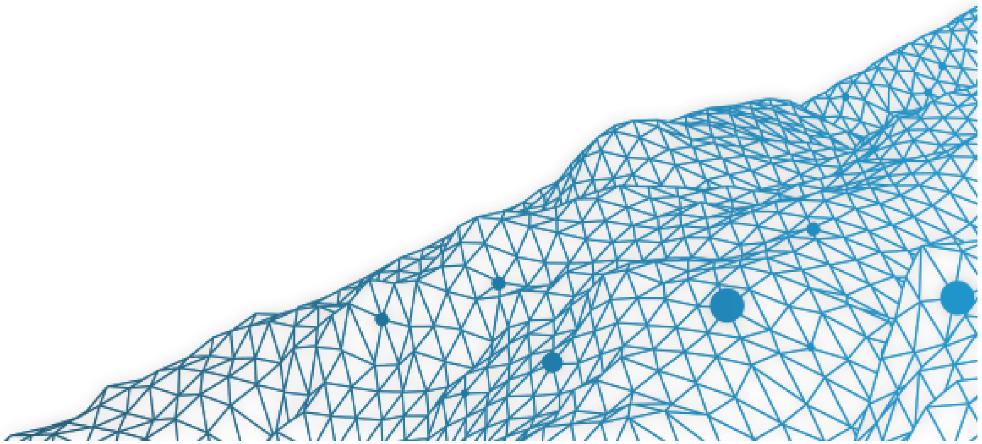
Advanced Econometrics #3: Model & Variable Selection

A. Charpentier (Université de Rennes 1)

Université de Rennes 1,

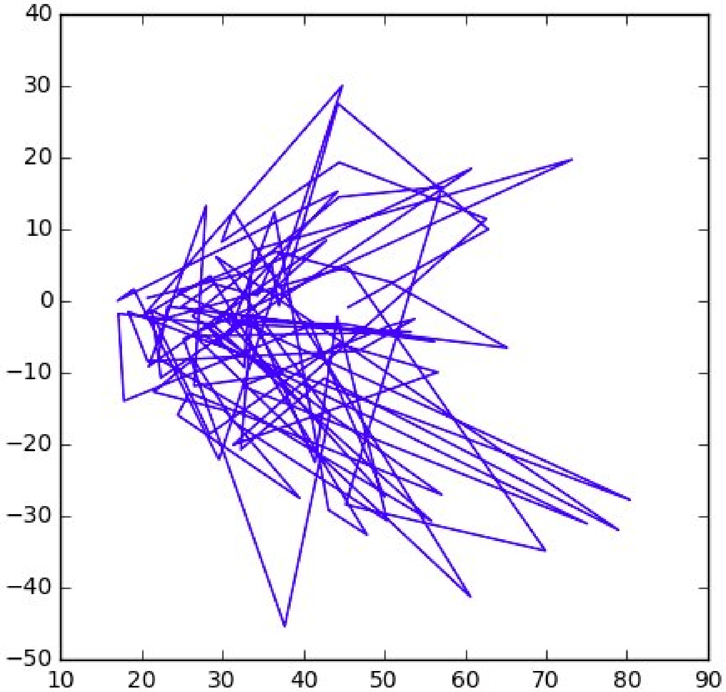
Graduate Course, 2017.



"Great plot.

Now need to find the theory that explains it"

Deville (2017) http://twitter.com



Preliminary Results: Numerical Optimization

Problem : $\boldsymbol{x}^\star \in \text{argmin}\{f(\boldsymbol{x});\ \boldsymbol{x} \in \mathbb{R}^d\}$

Gradient descent : $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \eta\nabla f(\boldsymbol{x}_k)$ starting from some $\boldsymbol{x}_0$

Problem : $\boldsymbol{x}^\star \in \text{argmin}\{f(\boldsymbol{x});\ \boldsymbol{x} \in \mathcal{X} \subset \mathbb{R}^d\}$

Projected descent : $\boldsymbol{x}_{k+1} = \Pi_{\mathcal{X}}\big(\boldsymbol{x}_k - \eta\nabla f(\boldsymbol{x}_k)\big)$ starting from some $\boldsymbol{x}_0$

A constrained problem is said to be convex if

$$\begin{cases} \min\{f(\boldsymbol{x})\} & \text{with } f \text{ convex} \\ \text{s.t. } g_i(\boldsymbol{x}) = 0,\ \forall i = 1, \cdots, n & \text{with } g_i \text{ linear} \\ \qquad h_i(\boldsymbol{x}) \leq 0,\ \forall i = 1, \cdots, m & \text{with } h_i \text{ convex} \end{cases}$$

Lagrangian : $\mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = f(\boldsymbol{x}) + \sum_{i=1}^n \lambda_i g_i(\boldsymbol{x}) + \sum_{i=1}^m \mu_i h_i(\boldsymbol{x})$ where $\boldsymbol{x}$ are primal variables and $(\boldsymbol{\lambda}, \boldsymbol{\mu})$ are dual variables.

Remark $\mathcal{L}$ is an affine function in $(\boldsymbol{\lambda}, \boldsymbol{\mu})$

Preliminary Results: Numerical Optimization

Karush–Kuhn–Tucker conditions : a convex problem has a solution $\boldsymbol{x}^\star$ if and only if there are $(\boldsymbol{\lambda}^\star, \boldsymbol{\mu}^\star)$ such that the following condition hold

- stationarity : $\nabla_{\boldsymbol{x}}\mathcal{L}(\boldsymbol{x},\boldsymbol{\lambda},\boldsymbol{\mu}) = \boldsymbol{0}$ at $(\boldsymbol{x}^\star,\boldsymbol{\lambda}^\star,\boldsymbol{\mu}^\star)$
- primal admissibility : $g_i(\boldsymbol{x}^\star) = 0$ and $h_i(\boldsymbol{x}^\star) \leq 0$, $\forall i$
- dual admissibility : $\boldsymbol{\mu}^\star \geq 0$

Let L denote the associated dual function $L(\boldsymbol{\lambda},\boldsymbol{\mu}) = \min_{\boldsymbol{x}}\{\mathcal{L}(\boldsymbol{x},\boldsymbol{\lambda},\boldsymbol{\mu})\}$

L is a convex function in $(\boldsymbol{\lambda},\boldsymbol{\mu})$ and the dual problem is $\max_{\boldsymbol{\lambda},\boldsymbol{\mu}}\{L(\boldsymbol{\lambda},\boldsymbol{\mu})\}$.

References

Motivation

Banerjee, A., Chandrasekhar, A.G., Duflo, E. & Jackson, M.O. (2016). Gossip: Identifying Central Individuals in a Social Networks.

References

Belloni, A. & Chernozhukov, V. 2009. Least squares after model selection in high-dimensional sparse models.

Hastie, T., Tibshirani, R. & Wainwright, M. 2015 Statistical Learning with Sparsity: The Lasso and Generalizations. CRC Press.

Preambule

Assume that $y = m(\boldsymbol{x}) + \varepsilon$, where ε is some idosyncatic impredictible noise.

The error $\mathbb{E}[(y - m(\boldsymbol{x}))^2]$ is the sume of three terms

- variance of the estimator : $\mathbb{E}[(y - \widehat{m}(\boldsymbol{x}))^2]$
- bias^2 of the estimator : $[m(\boldsymbol{x} - \widehat{m}(\boldsymbol{x})]^2$
- variance of the noise : $\mathbb{E}[(y - m(\boldsymbol{x}))^2]$

(the latter exists, even with a 'perfect' model).

Preambule

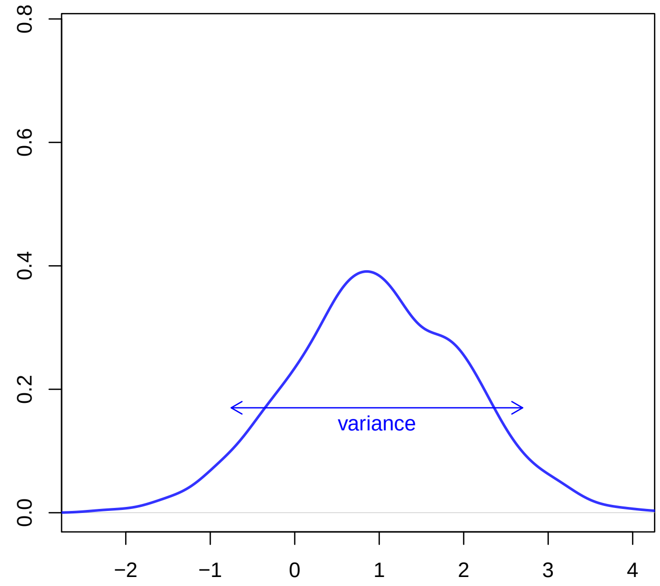
Consider a parametric model, with true (unkown) parameter θ, then

