Lasso regression

Wessel van Wieringen
w.n.van.wieringen@vu.nl

Department of Epidemiology and Biostatistics, VUmc
& Department of Mathematics, VU University
Amsterdam, The Netherlands

vrije Universiteit

VU medisch centrum

Lasso regression

Instead of ridge why not use a different penalty? E.g.:

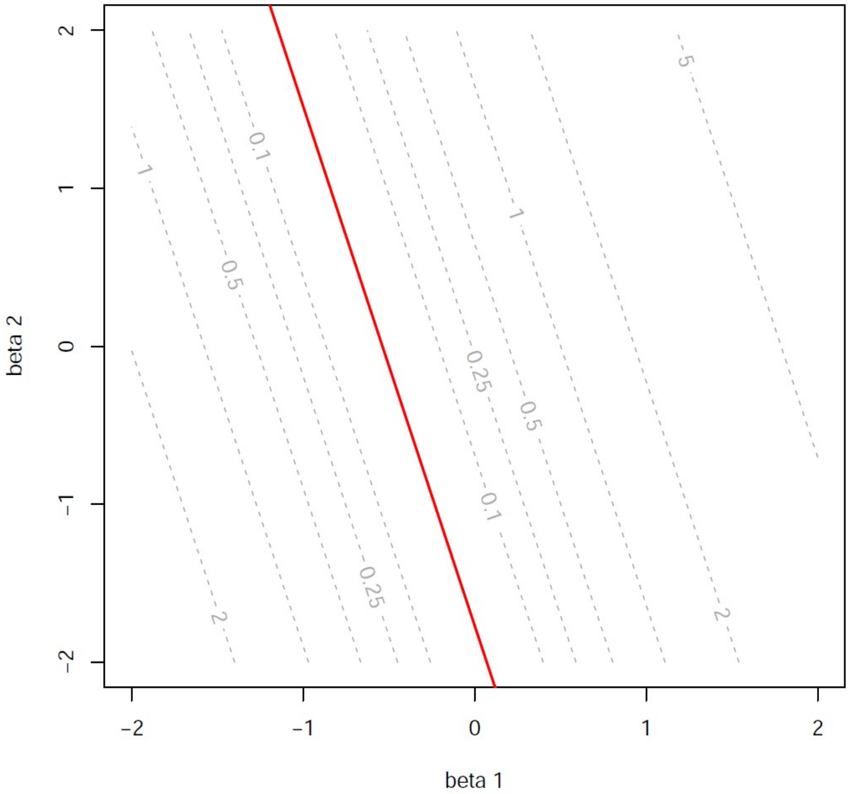
$$\begin{aligned}\mathcal{L}(\boldsymbol{\beta};\lambda) &= \|\mathbf{Y}-\mathbf{X}\,\boldsymbol{\beta}\|_2^2+\lambda_1\|\boldsymbol{\beta}\|_1\\ &= \underbrace{\sum_{i=1}^{n}(Y_i-\mathbf{X}_{i*}\,\boldsymbol{\beta})^2}_{\text{sum of squares}}+\lambda_1\underbrace{\sum_{j=1}^{p}|\beta_j|}_{\text{lasso penalty}}\end{aligned}$$

- $\lambda_1 \geq 0$ penalty parameter
- Penalty deals (super)-collinearity

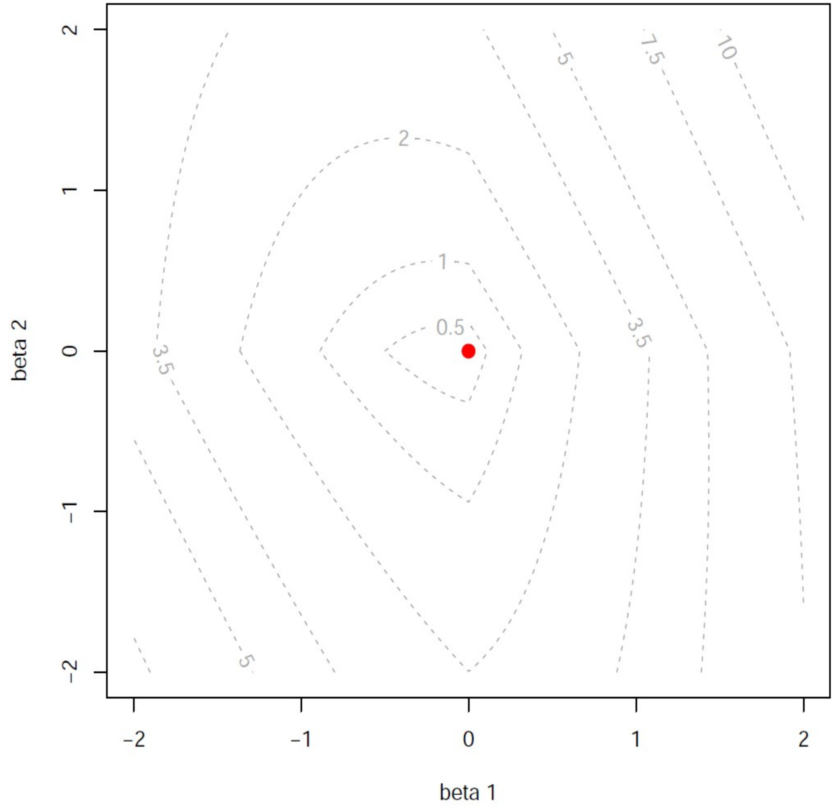
Lasso regression

Effect of the penalty on the loss function

unpenalized loss



penalized loss



The red line / dot represents the optimum (minimum) of the loss function.

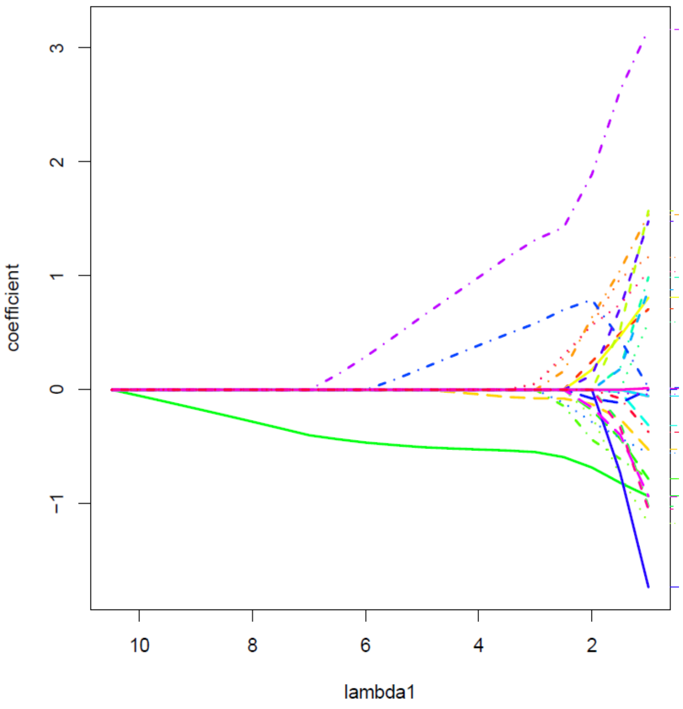
Lasso regression

Lasso regression fits the same linear regression model as ridge regression:

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon$$

The difference between ridge and lasso is in the estimators, confer the following theorem.

Theorem
The lasso loss function yields a piecewise linear (in λ_1) solution path $\beta(\lambda_1)$.



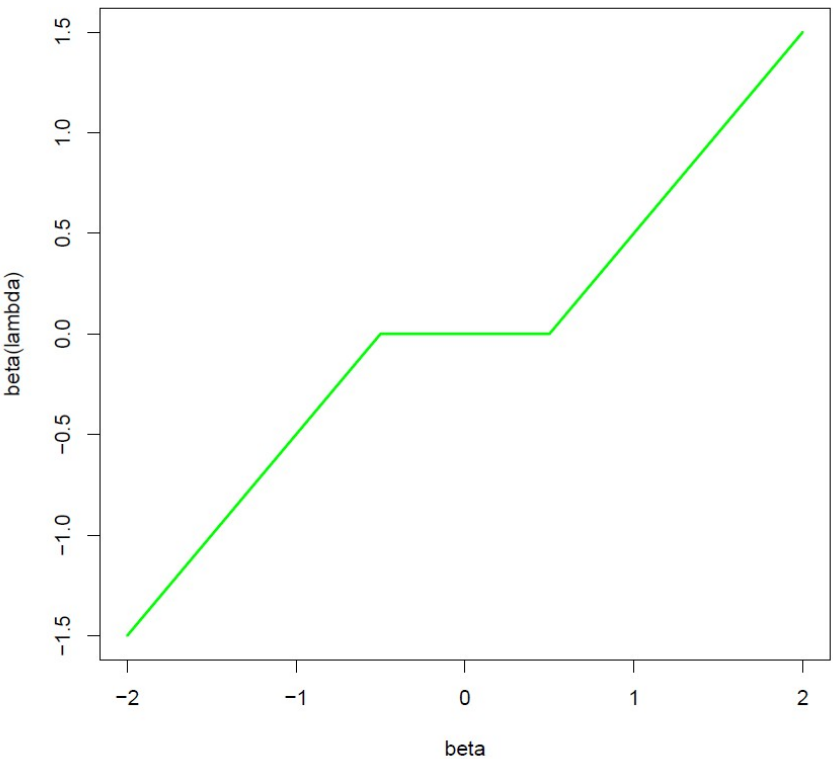
Lasso regression

In the orthonormal case, i.e. $\mathbf{X}^T\mathbf{X} = \mathbf{I} = (\mathbf{X}^T\mathbf{X})^{-1}$:

$$\hat{\beta}_j(\lambda_1) \quad = \quad \mathrm{sgn}(\hat{\beta}_j)\,(|\hat{\beta}_j| - \lambda_1/2)_+$$

Next slides for derivation.

That is, the lasso estimate is related to the OLS estimate via the so-called *soft threshold function* (depicted here for λ=1).



Lasso regression

In the orthonormal case, $\mathbf{X}^T\mathbf{X} = \mathbf{I} = (\mathbf{X}^T\mathbf{X})^{-1}$, rewrite:

$$
\begin{aligned}
\min_{\boldsymbol{\beta}} \|\mathbf{Y} - \mathbf{X}\,\boldsymbol{\beta}\|_2^2 + \lambda_1 \|\boldsymbol{\beta}\|_1 & \\
&= \min_{\boldsymbol{\beta}} \mathbf{Y}^{\mathrm{T}}\mathbf{Y} - \mathbf{Y}^{\mathrm{T}}\mathbf{X}\,\boldsymbol{\beta} - \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{Y} + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\,\boldsymbol{\beta} + \lambda_1 \sum_{j=1}^{p} |\beta_j| \\
&\propto \min_{\boldsymbol{\beta}} -[\hat{\boldsymbol{\beta}}^{\mathrm{OLS}}]^{\mathrm{T}}\,\boldsymbol{\beta} - \boldsymbol{\beta}^{\mathrm{T}}\,\hat{\boldsymbol{\beta}}^{\mathrm{OLS}} + \boldsymbol{\beta}^{\mathrm{T}}\,\boldsymbol{\beta} + \lambda_1 \sum_{j=1}^{p} |\beta_j| \\
&= \min_{\beta_1,\ldots,\beta_p} \sum_{j=1}^{p} \big(-2\hat{\beta}_j^{\mathrm{OLS}}\,\beta_j + \beta_j^2 + \lambda_1 |\beta_j| \big) \\
&= \sum_{j=1}^{p} \big(\min_{\beta_j} -2\hat{\beta}_j^{\mathrm{OLS}}\,\beta_j + \beta_j^2 + \lambda_1 |\beta_j| \big).
\end{aligned}
$$

Lasso regression

Minimization can be done per regression coefficient:

$$\min_{\beta_j} -2\hat{\beta}_j^{\text{OLS}}\,\beta_j + \beta_j^2 + \lambda_1|\beta_j| \quad =$$

$$\left\{ \begin{array}{ll} \min_{\beta_j} -2\hat{\beta}_j^{\text{OLS}}\,\beta_j + \beta_j^2 + \lambda_1\beta_j & \text{if} \quad \beta_j > 0 \\ \min_{\beta_j} -2\hat{\beta}_j^{\text{OLS}}\,\beta_j + \beta_j^2 - \lambda_1\beta_j & \text{if} \quad \beta_j < 0 \end{array} \right.$$

Solving the right-hand side yields:

$$\hat{\beta}_j^{\text{lasso}}(\lambda_1) \quad = \quad \left\{ \begin{array}{ll} \hat{\beta}_j^{\text{OLS}} - \frac{1}{2}\lambda_1 & \text{if} \quad \beta_j > 0 \\ \hat{\beta}_j^{\text{OLS}} + \frac{1}{2}\lambda_1 & \text{if} \quad \beta_j < 0 \end{array} \right.$$

Lasso regression

Convexity

Both the sum of squares and the lasso penalty are convex, and so is the lasso loss function. Consequently, there exist a global minimum. However, the lasso loss function is not strictly convex. Consequently, there may be multiple β's that minimize the lasso loss function.*

Problem

In general, there is no explicit solution that optimizes the lasso loss function.

Solution

Resort to numerical optimization procedures, e.g., gradient ascent.

* The possible existence of multiple β's was kindly pointed out to by José Pablo González.

Constrained estimation and the selection property

Constrained estimation

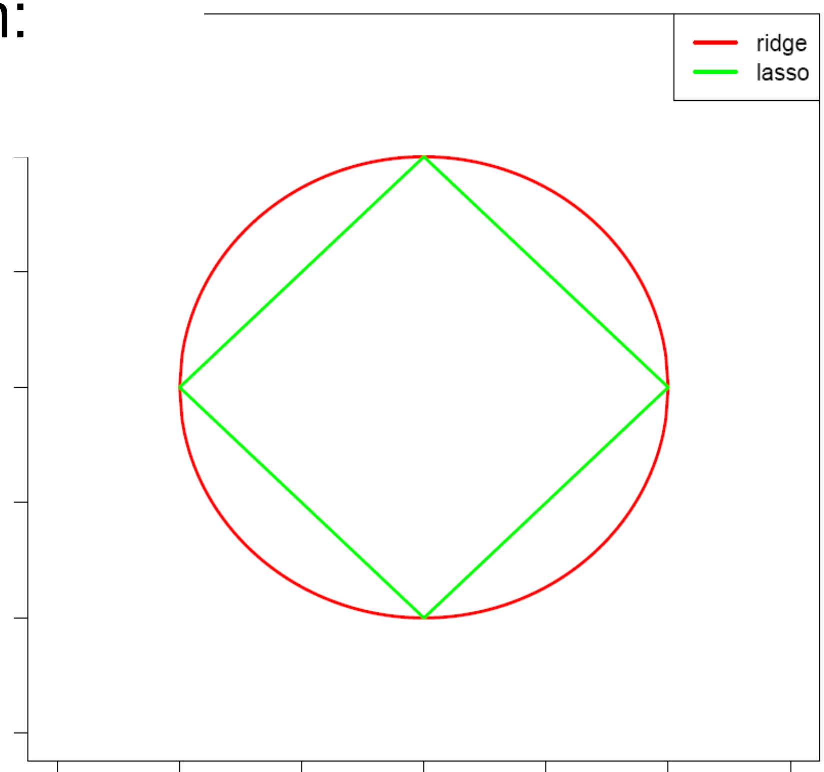
Lasso regression as constrained estimation

The method of Lagrange multipliers enables the reformulation of the penalized least square problem:

$$\min_{\boldsymbol{\beta}} \|\mathbf{Y} - \mathbf{X}\,\boldsymbol{\beta}\|_2^2 + \lambda_1 \|\boldsymbol{\beta}\|_1$$

as a constrained estimation problem:

$$\min_{\|\boldsymbol{\beta}\|_1 \leq \theta(\lambda)_1} \|\mathbf{Y} - \mathbf{X}\,\boldsymbol{\beta}\|_2^2$$



Ridge constraint:

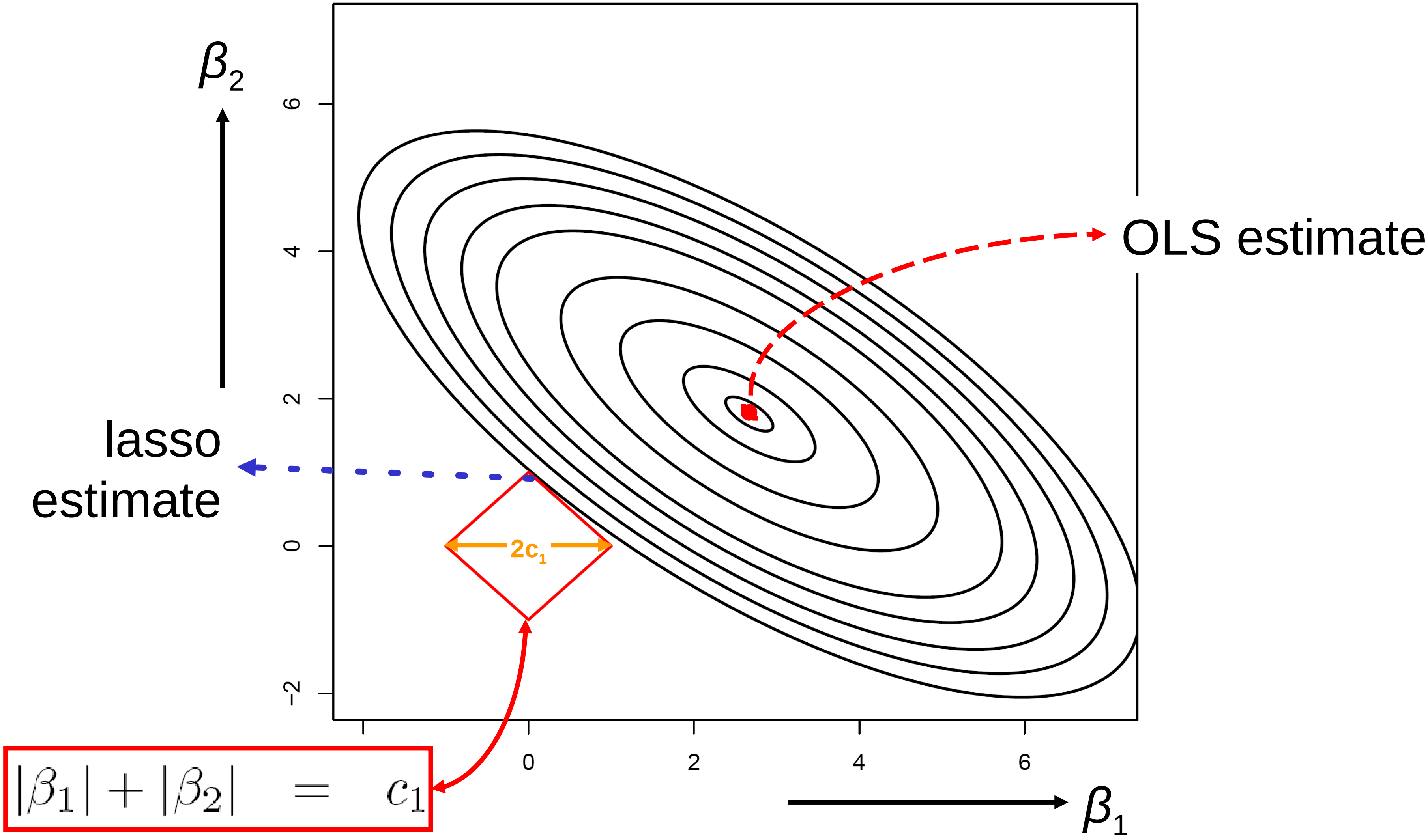
$$\beta_1^2 + \beta_2^2 \quad = \quad 1$$

Lasso constraint:

$$|\beta_1| + |\beta_2| \quad = \quad 1$$

Constrained estimation

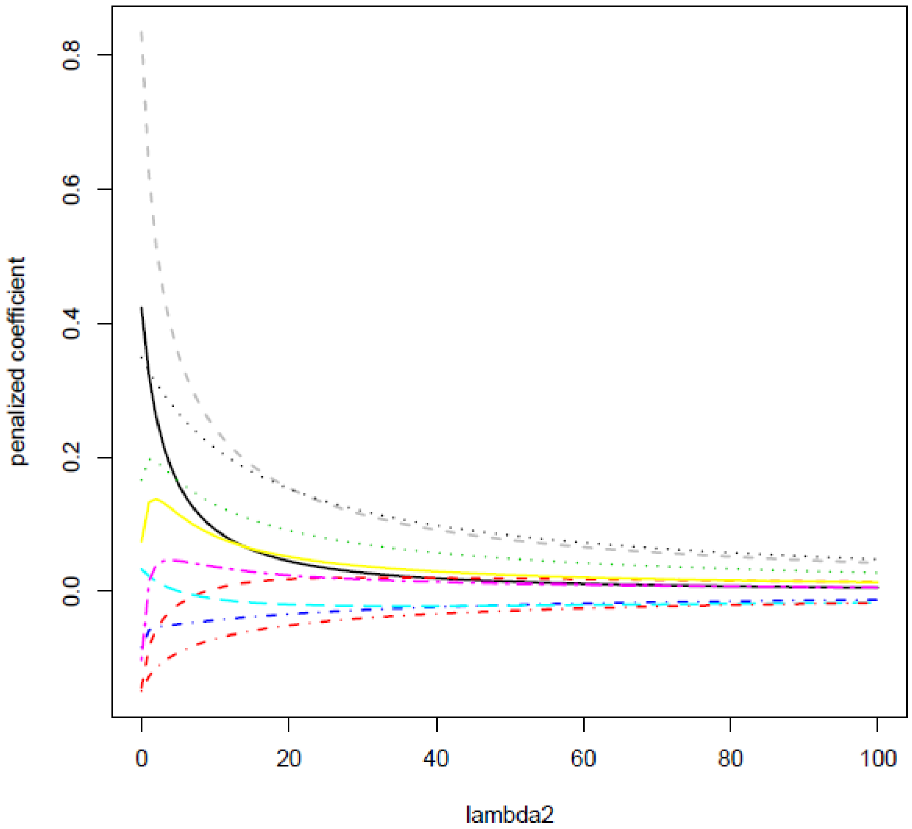
residual sum of squares: $\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2$



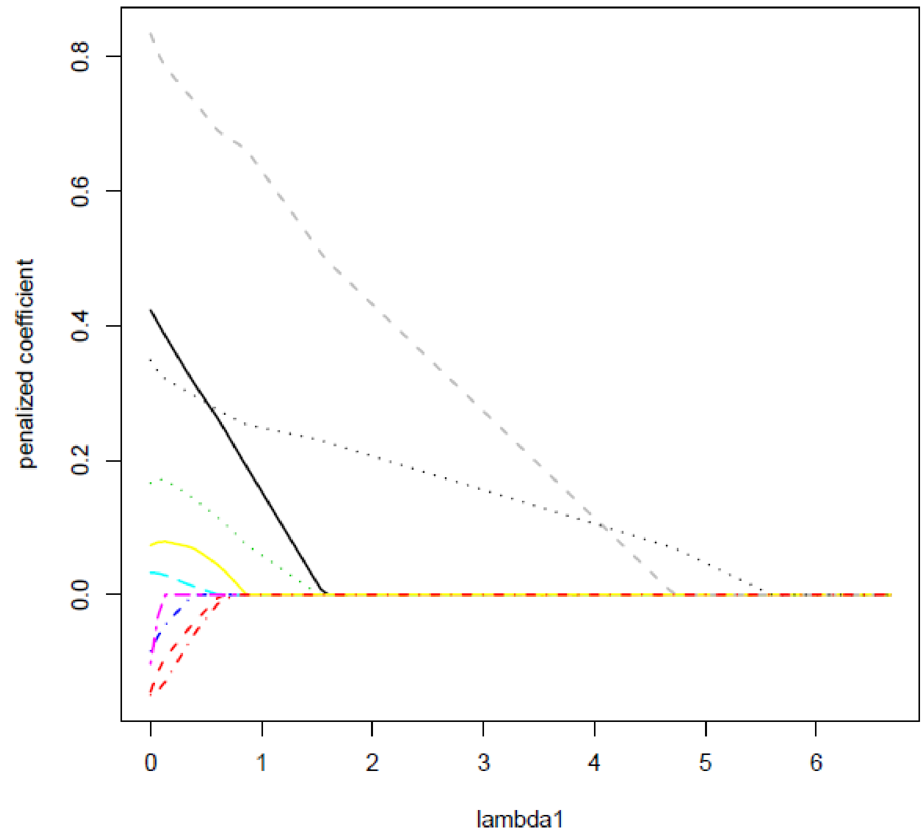
$$|\beta_1| + |\beta_2| = c_1$$

Selection

Ridge regularization path



Lasso regularization path



Question

What are the qualitative differences?

