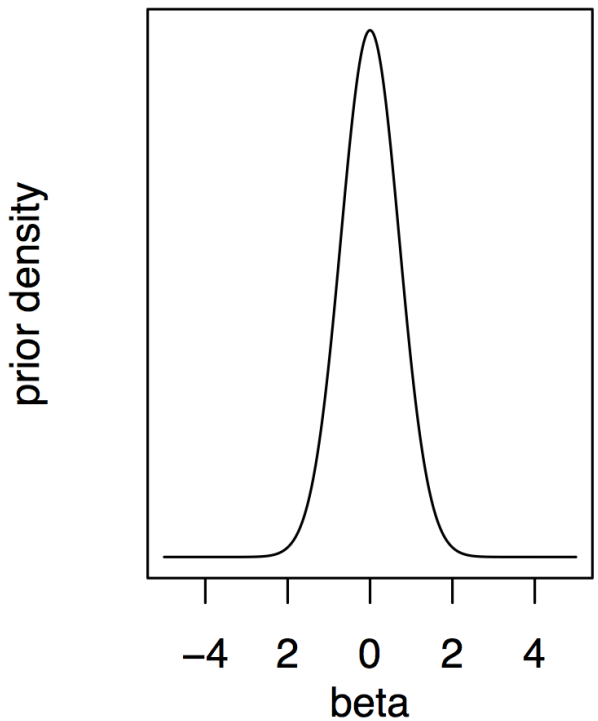
$$\begin{aligned} \text{posterior} &\propto \text{likelihood} \times \text{prior} \\ p(\beta|X, Y, \tau) &\propto p(Y|\beta, X)p(\beta|\tau) \\ \Leftrightarrow log(\beta|X, Y, \tau) &\propto \ell + \log p(\beta|\tau) \\ &\propto \ell - \lambda \sum_{i=1}^{p} \beta_i^2 - C \end{aligned}$$

# Bayesian interpretation of penalised regression

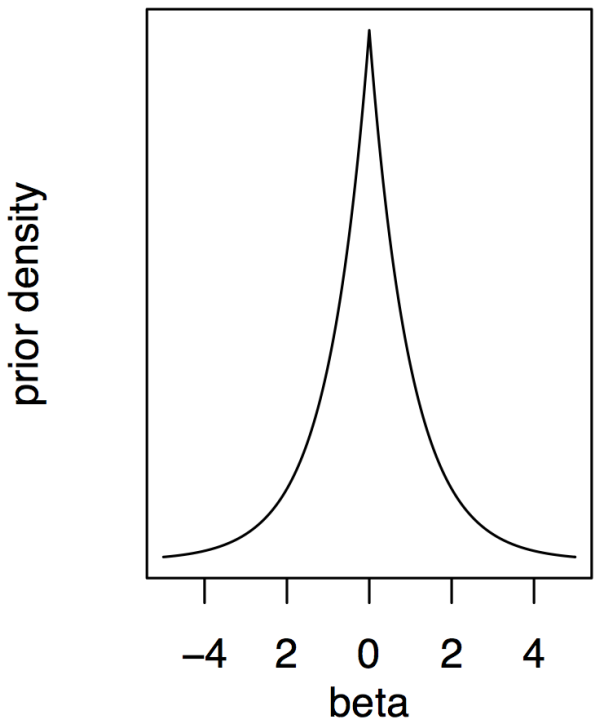
- Generalisation: Bridge regression (e. g. Frank and Friedman 1993)

$$\hat{\beta} = \arg\max_{\beta}(\ell - \lambda \sum_{i=1}^{p} |\beta_i|^q) \quad (q > 0)$$