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

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

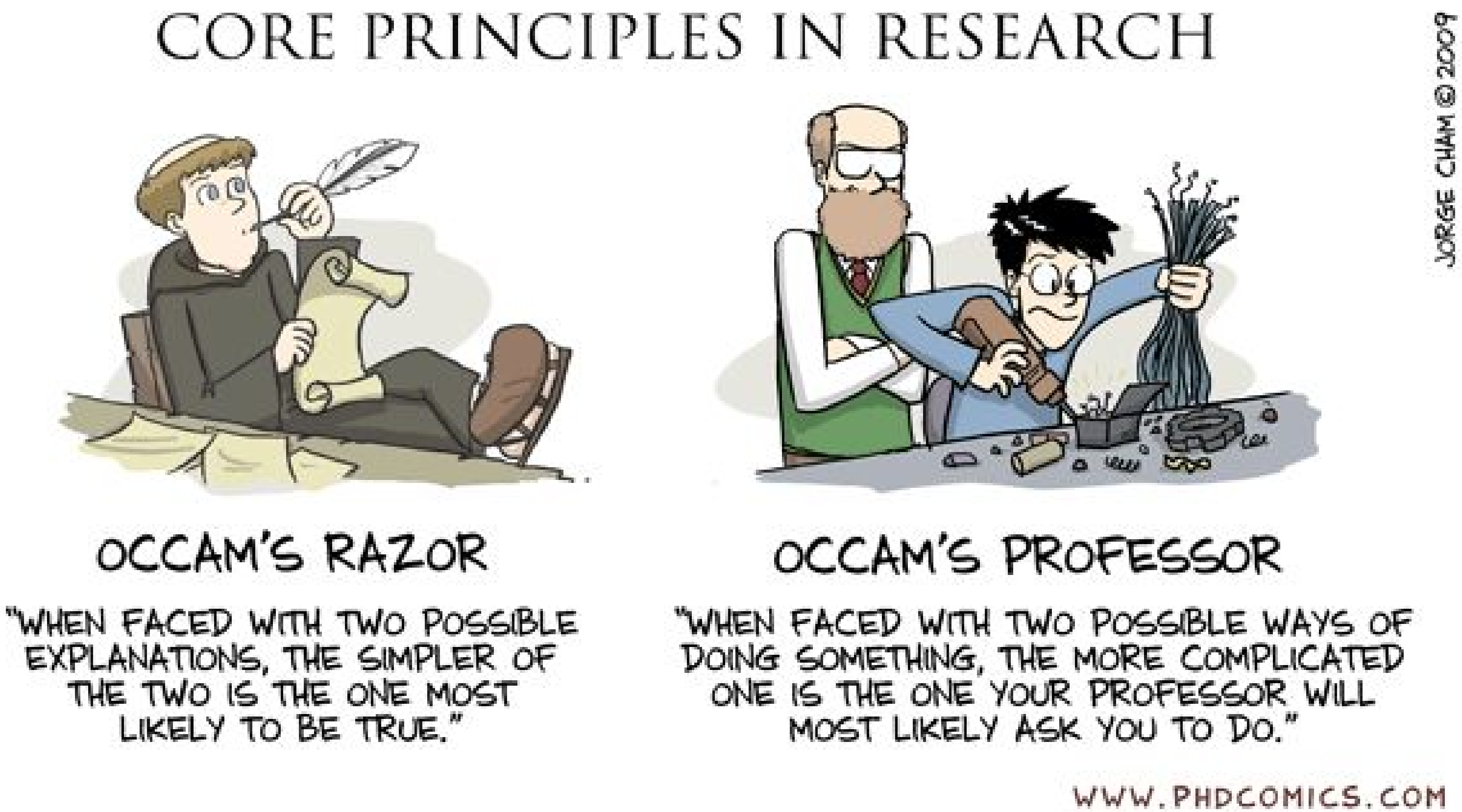
$$\hat\theta = \frac{\theta^2}{\theta^2+\text{mse}(\widetilde{\theta})}\cdot\widetilde{\theta} = \widetilde{\theta} - \underbrace{\frac{\text{mse}(\widetilde{\theta})}{\theta^2+\text{mse}(\widetilde{\theta})}\cdot\widetilde{\theta}}_{\text{penalty}}$$

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



Occam's Razor

The "law of parsimony", "*lex parsimoniæ*"



Penalize too complex models

James & Stein Estimator

Let $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \sigma^2\mathbb{I})$. We want to estimate $\boldsymbol{\mu}$.

$$\widehat{\boldsymbol{\mu}}_{\text{mle}} = \overline{X}_n \sim \mathcal{N}\left(\boldsymbol{\mu}, \frac{\sigma^2}{n}\mathbb{I}\right).$$

From James & Stein (1961) Estimation with quadratic loss

$$\widehat{\boldsymbol{\mu}}_{\text{JS}} = \left(1 - \frac{(d-2)\sigma^2}{n\|\overline{\boldsymbol{y}}\|^2}\right)\overline{\boldsymbol{y}}$$

where $\|\cdot\|$ is the Euclidean norm.

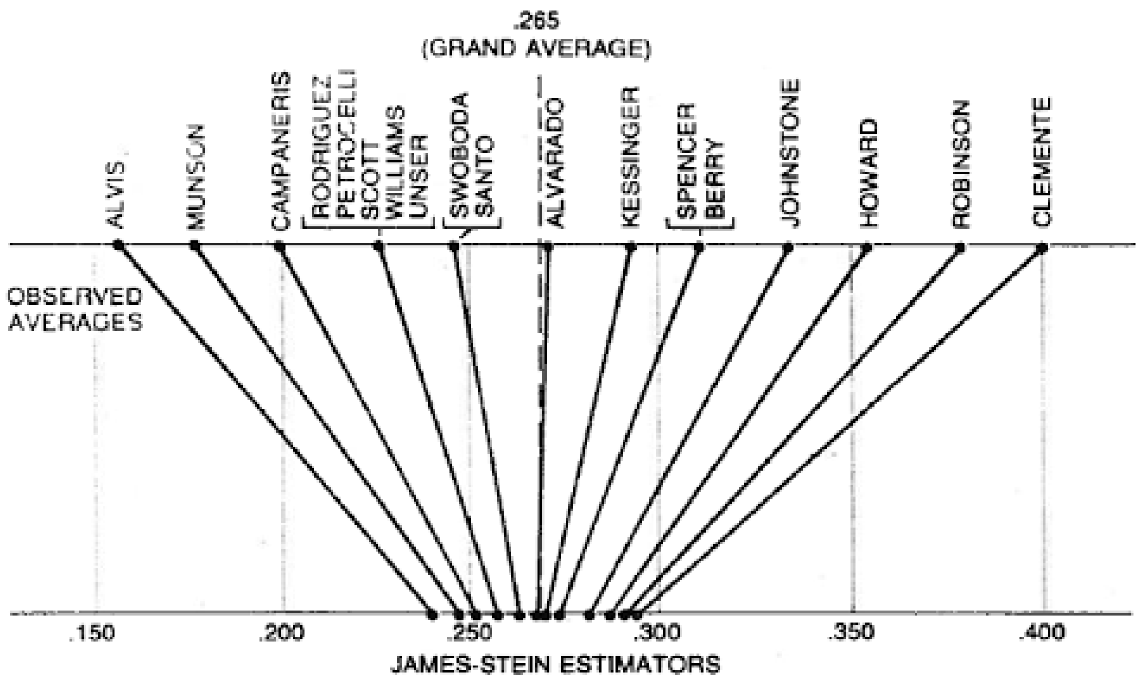
One can prove that if $d \geq 3$,

$$\mathbb{E}\left[\left(\widehat{\boldsymbol{\mu}}_{\text{JS}} - \widehat{\boldsymbol{\mu}}\right)^2\right] < \mathbb{E}\left[\left(\widehat{\boldsymbol{\mu}}_{\text{mle}} - \widehat{\boldsymbol{\mu}}\right)^2\right]$$

Samworth (2015) Stein's paradox, "*one should use the price of tea in China to obtain a better estimate of the chance of rain in Melbourne*".

James & Stein Estimator

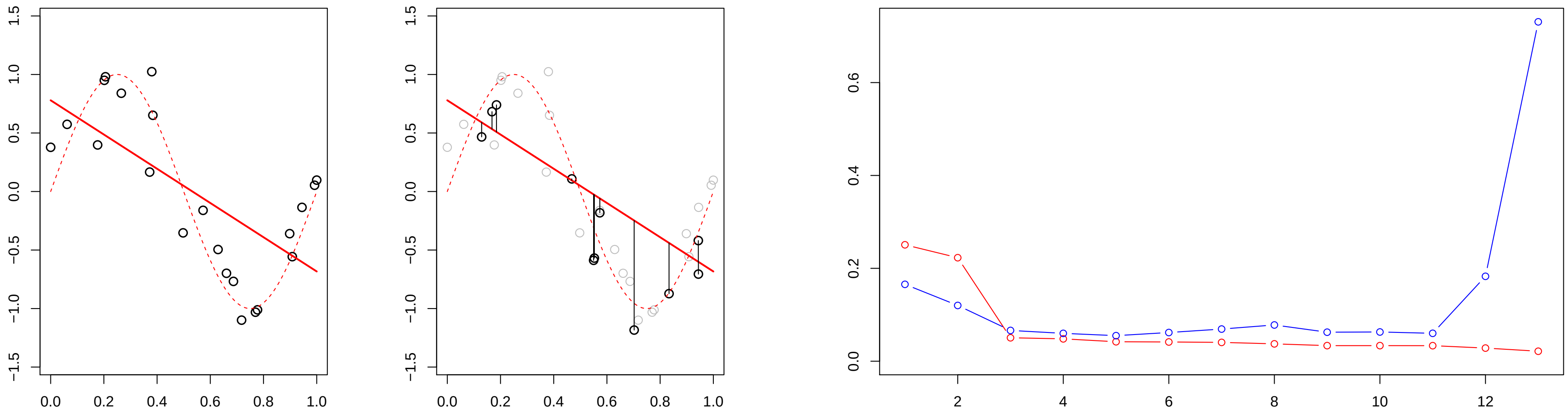
Heuristics : consider a biased estimator, to decrease the variance.



See Efron (2010) Large-Scale Inference

Motivation: Avoiding Overfit

Generalization : the model should perform well on new data (and not only on the training ones).



Reducing Dimension with PCA

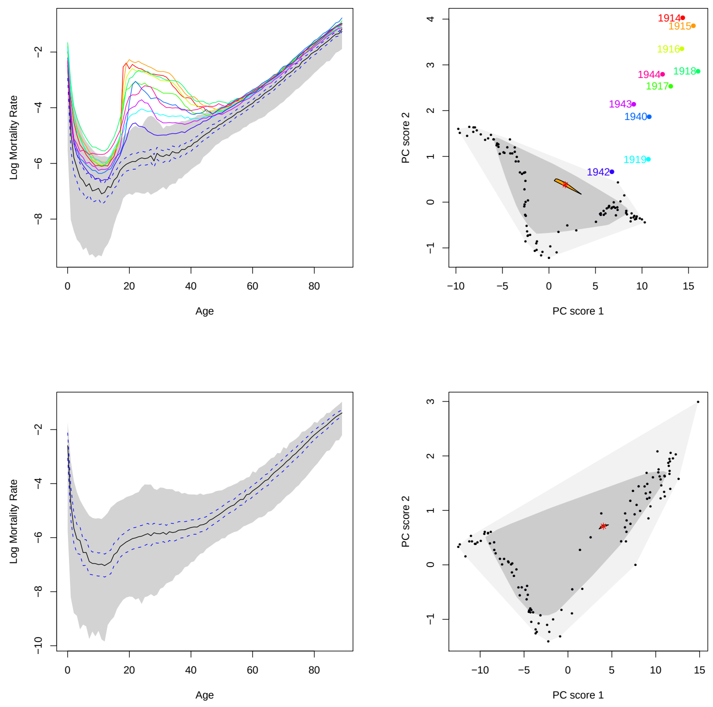
Use principal components to reduce dimension (on centered and scaled variables): we want d vectors $\boldsymbol{z}_1, \cdots, \boldsymbol{z}_d$ such that

First Compoment is $\boldsymbol{z}_1 = \boldsymbol{X}\boldsymbol{\omega}_1$ where

$$\boldsymbol{\omega}_1 = \underset{\|\boldsymbol{\omega}\|=1}{\text{argmax}}\left\{\|\boldsymbol{X}\cdot\boldsymbol{\omega}\|^2\right\} = \underset{\|\boldsymbol{\omega}\|=1}{\text{argmax}}\left\{\boldsymbol{\omega}^\top\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\omega}\right\}$$

Second Compoment is $\boldsymbol{z}_2 = \boldsymbol{X}\boldsymbol{\omega}_2$ where

$$\boldsymbol{\omega}_2 = \underset{\|\boldsymbol{\omega}\|=1}{\text{argmax}}\left\{\|\widetilde{\boldsymbol{X}}^{(1)}\cdot\boldsymbol{\omega}\|^2\right\}$$



with $\widetilde{\boldsymbol{X}}^{(1)} = \boldsymbol{X} - \underbrace{\boldsymbol{X}\boldsymbol{\omega}_1}_{\boldsymbol{z}_1}\boldsymbol{\omega}_1^\top$.