Selection

Simple example

Data have been generated in accordance with:

$$Y_i = X_{i1} + X_{i2} + \varepsilon_i$$

where $\varepsilon_i \sim \mathcal{N}(0, 1)$.

Fit lasso and ridge both with a penalty equal to 3:

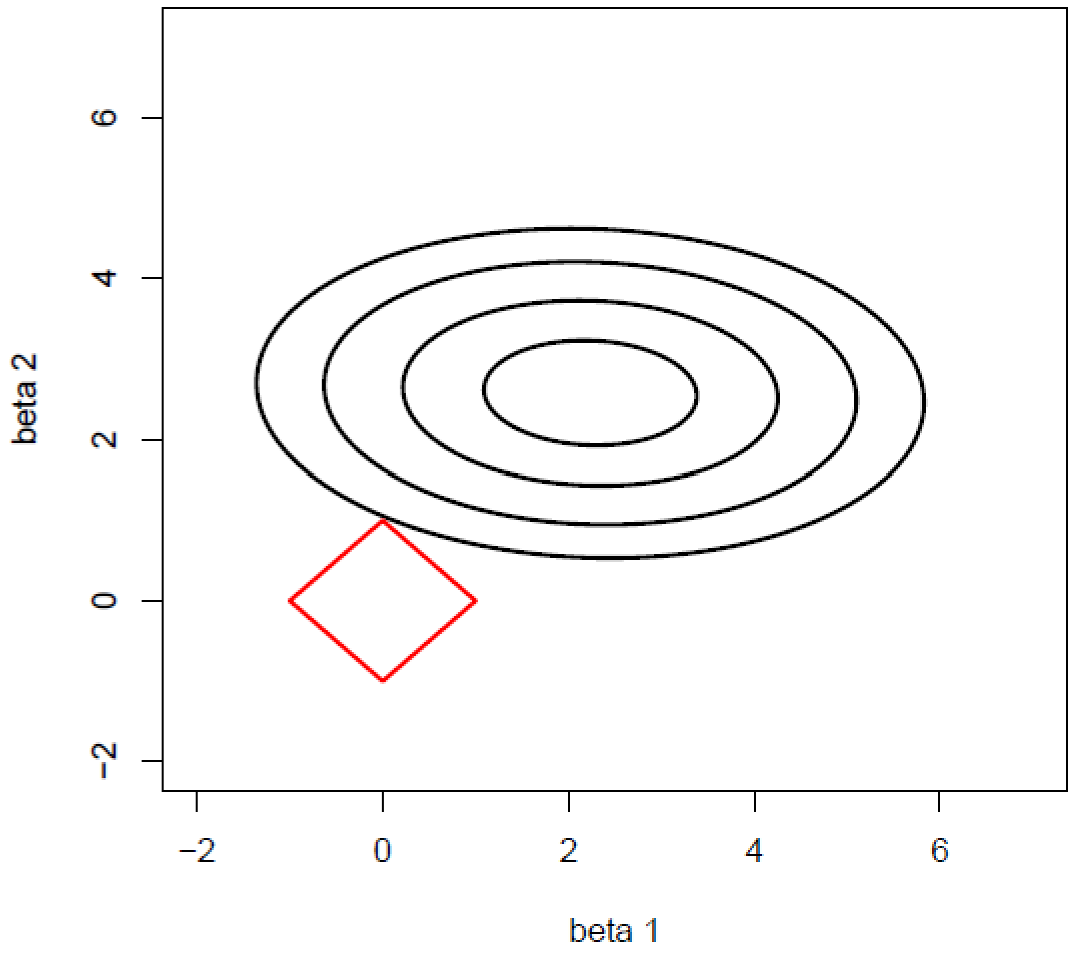
```
> # lasso
> coef(penalized(Y ~ X[,1] + X[,2], unpenalized=~0, lambda1=3), "all")
# nonzero coefficients: 1
     X[, 1]      X[, 2]
-0.02964444  0.00000000

> # ridge
> coef(penalized(Y ~ X[,1] + X[,2], unpenalized=~0, lambda2=3))

     X[, 1]      X[, 2]
-0.09712333 -0.04480700
```

Selection

Illustration of the sparsity of the lasso solution

In the 2-dim setting, for a point to lie on an axis, one coordinate needs to equal zero.

If the lasso estimate coincides with a corner of the diamond, one of the coordinates (estimated regression parameters) equals zero.

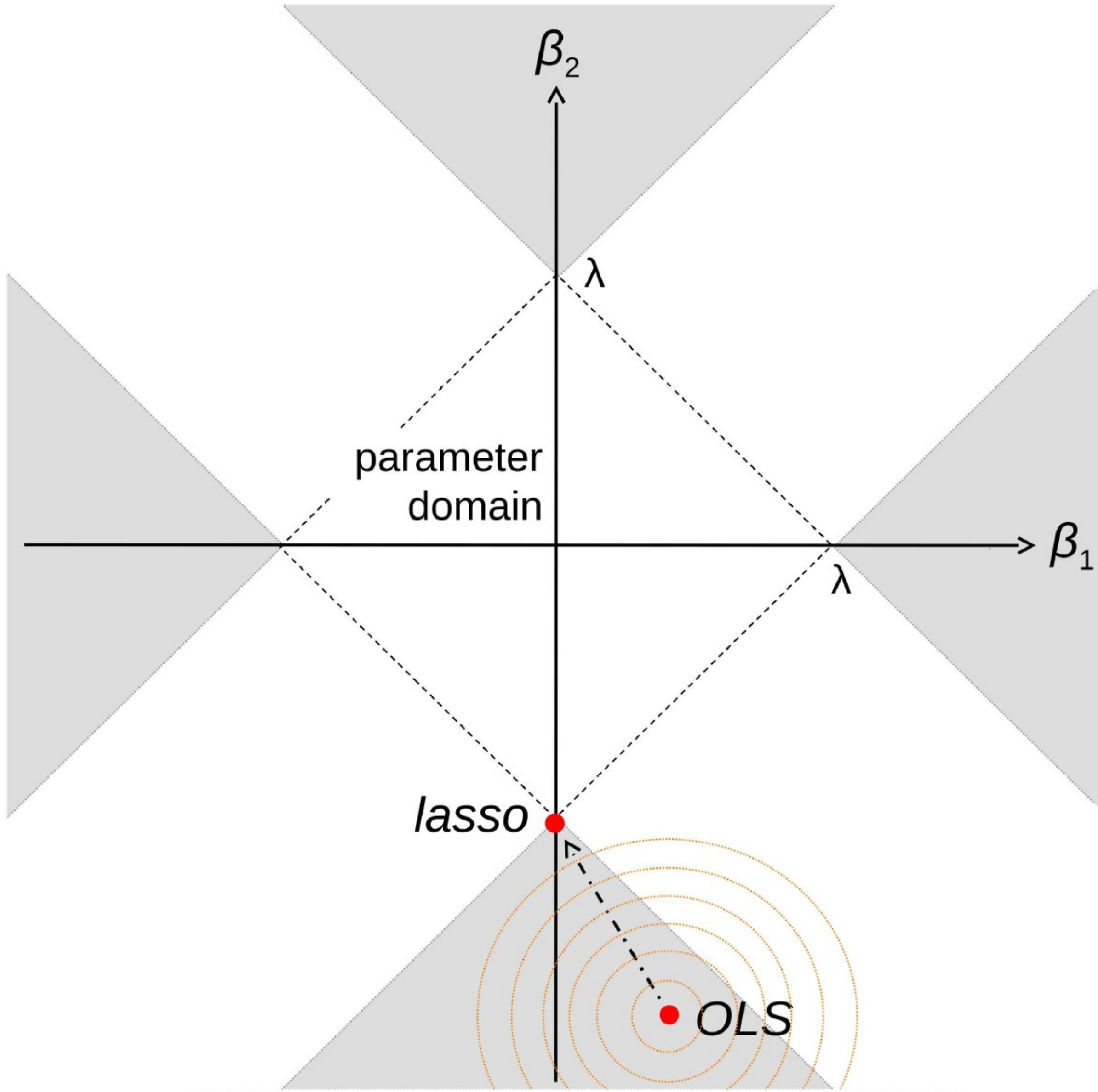


Selection

Suppose ***X*** is orthonormal.

Recall explicit expression for lasso estimate.

Grey domains yield sparse solution, at least for large enough lambda.



Selection

In summary
Lasso regression has the advantage (for the purpose of interpretation) of yielding a sparse solution, in which many parameters (β's) are equal to zero.

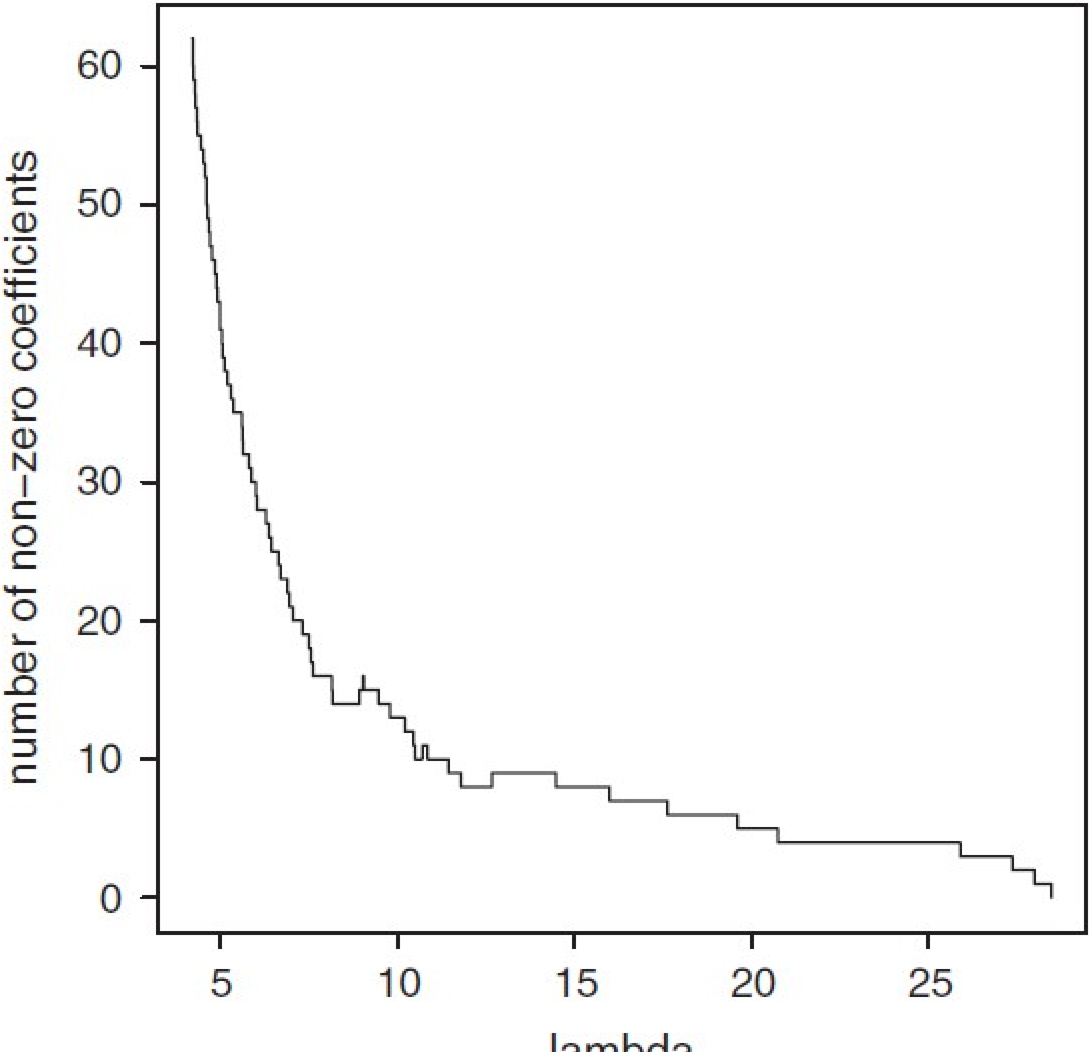
The true model may not be sparse in terms of containing many zero elements. A regularization method that shrinks the parameters proportionally may then be preferred.

Question
When is sparsity a reasonable assumption? Think about the gene expression data. How about astronomy data?

Selection

Lasso fit

The number of non-zero regression coefficients is not necessarily a monotone function of the penalty parameter.



Picture: Goeman (2010).

Number of non-zero parameters

"Every lasso estimated model has cardinality smaller or equal to min(n, p)." (B, vdG, 2011)

Proven in Osborne *et al*. (2000), and "obvious from the analysis of the LARS algorithm (Efron *et al*., 2004)." (Buhlmann, Van de Geer (2011).

When p large and n small, this implies a large dimension reduction.

A simple numerical illustration

```
> library(penalized)
> X <- matrix(rnorm(6), ncol=3)
> Y <- matrix(rnorm(2), ncol=1)
> coef(penalized(Y ~ X[,1] + X[,2] + X[,3],
unpenalized=~0, lambda1=0.0001), "all")
# nonzero coefficients: 2
     X[, 1]      X[, 2]      X[, 3]
  0.0000000  0.7327322 -1.0369745
```

Number of non-zero parameters

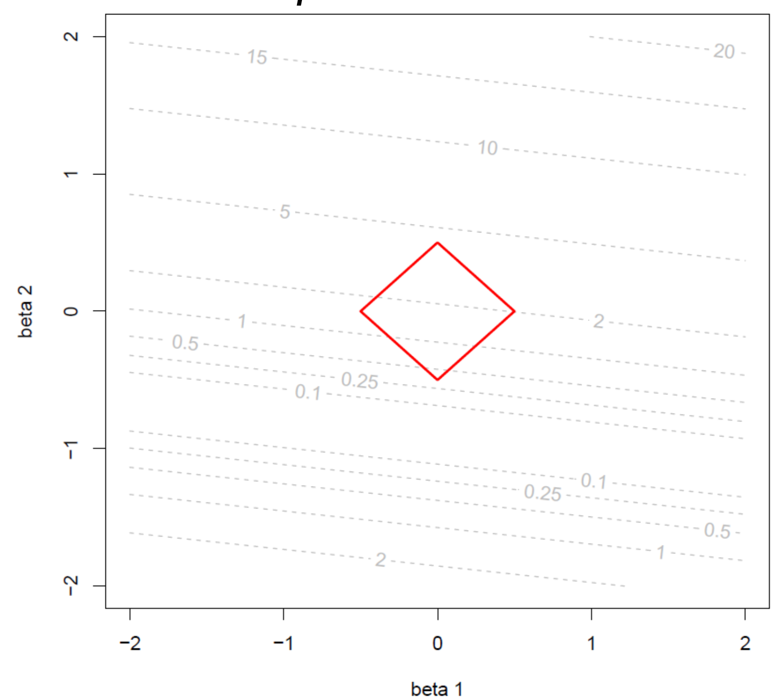
Some intuition

Assume $n < p$ and consider the lasso problem:

$$\min_{||\beta||_1 \leq c(\lambda_1)} ||\mathbf{Y} - \mathbf{X}\beta||_2^2$$

The canonical form of this quadratic problem has n nonzero, positive eigenvalues. This describes an ellipsoid in n dimensions.

Contour plot of the quadratic form for p=2 and n=1:



Consistency

Consider the high-dimensional prediction problem:

$$Y_i = \mathbf{X}_i\beta + \varepsilon_i$$

Let

→ S_0 : set of “true” covariates that contribute to Y.

→ λ_{cv} : cross-validated lasso penalty parameter

→ $S(\lambda_{cv})$: set of selected covariates for λ_{cv}.

Then,

→ with high probability $S(\lambda_{cv})$ contains S_0, or at least the most relevant covariates of S_0.

→ Under suitable assumption $S(\lambda_{optimal})$ contains with probability one S_0, asymptotically.

Parameter estimation

Parameter estimation - I

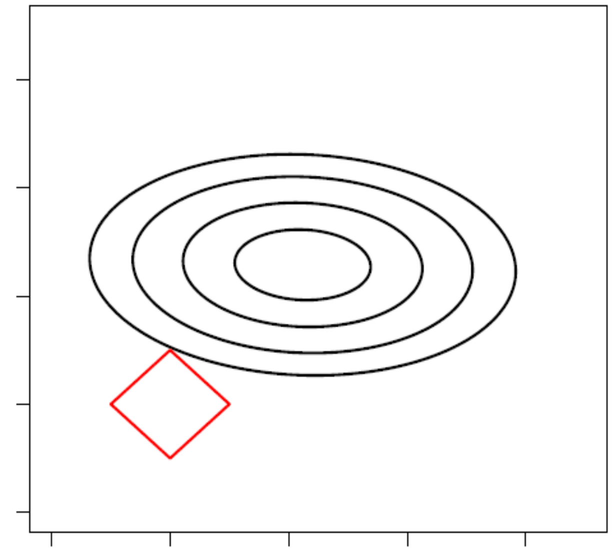
Quadratic programming

The constrained estimation problem of the lasso:

$$\arg \min_{\|\boldsymbol{\beta}\| \leq c(\lambda)} \|_1 \mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2$$

can be reformulated as a quadratic program (e.g. for *p*=2):

$$\arg \min_{\substack{\beta_1 + \beta_2 \leq c(\lambda) \\ \beta_1 - \beta_2 \leq c(\lambda) \\ -\beta_1 + \beta_2 \leq c(\lambda) \\ -\beta_1 - \beta_2 \leq c(\lambda)}} \frac{1}{2} \boldsymbol{\beta}^\top \mathbf{X}^\top \mathbf{X}\boldsymbol{\beta} - \mathbf{Y}^\top \mathbf{X}\boldsymbol{\beta}$$



Question

Why not feasible for large *p*?

Parameter estimation - II

The loss function of the lasso regression:

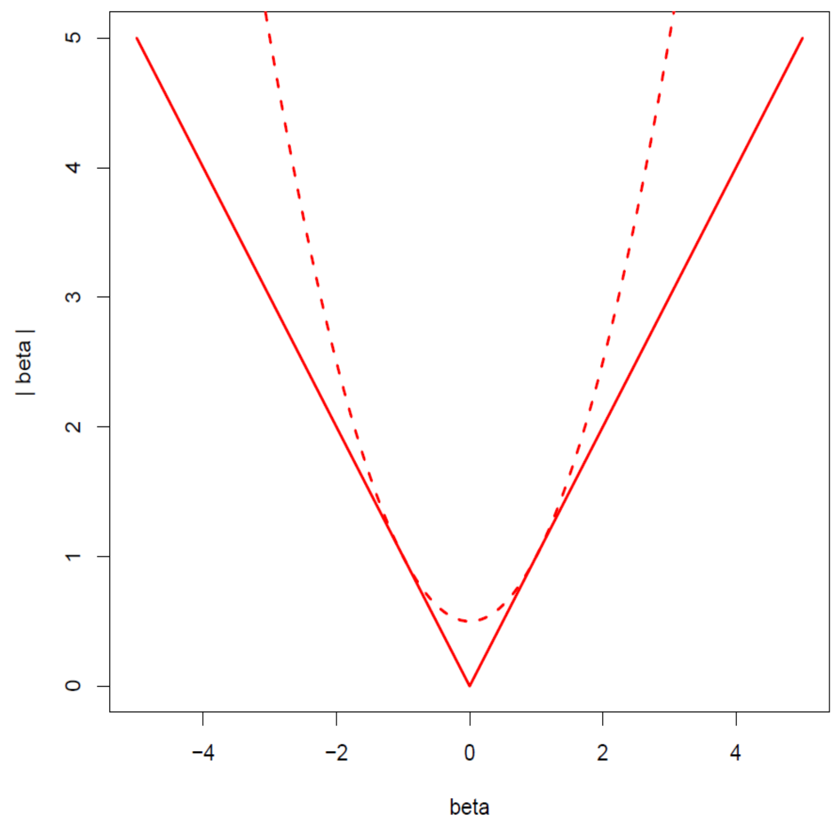
$$\mathcal{L}(\boldsymbol{\beta}; \lambda) \quad = \quad \|\mathbf{Y} - \mathbf{X}\,\boldsymbol{\beta}\|_2^2 + \lambda_1 \|\boldsymbol{\beta}\|_1$$

may be optimized by iteratively applying ridge regression.