**Ridge (q = 2): Gaussian prior**



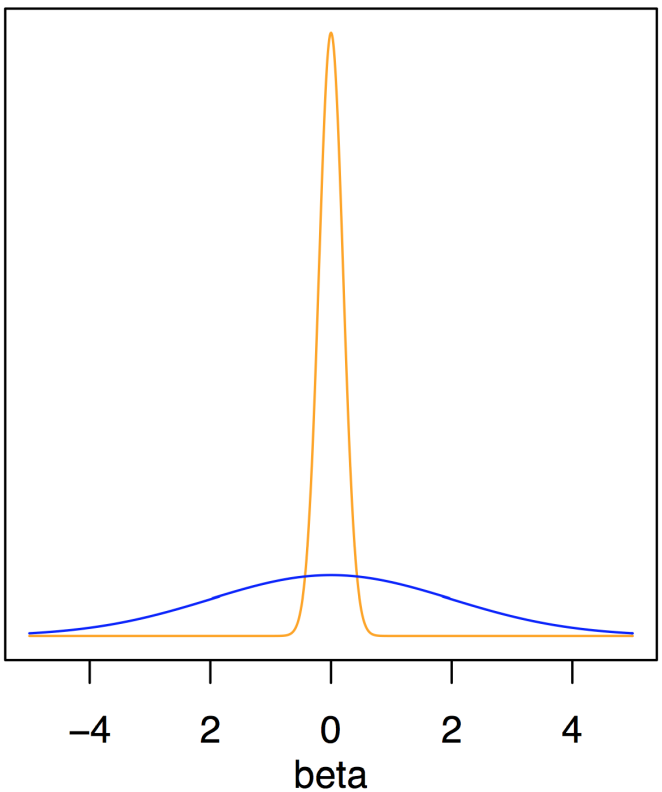

**Lasso (q = 1): Laplace prior**



# Bayesian hierarchical model for variable selection (BVS)

- Bayesian variable selection model with indicator variable

$$\gamma_i = \begin{cases} 1 & \text{variable i is included} \\ 0 & \text{variable i is excluded} \end{cases}$$




E. g. normal mixture prior (George and McCulloch 1993):

$$\beta_i|\gamma_i \sim (1-\gamma_i)N(0,\sigma^2)+\gamma_i N(0,g\sigma^2)$$

where $\sigma^2 > 0, g > 0$

# Bayesian hierarchical model for variable selection (BVS)

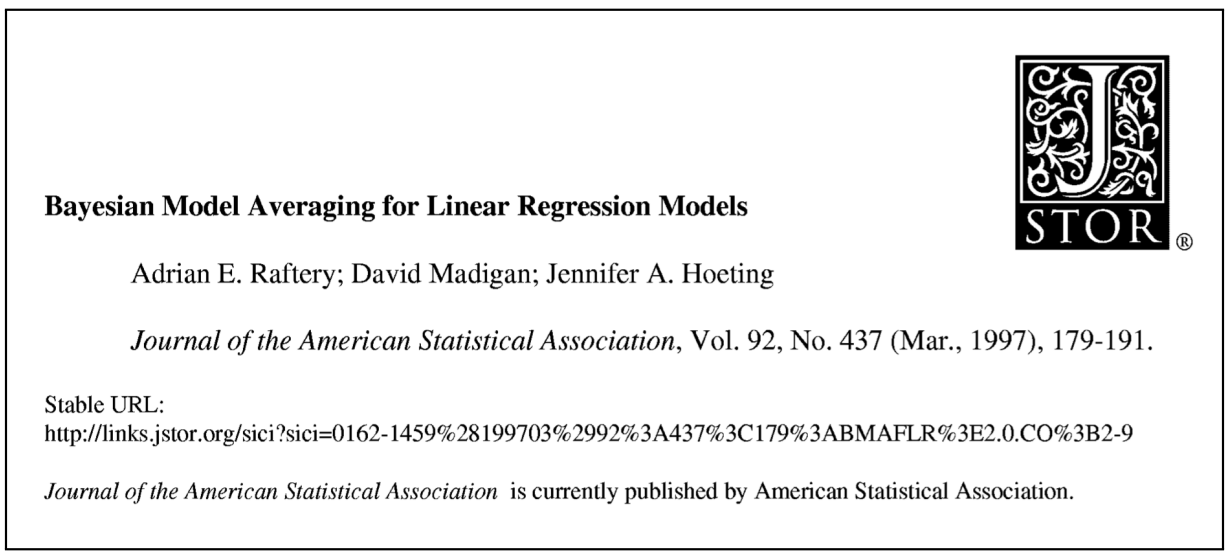
**Example (p=6):**

$$\begin{aligned} \gamma^1 &= (1,0,0,1,1,1) \quad \rightarrow \text{ variables 1, 4, 5, 6 included} \\ \gamma^2 &= (1,1,0,1,1,0) \quad \rightarrow \text{ variables 1, 2, 4, 5 included} \end{aligned}$$

- The model space is huge, of size $2^p$ (full exploration unfeasible).
- For high-dimensional data ($p >> n$) many alternative models have similar explanatory power.
- Use of MCMC methods as stochastic search algorithms.
- We favour sparse solutions via prior distribution for model size.