Reducing Dimension with PCA

A regression on (the d) principal components, $y = \boldsymbol{z}^\top \boldsymbol{\beta} + \boldsymbol{\eta}$ could be an interesting idea, unfortunatley, principal components have no reason to be correlated with y. First compoment was $\boldsymbol{z}_1 = \boldsymbol{X}\boldsymbol{\omega}_1$ where

$$\boldsymbol{\omega}_1 = \underset{\|\boldsymbol{\omega}\|=1}{\text{argmax}} \left\{\|\boldsymbol{X}\cdot\boldsymbol{\omega}\|^2\right\} = \underset{\|\boldsymbol{\omega}\|=1}{\text{argmax}} \left\{\boldsymbol{\omega}^\top\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\omega}\right\}$$

It is a non-supervised technique.

Instead, use partial least squares, introduced in Wold (1966) Estimation of Principal Components and Related Models by Iterative Least squares. First compoment is $\boldsymbol{z}_1 = \boldsymbol{X}\boldsymbol{\omega}_1$ where

$$\boldsymbol{\omega}_1 = \underset{\|\boldsymbol{\omega}\|=1}{\text{argmax}} \left\{\langle \boldsymbol{y}, \boldsymbol{X}\cdot\boldsymbol{\omega}\rangle\right\} = \underset{\|\boldsymbol{\omega}\|=1}{\text{argmax}} \left\{\boldsymbol{\omega}^\top\boldsymbol{X}^\top\boldsymbol{y}\boldsymbol{y}^\top\boldsymbol{X}\boldsymbol{\omega}\right\}$$

Terminology

Consider a dataset $\{y_i, \boldsymbol{x}_i\}$, assumed to be generated from $Y, \boldsymbol{X}$, from an unknown distribution $\mathbb{P}$.

Let $m_0(\cdot)$ be the "true" model. Assume that $y_i = m_0(\boldsymbol{x}_i) + \varepsilon_i$.

In a regression context (quadratic loss function function), the risk associated to m is

$$\mathcal{R}(m) = \mathbb{E}_{\mathbb{P}}\big[\big(Y - m(\boldsymbol{X})\big)^2\big]$$

An optimal model $m^\star$ within a class $\mathcal{M}$ satisfies

$$\mathcal{R}(m^\star) = \inf_{m \in \mathcal{M}} \big\{\mathcal{R}(m)\big\}$$

Such a model $m^\star$ is usually called oracle.

Observe that $m^\star(\boldsymbol{x}) = \mathbb{E}[Y|\boldsymbol{X} = \boldsymbol{x}]$ is the solution of

$$\mathcal{R}(m^\star) = \inf_{m \in \mathcal{M}} \big\{\mathcal{R}(m)\big\} \text{ where } \mathcal{M} \text{ is the set of measurable functions}$$

The empirical risk is

$$\mathcal{R}_n(m) = \frac{1}{n}\sum_{i=1}^{n}\big(y_i - m(\boldsymbol{x}_i)\big)^2$$

For instance, m can be a linear predictor, $m(\boldsymbol{x}) = \beta_0 + \boldsymbol{x}^\mathsf{T}\boldsymbol{\beta}$, where $\boldsymbol{\theta} = (\beta_0, \boldsymbol{\beta})$ should estimated (trained).

$\mathbb{E}\big[R_n(\widehat{m})\big] = \mathbb{E}\big[(\widehat{m}(\boldsymbol{X}) - Y)^2\big]$ can be expressed as

$$\begin{array}{rll} & \mathbb{E}\big[(\widehat{m}(\boldsymbol{X}) - \mathbb{E}[\widehat{m}(\boldsymbol{X})|\boldsymbol{X}])^2\big] & \text{variance of } \widehat{m} \\ + & \mathbb{E}\Big[\big(\mathbb{E}[\widehat{m}(\boldsymbol{X})|\boldsymbol{X}] - \underbrace{\mathbb{E}[Y|\boldsymbol{X}]}_{m_0(\boldsymbol{X})}\big)^2\Big] & \text{bias of } \widehat{m} \\ + & \mathbb{E}\Big[\big(Y - \underbrace{\mathbb{E}[Y|\boldsymbol{X}]}_{m_0(\boldsymbol{X})}\big)^2\Big] & \text{variance of the noise} \end{array}$$

The third term is the risk of the “optimal” estimator m, that cannot be decreased.

Mallows Penalty and Model Complexity

Consider a linear predictor (see #1), i.e. $\widehat{\boldsymbol{y}} = \widehat{m}(\boldsymbol{x}) = \boldsymbol{A}\boldsymbol{y}$.

Assume that $\boldsymbol{y} = m_0(\boldsymbol{x}) + \boldsymbol{\varepsilon}$, with $\mathbb{E}[\boldsymbol{\varepsilon}] = \boldsymbol{0}$ and $\text{Var}[\boldsymbol{\varepsilon}] = \sigma^2\mathbb{I}$.

Let $\|\cdot\|$ denote the Euclidean norm

Empirical risk : $\widehat{\mathcal{R}}_n(m) = \frac{1}{n}\|\boldsymbol{y} - m(\boldsymbol{x})\|^2$

Vapnik's risk : $\mathbb{E}[\widehat{\mathcal{R}}_n(m)] = \dfrac{1}{n}\|m_0(\boldsymbol{x} - m(\boldsymbol{x})\|^2 + \dfrac{1}{n}\mathbb{E}\big(\|\boldsymbol{y} - m_0(\boldsymbol{x}\|^2\big)$ with $m_0(\boldsymbol{x} = \mathbb{E}[Y|\boldsymbol{X} = \boldsymbol{x}]$.

Observe that

$$n\mathbb{E}\big[\widehat{\mathcal{R}}_n(\widehat{m})\big] = \mathbb{E}\big(\|\boldsymbol{y} - \widehat{m}(\boldsymbol{x})\|^2\big) = \|(\mathbb{I} - \boldsymbol{A})m_0\|^2 + \sigma^2\|\mathbb{I} - \boldsymbol{A}\|^2$$

while

$$= \mathbb{E}\big(\|m_0(\boldsymbol{x}) - \widehat{m}(\boldsymbol{x})\|^2\big) = \underbrace{\|(\mathbb{I} - \boldsymbol{A})m_0\|^2}_{\text{bias}} + \underbrace{\sigma^2\|\boldsymbol{A}\|^2}_{\text{variance}}$$

Mallows Penalty and Model Complexity

One can obtain

$$\mathbb{E}\big[\mathcal{R}_n(\widehat{m})\big] = \mathbb{E}\big[\widehat{\mathcal{R}}_n(\widehat{m})\big] + 2\frac{\sigma^2}{n}\text{trace}(\boldsymbol{A}).$$

If $\text{trace}(\boldsymbol{A}) \geq 0$ the empirical risk underestimate the true risk of the estimator.

The number of degrees of freedom of the (linear) predictor is related to $\text{trace}(\boldsymbol{A})$

$2\dfrac{\sigma^2}{n}\text{trace}(\boldsymbol{A})$ is called Mallow's penalty C_L.

If $\boldsymbol{A}$ is a projection matrix, $\text{trace}(\boldsymbol{A})$ is the dimension of the projection space, p, then we obtain Mallow's C_P, $2\dfrac{\sigma^2}{n}p$.

Remark : Mallows (1973) Some Comments on C_p introduced this penalty while focusing on the R^2.

Penalty and Likelihood

C_P is associated to a quadratic risk

an alternative is to use a distance on the (conditional) distribution of Y, namely Kullback-Leibler distance

discrete case: $D_{\text{KL}}(P\|Q)=\displaystyle\sum_i P(i)\log\frac{P(i)}{Q(i)}$

continuous case :

$$D_{\text{KL}}(P\|Q)=\int_{-\infty}^{\infty}p(x)\log\frac{p(x)}{q(x)}\,dx D_{\text{KL}}(P\|Q)=\textstyle\int_{-\infty}^{\infty}p(x)\log\frac{p(x)}{q(x)}\,dx$$

Let f denote the true (unknown) density, and f_θ some parametric distribution,

$$D_{\text{KL}}(f\|f_\theta)=\int_{-\infty}^{\infty}f(x)\log\frac{f(x)}{f_\theta(x)}\,dx=\int f(x)\log[f(x)]\,dx-\underbrace{\int f(x)\log[f_\theta(x)]\,dx}_{\text{relative information}}$$

Hence

$$\text{minimize}\,\{D_{\text{KL}}(f\|f_\theta)\}\quad\longleftrightarrow\quad\text{maximize}\,\{\mathbb{E}\big[\log[f_\theta(X)]\big]\}$$

Penalty and Likelihood

Akaike (1974) A new look at the statistical model identification observe that for n large enough

$$\mathbb{E}\Big[\log[f_\theta(X)]\Big] \sim \log[\mathcal{L}(\widehat{\theta})] - \dim(\theta)$$

Thus

$$AIC = -2\log\mathcal{L}(\widehat{\theta}) + 2\dim(\theta)$$

Example : in a (Gaussian) linear model, $y_i = \beta_0 + \boldsymbol{x}_i^\mathsf{T}\boldsymbol{\beta} + \varepsilon_i$

$$AIC = n\log\left(\frac{1}{n}\sum_{i=1}^n \widehat{\varepsilon}_i\right) + 2[\dim(\boldsymbol{\beta}) + 2]$$

Penalty and Likelihood

Remark : this is valid for large sample (rule of thumb $n/\dim(\theta) > 40$), otherwise use a corrected AIC

$$AICc = AIC + \underbrace{\frac{2k(k+1)}{n-k-1}}_{\text{bias correction}} \quad \text{where } k = \dim(\theta)$$

see Sugiura (1978) Further analysis of the data by Akaike's information criterion and the finite corrections second order AIC.

Using a Bayesian interpretation, Schwarz (1978) Estimating the dimension of a model obtained

$$BIC = -2\log\mathcal{L}(\widehat{\theta}) + \log(n)\dim(\theta).$$

Observe that the criteria considered is

$$\text{criteria} = -\text{function}\big(\mathcal{L}(\widehat{\theta})\big) + \text{penality}\big(\text{complexity}\big)$$

Estimation of the Risk

Consider a naive bootstrap procedure, based on a bootstrap sample $\mathcal{S}_b = \{(y_i^{(b)}, \boldsymbol{x}_i^{(b)})\}$.

