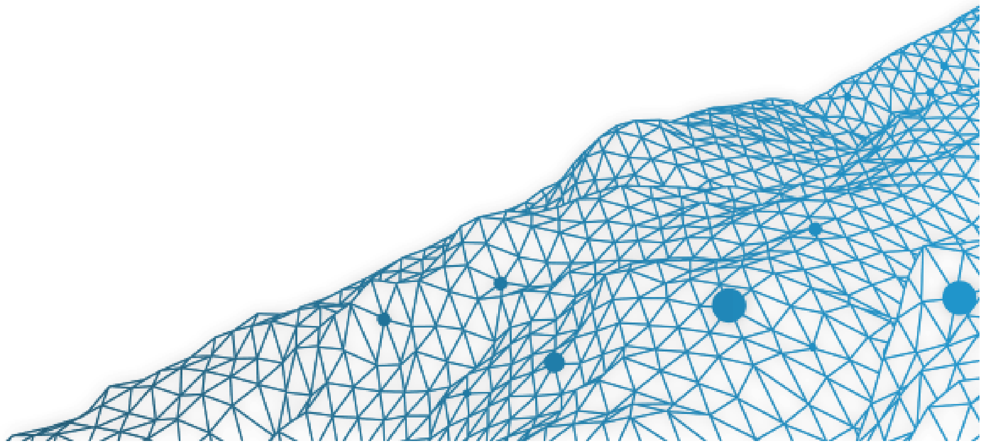
3 Regularization & Penalized Regression

Arthur Charpentier (Université du Québec à Montréal)

Machine Learning & Econometrics

SIDE Summer School - July 2019



Linear Model and Variable Selection

Let s denote a subset of $\{0,1,\cdots,p\}$, with cardinal $|s|$.

$\boldsymbol{X}_s$ is the matrice with columns $\boldsymbol{x}_j$ where $j\in s$.

Consider the model $\boldsymbol{Y}=\boldsymbol{X}_s\boldsymbol{\beta}_s+\boldsymbol{\eta}$, so that $\widehat{\boldsymbol{\beta}}_s=\big(\boldsymbol{X}_s^\top\boldsymbol{X}_s\big)^{-1}\boldsymbol{X}_s^\top\boldsymbol{y}$

In general, $\widehat{\boldsymbol{\beta}}_s\neq(\widehat{\boldsymbol{\beta}})_s$

R^2 is usually not a good measure since $R^2(s)\leq R^2(t)$ when $s\subset t$.

Some use the adjusted R^2, $\overline{R}^2(s)=1-\dfrac{n-1}{n-|s|}\big(1-R^2(s)\big)$

The mean square error is

$$\text{mse}(s)=\mathbb{E}\big[(\boldsymbol{X}\boldsymbol{\beta}-\boldsymbol{X}_s\widehat{\boldsymbol{\beta}}_s)^2\big]=\mathbb{E}\big[RSS(s)\big]-n\sigma^2+2|s|\sigma^2$$

Define Mallows'C_p as $C_p(s)=\dfrac{RSS(s)}{\widehat{\sigma}^2}-n+2|s|$

Rule of thumb: model with variables s is valid if $C_p(s)\leq|s|$

Linear Model and Variable Selection

In a linear model,

$$\log \mathcal{L}(\boldsymbol{\beta},\sigma^2) = -\frac{n}{2}\log\sigma^2 - \frac{n}{2}\log(2\pi) - \frac{1}{2\sigma^2}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|^2$$

and

$$\log \mathcal{L}(\widehat{\boldsymbol{\beta}}_s,\widehat{\sigma}^2_s) = -\frac{n}{2}\log\frac{RSS(s)}{n} - \frac{n}{2}[1+\log(2\pi)]$$

It is necessary to penalize too complex models

Akaike's *AIC* : $AIC(s) = \frac{n}{2}\log\frac{RSS(s)}{n} + \frac{n}{2}[1+\log(2\pi)] + 2|s|$

Schwarz's *BIC* : $BIC(s) = \frac{n}{2}\log\frac{RSS(s)}{n} + \frac{n}{2}[1+\log(2\pi)] + |s|\log n$

Exhaustive search of all models, 2^{p+1}... too complicated.

Stepwise procedure, forward or backward... not very stable and satisfactory.

Linear Model and Variable Selection

For variable selection, use the classical `stats::step` function, or `leaps::regsubset` for best subset, forward stepwide and backward stepwise.

For leave-one-out-cross validation, we can write

$$\frac{1}{n}\sum_{i=1}^n (y_i-\widehat{y}_{(i)})^2=\frac{1}{n}\sum_{i=1}^n\left(\frac{y_i-\widehat{y}_i}{1-\boldsymbol{H}_{i,i}}\right)^2$$

Heuristically, $(y_i-\widehat{y}_i)^2$ underestimates the true prediction error

High underestimation if the correlation between y_i and $\widehat{y}_i$ is high

One can use $\text{Cov}[\boldsymbol{y},\widehat{\boldsymbol{y}}]$, e.g. in Mallows' C_p,

$$C_p=\frac{1}{n}\sum_{i=1}^n (y_i-\widehat{y}_i)^2+\frac{2}{n}\sigma^2 p \text{ where } p=\frac{1}{\sigma^2}\text{trace}[\text{Cov}(\boldsymbol{y},\widehat{\boldsymbol{y}}])$$

with Gaussian errors, AIC and Mallows' C_p are asymptotically equivalent.

Penalized Inference and Shrinkage

Consider a parametric model, with true (unknown) parameter θ, then

$$\text{mse}(\hat{\theta}) = \mathbb{E}\left[(\hat{\theta}-\theta)^2\right] = \underbrace{\mathbb{E}\left[(\hat{\theta}-\mathbb{E}\left[\hat{\theta}\right])^2\right]}_{\text{variance}} + \underbrace{\mathbb{E}\left[(\mathbb{E}\left[\hat{\theta}\right]-\theta)^2\right]}_{\text{bias}^2}$$

One can think of a shrinkage of an unbiased estimator,

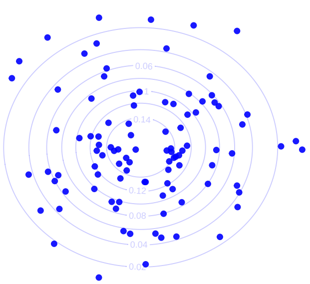
Let $\widetilde{\theta}$ denote an unbiased estimator of θ. Then

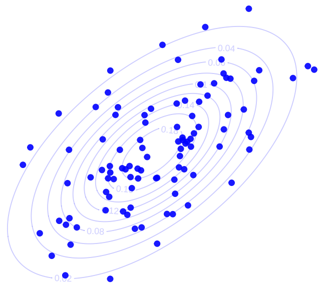
$$\hat{\theta} = \frac{\theta^2}{\theta^2+\text{mse}(\widetilde{\theta})}\cdot\widetilde{\theta} = \widetilde{\theta} - \underbrace{\frac{\text{mse}(\widetilde{\theta})}{\theta^2+\text{mse}(\widetilde{\theta})}}_{\text{penalty}}\cdot\widetilde{\theta}$$

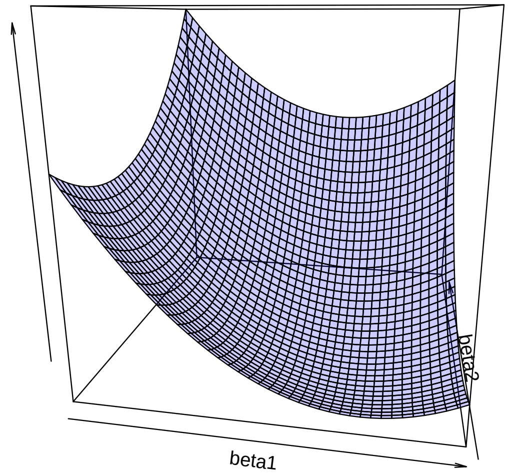
satisfies $\text{mse}(\hat{\theta}) \leq \text{mse}(\widetilde{\theta})$.

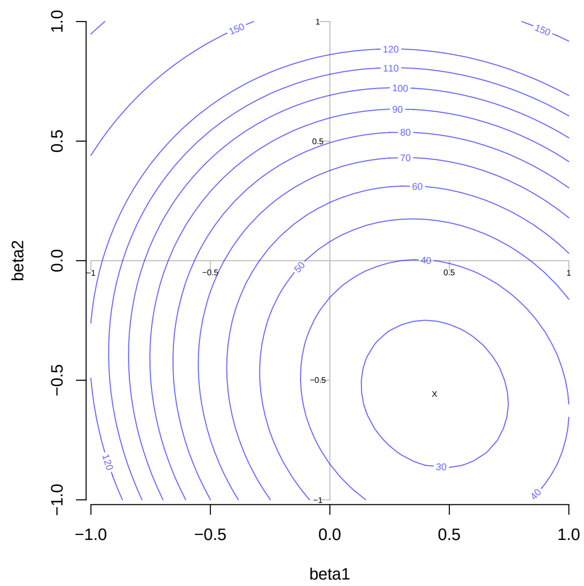
Normalization : Euclidean ℓ_2 vs. Mahalonobis

We want to penalize complicated models :
if β_k is "too small", we prefer to have $\beta_k = 0$.









Instead of $d(\boldsymbol{x},\boldsymbol{y}) = (\boldsymbol{x}-\boldsymbol{y})^{\mathsf{T}}(\boldsymbol{x}-\boldsymbol{y})$
use $d_{\boldsymbol{\Sigma}}(\boldsymbol{x},\boldsymbol{y}) = \sqrt{(\boldsymbol{x}-\boldsymbol{y})^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{y})}$

Linear Regression Shortcoming

Least Squares Estimator $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y}$

Unbiased Estimator $\mathbb{E}[\widehat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$

Variance $\text{Var}[\widehat{\boldsymbol{\beta}}] = \sigma^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1}$

which can be (extremely) large when $\det[(\boldsymbol{X}^\top\boldsymbol{X})] \sim 0$.

$$\boldsymbol{X} = \begin{bmatrix} 1 & -1 & 2 \\ 1 & 0 & 1 \\ 1 & 2 & -1 \\ 1 & 1 & 0 \end{bmatrix} \text{ then } \boldsymbol{X}^\top\boldsymbol{X} = \begin{bmatrix} 4 & 2 & 2 \\ 2 & 6 & -4 \\ 2 & -4 & 6 \end{bmatrix} \text{ while } \boldsymbol{X}^\top\boldsymbol{X}+\mathbb{I} = \begin{bmatrix} 5 & 2 & 2 \\ 2 & 7 & -4 \\ 2 & -4 & 7 \end{bmatrix}$$

$$\text{eigenvalues : } \quad \{10, 6, 0\} \qquad\qquad \{11, 7, 1\}$$

Ad-hoc strategy: use $\boldsymbol{X}^\top\boldsymbol{X} + \lambda\mathbb{I}$

Ridge Regression

... like the least square, but it shrinks estimated coefficients towards 0.

$$\widehat{\boldsymbol{\beta}}_\lambda^{\textsf{ridge}} = \text{argmin}\left\{\sum_{i=1}^n (y_i - \boldsymbol{x}_i^\top\boldsymbol{\beta})^2 + \lambda\sum_{j=1}^p \beta_j^2\right\}$$

$$\widehat{\boldsymbol{\beta}}_\lambda^{\textsf{ridge}} = \text{argmin}\left\{\underbrace{\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_{\ell_2}^2}_{=\text{criteria}} + \underbrace{\lambda\|\boldsymbol{\beta}\|_{\ell_2}^2}_{=\text{penalty}}\right\}$$

$\lambda \geq 0$ is a tuning parameter.