# Bayesian hierarchical model for variable selection (BVS)

- Flexibility in penalization through large variety of possible prior distributions for $\beta$ (hierarchical models)
- Full posterior distributions, including probabilities for the selection of variables and posterior distributions for $\beta_i$
- Improve prediction performance by Bayesian Model Averaging (**ensembling**)



**Bayesian Model Averaging for Linear Regression Models**

Adrian E. Raftery; David Madigan; Jennifer A. Hoeting

*Journal of the American Statistical Association*, Vol. 92, No. 437 (Mar., 1997), 179-191.

Stable URL:
http://links.jstor.org/sici?sici=0162-1459%28199703%2992%3A437%3C179%3ABMAFLR%3E2.0.CO%3B2-9

*Journal of the American Statistical Association* is currently published by American Statistical Association.

# Key points

- Specify priors instead of penalty terms for $\beta \rightarrow$
- Higher flexibility

- Posterior probabilities for selection of features:
  provide info on stability of selected models
- Bayesian Model Averaging as model ensembling to improve prediction

# Support vector machines as a penalisation method

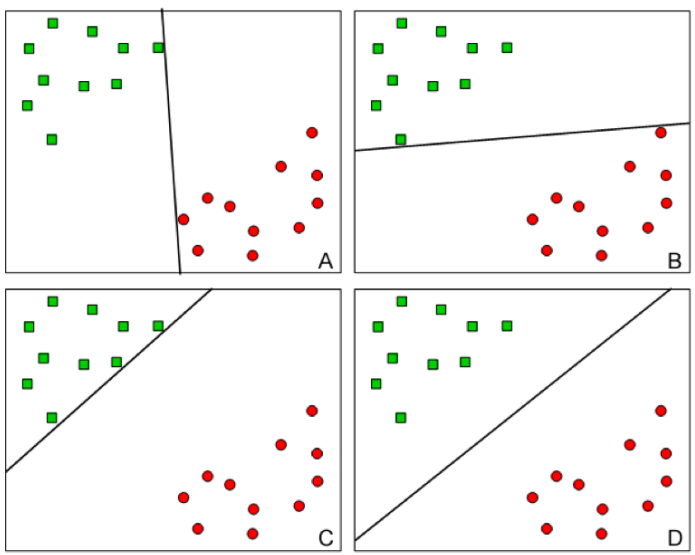
# Support Vector Machine

**Goal**

Build a classification and prediction model using relevant features (n samples << p features).

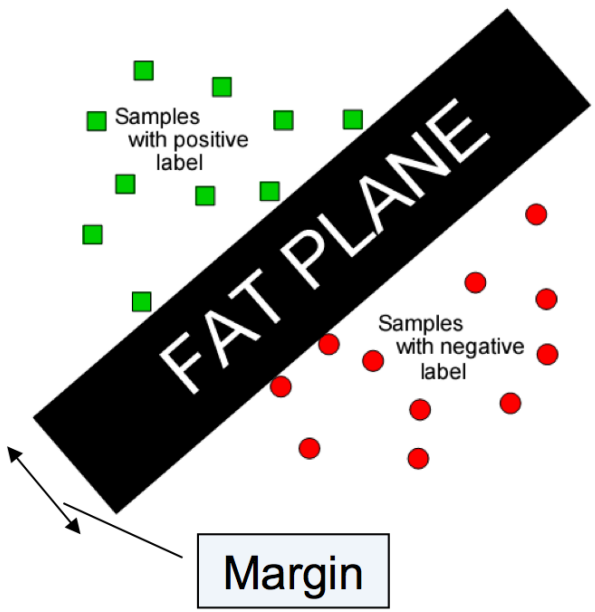
Vladimir Vapnik (1974)

**Which hyperplane is the best?**




( from F. Markowetz )