The plug-in estimator of the empirical risk is

$$\widehat{\mathcal{R}}_n(\widehat{m}^{(b)}) = \frac{1}{n}\sum_{i=1}^{n}\big(y_i - \widehat{m}^{(b)}(\boldsymbol{x}_i)\big)^2$$

and then

$$\widehat{\mathcal{R}}_n = \frac{1}{B}\sum_{b=1}^{B}\widehat{\mathcal{R}}_n(\widehat{m}^{(b)}) = \frac{1}{B}\sum_{b=1}^{B}\frac{1}{n}\sum_{i=1}^{n}\big(y_i - \widehat{m}^{(b)}(\boldsymbol{x}_i)\big)^2$$

Estimation of the Risk

One might improve this estimate using a out-of-bag procedure

$$\widehat{\mathcal{R}}_n = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{\#\mathcal{B}_i}\sum_{b\in\mathcal{B}_i}\big(y_i - \widehat{m}^{(b)}(\boldsymbol{x}_i)\big)^2$$

where $\mathcal{B}_i$ is the set of all boostrap sample that contain $(y_i, \boldsymbol{x}_i)$.

Remark: $\mathbb{P}\left((y_i, \boldsymbol{x}_i)\notin\mathcal{S}_b\right) = \left(1-\frac{1}{n}\right)^n \sim e^{-1} = 36,78\%.$

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 $\boldsymbol{X}^\top\boldsymbol{X} + \lambda\mathbb{I}$

Linear Regression Shortcoming

Evolution of $(\beta_1, \beta_2) \mapsto \displaystyle\sum_{i=1}^{n}[y_i - (\beta_1 x_{1,i} + \beta_2 x_{2,i})]^2$

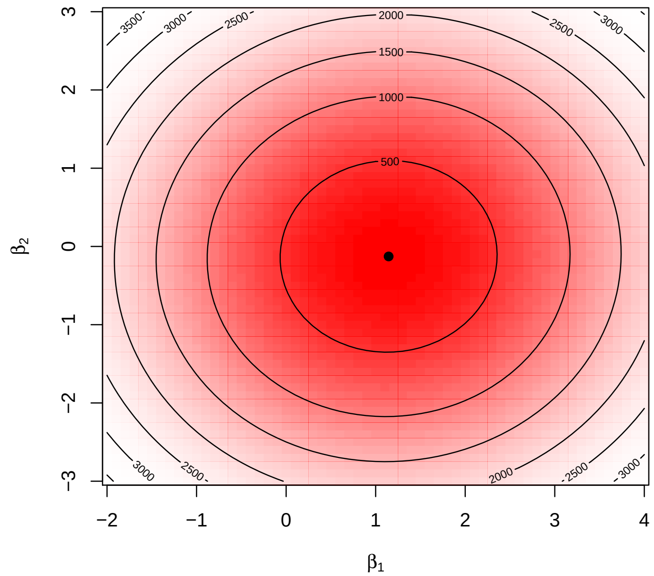
when $\text{cor}(X_1, X_2) = r \in [0, 1]$, on top.

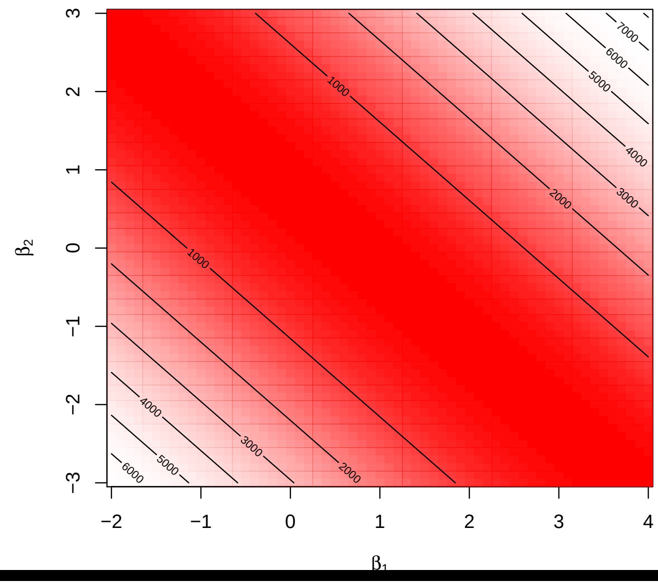
Below, Ridge regression

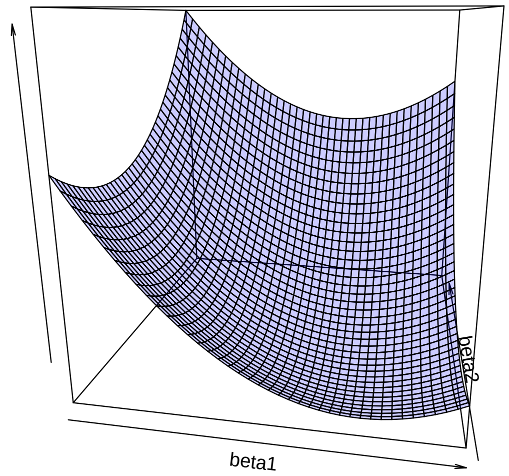
$$(\beta_1, \beta_2) \mapsto \sum_{i=1}^{n}[y_i - (\beta_1 x_{1,i} + \beta_2 x_{2,i})]^2 + \lambda(\beta_1^2 + \beta_2^2)$$

where $\lambda \in [0, \infty)$, below,

when $\text{cor}(X_1, X_2) \sim 1$ (colinearity).



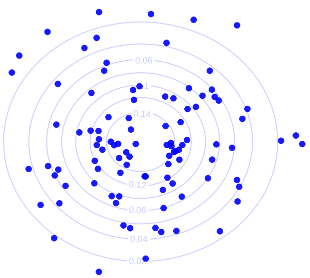


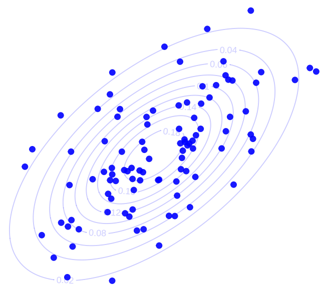


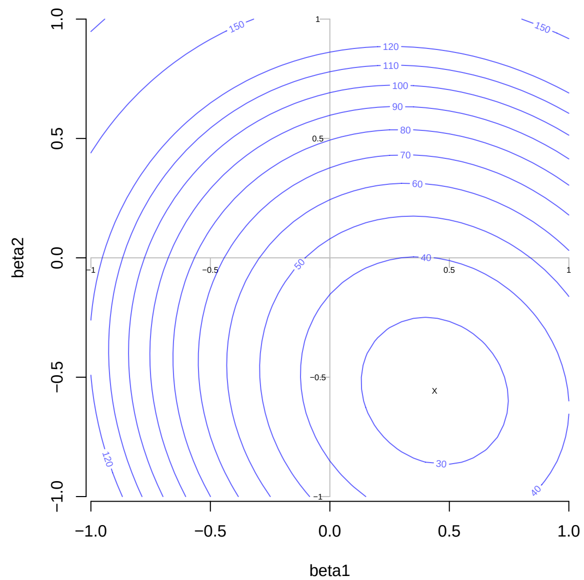
Normalization : Euclidean ℓ_2 vs. Mahalonobis

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

Ridge Regression

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

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

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

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

The constant is usually unpenalized. The *true* equation is

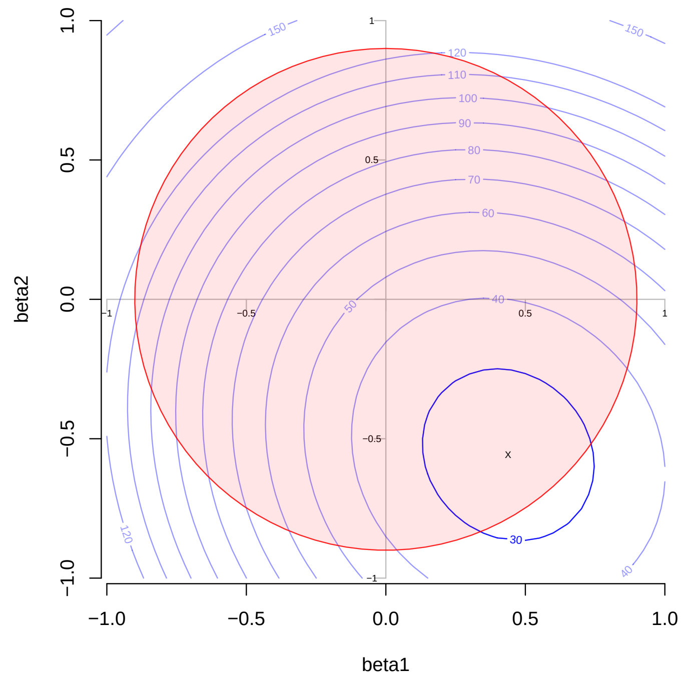
$$\widehat{\boldsymbol{\beta}}_{\lambda}^{\text{ridge}} = \text{argmin}\left\{\underbrace{\|\boldsymbol{y} - (\beta_0 + \boldsymbol{X}\boldsymbol{\beta})\|_{\ell_2}^2}_{=\text{criteria}} + \underbrace{\lambda\|\boldsymbol{\beta}\|_{\ell_2}^2}_{=\text{penalty}}\right\}$$