Ridge Regression

an Wieringen (2018 Lecture notes on ridge regression

Ridge Estimator (OLS)

$$\widehat{\boldsymbol{\beta}}_\lambda^{\text{ridge}} = \text{argmin}\left\{\sum_{i=1}^n (y_i - \boldsymbol{x}_i^\top\boldsymbol{\beta})^2 + \lambda\sum_{j=1}^p \beta_j^2\right\}$$

Ridge Estimator (GLM)

$$\widehat{\boldsymbol{\beta}}_\lambda^{\text{ridge}} = \text{argmin}\left\{-\sum_{i=1}^n \log f(y_i|\mu_i = g^{-1}(\boldsymbol{x}_i^\top\boldsymbol{\beta})) + \frac{\lambda}{2}\sum_{j=1}^p \beta_j^2\right\}$$

Ridge Regression

$$\widehat{\boldsymbol{\beta}}_\lambda^{\text{ridge}} = \text{argmin}\left\{\left\|\boldsymbol{y}-(\beta_0+\boldsymbol{X}\boldsymbol{\beta})\right\|_{\ell_2}^2+\lambda\left\|\boldsymbol{\beta}\right\|_{\ell_2}^2\right\}$$

can be seen as a constrained optimization problem

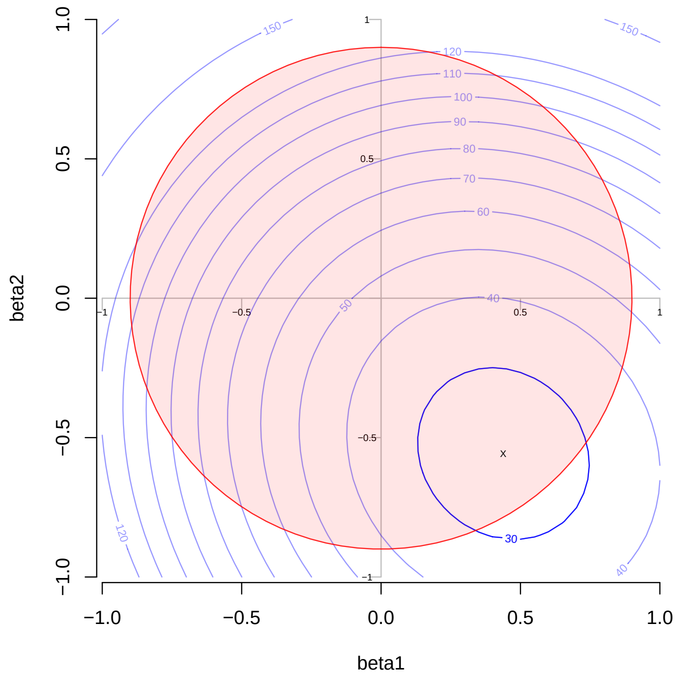
$$\widehat{\boldsymbol{\beta}}_\lambda^{\text{ridge}} = \underset{\|\boldsymbol{\beta}\|_{\ell_2}^2\leq h_\lambda}{\text{argmin}}\left\{\left\|\boldsymbol{y}-(\beta_0+\boldsymbol{X}\boldsymbol{\beta})\right\|_{\ell_2}^2\right\}$$

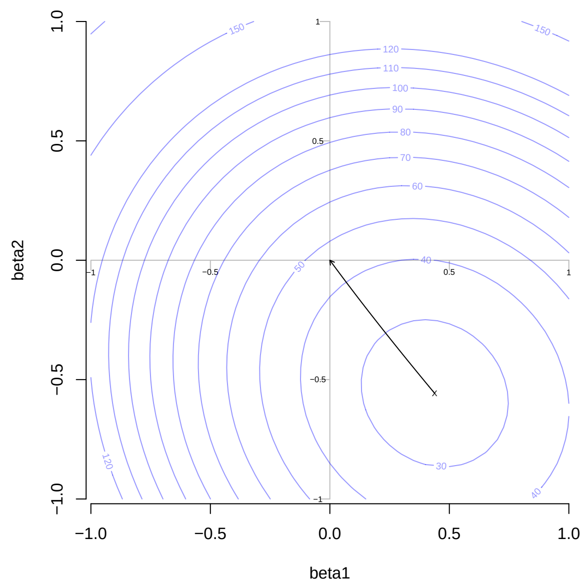
Explicit solution

$$\widehat{\boldsymbol{\beta}}_\lambda = (\boldsymbol{X}^\mathsf{T}\boldsymbol{X}+\lambda\mathbb{I})^{-1}\boldsymbol{X}^\mathsf{T}\boldsymbol{y}$$

If $\lambda\to 0$, $\widehat{\boldsymbol{\beta}}_0^{\text{ridge}} = \widehat{\boldsymbol{\beta}}^{\text{ols}}$

If $\lambda\to\infty$, $\widehat{\boldsymbol{\beta}}_\infty^{\text{ridge}} = \boldsymbol{0}$.





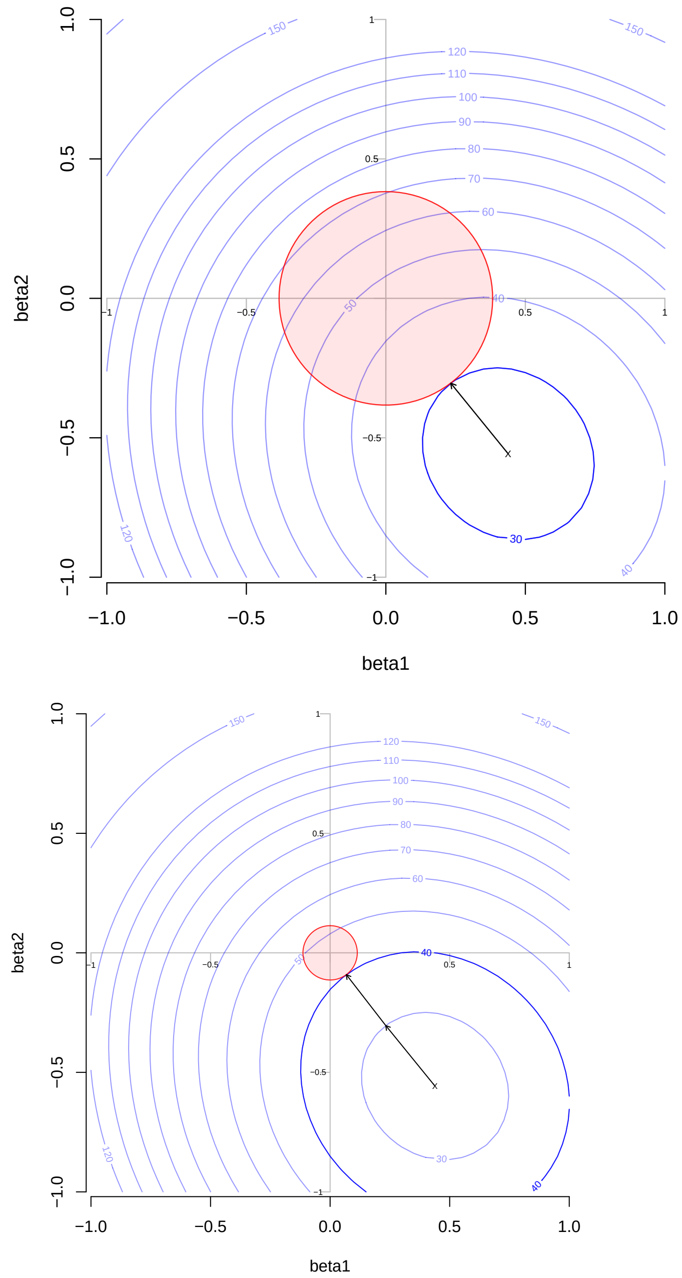
Ridge Regression

This penalty can be seen as rather unfair if components of $\boldsymbol{x}$ are not expressed on the same scale

- center: $\overline{\boldsymbol{x}}_j = 0$, then $\widehat{\beta}_0 = \overline{\boldsymbol{y}}$
- scale: $\boldsymbol{x}_j^\top \boldsymbol{x}_j = 1$

Then compute

$$\widehat{\boldsymbol{\beta}}_\lambda^{\text{ridge}} = \text{argmin}\left\{\underbrace{\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|_{\ell_2}^2}_{=\text{loss}} + \underbrace{\lambda\|\boldsymbol{\beta}\|_{\ell_2}^2}_{=\text{penalty}}\right\}$$



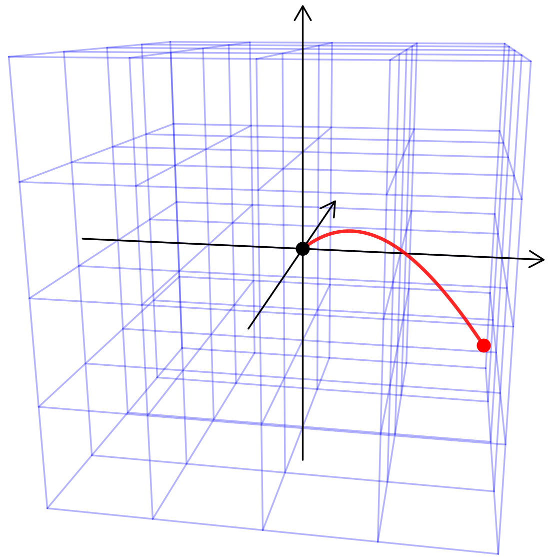
Ridge Regression

Observe that if $\boldsymbol{x}_{j_1} \perp \boldsymbol{x}_{j_2}$, then

$$\widehat{\boldsymbol{\beta}}_{\lambda}^{\text{ridge}} = [1+\lambda]^{-1}\widehat{\boldsymbol{\beta}}_{\lambda}^{\text{ols}}$$

which explain relationship with shrinkage.

But generally, it is not the case...