**No sharp knive, but a fat plane**



# Support Vector Machine

**Data:**
Sample x with class label y={-1,1}

**Goal:**
classification and prediction model

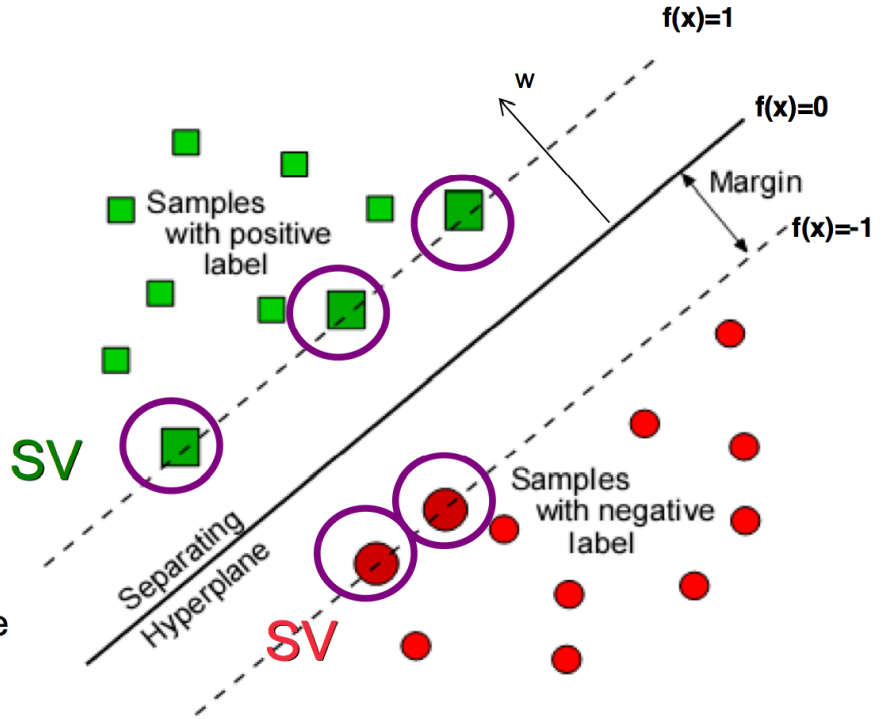
**Mathematical task**:
Separate samples by a hyperplane f(x) with maximal margin

**Prediction** :
sign of the hyperplane

$$y_{new} = Sign[f(x_{new})]$$

**Why SVM?**
Support Vectors – samples lying on the margin. Only SVs have an impact on the hyperline.

$$f(x) = w_1 x_1 + w_2 x_2 + ... + w_p x_p + b$$




( from F. Markowetz )

# Support Vector Machine

## SVM: linearly non-separable model:

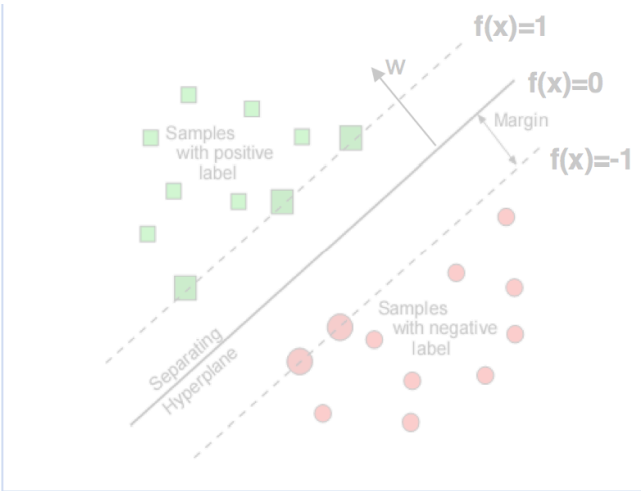
**Goal** : maximise Margin ↔ minimise ||w||

$$\min_{b,w} \|w\|_2^2 + \gamma \sum_{i=1}^{n} \xi_i,$$

$$s.t.$$

$$y_i f(x_i) \geq 1 - \xi_i,$$

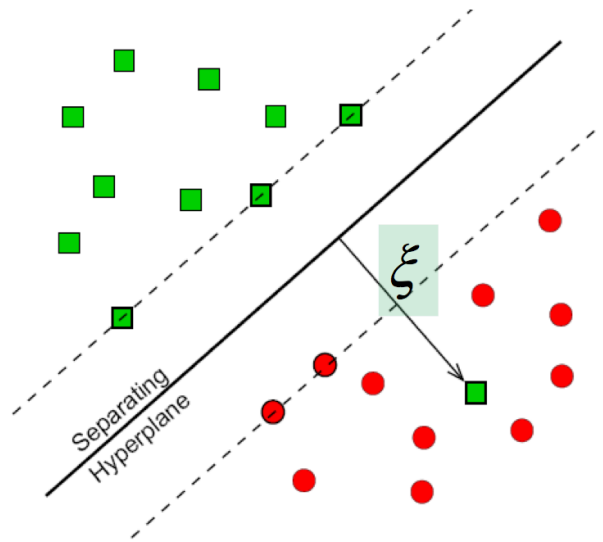
$$\xi_i > 0, \quad \text{for i} = 1,\ldots,\text{n}$$