Key observation

Given some initial parameter value, the lasso penalty is approximated by:

$$|\beta| = |\beta_0| + \frac{1}{2|\beta_0|}(\beta^2 - \beta_0^2)$$



Source: Fan & Li (2001).

Parameter estimation - II

Plug the approximation into the lasso loss function:

$$
\begin{aligned}
\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}^{(k+1)}\| + \lambda_1 \|\boldsymbol{\beta}^{(k+1)}\|_1 \\
\approx \quad & \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}^{(k+1)}\| + \lambda_1 \|\boldsymbol{\beta}^{(k)}\|_1 \\
& + \frac{\lambda_1}{2} \sum_{j}^{p} \frac{1}{|\beta_j^{(k)}|} \big[\beta_j^{(k+1)}\big]^2 - \frac{\lambda_1}{2} \sum_{j}^{p} \frac{1}{|\beta_j^{(k)}|} \big[\beta_j^{(k)}\big]^2 \\
\propto \quad & \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}^{(k+1)}\| + \frac{\lambda_1}{2} \sum_{j}^{p} \frac{1}{|\beta_j^{(k)}|} \big[\beta_j^{(k+1)}\big]^2
\end{aligned}
$$

The loss function now contains a weighted ridge penalty.

Parameter estimation - II

Analogous to the derivation of the ridge estimator, the approximated lasso loss function is optimized by:

$$\boldsymbol{\beta}^{(k+1)} \quad = \quad \{\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda_1 \boldsymbol{\Psi}[\boldsymbol{\beta}^{(k)}]\}^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y}$$

where

$$\begin{aligned} &\mathrm{diag}\{\boldsymbol{\Psi}[\boldsymbol{\beta}^{(k)}]\} \\ &\quad = \quad (1/|\beta_1^{(k)}|, 1/|\beta_2^{(k)}|, \ldots, 1/|\beta_p^{(k)}|) \end{aligned}$$

The solution above converges to the lasso estimator.

Parameter estimation - II

Gradient ascent approach (explained next):

```
> coef(penalized(Y ~ X[,1] + X[,2], unpenalized=~0,
lambda1=1), "all")
# nonzero coefficients: 1
     X[, 1]      X[, 2]
 0.00000000 -0.01405338
```

Iterative ridge:

```
Error in solve.default(...) :
  system is computationally singular: reciprocal
condition number = 2.15377e-16

      X[, 1]      X[, 2]
1.678667e-18 -0.01405338
```

The latter requires a modification to accommodate estimates that get very close to zero.

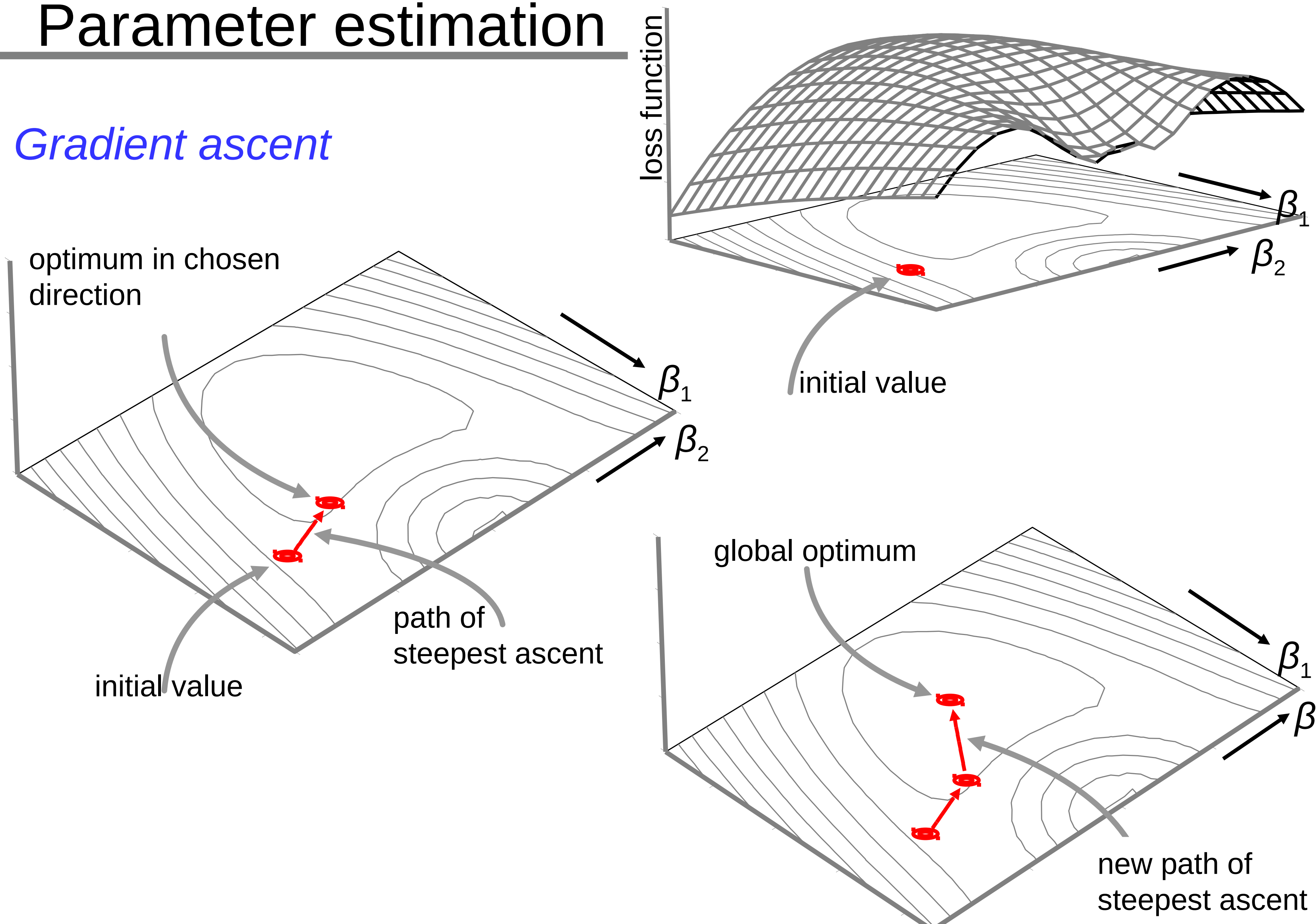
Parameter estimation - III

Gradient ascent (hill climbing)

1) Choose a starting value.
2) Calculate the derivative of the loss function, and determine the direction in which the loss function increases most. This direction is the *path of steepest ascent*.
3) Proceed in this direction, until the loss function no longer increases.
4) At this point recalculate the gradient to determine a new path of steepest ascent.
5) Repeat the above until the region around the optimum is found (usually: when a linear model is no longer adequate).

Parameter estimation

Gradient ascent



Parameter estimation - III

Gradient ascent

Recall: $f(x) = |x|$ is not differentiable at $x=0$. Consequently, so is the lasso loss function. Solution: employ the Gateaux derivative, which is properly defined at $x=0$.

The Gateaux derivative of $f:\mathbf{R}^p \rightarrow \mathbf{R}$ at $\mathbf{x}$ in $\mathbf{R}^p$ in the direction of $\mathbf{v}$ in $\mathbf{R}^p$ as:

$$f'(\mathbf{x}) = \lim_{\tau \downarrow 0} \frac{1}{\tau} \big[f(\mathbf{x} + \tau \mathbf{v}) - f(\mathbf{x}) \big]$$

To uniquely define this derivative the directional vectors $\mathbf{v}$ are limited to

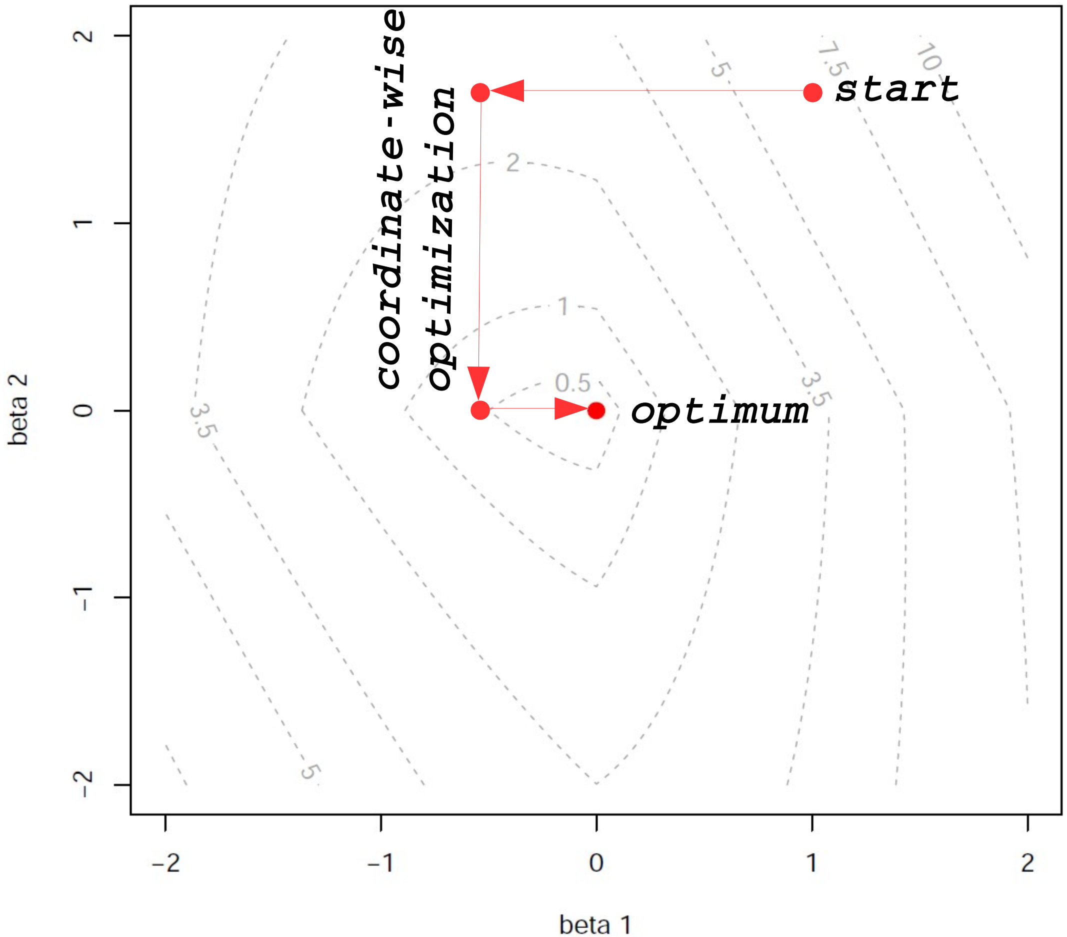
→ those with unit length, and

→ the direction of steepest ascent.

Parameter estimation - IV

Coordinate descent

Due to the convexity of the loss function, parameter-by-parameter optimization converges to the lasso estimate.



Parameter estimation - IV

Coordinate descent

Thus, solve:

$$\arg\min_{\beta_j} \| \mathbf{Y} - \mathbf{X}_{*,\setminus j} \boldsymbol{\beta}_{\setminus j} - \mathbf{X}_{*,j} \beta_j \|_2^2 + \lambda_1 \| \boldsymbol{\beta} \|_1$$

This is equivalent to:

$$\arg\min_{\beta_j} \| \tilde{\mathbf{Y}} - \mathbf{X}_{*,j} \beta_j \|_2^2 + \lambda_1 \| \beta_j \|_1$$

which (assuming $\mathbf{X}_{*,j}^\top \mathbf{X}_{*,j} = 1$) has an explicit solution:

$$\hat{\beta}_j^{(\mathrm{update})}(\lambda_1) = \mathrm{sign}(\mathbf{X}_{*,j}^\top \tilde{\mathbf{Y}})(|\mathbf{X}_{*,j}^\top \tilde{\mathbf{Y}}| - \tfrac{1}{2}\lambda_1)_+$$

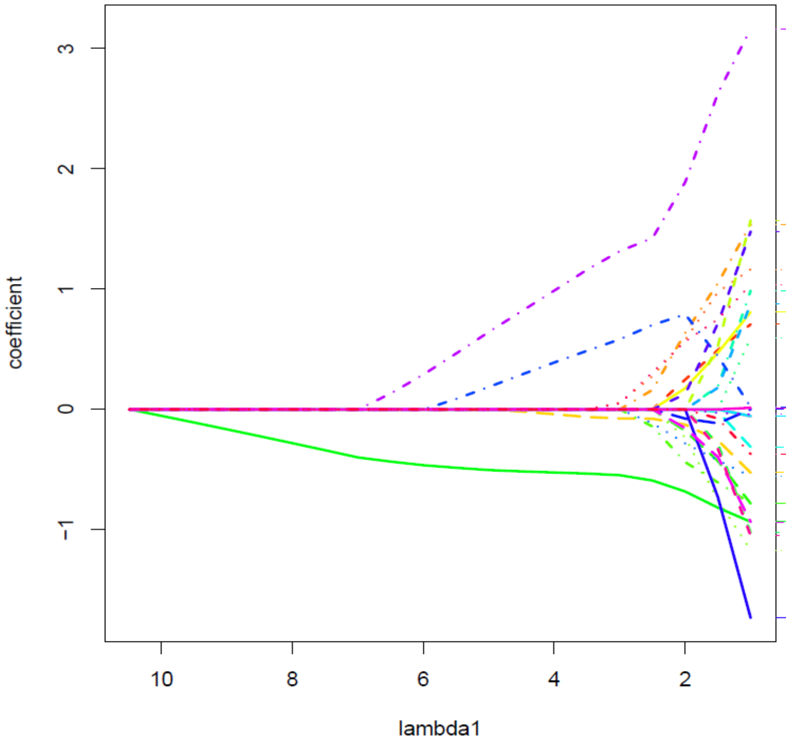
Finally, run over the parameters until convergence to arrive at the lasso estimate.

Parameter estimation - V

LARS

The LARS (Least Angular Regression) algorithm solves the lasso problem over the whole domain of the penalty parameter.

This yields the full piecewise linear solution path of the regression coefficients.



Parameter estimation - V

LARS

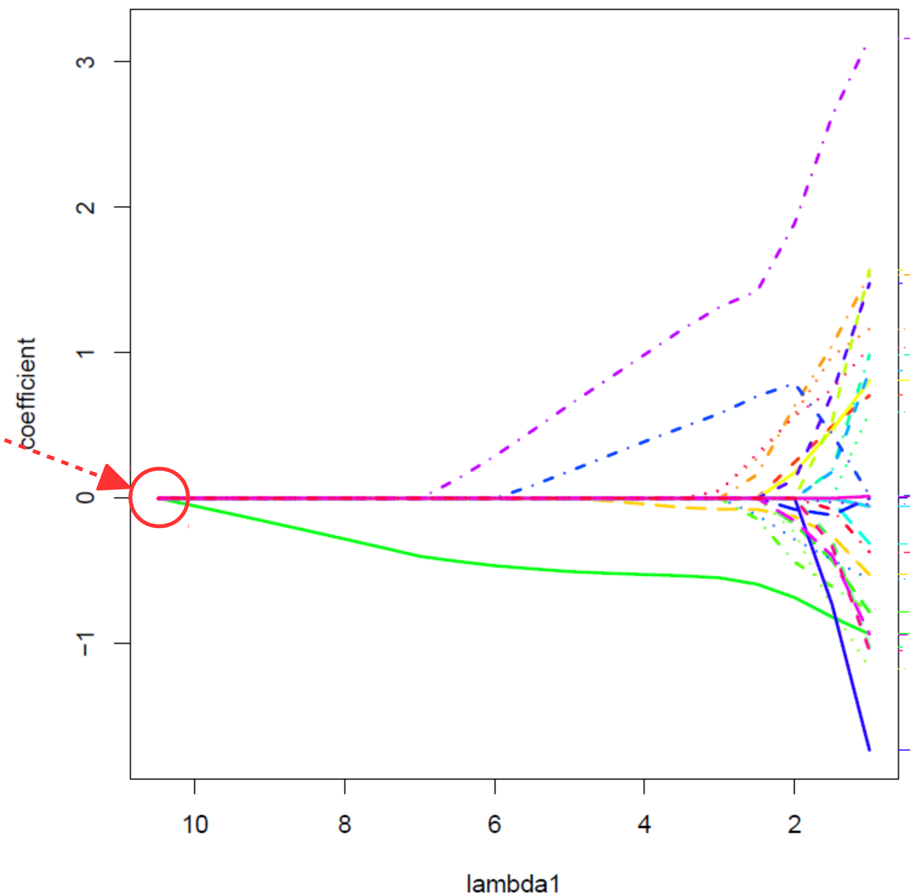
Covariates with nonzero coefficients form the *active set*.

Algorithm

→ initiate with an empty active set ($\lambda_1 = \infty$),

→ determine largest λ_1 for which active set is non-empty.

→ at this λ_1 determine for covariates in active set the optimal direction direction of β.



Parameter estimation - V

LARS

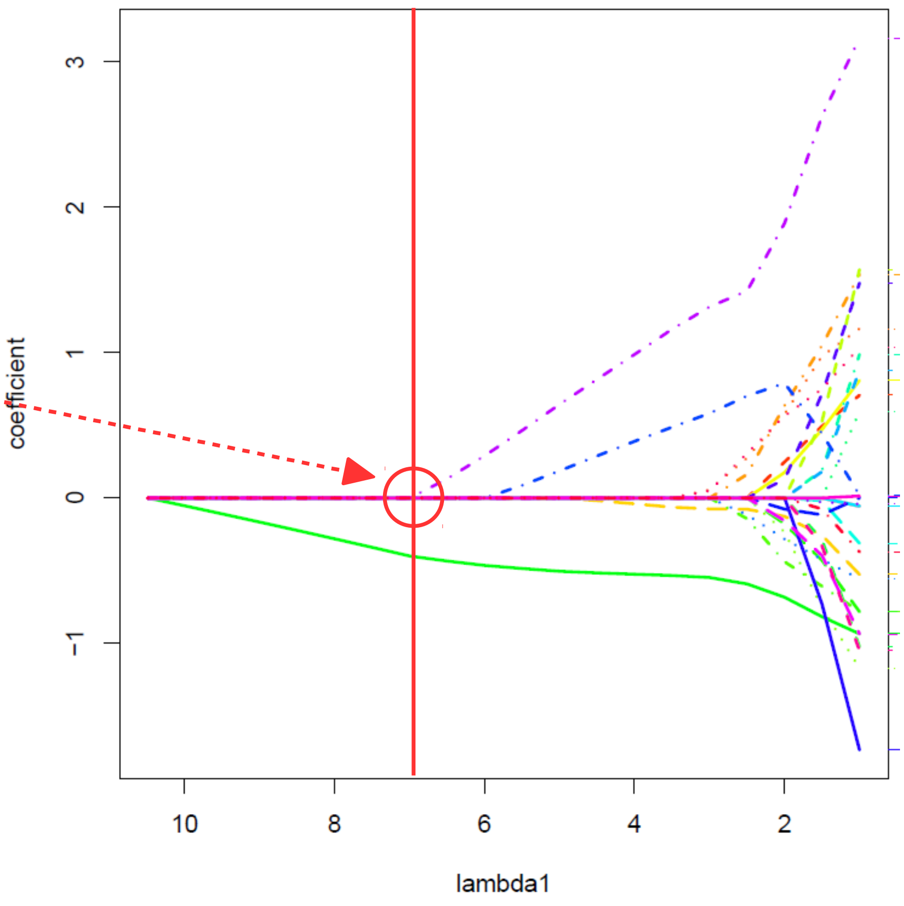
Covariates with nonzero coefficients form the *active set*.

Algorithm (*continued*)

→ decrease λ_1 and determine when active set changes,

→ at this λ_1 determine for covariate in active set the optimal direction direction of β.