Smaller mse

There exists λ such that $\text{mse}[\widehat{\boldsymbol{\beta}}_{\lambda}^{\text{ridge}}] \leq \text{mse}[\widehat{\boldsymbol{\beta}}_{\lambda}^{\text{ols}}]$

Ridge Regression

$$\mathcal{L}_\lambda(\boldsymbol{\beta})=\sum_{i=1}^n(y_i-\beta_0-\boldsymbol{x}_i^\top\boldsymbol{\beta})^2+\lambda\sum_{j=1}^p\beta_j^2$$

$$\frac{\partial\mathcal{L}_\lambda(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}=-2\boldsymbol{X}^\top\boldsymbol{y}+2(\boldsymbol{X}^\top\boldsymbol{X}+\lambda\mathbb{I})\boldsymbol{\beta}$$

$$\frac{\partial^2\mathcal{L}_\lambda(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^\top}=2(\boldsymbol{X}^\top\boldsymbol{X}+\lambda\mathbb{I})$$

where $\boldsymbol{X}^\top\boldsymbol{X}$ is a semi-positive definite matrix, and $\lambda\mathbb{I}$ is a positive definite matrix, and

$$\widehat{\boldsymbol{\beta}}_\lambda=(\boldsymbol{X}^\top\boldsymbol{X}+\lambda\mathbb{I})^{-1}\boldsymbol{X}^\top\boldsymbol{y}$$

The Bayesian Interpretation

From a Bayesian perspective,

$$\underbrace{\mathbb{P}[\boldsymbol{\theta}|\boldsymbol{y}]}_{\text{posterior}} \propto \underbrace{\mathbb{P}[\boldsymbol{y}|\boldsymbol{\theta}]}_{\text{likelihood}} \cdot \underbrace{\mathbb{P}[\boldsymbol{\theta}]}_{\text{prior}} \quad \text{i.e.} \quad \log \mathbb{P}[\boldsymbol{\theta}|\boldsymbol{y}] = \underbrace{\log \mathbb{P}[\boldsymbol{y}|\boldsymbol{\theta}]}_{\text{log likelihood}} + \underbrace{\log \mathbb{P}[\boldsymbol{\theta}]}_{\text{penalty}}$$

If $\boldsymbol{\beta}$ has a prior $\mathcal{N}(\mathbf{0}, \tau^2\mathbb{I})$ distribution, then its posterior distribution has mean

$$\mathbb{E}[\boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{X}] = \left(\boldsymbol{X}^\top\boldsymbol{X} + \frac{\sigma^2}{\tau^2}\mathbb{I}\right)^{-1}\boldsymbol{X}^\top\boldsymbol{y}.$$

Properties of the Ridge Estimator

$$\widehat{\boldsymbol{\beta}}_\lambda = (\boldsymbol{X}^\top\boldsymbol{X} + \lambda\mathbb{I})^{-1}\boldsymbol{X}^\top\boldsymbol{y}$$

$$\mathbb{E}[\widehat{\boldsymbol{\beta}}_\lambda] = \boldsymbol{X}^\top\boldsymbol{X}(\lambda\mathbb{I} + \boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{\beta}.$$

i.e. $\mathbb{E}[\widehat{\boldsymbol{\beta}}_\lambda] \neq \boldsymbol{\beta}$.

Observe that $\mathbb{E}[\widehat{\boldsymbol{\beta}}_\lambda] \to \mathbf{0}$ as $\lambda \to \infty$.

Ridge & Shrinkage

Assume that $\boldsymbol{X}$ is an orthogonal design matrix, i.e. $\boldsymbol{X}^\top\boldsymbol{X} = \mathbb{I}$, then

$$\widehat{\boldsymbol{\beta}}_\lambda = (1+\lambda)^{-1}\widehat{\boldsymbol{\beta}}^{\mathsf{ols}}.$$

Properties of the Ridge Estimator

Set $\boldsymbol{W}_\lambda = (\mathbb{I} + \lambda[\boldsymbol{X}^\mathsf{T}\boldsymbol{X}]^{-1})^{-1}$. One can prove that

$$\boldsymbol{W}_\lambda \widehat{\boldsymbol{\beta}}^{\mathsf{ols}} = \widehat{\boldsymbol{\beta}}_\lambda.$$

Thus,

$$\text{Var}[\widehat{\boldsymbol{\beta}}_\lambda] = \boldsymbol{W}_\lambda \text{Var}[\widehat{\boldsymbol{\beta}}^{\mathsf{ols}}]\boldsymbol{W}_\lambda^\mathsf{T}$$

and

$$\text{Var}[\widehat{\boldsymbol{\beta}}_\lambda] = \sigma^2(\boldsymbol{X}^\mathsf{T}\boldsymbol{X} + \lambda\mathbb{I})^{-1}\boldsymbol{X}^\mathsf{T}\boldsymbol{X}[(\boldsymbol{X}^\mathsf{T}\boldsymbol{X} + \lambda\mathbb{I})^{-1}]^\mathsf{T}.$$

Observe that

$$\text{Var}[\widehat{\boldsymbol{\beta}}^{\mathsf{ols}}] - \text{Var}[\widehat{\boldsymbol{\beta}}_\lambda] = \sigma^2\boldsymbol{W}_\lambda[2\lambda(\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-2} + \lambda^2(\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-3}]\boldsymbol{W}_\lambda^\mathsf{T} \geq \boldsymbol{0}.$$

Properties of the Ridge Estimator

Hence, the confidence ellipsoid of ridge estimator is indeed smaller than the OLS,
If $\boldsymbol{X}$ is an orthogonal design matrix,

$$\text{Var}[\widehat{\boldsymbol{\beta}}_\lambda] = \sigma^2(1+\lambda)^{-2}\mathbb{I}.$$

$$\text{mse}[\widehat{\boldsymbol{\beta}}_\lambda] = \sigma^2\text{trace}(\boldsymbol{W}_\lambda(\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1}\boldsymbol{W}_\lambda^\mathsf{T}) + \boldsymbol{\beta}^\mathsf{T}(\boldsymbol{W}_\lambda - \mathbb{I})^\mathsf{T}(\boldsymbol{W}_\lambda - \mathbb{I})\boldsymbol{\beta}.$$

If $\boldsymbol{X}$ is an orthogonal design matrix,

$$\text{mse}[\widehat{\boldsymbol{\beta}}_\lambda] = \frac{p\sigma^2}{(1+\lambda)^2} + \frac{\lambda^2}{(1+\lambda)^2}\boldsymbol{\beta}^\mathsf{T}\boldsymbol{\beta}$$

Properties of the Ridge Estimator

$$\text{mse}[\widehat{\boldsymbol{\beta}}_\lambda] = \frac{p\sigma^2}{(1+\lambda)^2} + \frac{\lambda^2}{(1+\lambda)^2}\boldsymbol{\beta}^\top\boldsymbol{\beta}$$

is minimal for

$$\lambda^\star = \frac{p\sigma^2}{\boldsymbol{\beta}^\top\boldsymbol{\beta}}$$

Note that there exists $\lambda > 0$ such that $\text{mse}[\widehat{\boldsymbol{\beta}}_\lambda] < \text{mse}[\widehat{\boldsymbol{\beta}}_0] = \text{mse}[\widehat{\boldsymbol{\beta}}^{\text{ols}}]$.

Ridge regression is obtained using `glmnet::glmnet(..., alpha = 0)` - and `glmnet::cv.glmnet` for cross validation

SVD decomposition

For any matrix A, $m \times n$, there are orthogonal matrices U $(m \times m)$, V $(n \times n)$ and a "diagonal" matrix Σ $(m \times n)$ such that $A = U\Sigma V^\top$, or $AV = U\Sigma$.

Hence, there exists a special orthonormal set of vectors (i.e. the columns of V), that is mapped by the matrix A into an orthonormal set of vectors (i.e. the columns of U).

Let $r = \text{rank}(A)$, then $A = \displaystyle\sum_{i=1}^{r} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^\top$ (called the dyadic decomposition of A).

Observe that it can be used to compute (e.g.) the Frobenius norm of A, $\|A\| = \sum a_{i,j}^2 = \sqrt{\sigma_1^2 + \cdots + \sigma_{\min\{m,n\}}^2}$.

Further $A^\top A = V\Sigma^\top\Sigma V^\top$ while $AA^\top = U\Sigma\Sigma^\top U^\top$.

Hence, σ_i^2's are related to eigenvalues of $A^\top A$ and $AA^\top$, and $\boldsymbol{u}_i, \boldsymbol{v}_i$ are associated eigenvectors.

Golub & Reinsh (1970, Singular Value Decomposition and Least Squares Solutions)

SVD decomposition

Consider the singular value decomposition of $\boldsymbol{X}$, $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top$.

Then

$$\widehat{\beta}^{\mathsf{ols}} = \boldsymbol{V}\underbrace{\boldsymbol{D}^{-2}\boldsymbol{D}}\boldsymbol{U}^\top\boldsymbol{y}$$

$$\widehat{\beta}_\lambda = \boldsymbol{V}\underbrace{(\boldsymbol{D}^2 + \lambda\mathbb{I})^{-1}\boldsymbol{D}}\boldsymbol{U}^\top\boldsymbol{y}$$

Observe that

$$\boldsymbol{D}_{i,i}^{-1} \geq \frac{\boldsymbol{D}_{i,i}}{\boldsymbol{D}_{i,i}^2 + \lambda}$$

hence, the ridge penalty shrinks singular values.

Set now $\boldsymbol{R} = \boldsymbol{U}\boldsymbol{D}$ ($n \times n$ matrix), so that $\boldsymbol{X} = \boldsymbol{R}\boldsymbol{V}^\top$,

$$\widehat{\beta}_\lambda = \boldsymbol{V}(\boldsymbol{R}^\top\boldsymbol{R} + \lambda\mathbb{I})^{-1}\boldsymbol{R}^\top\boldsymbol{y}$$

Hat matrix and Degrees of Freedom

Recall that $\widehat{\boldsymbol{Y}} = \boldsymbol{H}\boldsymbol{Y}$ with

$$\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top$$

Similarly

$$\boldsymbol{H}_\lambda = \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X} + \lambda\mathbb{I})^{-1}\boldsymbol{X}^\top$$

$$\text{trace}[\boldsymbol{H}_\lambda] = \sum_{j=1}^{p} \frac{d_{j,j}^2}{d_{j,j}^2 + \lambda} \rightarrow 0, \text{ as } \lambda \rightarrow \infty.$$

Sparsity Issues

In several applications, k can be (very) large, but a lot of features are just noise: $\beta_j = 0$ for many j's. Let s denote the number of relevant features, with $s << k$, cf Hastie, Tibshirani & Wainwright (2015, Statistical Learning with Sparsity),

$$s = \text{card}\{\mathcal{S}\} \text{ where } \mathcal{S} = \{j; \beta_j \neq 0\}$$

The model is now $y = \boldsymbol{X}_{\mathcal{S}}^\top \boldsymbol{\beta}_{\mathcal{S}} + \varepsilon$, where $\boldsymbol{X}_{\mathcal{S}}^\top \boldsymbol{X}_{\mathcal{S}}$ is a full rank matrix.