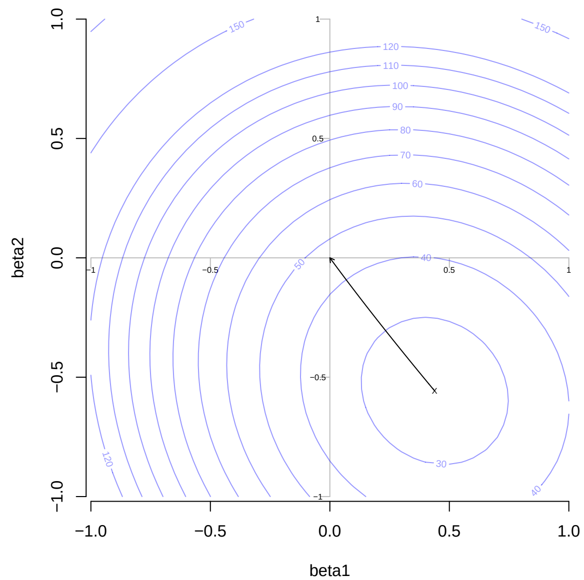
Ridge Regression

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

can be seen as a constrained optimization problem

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





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

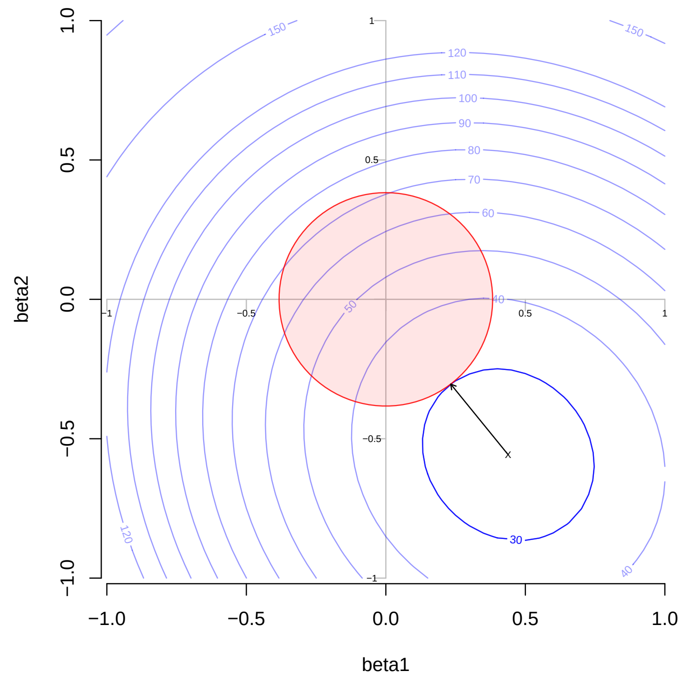
Ridge Regression

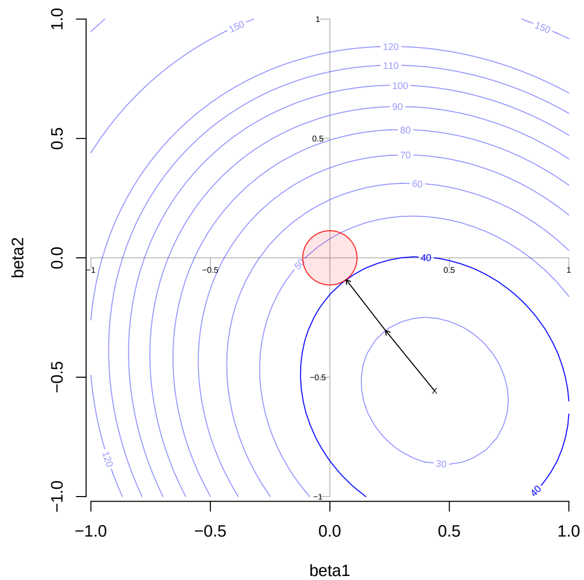
This penalty can be seen as rather unfair if components of $\boldsymbol{x}$ are not expressed on the same scale

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

Then compute

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



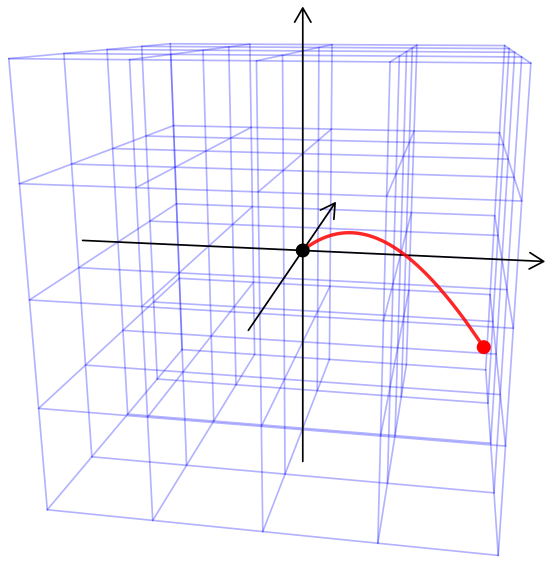


Ridge Regression

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...



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

Ridge Regression

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

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

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

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

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

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 $\lambda \to \infty$.

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

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

Properties of the Ridge Estimator

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

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

Thus,

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

and

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

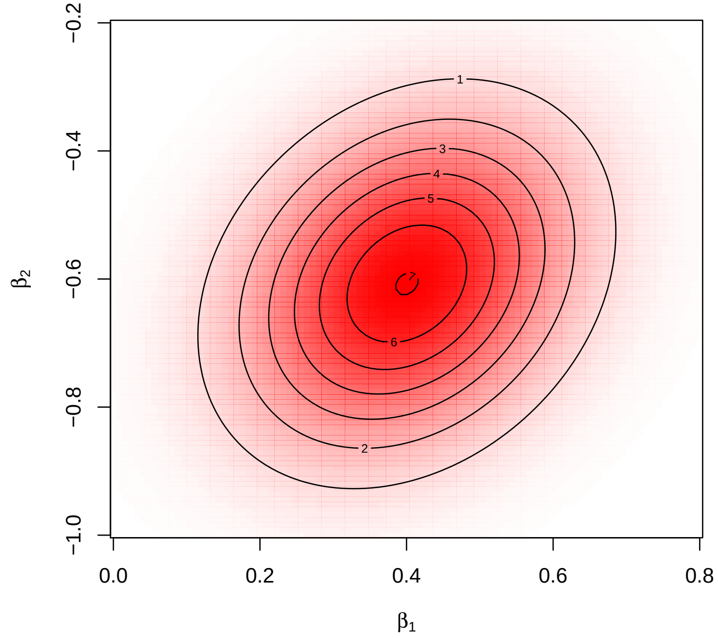
Observe that

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

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

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

is minimal for

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

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

SVD decomposition

Consider the singular value decomposition $\boldsymbol{X}=\boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^{\mathsf{T}}$. Then

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

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

Observe that

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

hence, the ridge penality shrinks singular values.

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

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

Hat matrix and Degrees of Freedom

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

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

Similarly

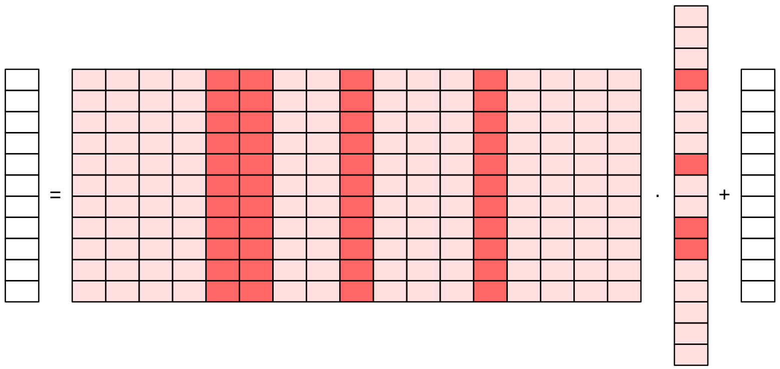
$$\boldsymbol{H}_\lambda = \boldsymbol{X}(\boldsymbol{X}^\mathsf{T}\boldsymbol{X} + \lambda\mathbb{I})^{-1}\boldsymbol{X}^\mathsf{T}$$

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

Sparsity Issues

In severall 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 relevent 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 $y = \boldsymbol{X}_{\mathcal{S}}^{\mathsf{T}}\boldsymbol{\beta}_{\mathcal{S}} + \varepsilon$, where $\boldsymbol{X}_{\mathcal{S}}^{\mathsf{T}}\boldsymbol{X}_{\mathcal{S}}$ is a full rank matrix.

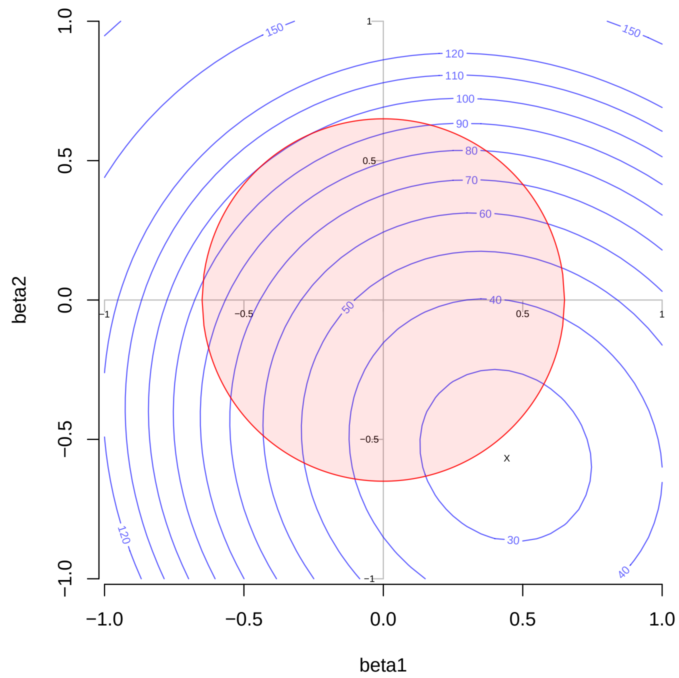
Going further on sparcity issues

The Ridge regression problem was to solve

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

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

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



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

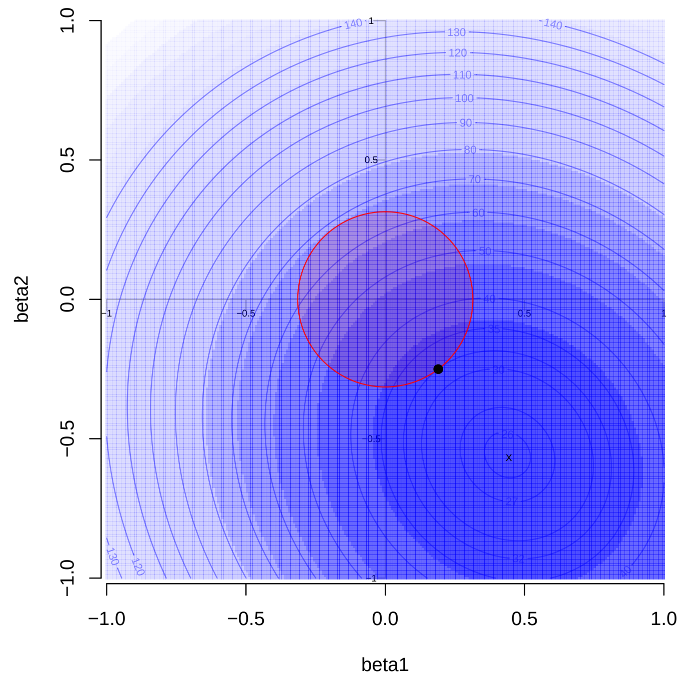
Going further on sparcity issues

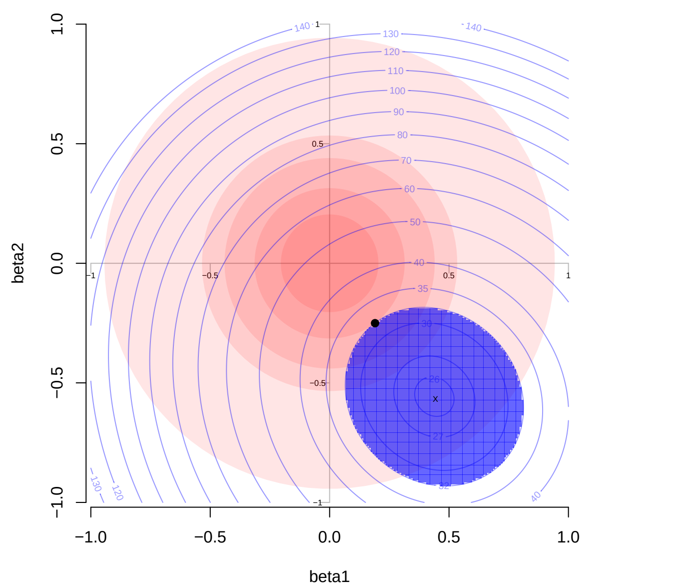
In a convex problem, solve the dual problem, e.g. in the Ridge regression : primal problem