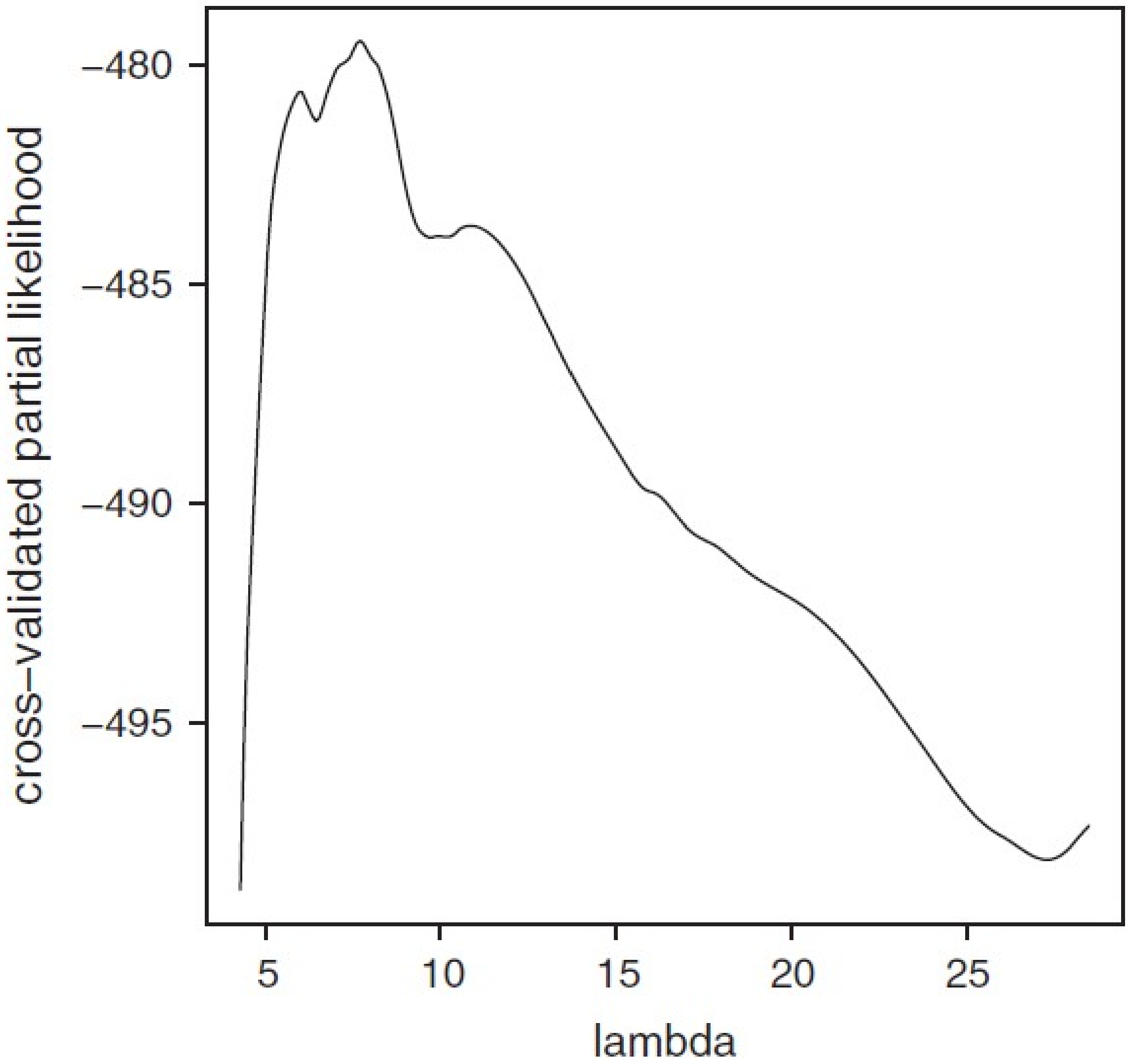
→ iterate last 2 steps.



Parameter estimation

Penalty parameter

The cross-validated (partial) likelihood has several local maxima. This is a typical feature of lasso fits. Hence, always check for global optimality.



Picture: Goeman (2010).

Moments of the lasso estimator

Moments of the lasso estimator

Summary

In contrast to ridge regression, there are no explicit expressions for the bias and variance of the lasso estimator.

Approximations of the variance of the lasso estimates can be found in Tibshirani (1996) and in Osborne et al. (2000). Discussed on the next slides.

As with the ridge estimator:
→ the bias of lasso estimator increases and
→ the variance of the lasso estimator decreases
as the lasso penalty parameter increases.

Moments of the lasso estimator

Moment approximations

Approximate the lasso penalty quadratically around the lasso:

$$\begin{aligned} & \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda_1 \|\boldsymbol{\beta}\|_1 \\ \approx \quad & \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \frac{\lambda_1}{2} \sum_{j=1}^{p} \frac{1}{|\hat{\beta}(\lambda_1)|} \beta_j^2 \end{aligned}$$

Optimization of this loss function gives a 'ridge approximation' to the lasso estimate:

$$\hat{\boldsymbol{\beta}}(\lambda_1) \quad \approx \quad \{\mathbf{X}^\top \mathbf{X} + \lambda_1 \boldsymbol{\Psi}[\hat{\boldsymbol{\beta}}(\lambda_1)]\}^{-1} \mathbf{X}^\top \mathbf{Y}$$

where $\boldsymbol{\Psi}$ diagonal with $(\boldsymbol{\Psi}[\hat{\boldsymbol{\beta}}(\lambda_1)])_{jj} = 1/|\hat{\beta}_j(\lambda_1)|$ if $\hat{\beta}_j(\lambda_1) \neq 0$ and zero otherwise.

Moments of the lasso estimator

Moment approximations

Analogous to moment derivation of the ridge estimator, one obtains:

$$\mathbb{E}[\hat{\boldsymbol{\beta}}(\lambda_1)] \approx \{\mathbf{X}^\top\mathbf{X} + \lambda_1\boldsymbol{\Psi}[\hat{\boldsymbol{\beta}}(\lambda_1)]\}^{-1}\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta}$$

and

$$\begin{aligned}\mathrm{Var}[\hat{\boldsymbol{\beta}}(\lambda_1)] \approx \sigma^2\{\mathbf{X}^\top\mathbf{X} + \lambda_1\boldsymbol{\Psi}[\hat{\boldsymbol{\beta}}(\lambda_1)]\}^{-1} \\ \times\mathbf{X}^\top\mathbf{X}\{\mathbf{X}^\top\mathbf{X} + \lambda_1\boldsymbol{\Psi}[\hat{\boldsymbol{\beta}}(\lambda_1)]\}^{-1}\end{aligned}$$

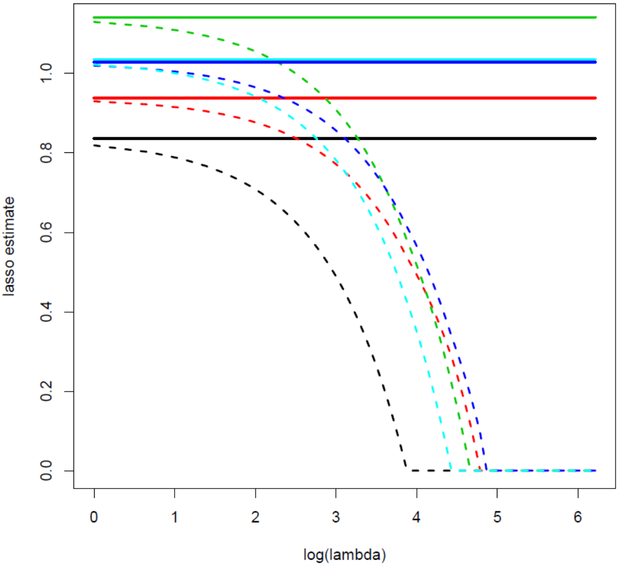
where σ^2 is the residual variance.

The design matrix **X** should be of full rank to warrant the existence of the variance matrix estimate.

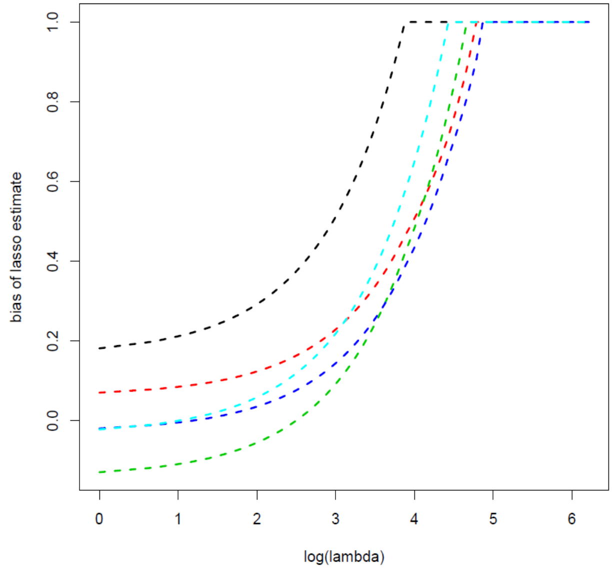
Osborne et al (2000) improves on these approximations.

Moments of the lasso estimator

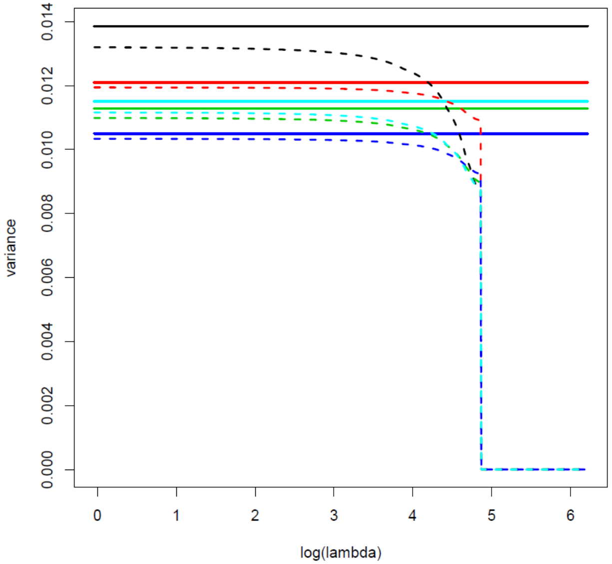
OLS and lasso estimates



Bias of lasso estimates



Variance of estimates



Questions

The (approximated) variance of the lasso estimates may equal zero. Interpretation? Realistic?

How about the MSE? *Hint*: Contrast a truly sparse model vs. a full model.

A Bayesian interpretation

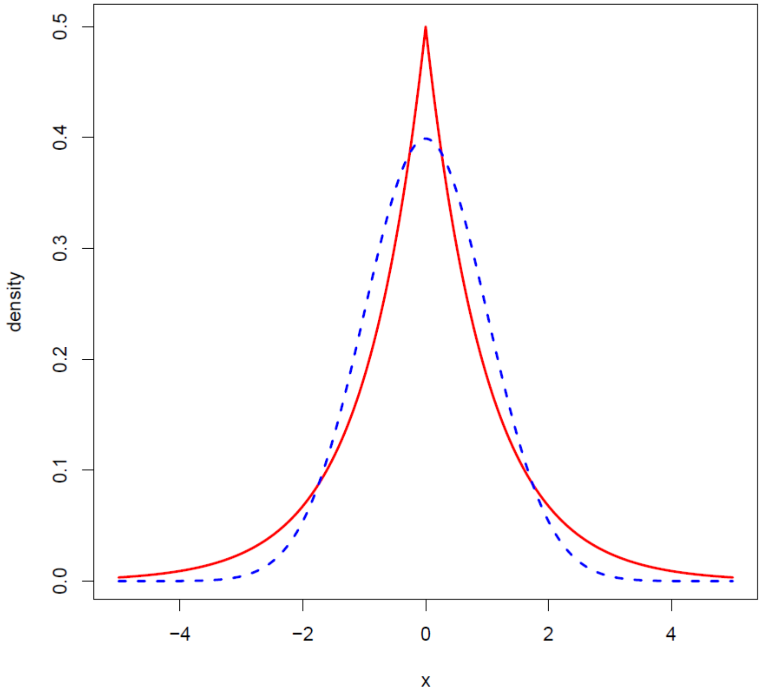
A Bayesian interpretation

Recall, the ridge regression estimator can be viewed as a Bayesian estimate of β when imposing a Gaussian prior.

Similarly, the lasso regression estimator can be viewed as a Bayesian estimate when imposing a Laplacian (or double exponential) prior:

$$f(\beta_j) \quad = \quad \frac{1}{2}\lambda_1 \exp(-\lambda_1 |\beta_j|)$$

The lasso loss function suggests form of the prior.

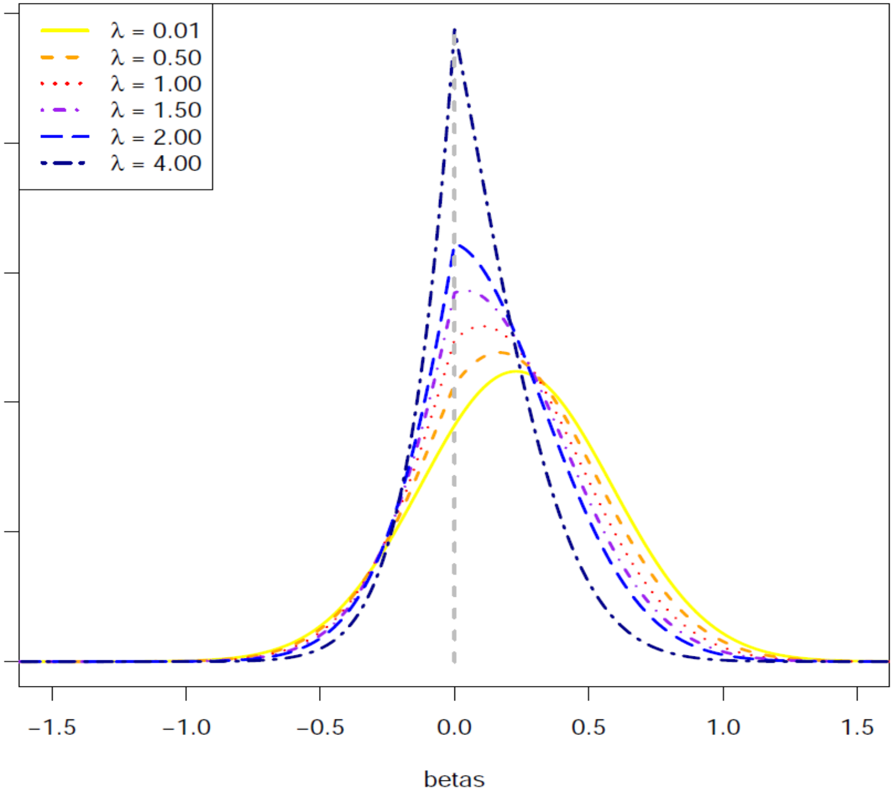


The lasso prior puts more mass close to zero and in the tails than the ridge prior. Hence, the tendency of the lasso to produce either zero or large estimates.

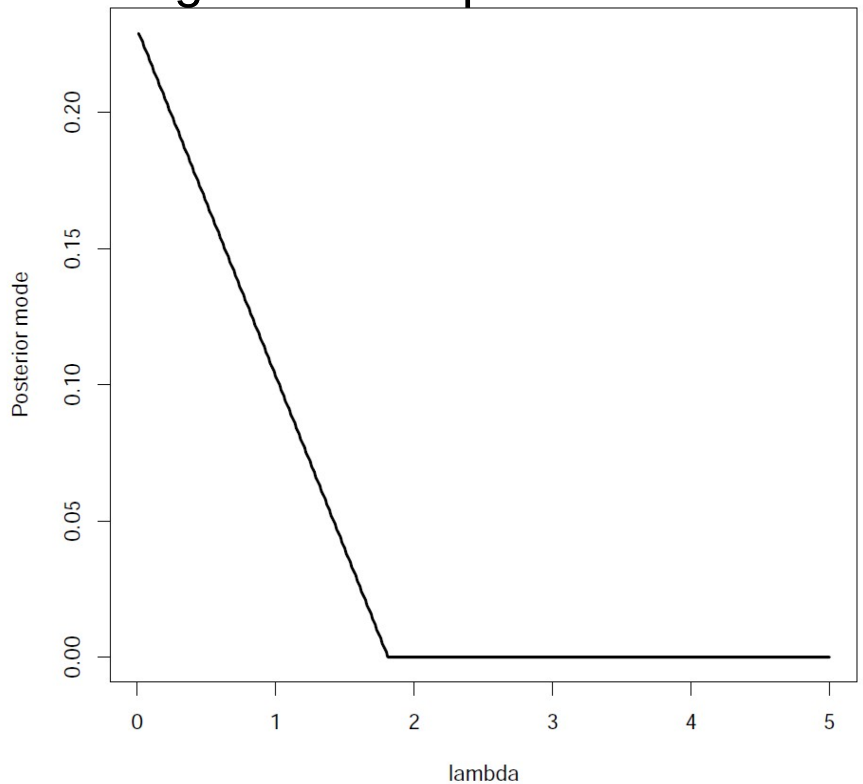
A Bayesian interpretation

The lasso regression estimates then correspond to the posterior mode estimate of β.

Posterior distribution



Regularization path



A Bayesian interpretation

Remarks

→ A “true Bayesian” also puts a prior on the penalty parameter (giving rise to Bayesian lasso regression, Casella, Park, 2004).

→ In high-dimensions, the Bayesian posterior need not concentrate on the “true” parameter (even though its mode is a good estimator of the regression parameter).

Cassella, Park (2008).
Castillo, Schmidt-Hieber, & Van der Vaart, (2015).

Stability selection

Stability selection

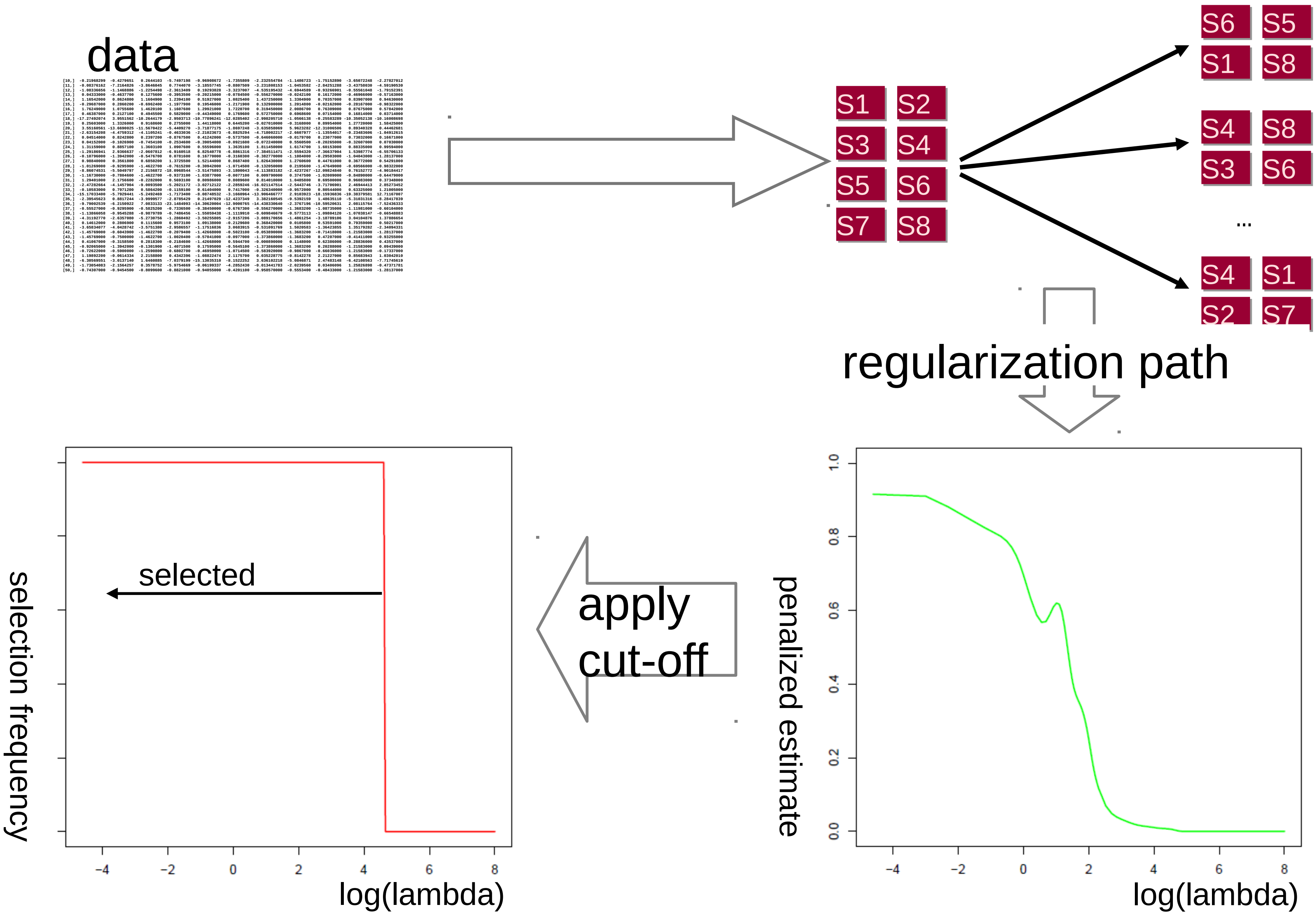
Which penalty parameter to use?

Problem:
→ Scale of the penalty parameter is meaningless.

Solution:
→ Map, by re-sampling, λ to a scale with a tangible interpretation.