Going further on sparsity issues

The Ridge regression problem was to solve

$$\widehat{\beta} = \underset{\boldsymbol{\beta}\in\{\|\boldsymbol{\beta}\|_{\ell_2}\leq s\}}{\text{argmin}}\{\|\boldsymbol{Y}-\boldsymbol{X}^\top\beta\|_{\ell_2}^2\}$$

Define $\|\boldsymbol{a}\|_{\ell_0} = \sum \mathbf{1}(|a_i|>0)$.
Here $\dim(\boldsymbol{\beta})=k$ but $\|\boldsymbol{\beta}\|_{\ell_0}=s$.
We wish we could solve

$$\widehat{\beta} = \underset{\boldsymbol{\beta}\in\{\|\boldsymbol{\beta}\|_{\ell_0}= s\}}{\text{argmin}}\{\|\boldsymbol{Y}-\boldsymbol{X}^\top\beta\|_{\ell_2}^2\}$$

Problem: it is usually not possible to describe all possible constraints, since $\binom{s}{k}$ coefficients should be chosen here (with k (very) large).

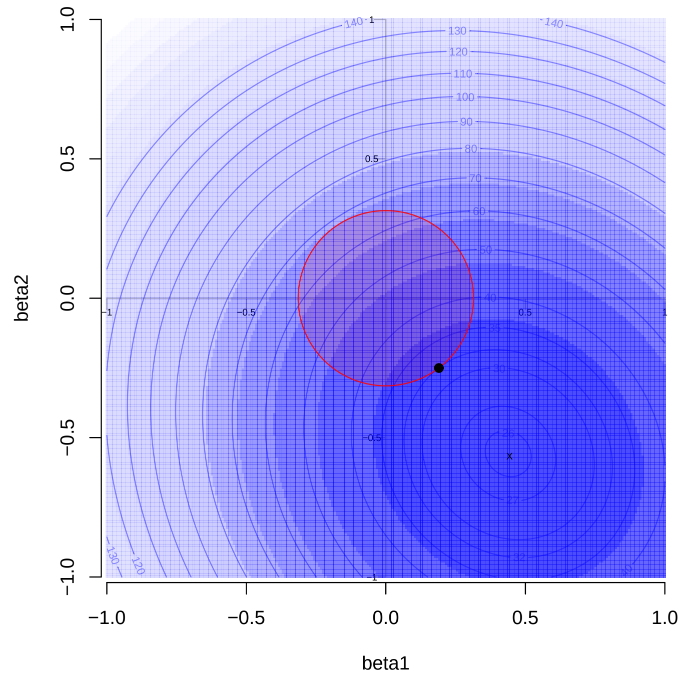
Going further on sparsity issues

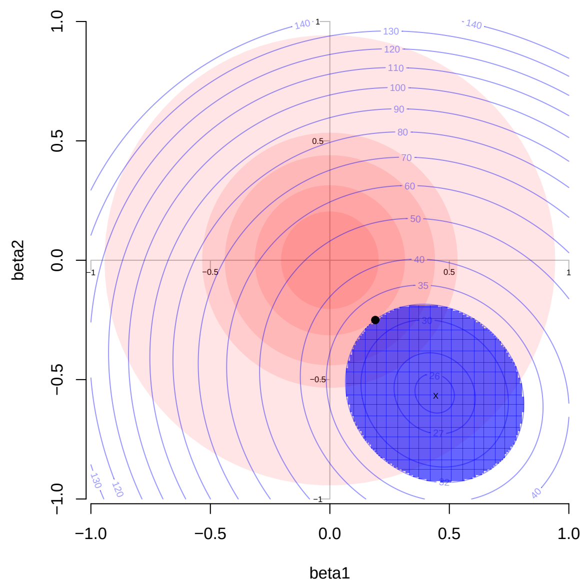
In a convex problem, solve the dual problem,
e.g. in the Ridge regression : primal problem

$$\min_{\boldsymbol{\beta}\in\{\|\boldsymbol{\beta}\|_{\ell_2}\leq s\}}\{\|\boldsymbol{Y}-\boldsymbol{X}^{\mathsf{T}}\boldsymbol{\beta}\|_{\ell_2}^2\}$$

and the dual problem

$$\min_{\boldsymbol{\beta}\in\{\|\boldsymbol{Y}-\boldsymbol{X}^{\mathsf{T}}\boldsymbol{\beta}\|_{\ell_2}\leq t\}}\{\|\boldsymbol{\beta}\|_{\ell_2}^2\}$$



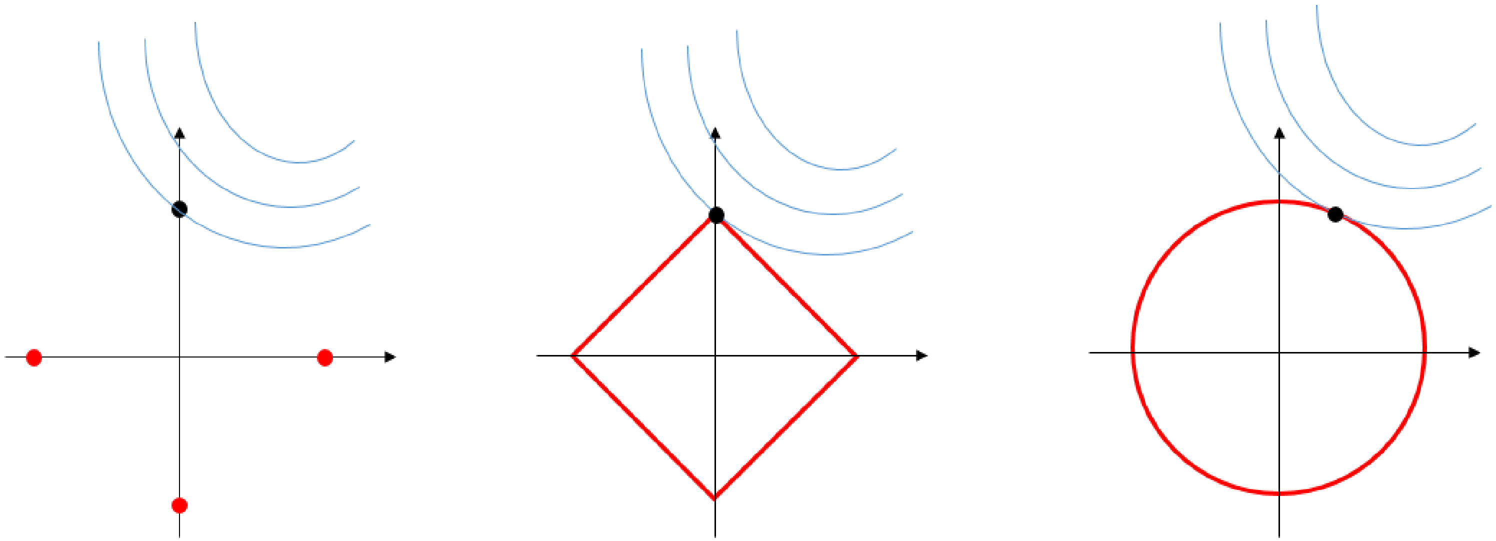


Going further on sparsity issues

Idea: solve the dual problem

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}\in\{\|\boldsymbol{Y}-\boldsymbol{X}^\top\boldsymbol{\beta}\|_{\ell_2}\leq h\}}{\text{argmin}} \{\|\boldsymbol{\beta}\|_{\ell_0}\}$$

where we might convexify the ℓ_0 norm, $\|\cdot\|_{\ell_0}$.



Going further on sparsity issues

On $[-1,+1]^k$, the convex hull of $\|\boldsymbol{\beta}\|_{\ell_0}$ is $\|\boldsymbol{\beta}\|_{\ell_1}$

On $[-a,+a]^k$, the convex hull of $\|\boldsymbol{\beta}\|_{\ell_0}$ is $a^{-1}\|\boldsymbol{\beta}\|_{\ell_1}$

Hence, why not solve

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta};\|\boldsymbol{\beta}\|_{\ell_1}\leq\tilde{s}}{\text{argmin}} \{\|\boldsymbol{Y}-\boldsymbol{X}^\top\boldsymbol{\beta}\|_{\ell_2}\}$$

which is equivalent (Kuhn-Tucker theorem) to the Lagragian optimization problem

$$\widehat{\boldsymbol{\beta}} = \text{argmin}\{\|\boldsymbol{Y}-\boldsymbol{X}^\top\boldsymbol{\beta}\|_{\ell_2}^2+\lambda\|\boldsymbol{\beta}\|_{\ell_1}\}$$

LASSO *Least Absolute Shrinkage and Selection Operator*

LASSO Estimator (OLS)

$$\widehat{\boldsymbol{\beta}}_\lambda^{\text{lasso}} = \text{argmin}\left\{\sum_{i=1}^n (y_i - \boldsymbol{x}_i^\top\boldsymbol{\beta})^2 + \lambda\sum_{j=1}^p |\beta_j|\right\}$$

LASSO Estimator (GLM)

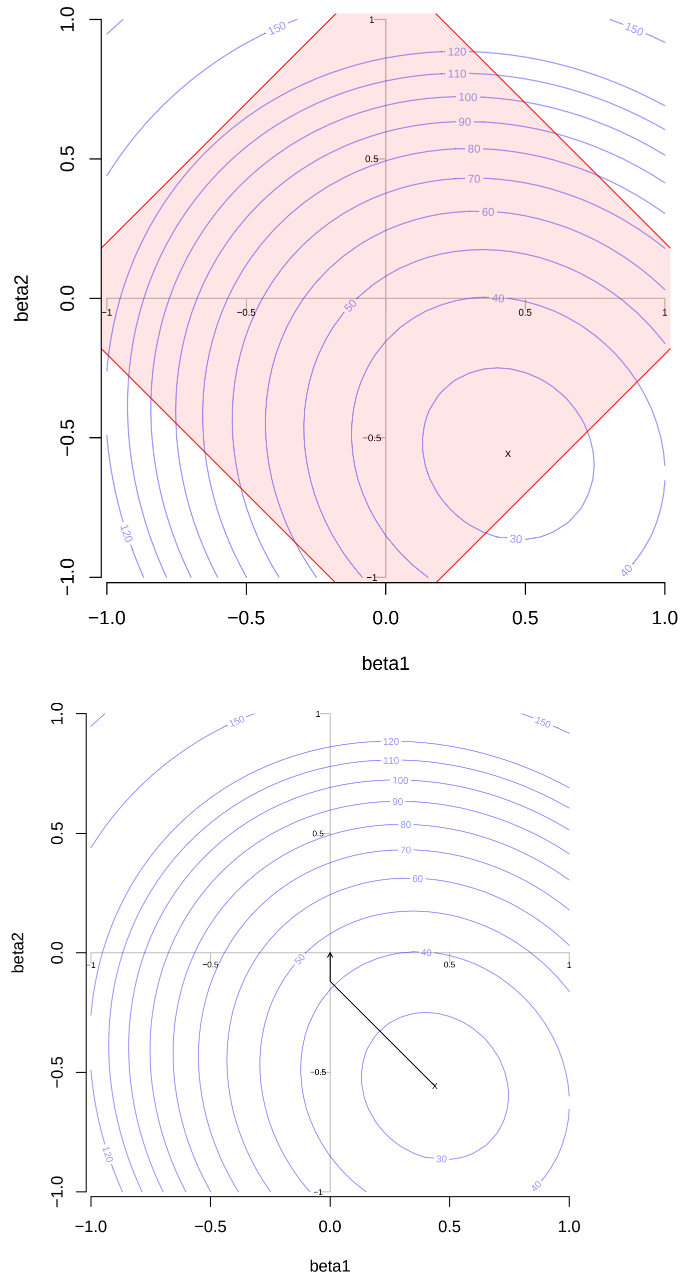
$$\widehat{\boldsymbol{\beta}}_\lambda^{\text{lasso}} = \text{argmin}\left\{-\sum_{i=1}^n \log f(y_i|\mu_i = g^{-1}(\boldsymbol{x}_i^\top\boldsymbol{\beta})) + \frac{\lambda}{2}\sum_{j=1}^p |\beta_j|\right\}$$

LASSO **Regression**

No explicit solution...