**Soft margin SVM**:

Use linear separation, penalise training errors

Penalty of error: distance to hyperplane ξ multiplied by error cost γ



# Support Vector Machine

## SVM as a penalisation method:

With $f(x) = h(x)^T\beta + \beta_0$, consider the optimization problem

$$\min_{\beta_0,\,\beta} \sum_{i=1}^{N} [1 - y_i f(x_i)]_+ + \frac{\lambda}{2}\|\beta\|^2 \qquad (12.25)$$

where the subscript "+" indicates positive part. This has the form *loss* + *penalty*, which is a familiar paradigm in function estimation. It is easy to show (Exercise 12.1) that the solution to (12.25), with $\lambda = 1/C$, is the same as that for (12.8).

# Support Vector Machine

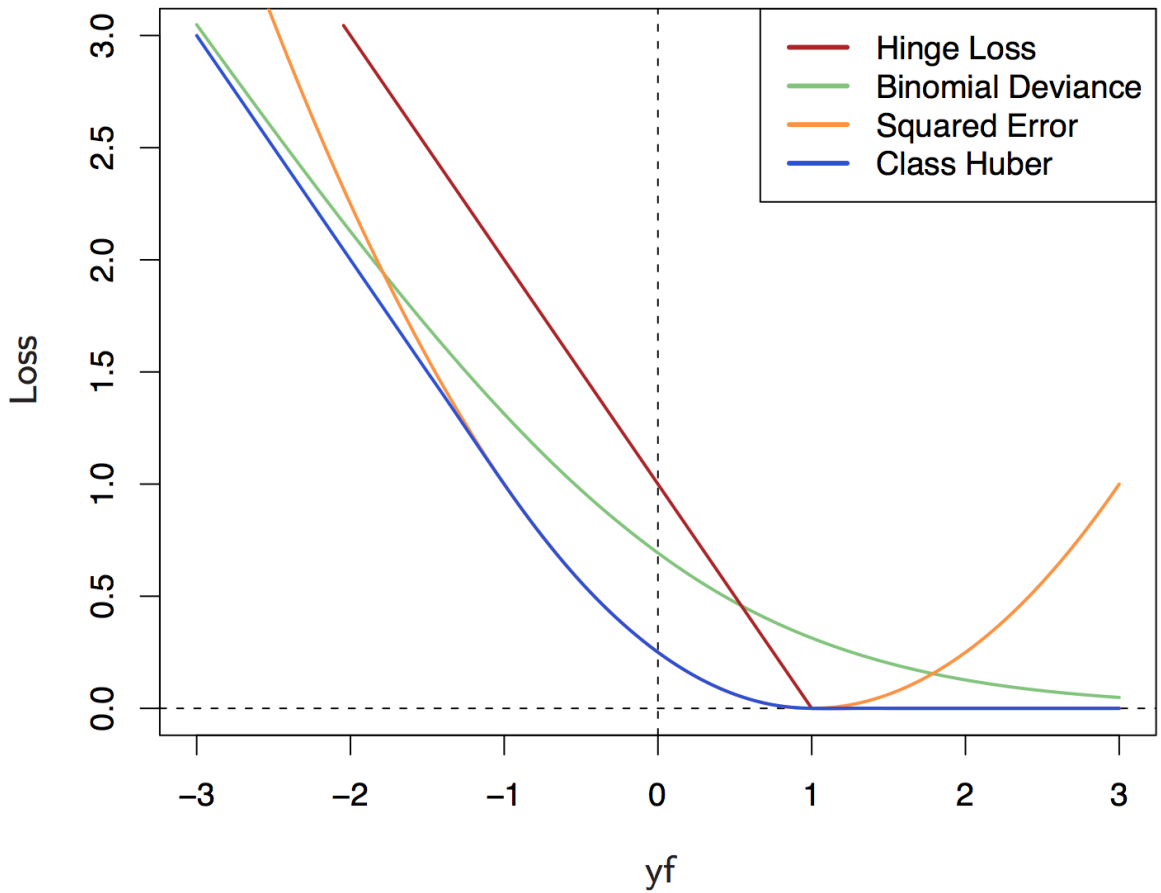
## SVM as a penalisation method:

| Loss Function | $L[y, f(x)]$ | Minimizing Function |
|---|---|---|
| Binomial Deviance | $\log[1 + e^{-yf(x)}]$ | $f(x) = \log \frac{\Pr(Y = +1\|x)}{\Pr(Y = \text{-}1\|x)}$ |
| SVM Hinge Loss | $[1 - yf(x)]_+$ | $f(x) = \text{sign}[\Pr(Y = +1\|x) - \frac{1}{2}]$ |
| Squared Error | $[y - f(x)]^2 = [1 - yf(x)]^2$ | $f(x) = 2\Pr(Y = +1\|x) - 1$ |