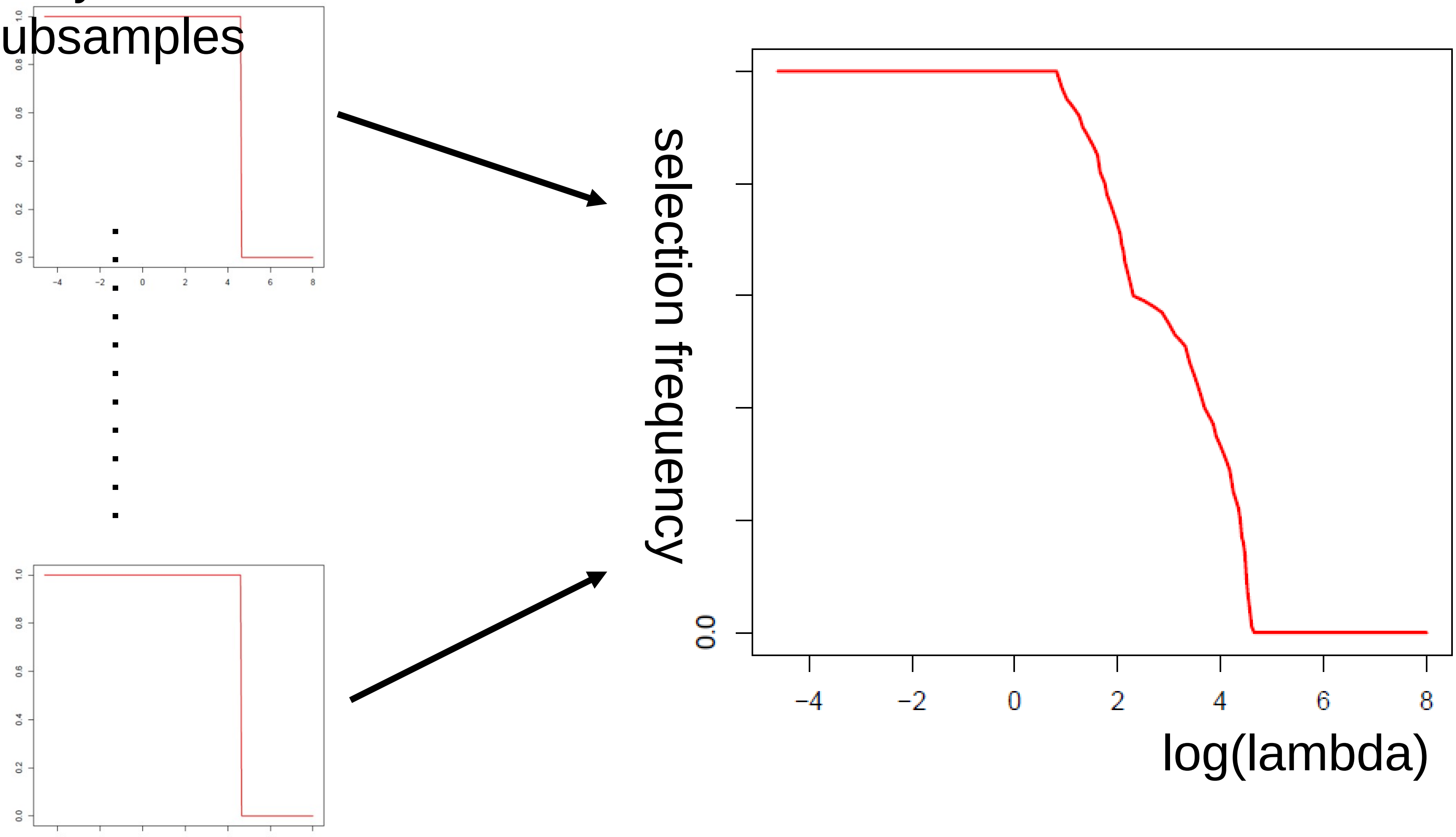
Selection frequency
→ number of times a parameter is included in the model.
→ directly related to λ,
→ used to determine the amount of penalization.

Stability selection



Stability selection

many
subsamples



Stability selection

Stability selection (Meinshausen, Bühlman, 2009)

→ Given a selection frequency cut-off: upperbound on the expected number of falsely selected parameters.

→ The upperbound further only depends on the average number of selected parameters, a quantity directly determined by λ.

→ Having specied the selection frequency cut-off, the desired error rate is achieved by chosen the appropriate penalty parameter.

Ridge vs. lasso I

---

shrinkage

Ridge vs. lasso I

Recall in the orthonormal case the ridge estimator equals:

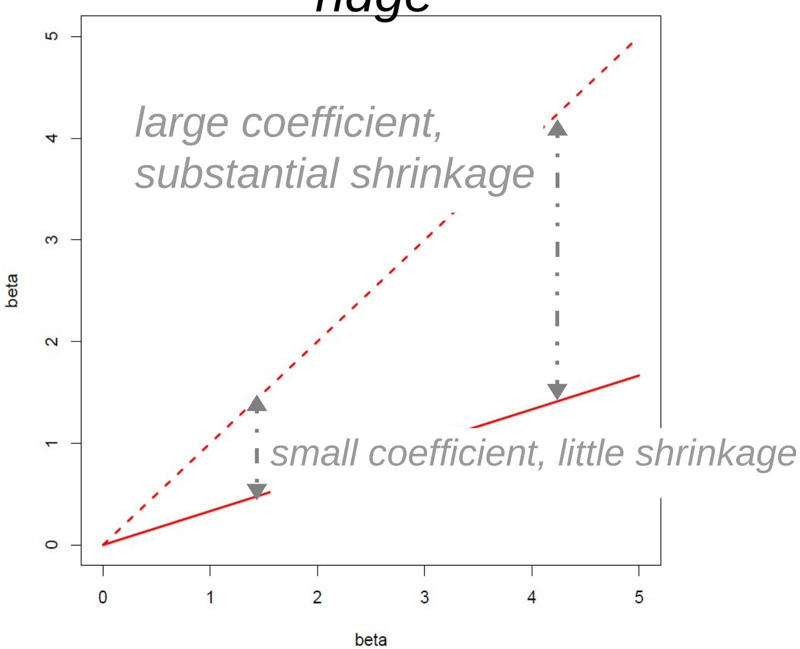
$$\hat{\beta}_j(\lambda_2) \quad = \quad (1+\lambda_2)^{-1}\hat{\beta}_j$$

and the lasso estimator:

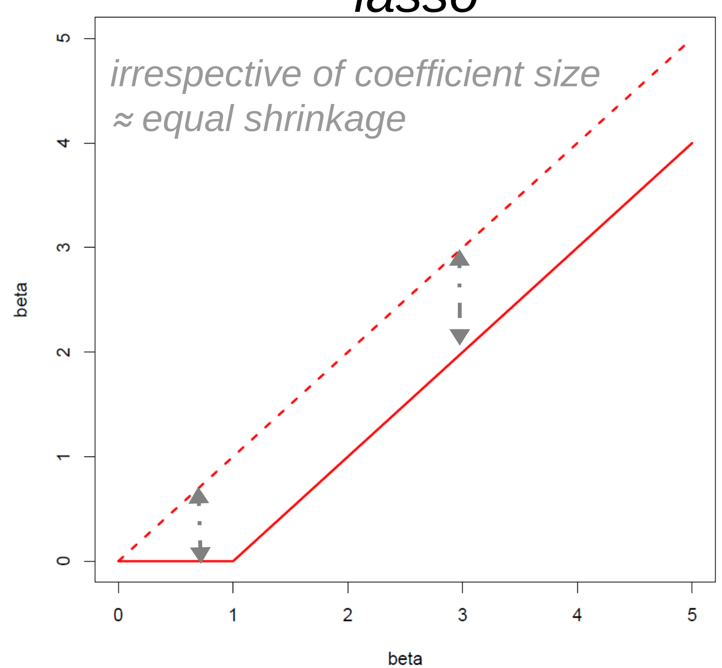
$$\hat{\beta}_j(\lambda_1) \quad = \quad \operatorname{sgn}(\hat{\beta}_j)\,(|\hat{\beta}_j| - \lambda_1/2)_+$$

Ridge scales and whereas lasso translates:

ridge

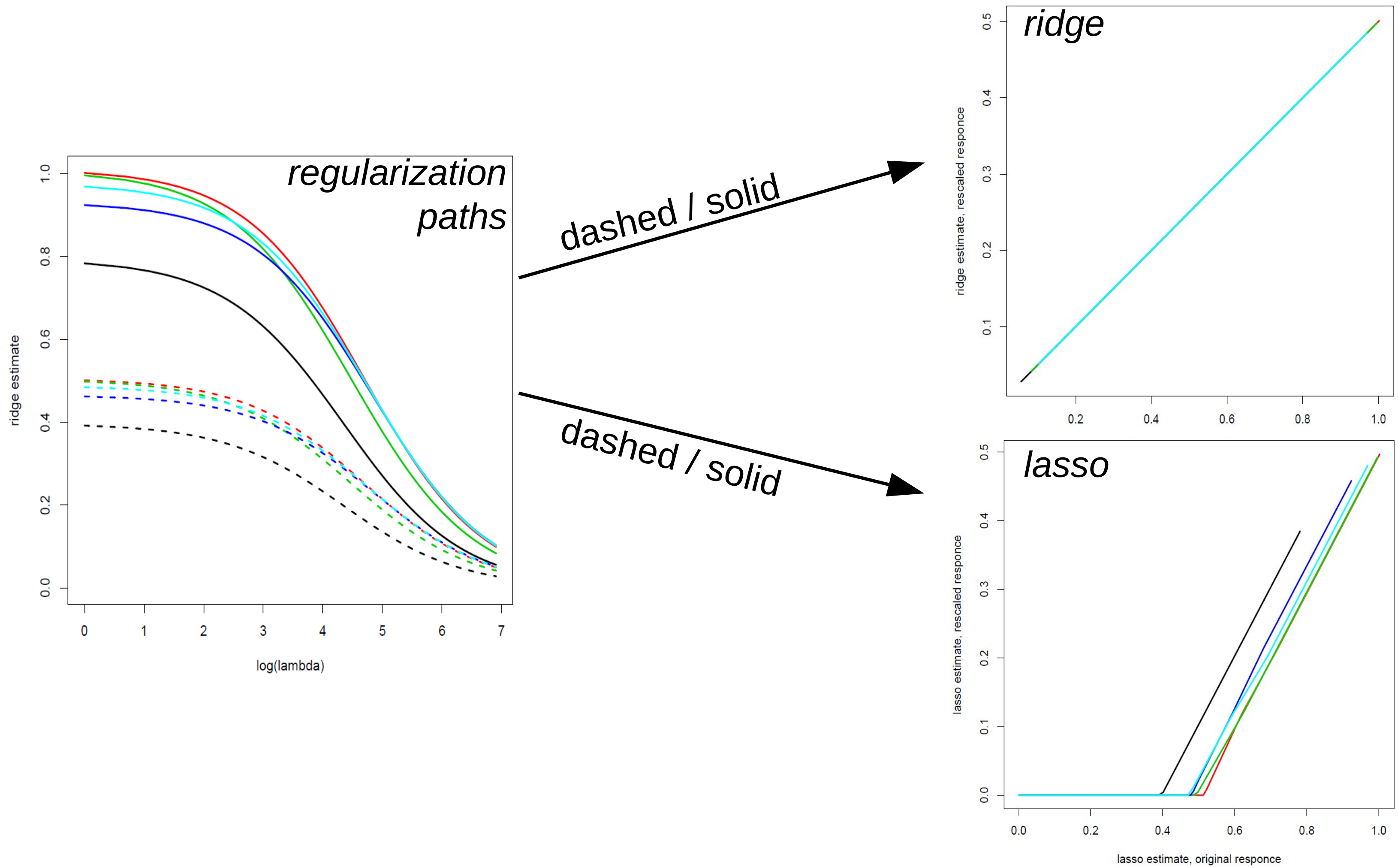


lasso



Ridge vs. lasso I

Ridge estimator is *linear* in the response, while lasso is not.
Compar fit of $Y = X\beta + \varepsilon$ (*solid*) and $Y/2 = X\beta + \varepsilon$ (*dashed*).



Ridge vs. lasso II

Simulations

Simulation I

Ridge vs. lasso estimation

Consider a set of 50 genes. Their expression levels follow a standard multivariate normal law.

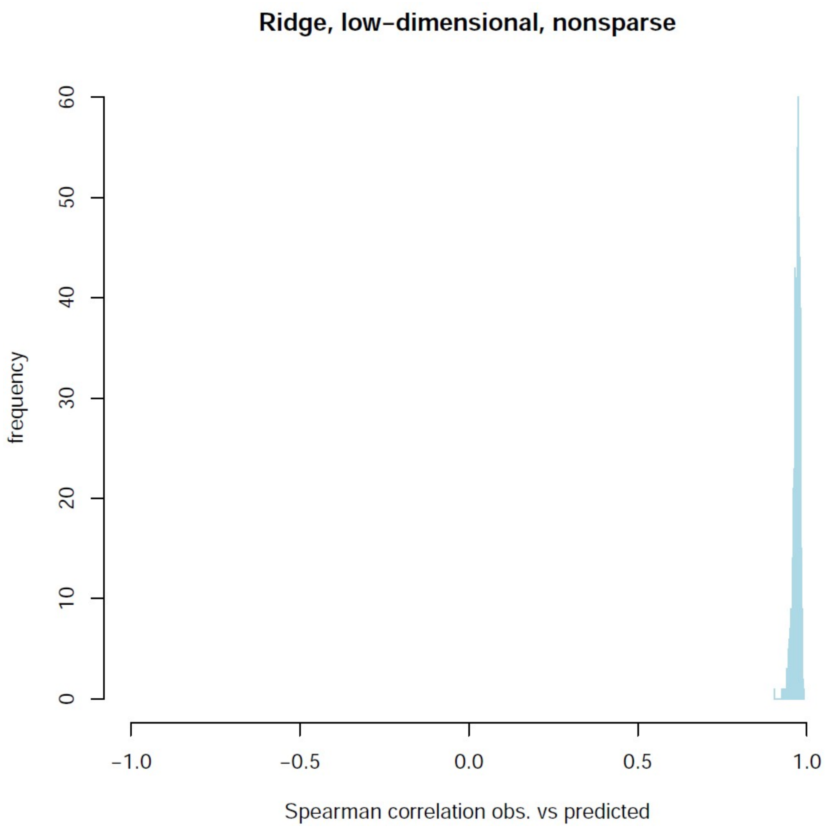
Together they regulate a 51th gene through: $Y_i = \mathbf{X}_{i*}\boldsymbol{\beta} + \varepsilon_i$ with $\varepsilon_i \sim \mathcal{N}(0,1)$ and regression coefficients $\boldsymbol{\beta} = \mathbf{1}_{50}$. Hence, the 50 genes contribute equally.

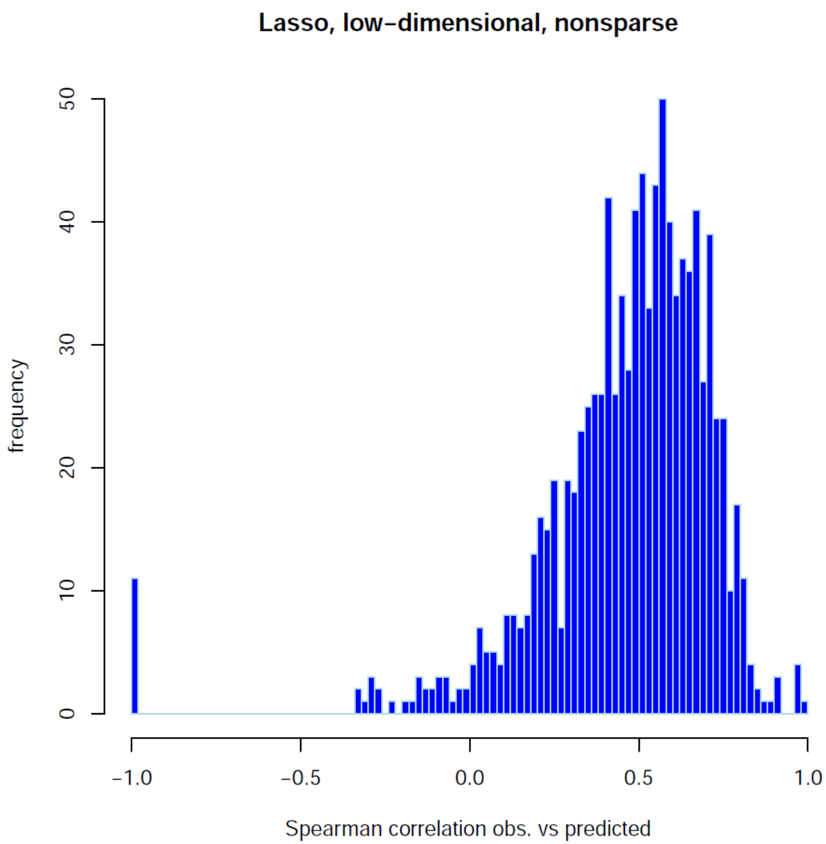
→ Fit a linear regression model with ridge and lasso.
→ Penalty parameters chosen through cross-validation.
→ With these penalty parameters, penalized regression parameters and linear predictors are obtained.
→ The linear predictor is compared to the observations.

Simulation II

Ridge vs. lasso estimation (n=100, p=50)

Spearman's correlations of observation vs. model prediction

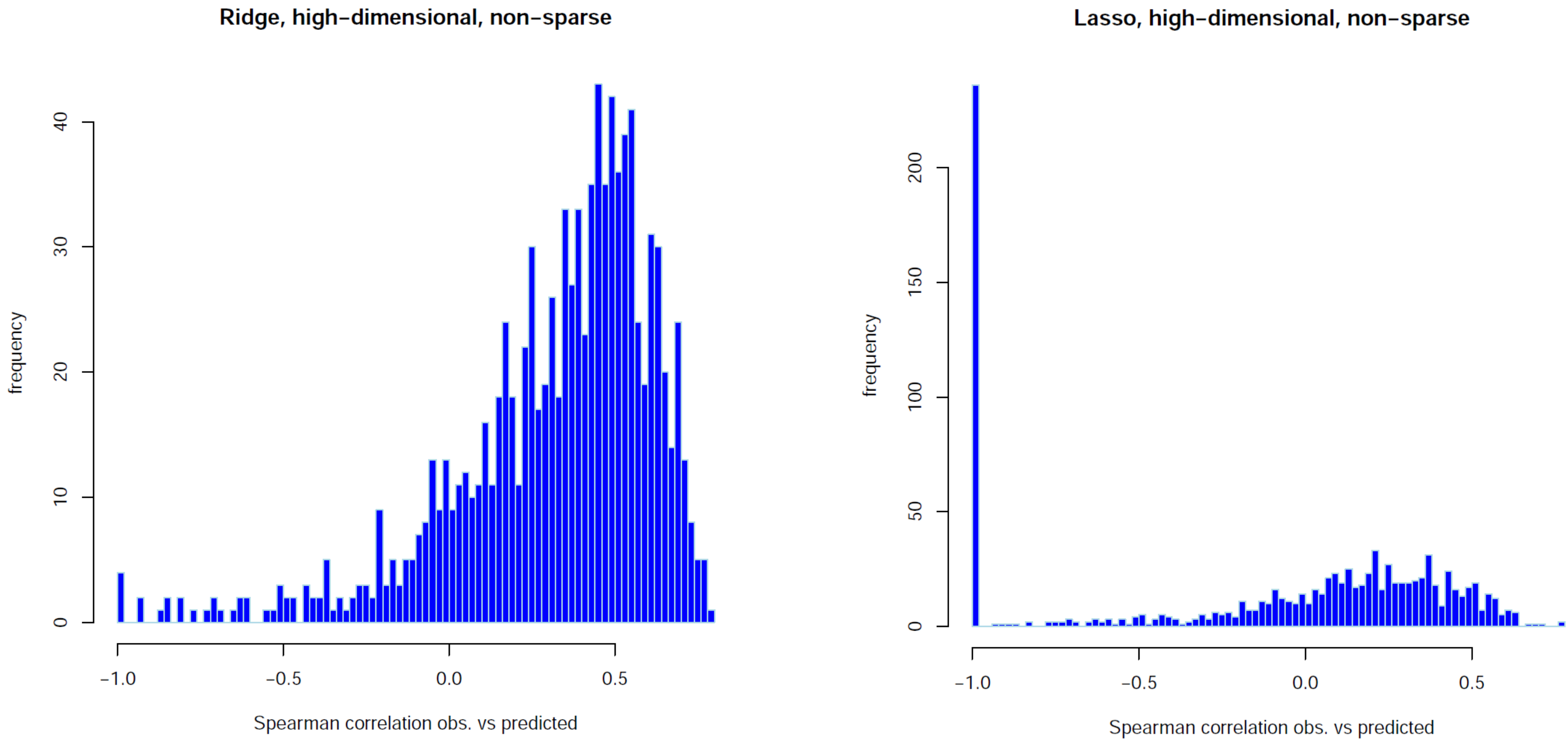




Simulation II

Ridge vs. lasso estimation (n=50, p=100)

Spearman's correlations of observation vs. model prediction



Simulation II

Ridge vs. lasso estimation

Consider a set of 50 genes. Their expression levels follow a standard multivariate normal law.

Together they regulate a 51th gene, in accordance with the following relationship:

$$Y_i = \mathbf{X}_{i*}\boldsymbol{\beta} + \varepsilon_i \quad \text{with} \quad \varepsilon \sim \mathcal{N}(0, 1)$$

The regression coefficients are

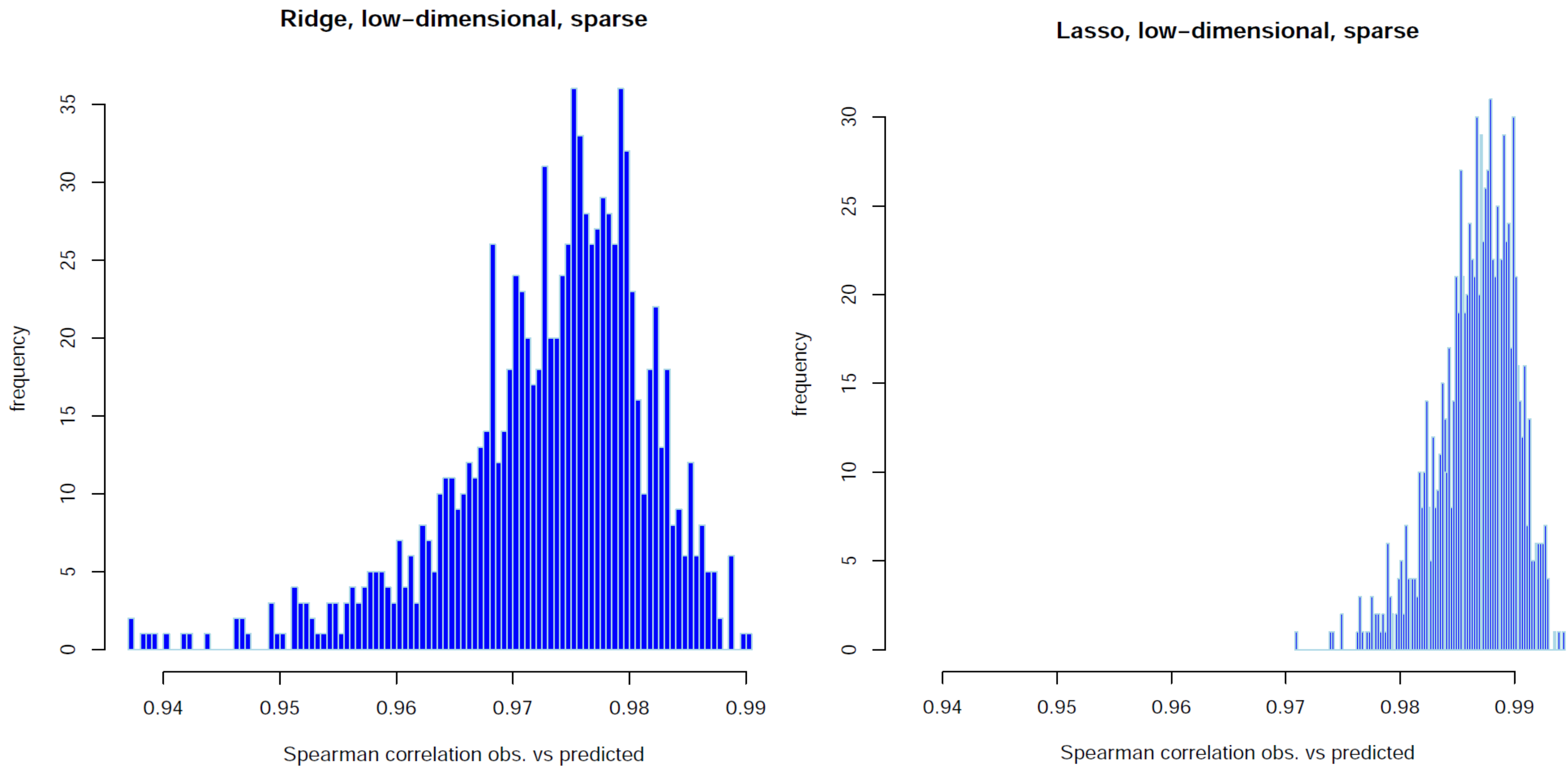
$$\beta_j = \begin{cases} j & \text{if} \quad j = 1, 2, \ldots, 5 \\ 0 & \text{if} \quad j > 5 \end{cases}$$

Hence, only five genes contribute.