If $\lambda \to 0$, $\widehat{\beta}_0^{\text{lasso}} = \widehat{\beta}^{\text{ols}}$

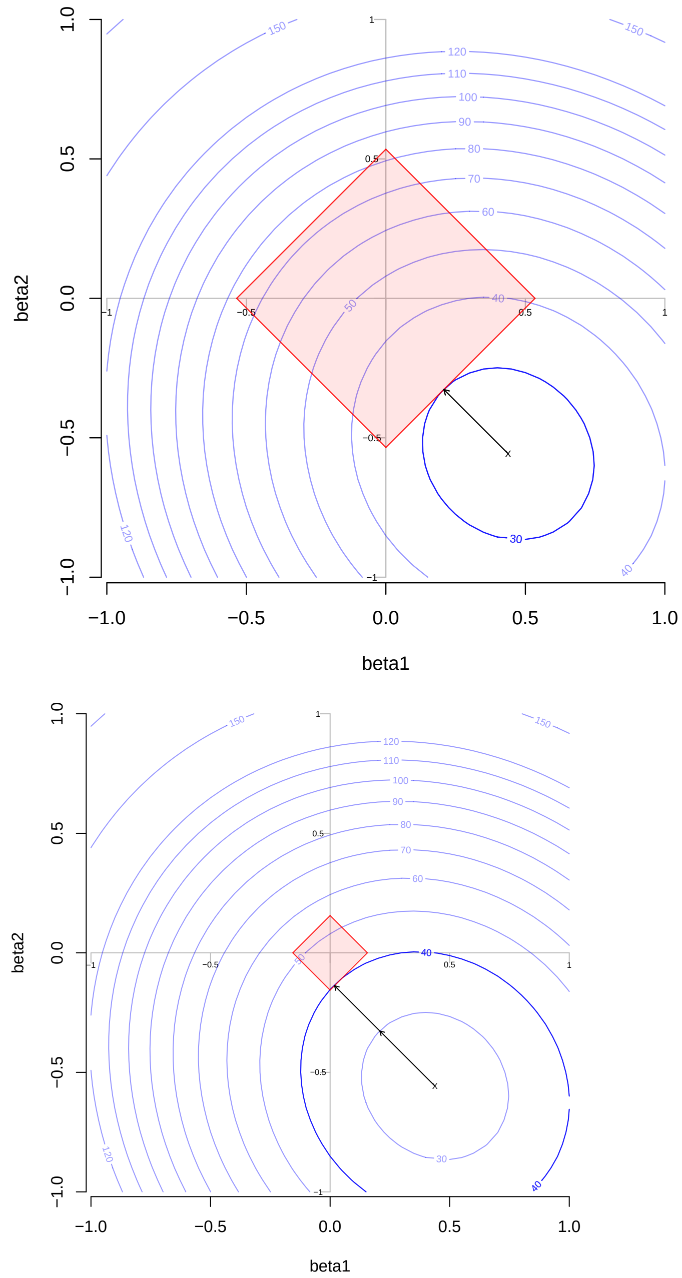
If $\lambda \to \infty$, $\widehat{\beta}_\infty^{\text{lasso}} = \mathbf{0}$.



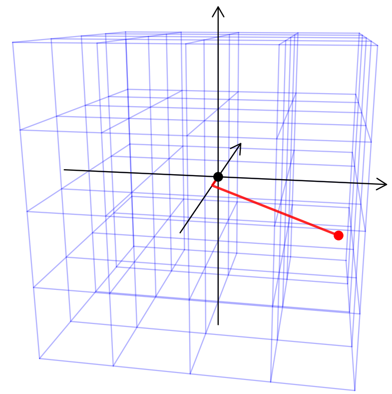
LASSO Regression

For some λ, there are k's such that $\widehat{\boldsymbol{\beta}}^{\text{lasso}}_{k,\lambda} = 0$.

Further, $\lambda \mapsto \widehat{\boldsymbol{\beta}}^{\text{lasso}}_{k,\lambda}$ is piecewise linear



LASSO Regression



In the orthogonal case, $\boldsymbol{X}^\top\boldsymbol{X} = \mathbb{I}$,

$$\widehat{\boldsymbol{\beta}}_{k,\lambda}^{\text{lasso}} = \text{sign}(\widehat{\boldsymbol{\beta}}_{k}^{\text{ols}}) \left(|\widehat{\boldsymbol{\beta}}_{k}^{\text{ols}}| - \frac{\lambda}{2} \right)$$

i.e. the LASSO estimate is related to the soft threshold function...

Optimal LASSO Penalty

Use cross validation, e.g. K-fold,

$$\widehat{\boldsymbol{\beta}}_{(-k)}(\lambda) = \text{argmin}\left\{\sum_{i\notin\mathcal{I}_k}[y_i - \boldsymbol{x}_i^\top\boldsymbol{\beta}]^2 + \lambda\|\boldsymbol{\beta}\|_{\ell_1}\right\}$$

then compute the sum of the squared errors,

$$Q_k(\lambda) = \sum_{i\in\mathcal{I}_k}[y_i - \boldsymbol{x}_i^\top\widehat{\boldsymbol{\beta}}_{(-k)}(\lambda)]^2$$

and finally solve

$$\lambda^\star = \text{argmin}\left\{\overline{Q}(\lambda) = \frac{1}{K}\sum_k Q_k(\lambda)\right\}$$

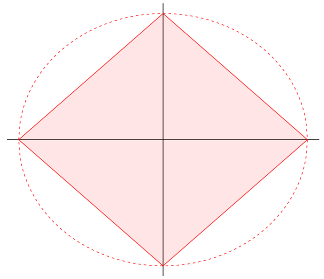
Optimal LASSO Penalty

Note that this might overfit, so Hastie, Tibshiriani & Friedman (2009, Elements of Statistical Learning) suggest the largest λ such that

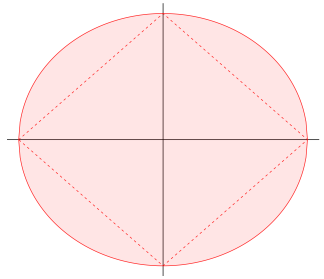
$$\overline{Q}(\lambda) \leq \overline{Q}(\lambda^\star) + \text{se}[\lambda^\star] \text{ with } \text{se}[\lambda]^2 = \frac{1}{K^2}\sum_{k=1}^{K}[Q_k(\lambda) - \overline{Q}(\lambda)]^2$$

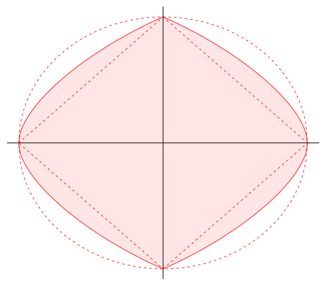
LASSO regression is obtained using `glmnet::glmnet(..., alpha = 1)` - and `glmnet::cv.glmnet` for cross validation.

LASSO and Ridge, with R



```
> library(glmnet)
> chicago=read.table("http://freakonometrics.free.fr/
    chicago.txt",header=TRUE,sep=";")
> standardize <-  function(x)  {(x-mean(x))/sd(x)}
> z0 <- standardize(chicago[, 1])
> z1 <- standardize(chicago[, 3])
> z2 <- standardize(chicago[, 4])
> ridge <-glmnet(cbind(z1, z2), z0, alpha=0, intercept=
    FALSE, lambda=1)
> lasso <-glmnet(cbind(z1, z2), z0, alpha=1, intercept=
    FALSE, lambda=1)
> elastic <-glmnet(cbind(z1, z2), z0, alpha=.5,
    intercept=FALSE, lambda=1)
```





Elastic net, $\lambda_1\|\boldsymbol{\beta}\|_{\ell_1}+\lambda_2\|\boldsymbol{\beta}\|_{\ell_2}^2$

LASSO and LAR (Least-Angle Regression)

LASSO estimation can be seen as an adaptation of LAR procedure

Least Angle Regression

(i) set (small) ϵ

(ii) start with initial residual $\boldsymbol{\varepsilon} = \boldsymbol{y}$, and $\boldsymbol{\beta} = \boldsymbol{0}$

(iii) find the predictor $\boldsymbol{x}_j$ with the highest correlation with $\boldsymbol{\varepsilon}$

(iv) update $\beta_j = \beta_j + \delta_j = \beta_j + \epsilon \cdot \text{sign}[\boldsymbol{\varepsilon}^\top \boldsymbol{x}_j]$

(v) set $\boldsymbol{\varepsilon} = \boldsymbol{\varepsilon} - \delta_j \boldsymbol{x}_j$ and go to (iii)

see Efron *et al.* (2004, Least Angle Regression)

Going further, ℓ_0, ℓ_1 and ℓ_2 penalty

Define

$$\|\boldsymbol{a}\|_{\ell_0}=\sum_{i=1}^d \mathbf{1}(a_i\neq 0),\quad \|\boldsymbol{a}\|_{\ell_1}=\sum_{i=1}^d |a_i|\quad\text{and}\quad \|\boldsymbol{a}\|_{\ell_2}=\left(\sum_{i=1}^d a_i^2\right)^{1/2},\ \text{for } \boldsymbol{a}\in\mathbb{R}^d.$$

constrained optimization	penalized optimization	
$\underset{\boldsymbol{\beta};\|\boldsymbol{\beta}\|_{\ell_0}\leq s}{\text{argmin}}\left\{\sum_{i=1}^n \ell(y_i,\beta_0+\boldsymbol{x}^\top\boldsymbol{\beta})\right\}$	$\underset{\boldsymbol{\beta},\lambda}{\text{argmin}}\left\{\sum_{i=1}^n \ell(y_i,\beta_0+\boldsymbol{x}^\top\boldsymbol{\beta})+\lambda\|\boldsymbol{\beta}\|_{\ell_0}\right\}$	$(\ell 0)$
$\underset{\boldsymbol{\beta};\|\boldsymbol{\beta}\|_{\ell_1}\leq s}{\text{argmin}}\left\{\sum_{i=1}^n \ell(y_i,\beta_0+\boldsymbol{x}^\top\boldsymbol{\beta})\right\}$	$\underset{\boldsymbol{\beta},\lambda}{\text{argmin}}\left\{\sum_{i=1}^n \ell(y_i,\beta_0+\boldsymbol{x}^\top\boldsymbol{\beta})+\lambda\|\boldsymbol{\beta}\|_{\ell_1}\right\}$	$(\ell 1)$
$\underset{\boldsymbol{\beta};\|\boldsymbol{\beta}\|_{\ell_2}\leq s}{\text{argmin}}\left\{\sum_{i=1}^n \ell(y_i,\beta_0+\boldsymbol{x}^\top\boldsymbol{\beta})\right\}$	$\underset{\boldsymbol{\beta},\lambda}{\text{argmin}}\left\{\sum_{i=1}^n \ell(y_i,\beta_0+\boldsymbol{x}^\top\boldsymbol{\beta})+\lambda\|\boldsymbol{\beta}\|_{\ell_2}\right\}$	$(\ell 2)$

Assume that ℓ is the quadratic norm.