Hastie et al. (2009), Table 12.1

# Support Vector Machine

## SVM as a penalisation method:



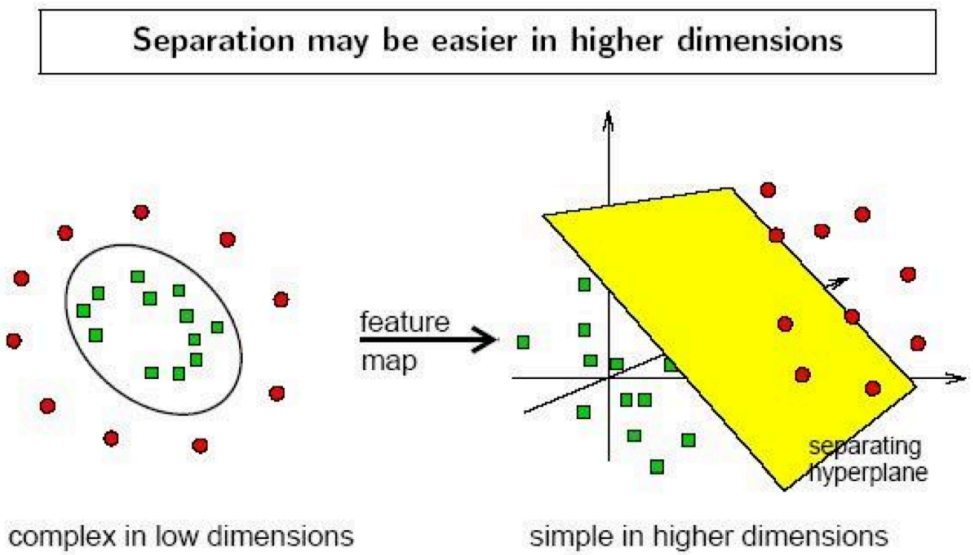

Hastie et al. (2009), Figure 12.4

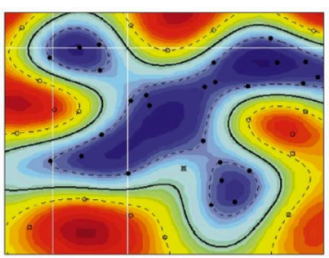
# Support Vector Machine

SVM: linearly non-separable model $\rightarrow$ kernel mapping
(applies to all penalised regressions, e.g. kernel ridge regression)

**In case of complex non-linear separation:**

- Use kernel mapping function
- Create higher dimension, where two classes can be separated by a linear hyperplane



# Key points

- SVM as a penalisation method with
- Loss function: Hinge loss
- Penalty: L2-norm

# Some literature

**Benner A, Zucknick M, Hielscher T, Ittrich C, Mansmann U (2010)**, High-dimensional Cox models: the choice of penalty as part of the model building process. *Biom J*, 52, 50–69

**Fan J, Li R (2001)**, Variable selection via nonconcave penalized likelihood and its oracle properties, *JASA*, 96, 1348–1360

**Hoerl AE, Kennard RW (1970)**, Ridge regression: biased estimation for nonorthogonal problems, *Technometrics*, 12, 55–67

**O'Hara R, Sillanpää M (2009)**, A review of Bayesian variable selection methods: what, how and which. *Bayesian Analysis* 4, 85–117

**Tibshirani R (1996)**, Regression shrinkage and selection via the lasso, *JRSSB*, 58, 267–288

**Van de Wiel M, et al. (2016)**, Better prediction by use of co-data: Adaptive group-regularized ridge regression. *Stat Med.* 35, 368–381

**Zou H, Hastie T (2005)**, Regularization and variable selection via the elastic net, *JRSSB*, 67, 301–320

**Zou H (2006)**, The adaptive lasso and its oracle properties, *JASA*, 101, 1418–1429