Simulation II

Ridge vs. lasso estimation (n=100, p=50)

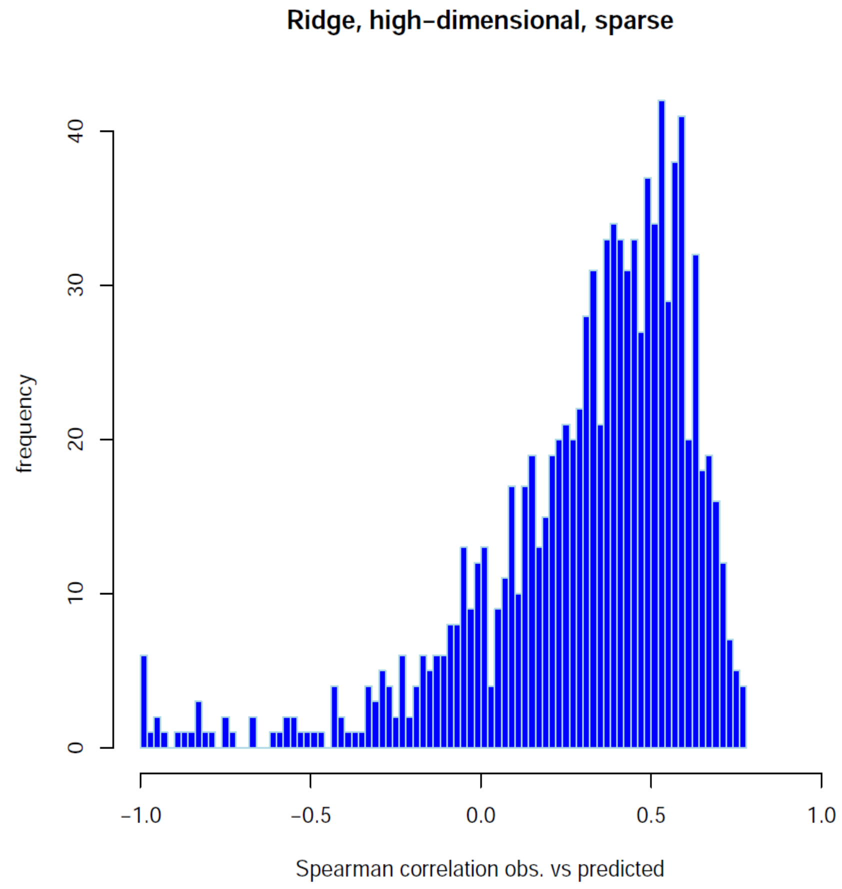
Spearman's correlations of observation vs. model prediction

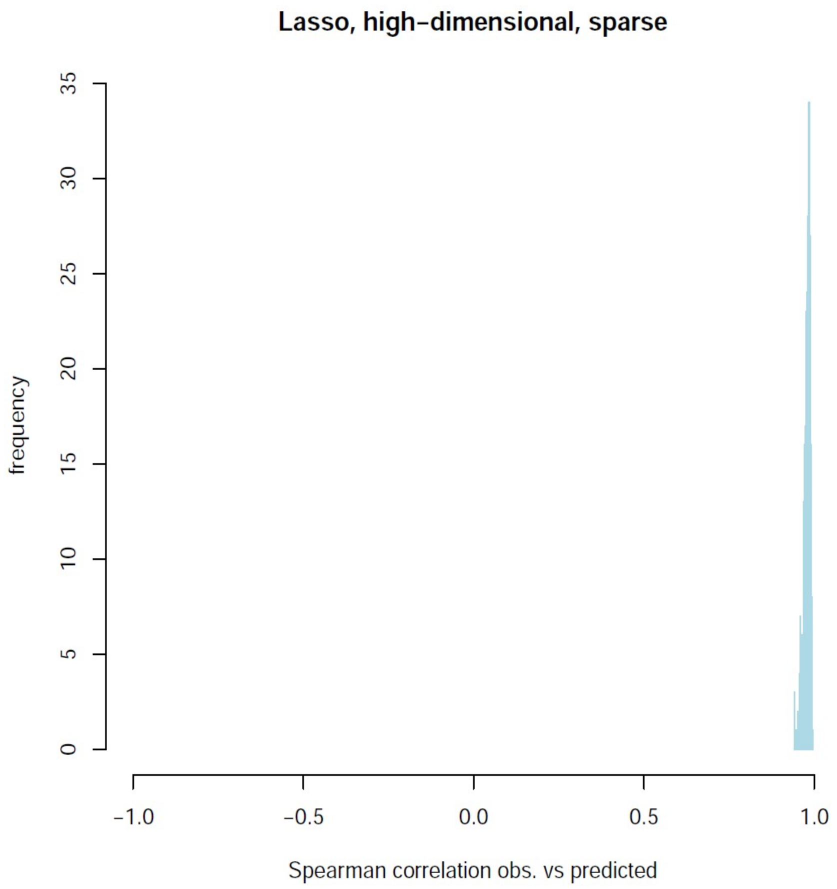


Simulation II

Ridge vs. lasso estimation (n=50, p=100)

Spearman's correlations of observation vs. model prediction





Simulations

Simulations

Simulation I and II suggest:

→In the presence of many small or medium effect sizes ridge is to be preferred.

→In only a few variables have a medium to large effect, the lasso is the method of choice.

However, simulations do not take into account collinearity.
A second run of these simulations, incorporating collinearities, indicates that ridge regression appear to profit more from collinearity.

Simulation III

Effect of lasso estimation

Consider a set of 50 genes. Their expression levels followa multivariate normal law with mean zero and covariance:

$$\mathbf{\Sigma} \quad = \begin{pmatrix} \mathbf{\Sigma}_{11} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{\Sigma}_{22} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{\Sigma}_{33} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{\Sigma}_{44} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{\Sigma}_{55} \end{pmatrix}$$

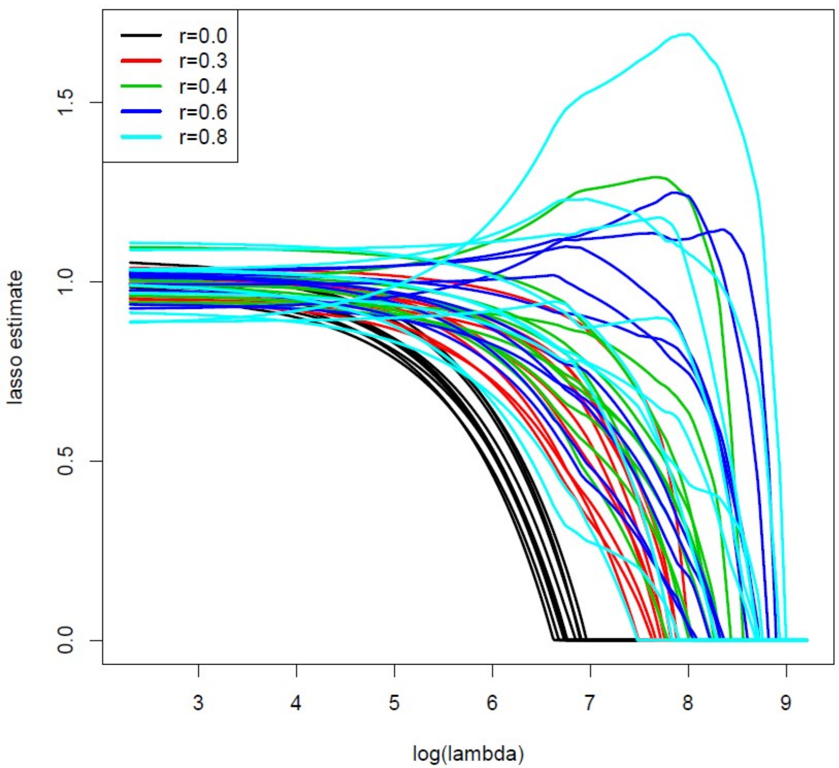
where $\mathbf{\Sigma}_{bb} = \frac{b-1}{5}\mathbf{1}_{10\times 10} + \frac{6-b}{5}\mathbf{I}_{10\times 10}$.

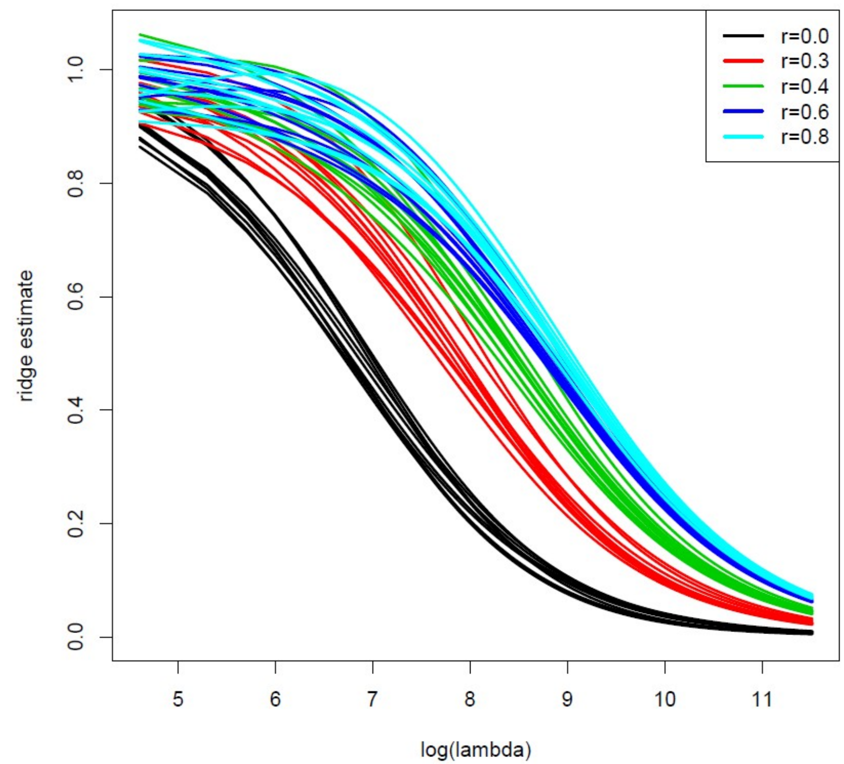
Together they regulate a 51th gene through: $Y_i = \mathbf{X}_{i*}\boldsymbol{\beta} + \varepsilon_i$ with $\varepsilon_i \sim \mathcal{N}(0,1)$ and regression coefficients $\boldsymbol{\beta} = \mathbf{1}_{50}$.
Hence, the 50 genes contribute equally.

Simulation III

Effect of lasso estimation

Whereas ridge regression shrinks coefficients of collinear covariates towards each other, lasso regression is somewhat indifferent to very correlated predictors and tends to pick one covariate and ignore the rest.

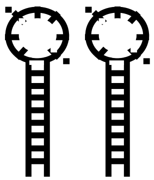




Example

Regulation of mRNA by microRNA

Example: microRNA-mRNA regulation



microRNAs

Recently, a new class of RNA was discovered: MicroRNA (mir). Mirs are non-coding RNAs of approx. 22 nucleotides. Like mRNAs, mirs are encoded in and transcribed from the DNA.

Mirs down-regulate gene expression by either of two post-transcriptional mechanisms: mRNA cleavage or transcriptional repression. Both depend on the degree of complementarity between the mir and the target.

A single mir can bind to and regulate many different mRNA targets and, conversely, several mirs can bind to and cooperatively control a single mRNA target.

Example: mir-mRNA regulation

Aim

Model microRNA regulation of mRNA expression levels.

Data

→ 90 prostate cancers

→ expression of 735 mirs

→ mRNA expression of the MCM7 gene

Motivation

→ MCM7 involved in prostate cancer.

→ mRNA levels of MCM7 reportedly affected by mirs.

Not part of the objective: *feature selection* ≈ understanding the basis of this prediction by identifying features (mirs) that characterize the mRNA expression.

Example: microRNA-mRNA regulation

Analysis

Find:

$$\texttt{mrna expr.} = f(\texttt{mir expression})$$

$$= \beta_0 + \beta_1 * \texttt{mir}_1 + \beta_2 * \texttt{mir}_2 + \ldots + \beta_p * \texttt{mir}_p + \texttt{error}$$

However, $p > n$: lasso regression. Having found the optimal λ, we obtain the lasso estimates for the coefficients: $b_j(\lambda)$.

With these estimates we calculate the linear predictor:

$$b_0 + b_1(\lambda) * \texttt{mir}_1 + \ldots + b_p(\lambda) * \texttt{mir}_p$$

Finally, we obtain the predicted survival:

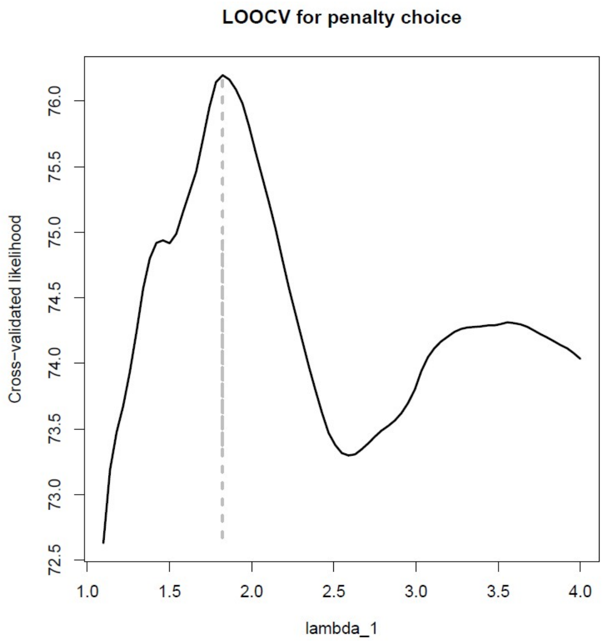
$$\texttt{pred. mrna expr.} = f(\texttt{linear predictor})$$

$$= b_0 + b_1(\lambda) * \texttt{mir}_1 + \ldots + b_p(\lambda) * \texttt{mir}_p$$

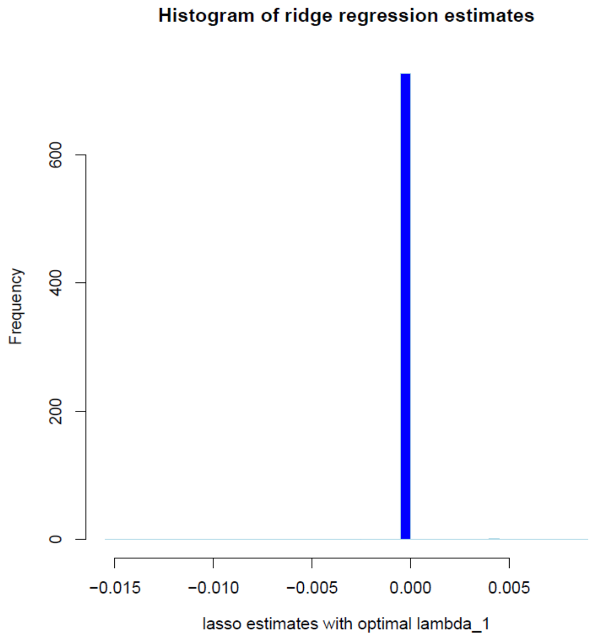
Compare observed and predicted mRNA expression.

Example: microRNA-mRNA regulation

Penalty parameter choice



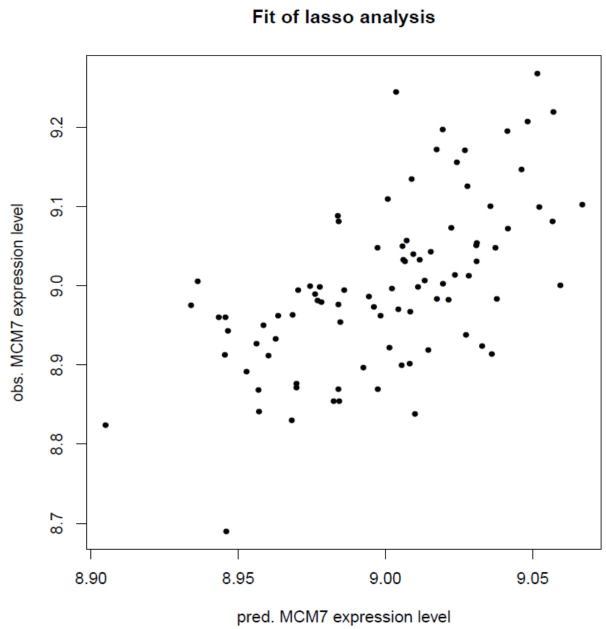
Beta hat distribution



$\#(\beta \mathrel{!=} 0) =$
8 (out of 735)

$\#(\beta < 0) =$
3 (out of 735)

Obs. vs. pred. mRNA expression



$\rho_{sp} = 0.626$
$R^2 = 0.372$

Example: microRNA-mRNA regulation

Biological dogma

MicroRNAs down-regulate mRNA levels: negative regression coefficients prevail. Re-analyze the data with sign parameter constraints.

Are the microRNAs identified to down-regulate MCM7 expression levels also reported by prediction tools?

Contingency table

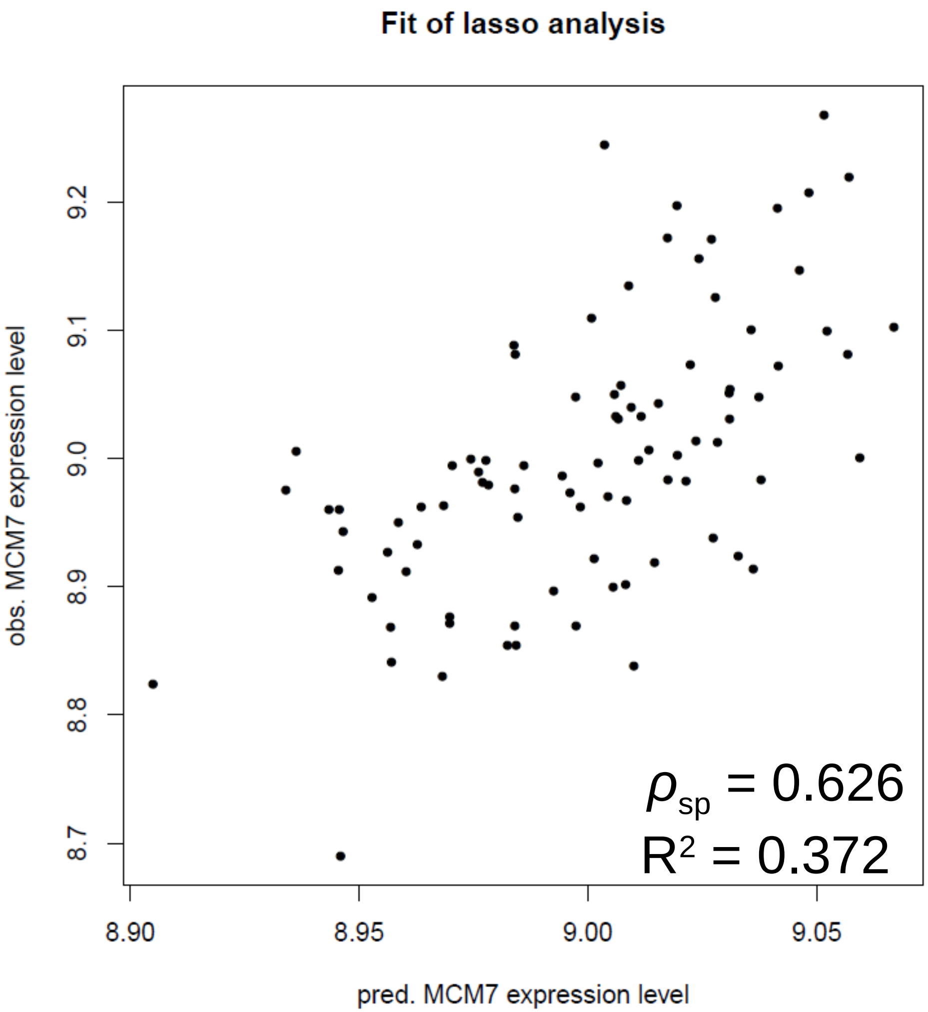
```
                           prediction tool
ridge regression    no-mir2MCM7  mir2MCM7
       β = 0               705        22
       β < 0                 8         0
```

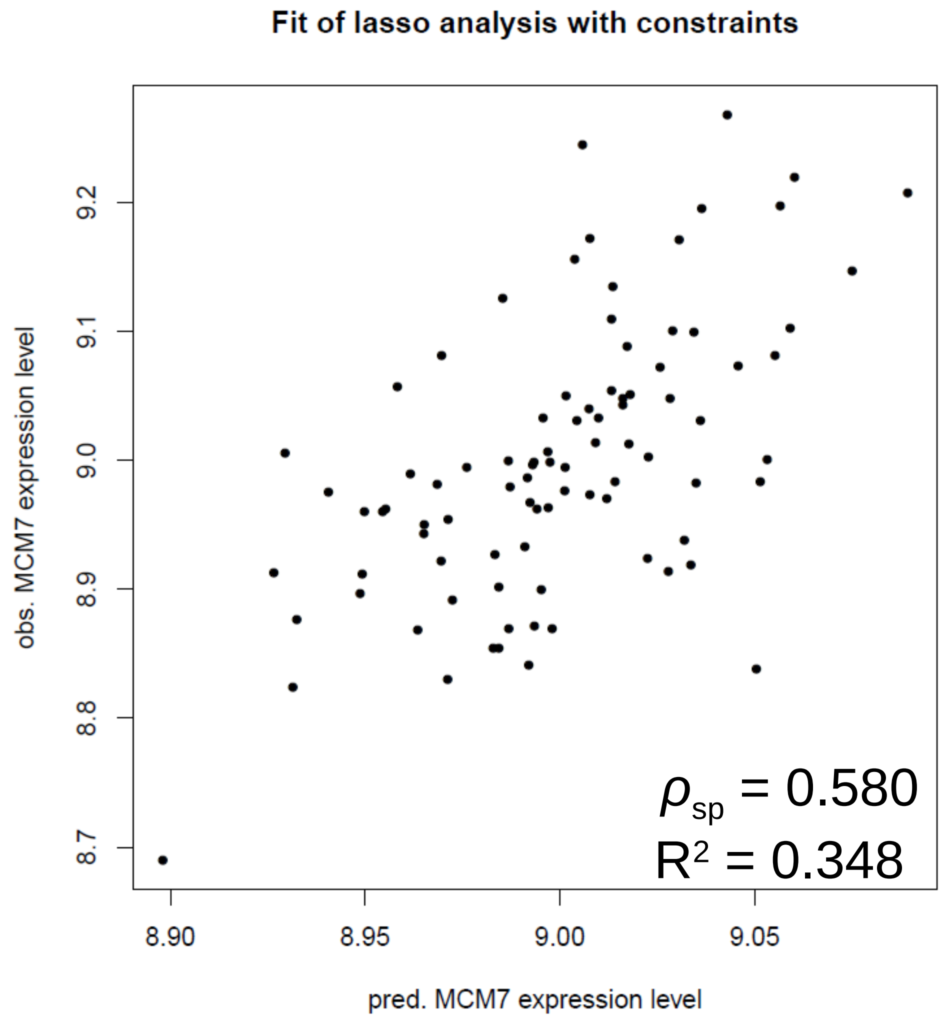
Chi-square test

```
Pearson's Chi-squared test with Yates' continuity correction

data:  table(nonzeroBetas, nonzeroPred)
X-squared = 0, df = 1, p-value = 1
```

Example: microRNA-mRNA regulation

Observed vs. predicted mRNA expression for both analyses.