Going further, ℓ_0, ℓ_1 and ℓ_2 penalty

The two problems ($\ell 2$) are equivalent : $\forall(\boldsymbol{\beta}^\star, s^\star)$ solution of the left problem, $\exists\lambda^\star$ such that $(\boldsymbol{\beta}^\star, \lambda^\star)$ is solution of the right problem. And conversely.

The two problems ($\ell 1$) are equivalent : $\forall(\boldsymbol{\beta}^\star, s^\star)$ solution of the left problem, $\exists\lambda^\star$ such that $(\boldsymbol{\beta}^\star, \lambda^\star)$ is solution of the right problem. And conversely. Nevertheless, if there is a theoretical equivalence, there might be numerical issues since there is not necessarily unicity of the solution.

The two problems ($\ell 0$) are not equivalent : if $(\boldsymbol{\beta}^\star, \lambda^\star)$ is solution of the right problem, $\exists s^\star$ such that $\boldsymbol{\beta}^\star$ is a solution of the left problem. But the converse is not true.

More generally, consider a ℓ_p norm,

- sparsity is obtained when $p \leq 1$
- convexity is obtained when $p \geq 1$

Going further, ℓ_0, ℓ_1 and ℓ_2 penalty

Foster & George (1994) the risk inflation criterion for multiple regression tried to solve directly the penalized problem of ($\ell 0$).

But it is a complex combinatorial problem in high dimension (Natarajan (1995) sparse approximate solutions to linear systems proved that it was a NP-hard problem)

One can prove that if $\lambda \sim \sigma^2 \log(p)$, alors

$$\mathbb{E}\big([\boldsymbol{x}^\top\widehat{\boldsymbol{\beta}}-\boldsymbol{x}^\top\boldsymbol{\beta}_0]^2\big)\leq\underbrace{\mathbb{E}\big([\boldsymbol{x}_{\mathcal{S}}{}^\top\widehat{\boldsymbol{\beta}}_{\mathcal{S}}-\boldsymbol{x}^\top\boldsymbol{\beta}_0]^2\big)}_{=\sigma^2\#\mathcal{S}}\cdot\big(4\log p+2+o(1)\big).$$

In that case

$$\widehat{\boldsymbol{\beta}}^{\text{sub}}_{\lambda,j}=\begin{cases}0 \text{ si } j\notin\mathcal{S}_\lambda(\boldsymbol{\beta})\\ \widehat{\boldsymbol{\beta}}^{\text{ols}}_j \text{ si } j\in\mathcal{S}_\lambda(\boldsymbol{\beta}),\end{cases}$$

where $\mathcal{S}_\lambda(\boldsymbol{\beta})$ is the set of non-null values in solutions of ($\ell 0$).

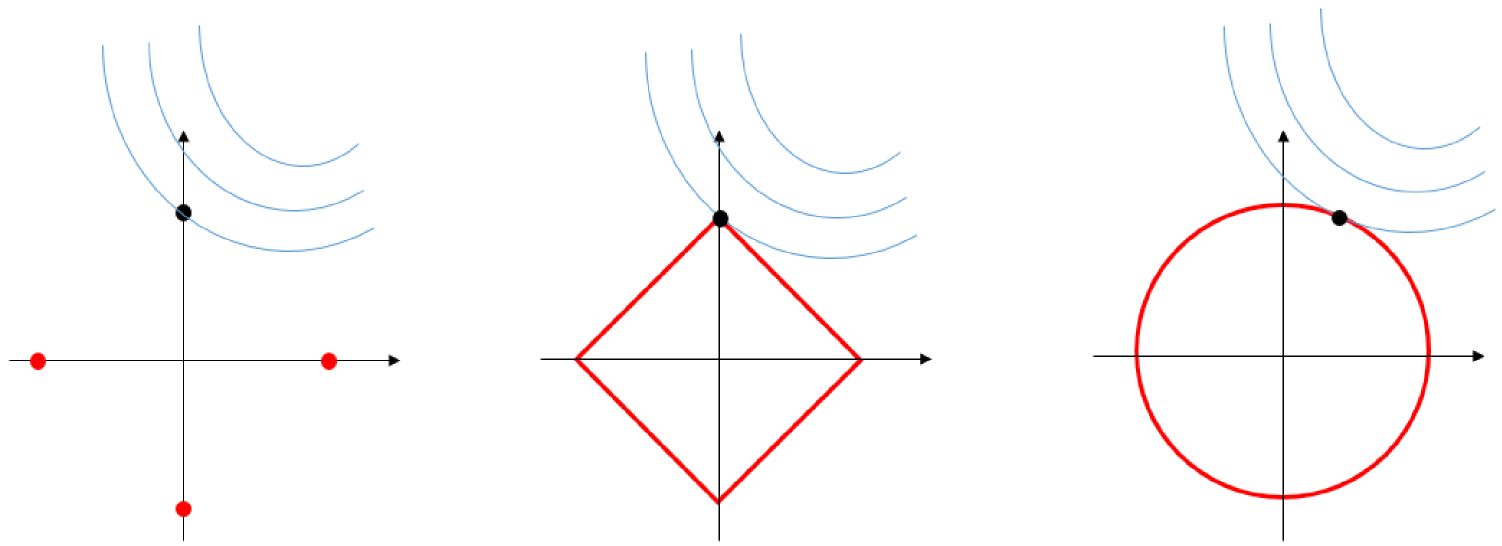
Going further, ℓ_0, ℓ_1 and ℓ_2 penalty

If ℓ is no longer the quadratic norm but ℓ_1, problem (ℓ1) is not always strictly convex, and optimum is not always unique (e.g. if $\boldsymbol{X}^\top\boldsymbol{X}$ is singular).

But in the quadratic case, ℓ is strictly convex, and at least $\boldsymbol{X}\widehat{\boldsymbol{\beta}}$ is unique.

Further, note that solutions are necessarily coherent (signs of coefficients) : it is not possible to have $\widehat{\beta}_j < 0$ for one solution and $\widehat{\beta}_j > 0$ for another one.

In many cases, problem (ℓ1) yields a corner-type solution, which can be seen as a "best subset" solution - like in (ℓ0).



Going further, ℓ_0, ℓ_1 and ℓ_2 penalty

Consider a simple regression $y_i = x_i\beta + \varepsilon$, with ℓ_1-penalty and a ℓ_2-loss fucntion. (ℓ1) becomes

$$\min\left\{\boldsymbol{y}^\top\boldsymbol{y} - 2\boldsymbol{y}^\top\boldsymbol{x}\beta + \beta\boldsymbol{x}^\top\boldsymbol{x}\beta + 2\lambda|\beta|\right\}$$

First order condition can be written

$$-2\boldsymbol{y}^\top\boldsymbol{x} + 2\boldsymbol{x}^\top\boldsymbol{x}\widehat{\beta}\pm 2\lambda = 0.$$

(the sign in $\pm$ being the sign of $\widehat{\beta}$). Assume that least-square estimate ($\lambda = 0$) is (strictly) positive, i.e. $\boldsymbol{y}^\top\boldsymbol{x} > 0$. If λ is not too large $\widehat{\beta}$ and $\widehat{\beta}^{\mathsf{ols}}$ have the same sign, and

$$-2\boldsymbol{y}^\top\boldsymbol{x} + 2\boldsymbol{x}^\top\boldsymbol{x}\widehat{\beta} + 2\lambda = 0.$$

with solution $\widehat{\beta}_\lambda^{\mathsf{lasso}} = \dfrac{\boldsymbol{y}^\top\boldsymbol{x} - \lambda}{\boldsymbol{x}^\top\boldsymbol{x}}$.

Going further, ℓ_0, ℓ_1 and ℓ_2 penalty

Increase λ so that $\widehat{\beta}_\lambda = 0$.

Increase slightly more, $\widehat{\beta}_\lambda$ cannot become negative, because the sign of the first order condition will change, and we should solve

$$-2\boldsymbol{y}^\top\boldsymbol{x} + 2\boldsymbol{x}^\top\boldsymbol{x}\widehat{\beta} - 2\lambda = 0.$$

and solution would be $\widehat{\beta}_\lambda^{\text{lasso}} = \dfrac{\boldsymbol{y}^\top\boldsymbol{x} + \lambda}{\boldsymbol{x}^\top\boldsymbol{x}}$. But that solution is positive (we assumed that $\boldsymbol{y}^\top\boldsymbol{x} > 0$), to we should have $\widehat{\beta}_\lambda < 0$.

Thus, at some point $\widehat{\beta}_\lambda = 0$, which is a corner solution.

In higher dimension, see Tibshirani & Wasserman (2016, a closer look at sparse regression) or Candès & Plan (2009, Near-ideal model selection by ℓ_1 minimization.)