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

and the dual problem

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



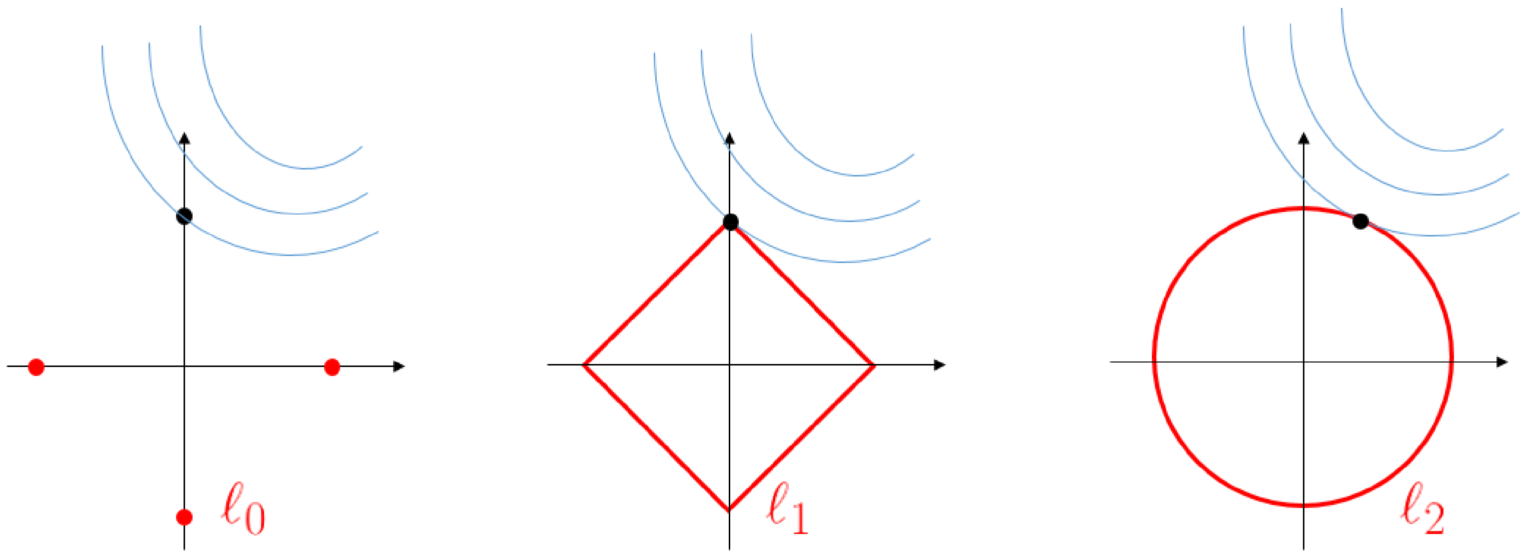


Going further on sparcity issues

Idea: solve the dual problem

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

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



Going further on sparcity 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

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

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

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

LASSO ***Least Absolute Shrinkage and Selection Operator***

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

is a convex problem (several algorithms*), but not strictly convex (no unicity of the minimum). Nevertheless, predictions $\widehat{\boldsymbol{y}} = \boldsymbol{x}^\mathsf{T}\widehat{\boldsymbol{\beta}}$ are unique.

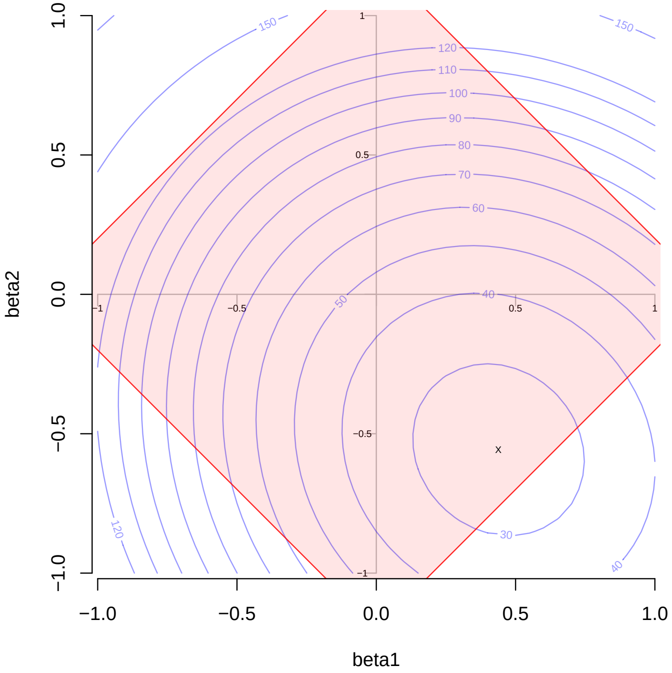
* MM, minimize majorization, coordinate descent Hunter & Lange (2003) A Tutorial on MM Algorithms.

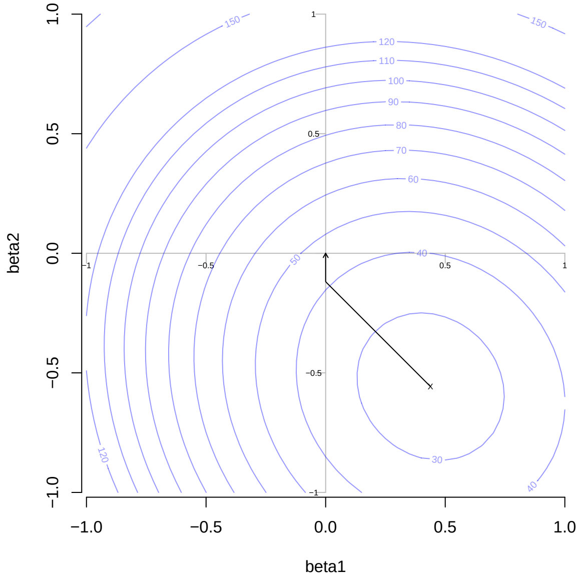
LASSO Regression

No explicit solution...

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

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

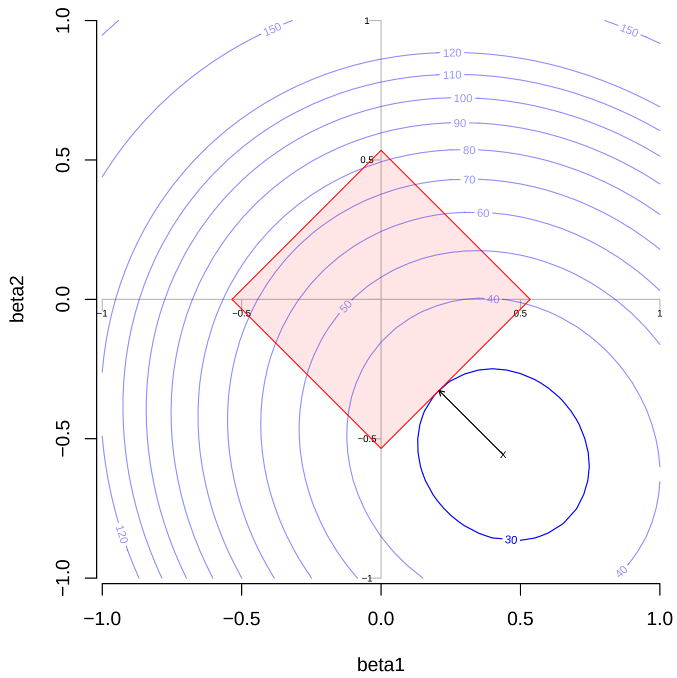


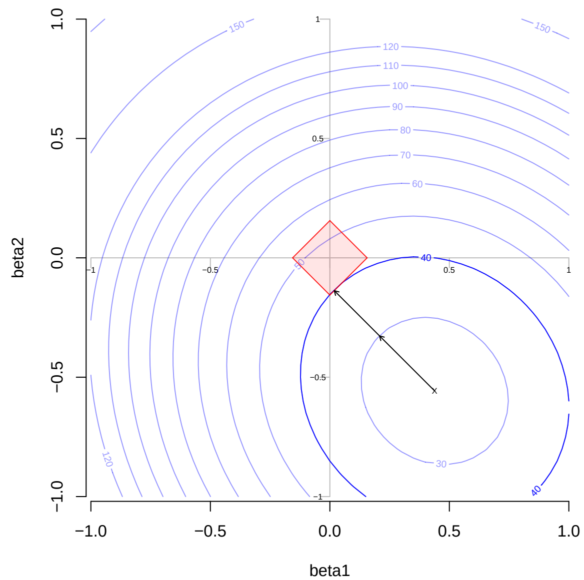


LASSO Regression

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



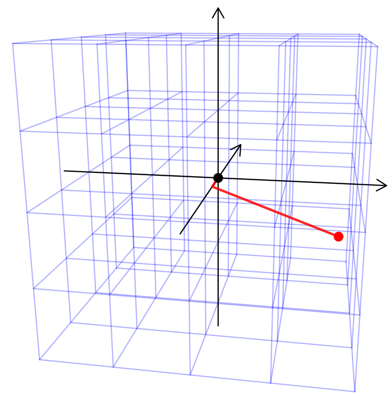


LASSO Regression

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

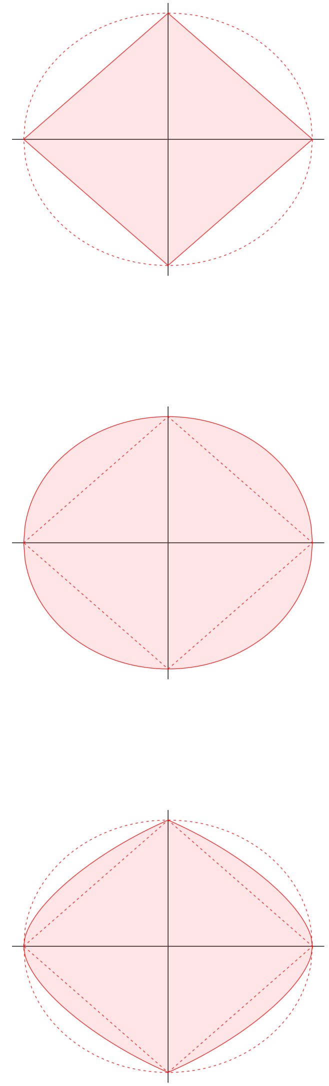
$$\lambda^\star = \text{argmin}\left\{\overline{Q}(\lambda) = \frac{1}{K}\sum_k Q_k(\lambda)\right\}$$