Example

Clinical outcome prediction

Example: clinical outcome prediction

Breast cancer data of Van ‘t Veer et al. (2004)

Study involves:

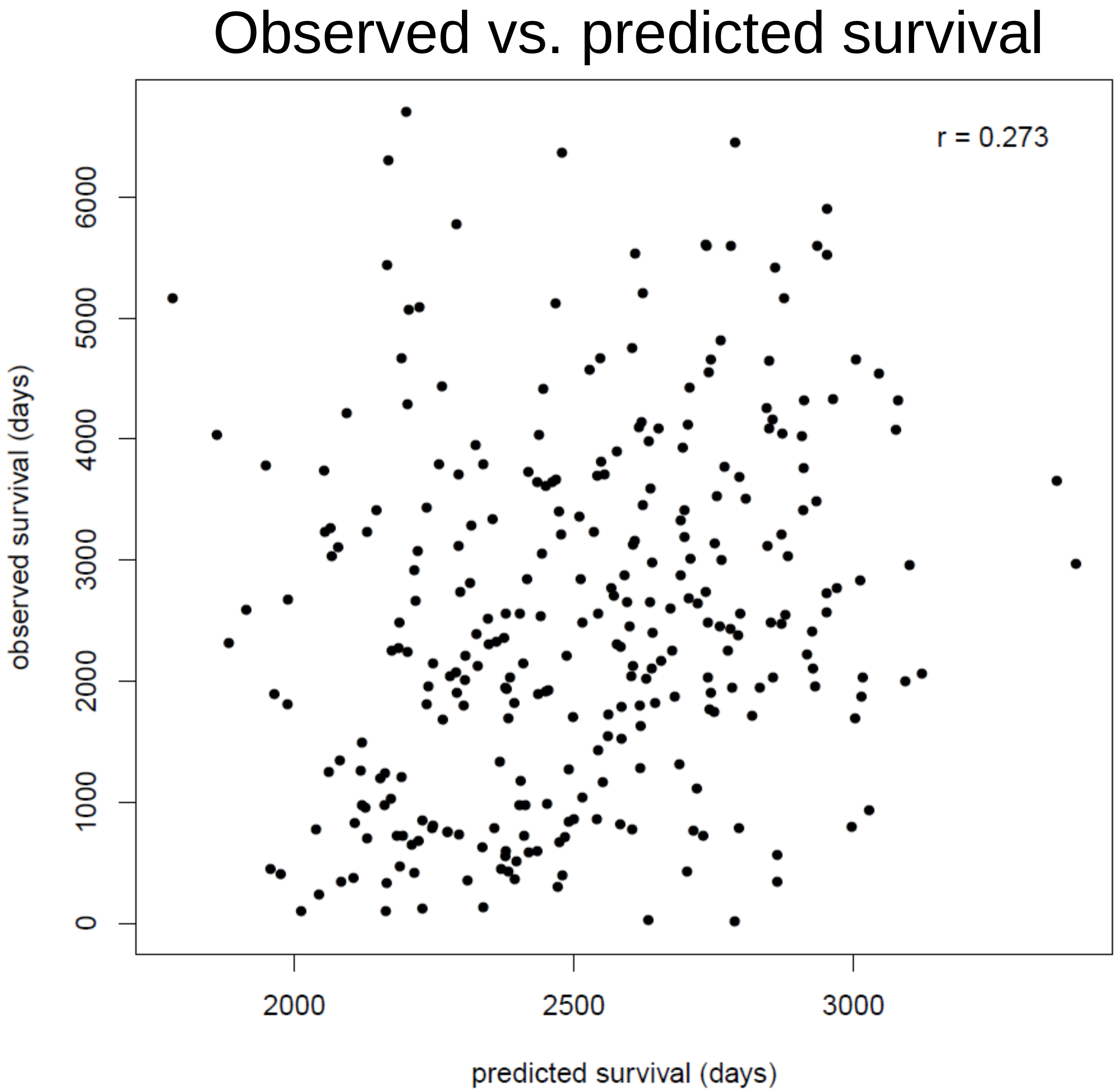
- 291 (after preprocessing) breast cancer samples,
- expression profile of 24158 genes for each sample, and
- survival data for each sample.

Question

Can we predict the survival time of a breast cancer patient on the basis of its gene expression data?

Now: lasso for the Cox model.

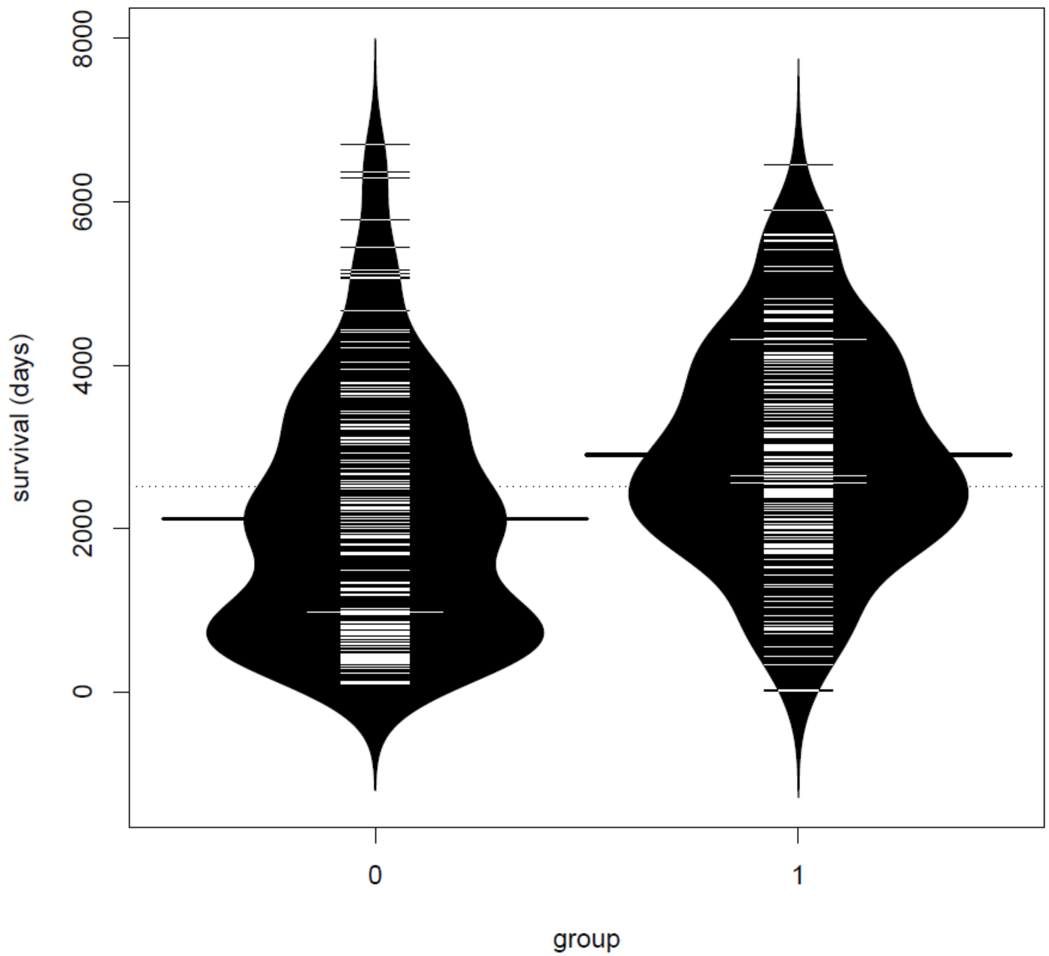
Example: clinical outcome prediction



Example: clinical outcome prediction

Analysis (continued)

Compare groups by means of violinplots.



median survival
-> group 0: 1937
-> group 1: 2726

mammaprint

Example: clinical outcome prediction

Analysis (continued)

Can we say anything about the underlying biology?
E.g., which genes contribute most to survival?

Solution

Look for non-zero regression coefficients.

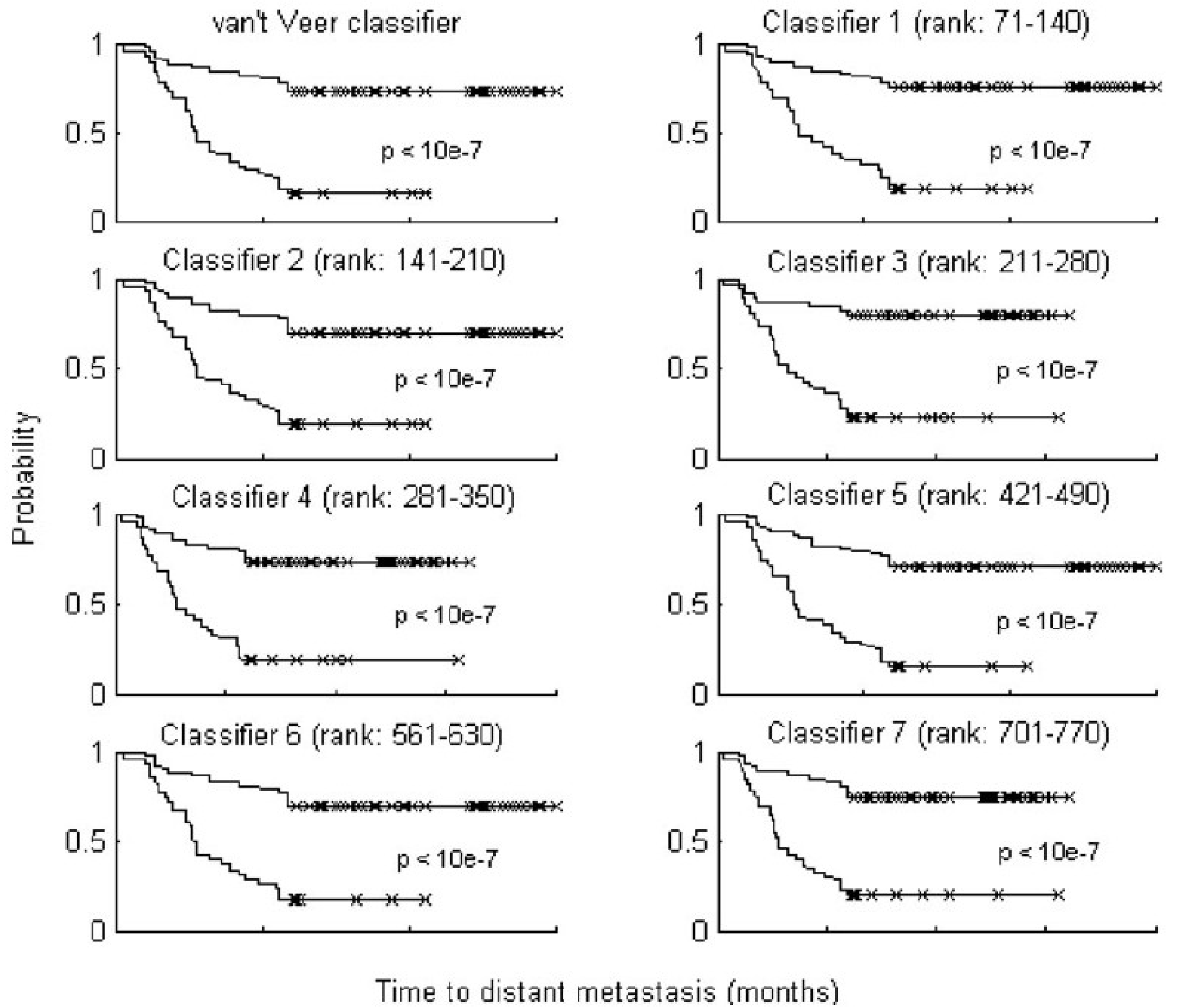
Lasso finds 8 genes with non-zero coefficients:

NM_000909	**NM_002411**	**AL117406**
NM_006115	**Contig48328_RC**	**NM_020974**
Contig14284_RC		**AF067420**

Example: clinical outcome prediction

Ein-Dor *et al*. (Bioinformatics, 2005) showed that predictor with non-overlapping gene sets may perform equally well.

Famous example in breast cancer: Amsterdam signature vs. Rotterdam signature



Example: clinical outcome prediction

Question

OPEN ACCESS Freely available online   PLoS COMPUTATIONAL BIOLOGY

Most Random Gene Expression Signatures Are Significantly Associated with Breast Cancer Outcome

David Venet[1], Jacques E. Dumont[2], Vincent Detours[2,3]*

1 IRIDIA-CoDE, Université Libre de Bruxelles (U.L.B.), Brussels, Belgium, **2** IRIBHM, Université Libre de Bruxelles (U.L.B.), Campus Erasme, Brussels, Belgium, **3** WELBIO, Université Libre de Bruxelles (U.L.B.), Campus Erasme, Brussels, Belgium

Explain the above title.
Note: size of signatures $p \approx 100$

Note
Ein-Dor *et al*. (PNAS, 2006) showed that a training set of thousands of samples is needed to produce a predictor with a stable gene set. That does not imply the predictor is any good.

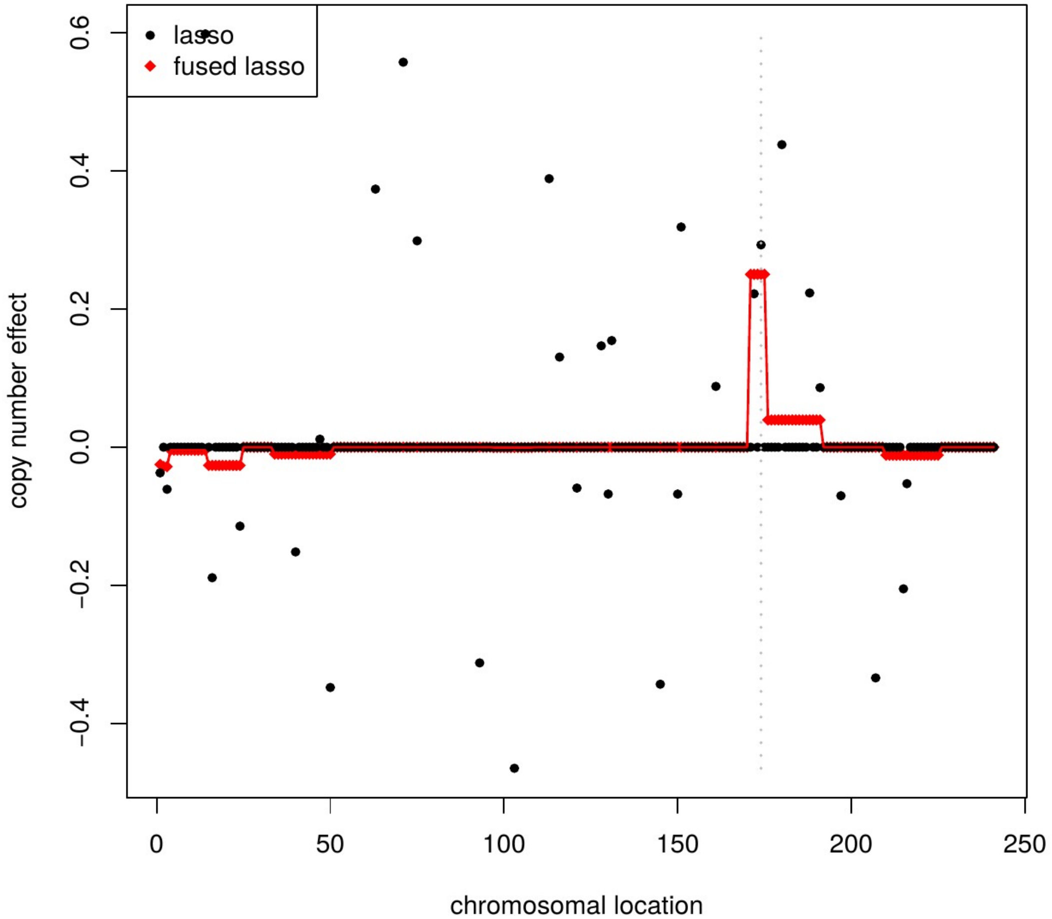
Lasso variants

Lasso variants - I

Fused lasso

The fused lasso estimator, using $\lambda_{1,f} \sum_{j=1}^{p-1} |\beta_{j+1} - \beta_j|$, penalizes differences instead of individual coefficients.

See exercises for its computation.



Application to the DNA copy number *trans*-effect of the fused ridge.

Lasso variants - II

Adaptive lasso

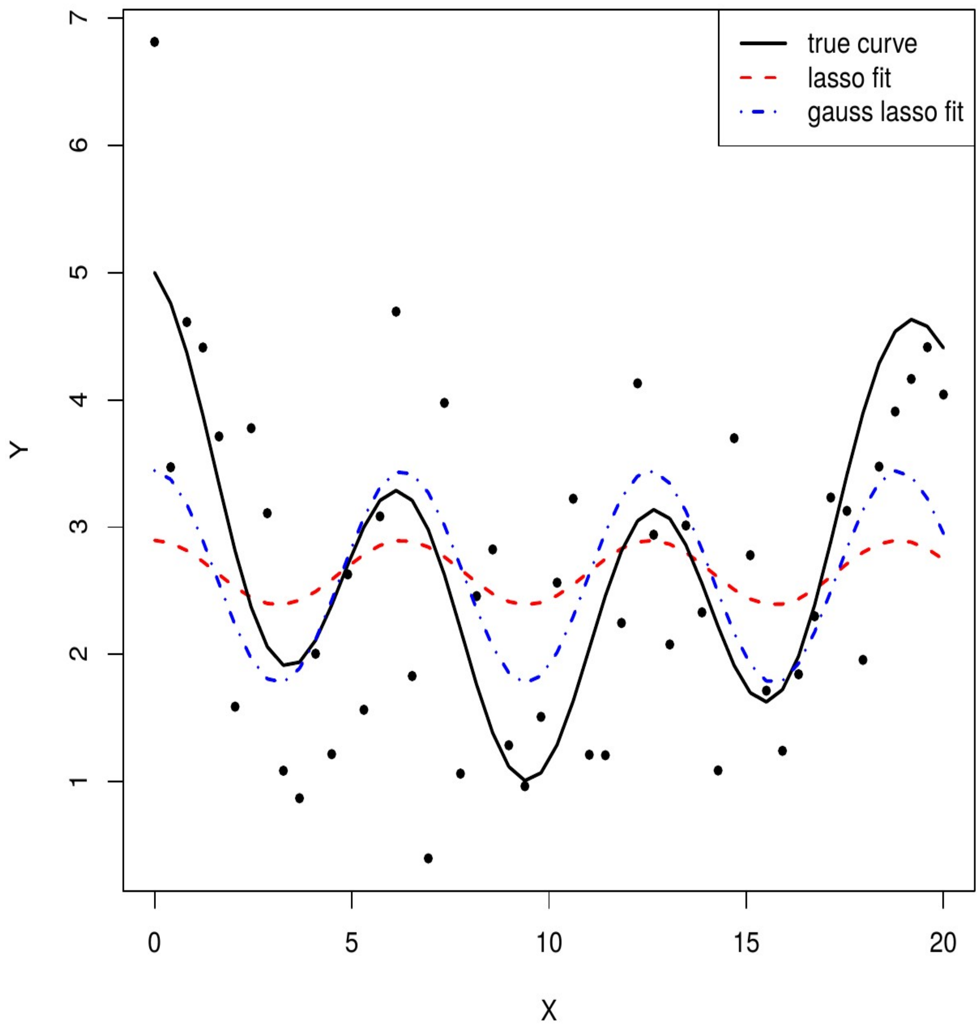
Lasso regression shrinks coefficients to zero.