With some additional technical assumption, that LASSO estimator is "sparsistent" in the sense that the support of $\widehat{\boldsymbol{\beta}}_\lambda^{\text{lasso}}$ is the same as $\boldsymbol{\beta}$,

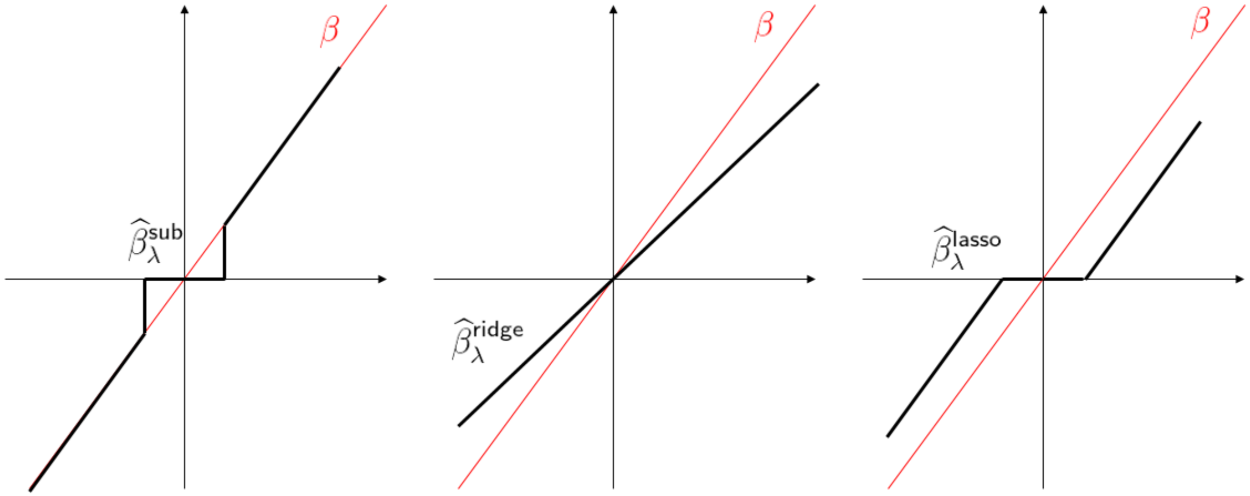
Going further, ℓ_0, ℓ_1 and ℓ_2 penalty

Thus, LASSO can be used for variable selection - see Hastie *et al.* (2001, The Elements of Statistical Learning).

Generally, $\widehat{\beta}_\lambda^{\text{lasso}}$ is a biased estimator but its variance can be small enough to have a smaller least squared error than the OLS estimate.

With orthonormal covariates, one can prove that

$$\widehat{\beta}^{\text{sub}}_{\lambda,j} = \widehat{\beta}^{\text{ols}}_{j}\mathbf{1}_{|\widehat{\beta}^{\text{sub}}_{\lambda,j}|>b},\quad \widehat{\beta}^{\text{ridge}}_{\lambda,j} = \frac{\widehat{\beta}^{\text{ols}}_{j}}{1+\lambda}\quad \text{and}\quad \widehat{\beta}^{\text{lasso}}_{\lambda,j} = \text{signe}[\widehat{\beta}^{\text{ols}}_{j}]\cdot(|\widehat{\beta}^{\text{ols}}_{j}|-\lambda)_{+}.$$



LASSO for Autoregressive Time Series

Consider some AR(p) autoregressive time series,

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_{p-1} X_{t-p+1} + \phi_p X_{t-p} + \varepsilon_t,$$

for some white noise (ε_t), with a causal type representation. Write $y = \boldsymbol{x}^\top \boldsymbol{\phi} + \varepsilon$.

The LASSO estimator $\widehat{\boldsymbol{\phi}}$ is a minimizer of

$$\frac{1}{2T} \| y = \boldsymbol{x}^\top \boldsymbol{\phi} \|^2 + \lambda \sum_{i=1}^{p} \lambda_i |\phi_i|,$$

for some tuning parameters $(\lambda, \lambda_1, \cdots, \lambda_p)$.

See Nardi & Rinaldo (2011, Autoregressive process modeling via the Lasso procedure).

LASSO and Non-Linearities

Consider knots $k_1, \cdots, k_m$, we want a function m which is a cubic polynomial between every pair of knots, continuous at each knot, and with ontinuous first and second derivatives at each knot.

We can write m as

$$m(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 (x - k_1)_+^3 + \cdots + \beta_{m+3}(x - k_m)_+^3$$

One strategy is the following

- fix the number of knots m ($m < n$)
- find the natural cubic spline $\widehat{m}$ which minimizes $\displaystyle\sum_{i=1}^{n}(y_i - m(x_i))^2$
- then choose m by cross validation

and alternative is to use a penalty based approach (Ridge type) to avoid overfit (since with $m = n$, the residual sum of square is null).

GAM, splines and Ridge regression

Consider a univariate nonlinear regression problem, so that $\mathbb{E}[Y|X=x]=m(x)$.

Given a sample $\{(y_1,x_1),\cdots,(y_n,x_n)\}$, consider the following penalized problem

$$m^\star=\underset{m\in\mathcal{C}^2}{\text{argmin}}\left\{\sum_{i=1}^n(y_i-m(x_i))^2+\lambda\int_{\mathbb{R}}m''(x)dx\right\}$$

with the Residual sum of squares on the left, and a penalty for the roughness of the function.

The solution is a natural cubic spline with knots at unique values of x (see Eubanks (1999, Nonparametric Regression and Spline Smoothing)

Consider some spline basis $\{h_1,\cdots,h_n\}$, and let $m(x)=\displaystyle\sum_{i=1}^n\beta_i h_i(x)$.

Let $\boldsymbol{H}$ and $\boldsymbol{\Omega}$ be the $n\times n$ matrices $H_{i,j}=h_j(x_i)$, and $\Omega_{i,j}=\displaystyle\int_{\mathbb{R}}h_i''(x)h_j''(x)dx$.

GAM, splines and Ridge regression

Then the objective function can be written

$$(\boldsymbol{y}-\boldsymbol{H}\boldsymbol{\beta})^\top(\boldsymbol{y}-\boldsymbol{H}\boldsymbol{\beta})+\lambda\boldsymbol{\beta}^\top\boldsymbol{\Omega}\boldsymbol{\beta}$$

Recognize here a generalized Ridge regression, with solution

$$\widehat{\boldsymbol{\beta}}_\lambda=\left(\boldsymbol{H}^\top\boldsymbol{H}+\lambda\Omega\right)^{-1}\boldsymbol{H}^\top\boldsymbol{y}.$$

Note that predicted values are linear functions of the observed value since

$$\widehat{\boldsymbol{y}}=\boldsymbol{H}\left(\boldsymbol{H}^\top\boldsymbol{H}+\lambda\Omega\right)^{-1}\boldsymbol{H}^\top\boldsymbol{y}=\boldsymbol{S}_\lambda\boldsymbol{y},$$

with degrees of freedom trace$(\boldsymbol{S}_\lambda)$.

One can obtain the so-called Reinsch form by considering the singular value decomposition of $\boldsymbol{H}=\boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top$.

GAM, splines and Ridge regression

Here $\boldsymbol{U}$ is orthogonal since $\boldsymbol{H}$ is square ($n \times n$), and $\boldsymbol{D}$ is here invertible. Then

$$\boldsymbol{S}_\lambda = (\mathbb{I} + \lambda \boldsymbol{U}^\top \boldsymbol{D}^{-1} \boldsymbol{V}^\top \boldsymbol{\Omega} \boldsymbol{V} \boldsymbol{D}^{-1} \boldsymbol{U})^{-1} = (\mathbb{I} + \lambda \boldsymbol{K})^{-1}$$

where $\boldsymbol{K}$ is a positive semidefinite matrix, $\boldsymbol{K} = \boldsymbol{B}\boldsymbol{\Delta}\boldsymbol{B}^\top$, where columns of $\boldsymbol{B}$ are know as the Demmler-Reinsch basis.

In that (orthonormal) basis, $\boldsymbol{S}_\lambda$ is a diagonal matrix,

$$\boldsymbol{S}_\lambda = \boldsymbol{B}\big(\mathbb{I} + \lambda \boldsymbol{\Delta}\big)^{-1} \boldsymbol{B}^\top$$

Observe that $\boldsymbol{S}_\lambda \boldsymbol{B}_k = \dfrac{1}{1 + \lambda \Delta_{k,k}} \boldsymbol{B}_k$.

Here again, eigenvalues are shrinkage coefficients of basis vectors.

With more covariates, consider an additive problem

$$(h_1, \cdots, h_p)^\star = \underset{h_1,\cdots,h_p \in \mathcal{C}^2}{\text{argmin}} \left\{ \sum_{i=1}^n \left(y_i - \sum_{j=1}^p m(x_{i,j}) \right)^2 + \lambda \sum_{j=1}^p \int_{\mathbb{R}} m_j''(x) dx \right\}$$

GAM, splines and Ridge regression

which can be written

$$\min\left\{(\boldsymbol{y}-\sum_{j=1}^{p}\boldsymbol{H}_j\boldsymbol{\beta}_j)^\mathsf{T}(\boldsymbol{y}-\sum_{j=1}^{p}\boldsymbol{H}_j\boldsymbol{\beta}_j)+\lambda\big(\boldsymbol{\beta}_1{}^\mathsf{T}\sum_{j=1}^{p}\boldsymbol{\Omega}_j\boldsymbol{\beta}_j\big)\right\}$$

where each matrix $\boldsymbol{H}_j$ is a Demmler-Reinsch basis for variable x_j.

Chouldechova & Hastie (2015, Generalized Additive Model Selection)

Assume that the mean function for the jth variable is $m_j(x)=\alpha_j x+\boldsymbol{m}_j(x)^\mathsf{T}\boldsymbol{\beta}_j$. One can write

$$\min\Big\{(\boldsymbol{y}-\alpha_0-\sum_{j=1}^{p}\alpha_j x_j-\sum_{j=1}^{p}\boldsymbol{H}_j\boldsymbol{\beta}_j)^\mathsf{T}(\boldsymbol{y}-\alpha_0-\sum_{j=1}^{p}\alpha_j x_j-\sum_{j=1}^{p}\boldsymbol{H}_j\boldsymbol{\beta}_j)$$

$$+\lambda\big(\gamma|\alpha_1|+(1-\gamma)\|\boldsymbol{\beta}_j\|_{\Omega_j}\big)+\big(\psi_1\boldsymbol{\beta}_1{}^\mathsf{T}\boldsymbol{\Omega}_1\boldsymbol{\beta}_1+\cdots+\psi_p\boldsymbol{\beta}_p{}^\mathsf{T}\boldsymbol{\Omega}_p\boldsymbol{\beta}_p\big)\Big\}$$

where $\|\boldsymbol{\beta}_j\|_{\Omega_j}=\sqrt{\boldsymbol{\beta}_j{}^\mathsf{T}\boldsymbol{\Omega}_j\boldsymbol{\beta}_j}$.