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 and Ridge, with R

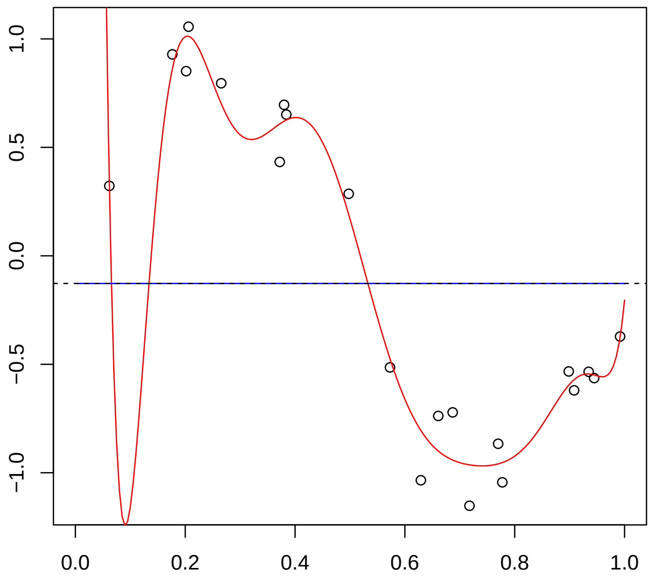
```
> library(glmnet)
> chicago=read.table("http://freakonometrics.free.fr/
    chicago.txt",header=TRUE,sep=";")
> standardize <-  function(x)  {(x-mean(x))/sd(x)}
> z0 <- standardize(chicago[, 1])
> z1 <- standardize(chicago[, 3])
> z2 <- standardize(chicago[, 4])
> ridge <-glmnet(cbind(z1, z2), z0, 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 Regression, Smoothing and Overfit

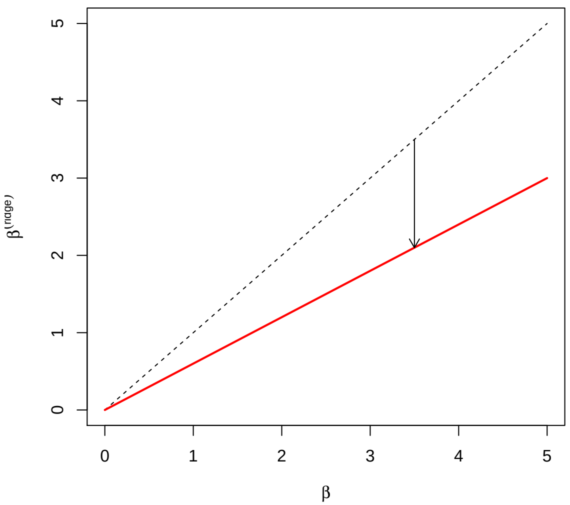
LASSO can be used to avoid overfit.

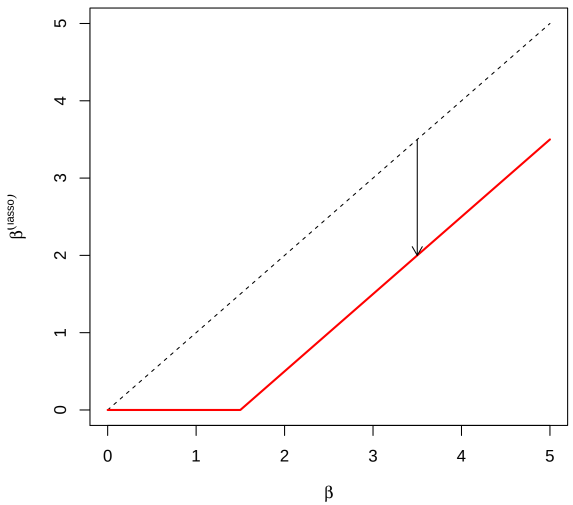


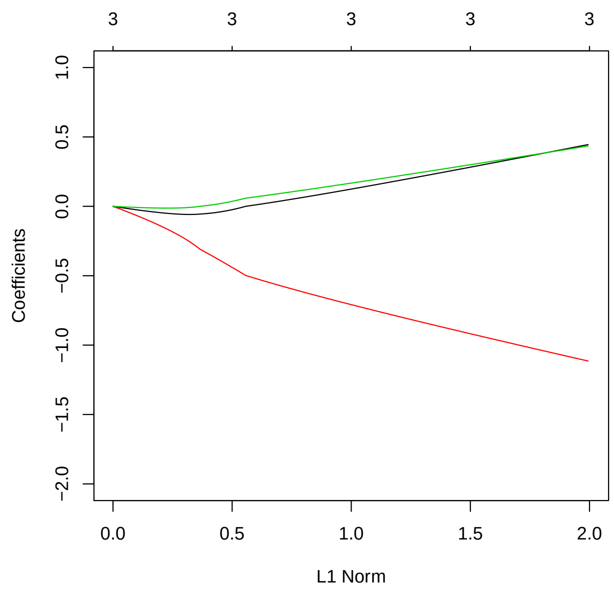
Ridge vs. LASSO

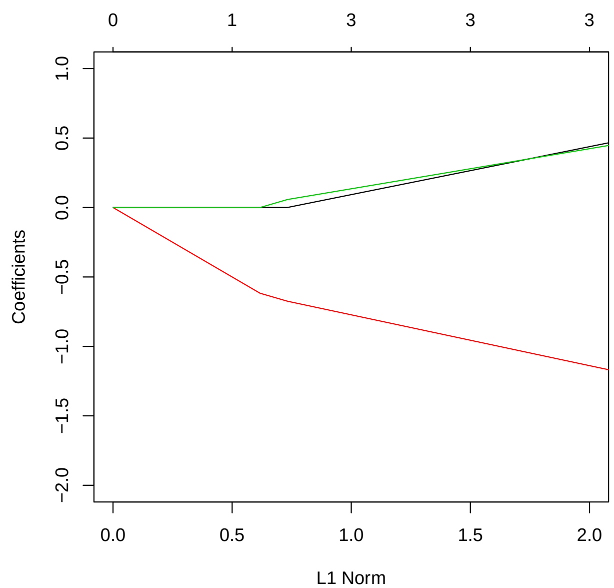
Consider simulated data (output on the right).

With orthogonal variables, shrinkage operators are









Optimization Heuristics

First idea: given some initial guess $\boldsymbol{\beta}_{(0)}$, $|\boldsymbol{\beta}| \sim |\boldsymbol{\beta}_{(0)}| + \dfrac{1}{2|\boldsymbol{\beta}_{(0)}|}(\boldsymbol{\beta}^2 - \boldsymbol{\beta}_{(0)}^2)$

LASSO estimate can probably be derived from iterated Ridge estimates

$$\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_{(k+1)}\|_{\ell_2}^2 + \lambda\|\boldsymbol{\beta}_{(k+1)}\|_{\ell_1} \sim \boldsymbol{X}\boldsymbol{\beta}_{(k+1)}\|_{\ell_2}^2 + \frac{\lambda}{2}\sum_j \frac{1}{|\boldsymbol{\beta}_{j,(k)}|}[\boldsymbol{\beta}_{j,(k+1)}]^2$$

which is a weighted ridge penalty function

Thus,

$$\boldsymbol{\beta}_{(k+1)} = \left(\boldsymbol{X}^\top\boldsymbol{X} + \lambda\boldsymbol{\Delta}_{(k)}\right)^{-1}\boldsymbol{X}^\top\boldsymbol{y}$$

where $\boldsymbol{\Delta}_{(k)} = \text{diag}[|\boldsymbol{\beta}_{j,(k)}|^{-1}]$. Then $\boldsymbol{\beta}_{(k)} \to \widehat{\boldsymbol{\beta}}^{\text{lasso}}$, as $k \to \infty$.

Properties of LASSO Estimate

From this iterative technique

$$\widehat{\boldsymbol{\beta}}_{\lambda}^{\text{lasso}} \sim \left(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X} + \lambda\boldsymbol{\Delta}\right)^{-1}\boldsymbol{X}^{\mathsf{T}}\boldsymbol{y}$$

where $\boldsymbol{\Delta} = \text{diag}[|\widehat{\boldsymbol{\beta}}_{j,\lambda}^{\text{lasso}}|^{-1}]$ if $\widehat{\boldsymbol{\beta}}_{j,\lambda}^{\text{lasso}} \neq 0$, 0 otherwise.

Thus,

$$\mathbb{E}[\widehat{\boldsymbol{\beta}}_{\lambda}^{\text{lasso}}] \sim \left(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X} + \lambda\boldsymbol{\Delta}\right)^{-1}\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X}\boldsymbol{\beta}$$

and

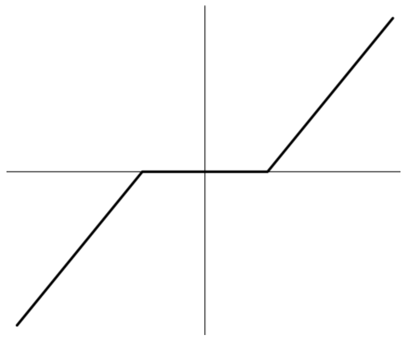
$$\text{Var}[\widehat{\boldsymbol{\beta}}_{\lambda}^{\text{lasso}}] \sim \sigma^2\left(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X} + \lambda\boldsymbol{\Delta}\right)^{-1}\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X}\left(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X} + \lambda\boldsymbol{\Delta}\right)^{-1}\boldsymbol{X}^{\mathsf{T}}$$

Optimization Heuristics

Consider here a simplified problem, $\min_{a\in\mathbb{R}}\Big\{\underbrace{\frac{1}{2}(a-b)^2+\lambda|a|}_{g(a)}\Big\}$ with $\lambda>0$.

Observe that $g'(0)=-b\pm\lambda$. Then

- if $|b|\leq\lambda$, then $a^\star=0$
- if $b\geq\lambda$, then $a^\star=b-\lambda$
- if $b\leq-\lambda$, then $a^\star=b+\lambda$



$$a^\star=\underset{a\in\mathbb{R}}{\text{argmin}}\Big\{\frac{1}{2}(a-b)^2+\lambda|a|\Big\}=S_\lambda(b)=\text{sign}(b)\cdot(|b|-\lambda)_+,$$

also called soft-thresholding operator.

Optimization Heuristics

Definition for any convex function h, define the proximal operator operator of h,

$$\text{proximal}_h(\boldsymbol{y}) = \underset{\boldsymbol{x}\in\mathbb{R}^d}{\text{argmin}}\left\{\frac{1}{2}\|\boldsymbol{x}-\boldsymbol{y}\|_{\ell_2}^2 + h(\boldsymbol{x})\right\}$$

Note that

$$\text{proximal}_{\lambda\|\cdot\|_{\ell_2}^2}(\boldsymbol{y}) = \frac{1}{1+\lambda}\boldsymbol{x} \quad \text{shrinkage operator}$$

$$\text{proximal}_{\lambda\|\cdot\|_{\ell_1}}(\boldsymbol{y}) = S_\lambda(\boldsymbol{y}) = \text{sign}(\boldsymbol{y})\cdot(|\boldsymbol{y}|-\lambda)_+$$

Optimization Heuristics

We want to solve here

$$\widehat{\boldsymbol{\theta}} \in \underset{\boldsymbol{\theta}\in\mathbb{R}^d}{\text{argmin}}\Big\{\underbrace{\frac{1}{n}\|\boldsymbol{y}-m_{\boldsymbol{\theta}}(\boldsymbol{x}))\|_{\ell_2}^2}_{f(\boldsymbol{\theta})}+\underbrace{\lambda\text{penalty}(\boldsymbol{\theta})}_{g(\boldsymbol{\theta})}\Big\}.$$

where f is convex and smooth, and g is convex, but not smooth...

1. Focus on f : descent lemma, $\forall\boldsymbol{\theta},\boldsymbol{\theta}'$

$$f(\boldsymbol{\theta})\leq f(\boldsymbol{\theta}')+\langle\nabla f(\boldsymbol{\theta}'),\boldsymbol{\theta}-\boldsymbol{\theta}'\rangle+\frac{t}{2}\|\boldsymbol{\theta}-\boldsymbol{\theta}'\|_{\ell_2}^2$$

Consider a gradient descent sequence $\boldsymbol{\theta}_k$, i.e. $\boldsymbol{\theta}_{k+1}=\boldsymbol{\theta}_k-t^{-1}\nabla f(\boldsymbol{\theta}_k)$, then

$$f(\boldsymbol{\theta})\leq \overbrace{f(\boldsymbol{\theta}_k)+\langle\nabla f(\boldsymbol{\theta}_k),\boldsymbol{\theta}-\boldsymbol{\theta}_k\rangle+\frac{t}{2}\|\boldsymbol{\theta}-\boldsymbol{\theta}_k\|_{\ell_2}^2}^{\varphi(\boldsymbol{\theta}):\ \boldsymbol{\theta}_{k+1}=\text{argmin}\{\varphi(\boldsymbol{\theta})\}}$$

Optimization Heuristics