Correction for shrinkage:

→ use lasso regression for variable selection, and

→ re-estimate parameters of selected variables by means of OLS.

This is referred to as the *Gauss-Lasso estimator*.



Lasso variants - II

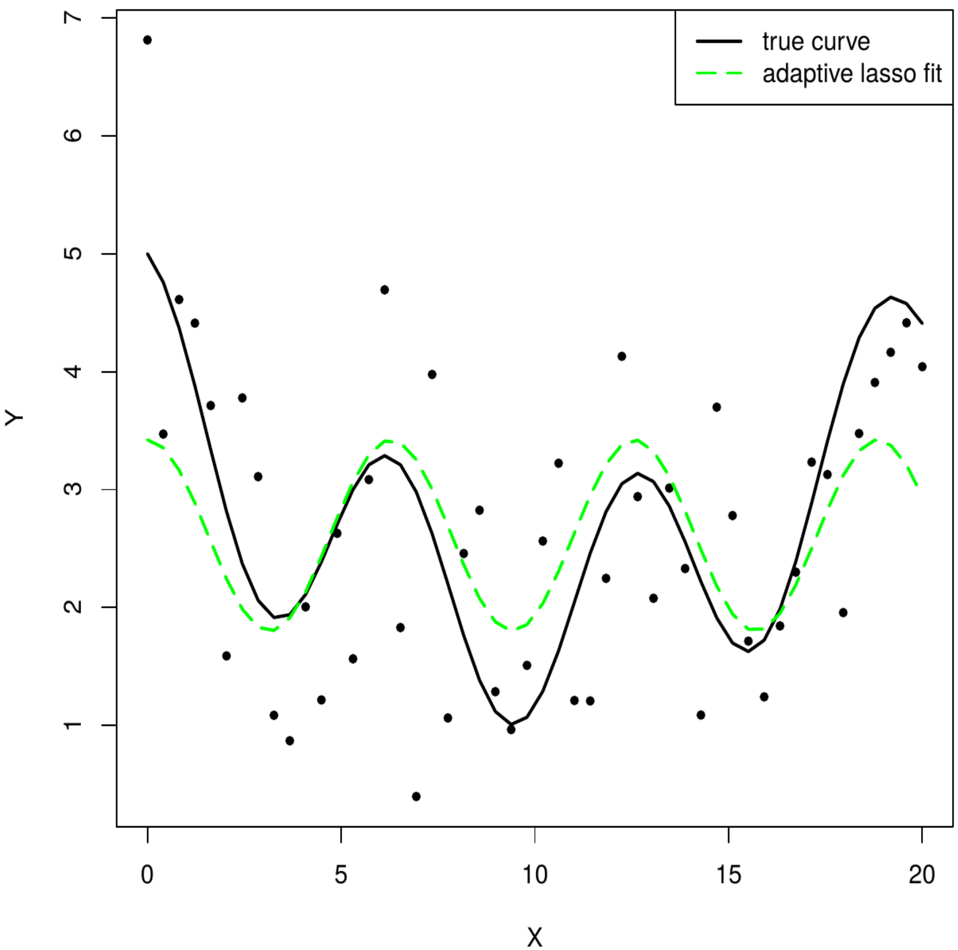
Adaptive lasso

As before but replace OLS estimator by lasso estimator with modified penalty:

$$\lambda_1 \sum_{j=1}^{p} \frac{|\beta_j|}{|[\hat{\beta}^{\text{Gauss-Lasso}}(\lambda)]_j|}$$

This yields the adaptive lasso estimator.

In similar fashion, a *Ridge-Lasso estimator* may be conceived.



Lasso variants - III

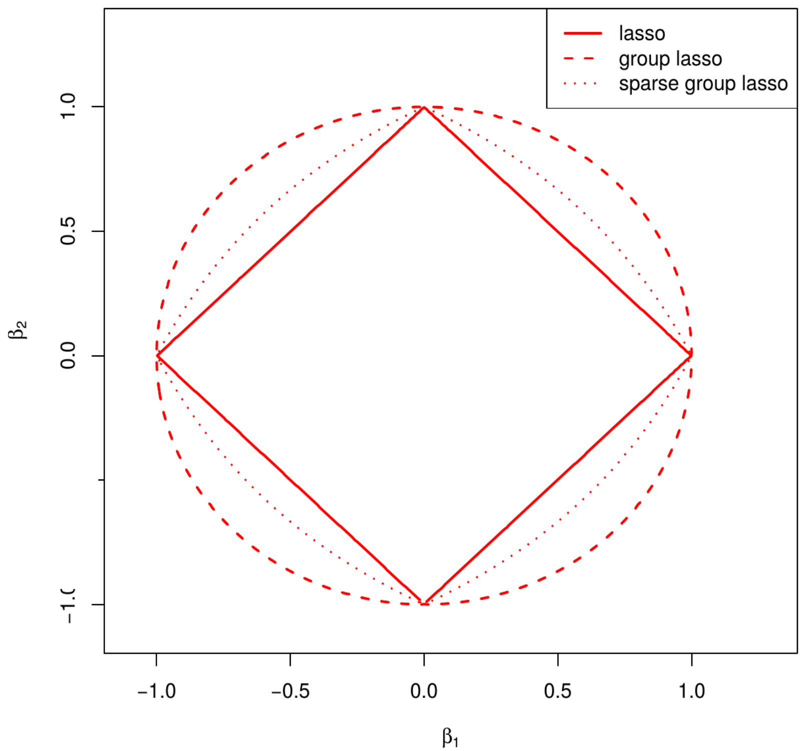
Sparse group lasso

Groups of variables may be discerned. To select at the group level employ the *group lasso penalty*:

$$\lambda_{1,G} \sum_{g=1}^{G} \sqrt{|J_g|} \|\boldsymbol{\beta}_g\|_2 = \lambda_{1,G} \sum_{g=1}^{G} \sqrt{|J_g|} \sqrt{\sum_{j \in J_g} \beta_j^2}$$

The group lasso estimator does not result in a sparse within-group estimate. This may be achieved through employment of the *sparse group lasso penalty*:

$$\lambda_1 \|\boldsymbol{\beta}\|_1 + \lambda_{1,G} \sum_{g=1}^{G} \sqrt{|J_g|} \|\boldsymbol{\beta}_g\|_2$$



Lasso variants - III

Sparse group lasso

The sparse group lasso estimator is found through exploitation of the convexity of the loss function:

→ group-wise optimization,

→ within-group parameter-wise optimization.

Much like the coordinate descent approach.

→ Does it work?

→ Show regularization paths.

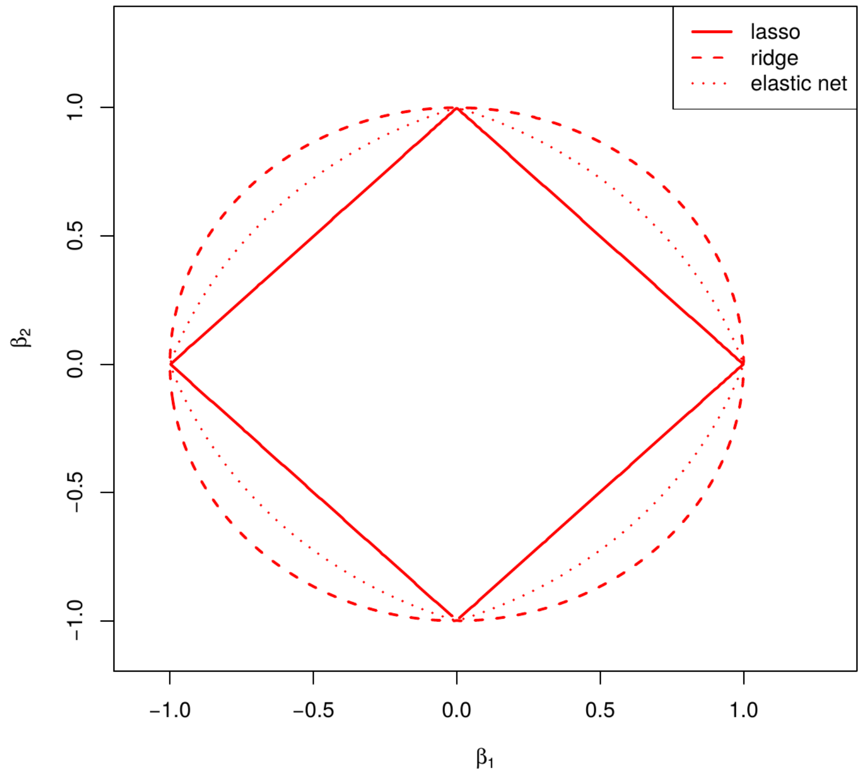
Lasso variants - IV

Elastic net penalty

Ridge regression shrinks coefficients of collinear covariates towards each other, while lasso regression is somewhat indifferent to correlated predictors and tends to pick one covariate and ignore the rest.

This drawback (?) of the lasso may be resolved by simply adding the two penalty, thus forming the elastic net penalty:

$$\lambda_1 \|\boldsymbol{\beta}\|_1 + \lambda_2 \|\boldsymbol{\beta}\|_2^2$$



Lasso variants - IV

Elastic net penalty

Consider a set of 50 genes. Their expression levels follow a multivariate normal law with mean zero and block diagonal covariance with $\mathbf{\Sigma}_{bb} = \frac{b-1}{5}\mathbf{1}_{10\times 10} + \frac{6-b}{5}\mathbf{I}_{10\times 10}$ for *b* = 1, … 5.

Together they regulate a 51th gene through: $Y_i = \mathbf{X}_{i*}\boldsymbol{\beta} + \varepsilon_i$ with $\varepsilon_i \sim \mathcal{N}(0, 1)$ and regression coefficients:

→ $\boldsymbol{\beta} = \mathbf{1}_{50}$,

→ $\left\{ \begin{array}{lll} \beta_j = 0 & \text{for} & j = 1, \ldots, 45, \\ \beta_j = 1 & \text{for} & j = 46, \ldots, 50. \end{array} \right.$

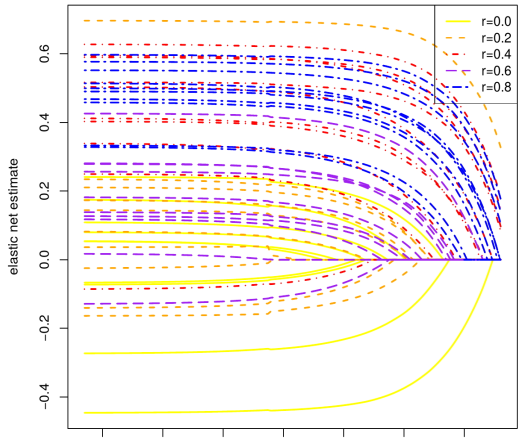
→ $\left\{ \begin{array}{lll} \beta_j = 0 & \text{for} & j = 1, \ldots, 25, 31, \ldots, 50 \\ \beta_j = 1 & \text{for} & j = 26, \ldots, 30. \end{array} \right.$

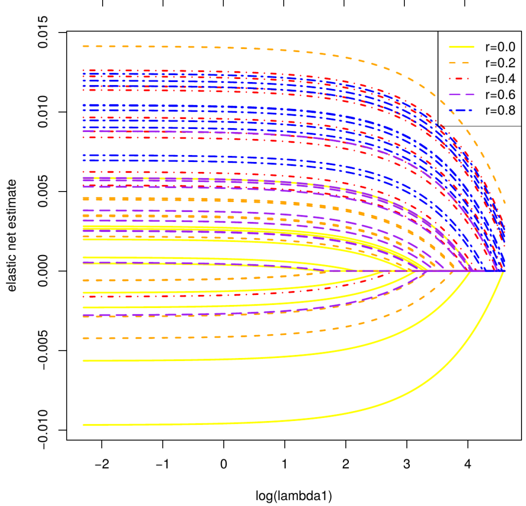
Evaluate (see exercises) the elastic net estimator with $\lambda_1 \in (0, 100]$ and either $\lambda_2 = 100$ or $\lambda_2 = 10000$.

Lasso variants - IV

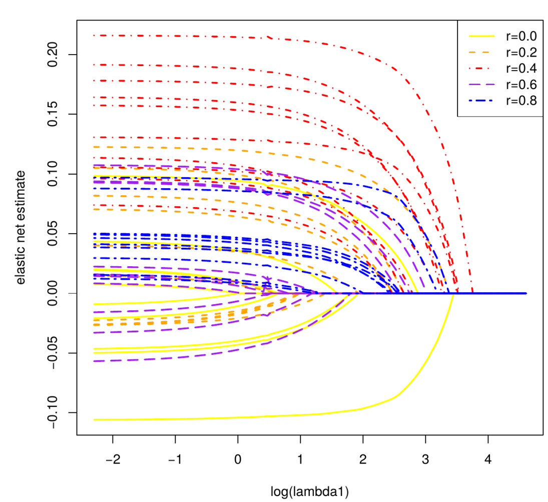
Elastic net penalty

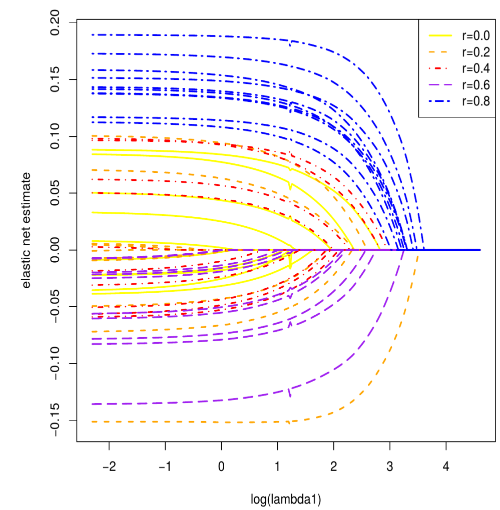
Non-sparse: need not have an obvious effect.





Sparse: tends to have an effect. High correlation + dominating ridge penalty preferred.





Lasso variants - IV

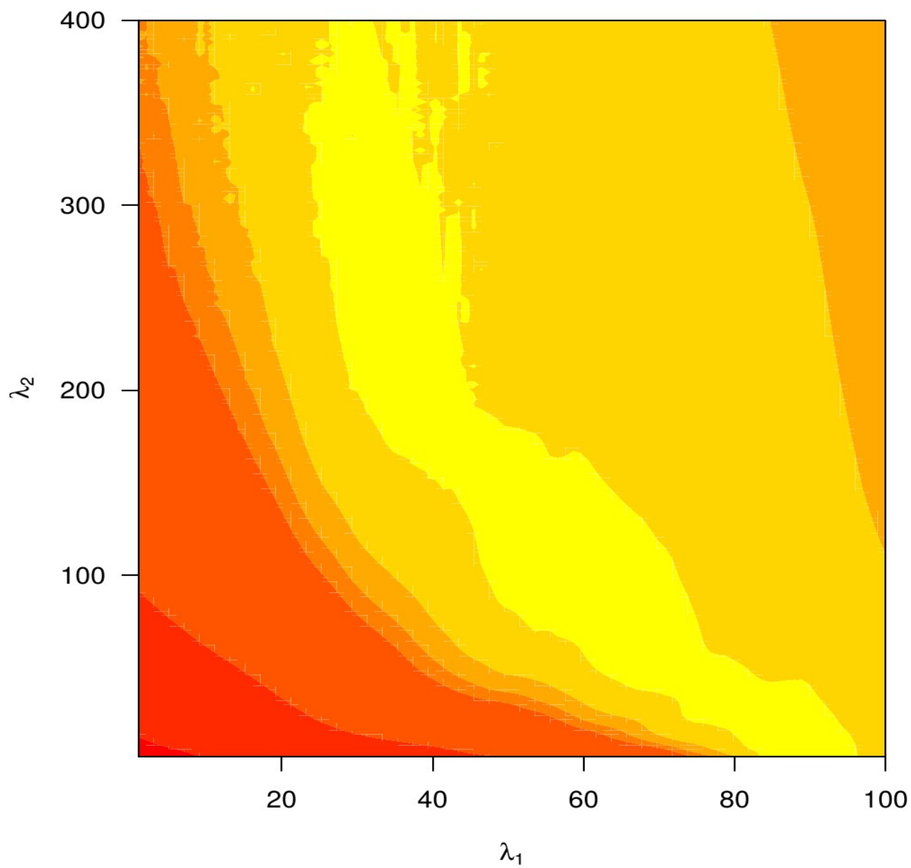
Elastic net penalty

Both penalty terms shrink the parameter estimates. These confounding shrinkage effects frustrate the choice of the penalty parameters when optimizing a prediction criterion.

CV-likelihood contour
→ red = low
→ yellow = high

Flat surface from orange to yellow.

Many penalty parameter combinations of yield the same CV-likelihood.



Lasso variants - V

Bridge penalty

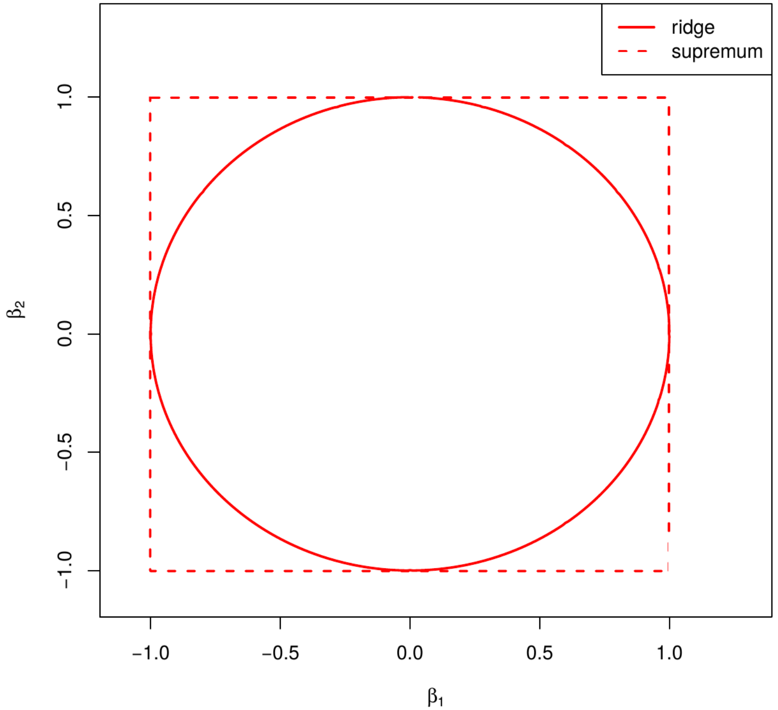
Large class of penalties, of which ridge and lasso are special cases.

Penalty:

$$\lambda_b \sum_{j=1}^{p} |\beta_j|^\gamma$$

Question

Supremum norm ($\gamma = \infty$) also yields corners in constraint. Why does the resulting estimator not select?

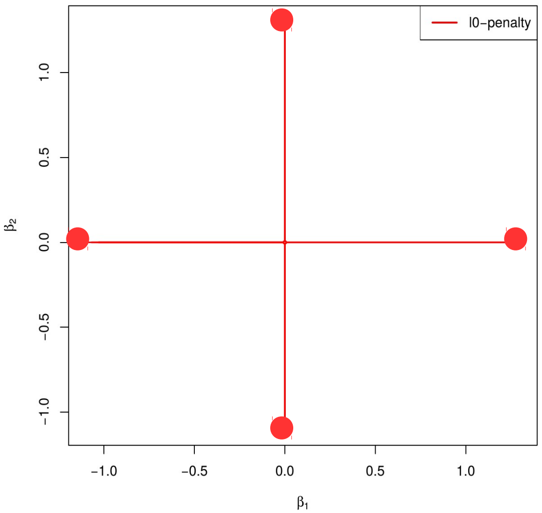


Lasso variants - VI

L_0 penalty

The ideal penalty would be the L_0-penalty: $$\lambda_0 \sum_{j=1}^{p} I_{\{\beta_j \neq 0\}}$$

This penalty thus punishes only the number of covariates that enters the model, not their regression coefficients (which are only surrogates).



This penalty is computationally too demanding: one searches over all possible subsets of the p covariates.

Question: can the adaptive lasso be viewed as a surrogate?

References & further reading

References & further reading

Buhlmann, P. Van der Geer, S. (2011), *Statistics for High-dimensional Data: Methods, Theory and Applications*, Springer.

Castillo, I., Schmidt-Hieber, J., & Van der Vaart, A. (2015). Bayesian linear regression with sparse priors. The Annals of Statistics, 43(5), 1986-2018.

Ein-Dor, Liat, et al. "Outcome signature genes in breast cancer: is there a unique set?." *Bioinformatics* 21.2 (2005): 171-178.

Fan, J., Li, R. (2001), "Variable selection via nonconcave penalized likelihood and its oracle properties", *JASA*, 96(456), 1348-1360.

Goeman, J.J. (2010), "L1 penalized estimation in the Cox proportional hazard model", *Biometrical Journal*, 52(1), 70-84.

Meinshausen, N., Buhlmann, P. (2010), "Stability selection", *JRSS B*, 74(4), 417-473.

Osborne, M.R., Presnell, B., Turlach, B.A. (2000), "On the LASSO and its dual", Journal of Computational and Graphical Statistics, 9(2), 319-337.

Park, T., & Casella, G. (2008). The Bayesian lasso. Journal of the American Statistical Association, 103(482), 681-686.

Tibshirani, R. (1996), "Regression shrinkage and selection via the Lasso", JRSS B, 58(1), 267-288.