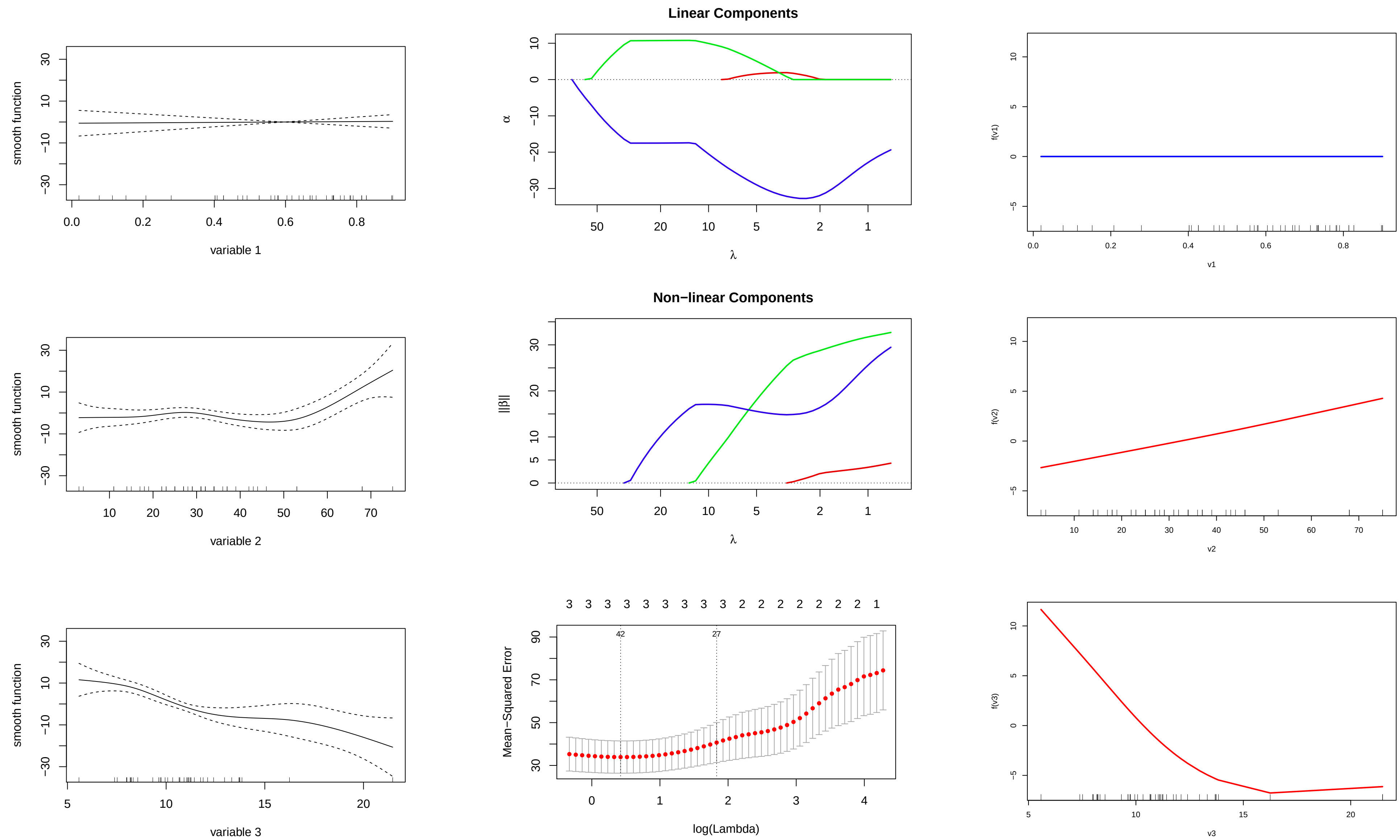
GAM, splines and Ridge regression

The second term is the selection penalty, with a mixture of ℓ_1 and ℓ_2 (type) norm-based penalty

The third term is the end-to-path penalty (GAM type when $\lambda = 0$).

For each predictor x_j, there are three possibilities

- zero, $\alpha_j = 0$ and $\boldsymbol{\beta}_j = \mathbf{0}$
- linear, $\alpha_j \neq 0$ and $\boldsymbol{\beta}_j = \mathbf{0}$
- nonlinear, $\boldsymbol{\beta}_j \neq \mathbf{0}$

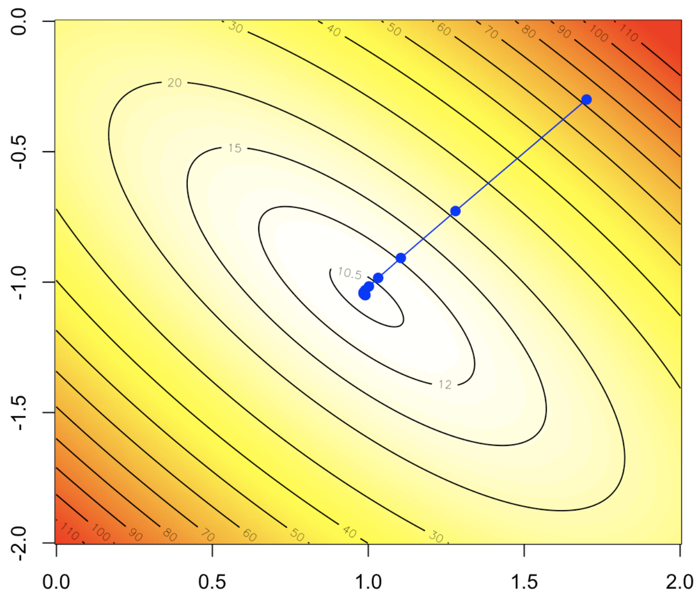


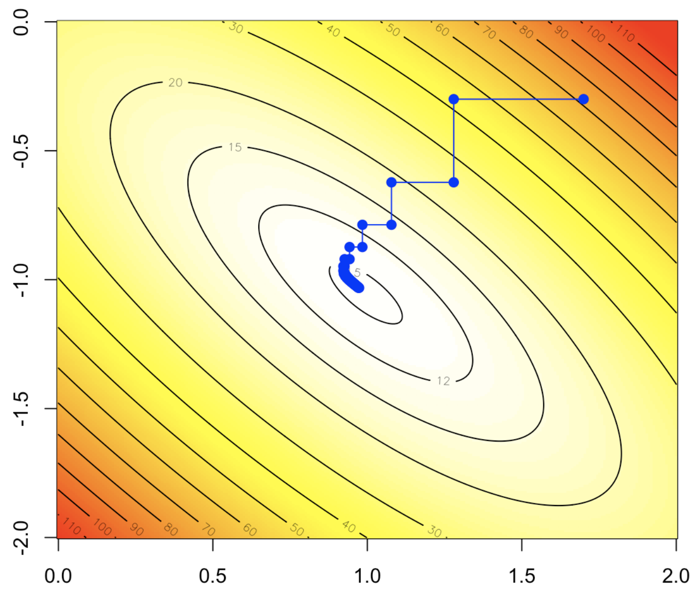
smooth function
variable 1
Linear Components
α
λ
f(v1)
v1
smooth function
variable 2
Non−linear Components
||β||
λ
f(v2)
v2
smooth function
variable 3
Mean−Squared Error
log(Lambda)
f(v3)
v3

Coordinate Descent

LASSO Coordinate Descent Algorithm

1. Set $\boldsymbol{\beta}_0 = \widehat{\boldsymbol{\beta}}$
2 . For $k = 1, \cdots$
   for $j = 1, \cdots, p$
   (i) compute $R_j = \boldsymbol{x}_j^\top \big(\boldsymbol{y} - \boldsymbol{X}_{-j}\boldsymbol{\beta}_{k-1(-j)}\big)$
   (ii) set $\boldsymbol{\beta}_{k,j} = R_j \cdot \left(1 - \dfrac{\lambda}{2|R_j|}\right)_+$
3. The final estimate $\boldsymbol{\beta}_\kappa$ is $\widehat{\boldsymbol{\beta}}_\lambda$





From LASSO to Dantzig Selection

Candès & Tao (2007, The Dantzig selector: Statistical estimation when p is much larger than n) defined

$$\widehat{\boldsymbol{\beta}}_\lambda^{\text{dantzig}} \in \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\text{argmin}}\left\{\|\boldsymbol{\beta}\|_{\ell_1}\right\} \text{ s.t. } \|\boldsymbol{X}^\top(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})\|_{\ell_\infty} \leq \lambda$$

From LASSO to Adaptative Lasso

Zou (2006, The Adaptive Lasso)

$$\widehat{\boldsymbol{\beta}}_{\lambda}^{\text{a-lasso}} \in \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}} \left\{ \|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|_{\ell_2}^2 + \lambda \sum_{j=1}^{p} \frac{|\beta_j|}{|\widehat{\beta}_{\lambda,j}^{\gamma\text{-lasso}}|} \right\}$$

where $\widehat{\boldsymbol{\beta}}_{\lambda}^{\gamma\text{-lasso}} = \Pi_{\boldsymbol{X}_{s(\lambda)}}\boldsymbol{y}$ where $s(\lambda)$ is the set of non null components $\widehat{\boldsymbol{\beta}}_{\lambda}^{\text{lasso}}$

See library `lqa` or `lassogrp`

From LASSO to Group Lasso

Assume that variables $\boldsymbol{x} \in \mathbb{R}^p$ can be grouped in L subgroups, $\boldsymbol{x} = (\boldsymbol{x}_1 \cdots, \boldsymbol{x}_L)$, where $\dim[\boldsymbol{x}_l] = p_l$.

Yuan & Lin (2007, Model selection and estimation in the Gaussian graphical model) defined, for some K_l matrices $n_l \times n_l$ definite positives

$$\widehat{\boldsymbol{\beta}}_\lambda^{\text{g-lasso}} \in \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\text{argmin}} \left\{ \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_{\ell_2}^2 + \lambda \sum_{l=1}^{L} \sqrt{\boldsymbol{\beta}_l^\top K_l \boldsymbol{\beta}_l} \right\}$$

or, if $K_l = p_l \mathbb{I}$

$$\widehat{\boldsymbol{\beta}}_\lambda^{\text{g-lasso}} \in \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\text{argmin}} \left\{ \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_{\ell_2}^2 + \lambda \sum_{l=1}^{L} p_l \|\boldsymbol{\beta}_l\|_{\ell_2} \right\}$$

See library `gglasso`

From LASSO to Sparse-Group Lasso

Assume that variables $\boldsymbol{x} \in \mathbb{R}^p$ can be grouped in L subgroups, $\boldsymbol{x} = (\boldsymbol{x}_1 \cdots, \boldsymbol{x}_L)$, where $\dim[\boldsymbol{x}_l] = p_l$.

Simon *et al.* (2013, A Sparse-Group LASSO) defined, for some K_l matrices $n_l \times n_l$ definite positives

$$\widehat{\boldsymbol{\beta}}^{\text{sg-lasso}}_{\lambda,\mu} \in \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\text{argmin}} \left\{ \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_{\ell_2}^2 + \lambda \sum_{l=1}^{L} \sqrt{\boldsymbol{\beta}_l^\top K_l \boldsymbol{\beta}_l} + \mu \|\boldsymbol{\beta}\|_{\ell_1} \right\}$$

See library `SGL`