2. Add function g

$$f(\boldsymbol{\theta})+g(\boldsymbol{\theta}) \leq \overbrace{f(\boldsymbol{\theta}_k) + \langle \nabla f(\boldsymbol{\theta}_k), \boldsymbol{\theta} - \boldsymbol{\theta}_k \rangle + \frac{t}{2}\|\boldsymbol{\theta} - \boldsymbol{\theta}_k\|_{\ell_2}^2 + g(\boldsymbol{\theta})}^{\psi(\boldsymbol{\theta})}$$

And one can proof that

$$\boldsymbol{\theta}_{k+1} = \underset{\boldsymbol{\theta}\in\mathbb{R}^d}{\text{argmin}}\left\{\psi(\boldsymbol{\theta})\right\} = \text{proximal}_{g/t}\left(\boldsymbol{\theta}_k - t^{-1}\nabla f(\boldsymbol{\theta}_k)\right)$$

so called proximal gradient descent algorithm, since

$$\text{argmin}\,\{\psi(\boldsymbol{\theta})\} = \text{argmin}\left\{\frac{t}{2}\left\|\boldsymbol{\theta} - \left(\boldsymbol{\theta}_k - t^{-1}\nabla f(\boldsymbol{\theta}_k)\right)\right\|_{\ell_2}^2 + g(\boldsymbol{\theta})\right\}$$

Coordinate-wise minimization

Consider some convex differentiable $f:\mathbb{R}^k\rightarrow\mathbb{R}$ function.

Consider $\boldsymbol{x}^\star\in\mathbb{R}^k$ obtained by minimizing along each coordinate axis, i.e.

$$f(x_1^\star,x_{i-1}^\star,x_i,x_{i+1}^\star,\cdots,x_k^\star)\geq f(x_1^\star,x_{i-1}^\star,x_i^\star,x_{i+1}^\star,\cdots,x_k^\star)$$

for all i. Is $\boldsymbol{x}^\star$ a global minimizer? i.e.

$$f(\boldsymbol{x})\geq f(\boldsymbol{x}^\star),\ \forall\boldsymbol{x}\in\mathbb{R}^k.$$

Yes. If f is convex and differentiable.

$$\nabla f(\boldsymbol{x})|_{\boldsymbol{x}=\boldsymbol{x}^\star}=\left(\frac{\partial f(\boldsymbol{x})}{\partial x_1},\cdots,\frac{\partial f(\boldsymbol{x})}{\partial x_k}\right)=\boldsymbol{0}$$

There might be problem if f is not differentiable (except in each axis direction).

If $f(\boldsymbol{x})=g(\boldsymbol{x})+\sum_{i=1}^k h_i(x_i)$ with g convex and differentiable, yes, since

$$f(\boldsymbol{x})-f(\boldsymbol{x}^\star)\geq\nabla g(\boldsymbol{x}^\star)^\mathsf{T}(\boldsymbol{x}-\boldsymbol{x}^\star)+\sum_i[h_i(x_i)-h_i(x_i^\star)]$$

Coordinate-wise minimization

$$f(\boldsymbol{x}) - f(\boldsymbol{x}^\star) \geq \sum_i \underbrace{[\nabla_i g(\boldsymbol{x}^\star)^\mathsf{T}(x_i - x_i^\star) h_i(x_i) - h_i(x_i^\star)]}_{\geq 0} \geq 0$$

Thus, for functions $f(\boldsymbol{x}) = g(\boldsymbol{x}) + \sum_{i=1}^k h_i(x_i)$ we can use coordinate descent to find a minimizer, i.e. at step j

$$x_1^{(j)} \in \underset{x_1}{\text{argmin}} f(x_1, x_2^{(j-1)}, x_3^{(j-1)}, \cdots x_k^{(j-1)})$$

$$x_2^{(j)} \in \underset{x_2}{\text{argmin}} f(x_1^{(j)}, x_2, x_3^{(j-1)}, \cdots x_k^{(j-1)})$$

$$x_3^{(j)} \in \underset{x_3}{\text{argmin}} f(x_1^{(j)}, x_2^{(j)}, x_3, \cdots x_k^{(j-1)})$$

Tseng (2001) Convergence of Block Coordinate Descent Method: if f is continuous, then $\boldsymbol{x}^\infty$ is a minimizer of f.

Application in Linear Regression

Let $f(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|^2$, with $\boldsymbol{y} \in \mathbb{R}^n$ and $\boldsymbol{A} \in \mathcal{M}_{n\times k}$. Let $\boldsymbol{A} = [\boldsymbol{A}_1, \cdots, \boldsymbol{A}_k]$.

Let us minimize in direction i. Let $\boldsymbol{x}_{-i}$ denote the vector in $\mathbb{R}^{k-1}$ without x_i. Here

$$0 = \frac{\partial f(\boldsymbol{x})}{\partial x_i} = \boldsymbol{A}_i^\top[\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}] = \boldsymbol{A}_i^\top[\boldsymbol{A}_i x_i + \boldsymbol{A}_{-i}\boldsymbol{x}_{-i} - \boldsymbol{y}]$$

thus, the optimal value is here

$$x_i^\star = \frac{\boldsymbol{A}_i^\top[\boldsymbol{A}_{-i}\boldsymbol{x}_{-i} - \boldsymbol{y}]}{\boldsymbol{A}_i^\top\boldsymbol{A}_i}$$

Application to LASSO

Let $f(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|^2 + \lambda\|\boldsymbol{x}\|_{\ell_1}$, so that the non-differentiable part is separable, since $\|\boldsymbol{x}\|_{\ell_1} = \sum_{i=1}^k |x_i|$.

Let us minimize in direction i. Let $\boldsymbol{x}_{-i}$ denote the vector in $\mathbb{R}^{k-1}$ without x_i. Here

$$0 = \frac{\partial f(\boldsymbol{x})}{\partial x_i} = \boldsymbol{A}_i^\mathsf{T}[\boldsymbol{A}_i x_i + \boldsymbol{A}_{-i}\boldsymbol{x}_{-i} - \boldsymbol{y}] + \lambda s_i$$

where $s_i \in \partial|x_i|$. Thus, solution is obtained by soft-thresholding

$$x_i^\star = S_{\lambda/\|\boldsymbol{A}_i\|^2}\left(\frac{\boldsymbol{A}_i^\mathsf{T}[\boldsymbol{A}_{-i}\boldsymbol{x}_{-i} - \boldsymbol{y}]}{\boldsymbol{A}_i^\mathsf{T}\boldsymbol{A}_i}\right)$$

Convergence rate for LASSO

Let $f(\boldsymbol{x}) = g(\boldsymbol{x}) + \lambda\|\boldsymbol{x}\|_{\ell_1}$ with

- g convex, ∇g Lipschitz with constant $L > 0$, and $Id - \nabla g/L$ monotone inscreasing in each component
- there exists $\boldsymbol{z}$ such that, componentwise, either $\boldsymbol{z} \geq S_\lambda(\boldsymbol{z} - \nabla g(\boldsymbol{z}))$ or $\boldsymbol{z} \leq S_\lambda(\boldsymbol{z} - \nabla g(\boldsymbol{z}))$

Saka & Tewari (2010), On the finite time convergence of cyclic coordinate descent methods proved that a coordinate descent starting from $\boldsymbol{z}$ satisfies

$$f(\boldsymbol{x}^{(j)}) - f(\boldsymbol{x}^\star) \leq \frac{L\|\boldsymbol{z} - \boldsymbol{x}^\star\|^2}{2j}$$

Graphical Lasso and Covariance Estimation

We want to estimate an (unknown) covariance matrix $\boldsymbol{\Sigma}$, or $\boldsymbol{\Sigma}^{-1}$.

An estimate for $\boldsymbol{\Sigma}^{-1}$ is $\boldsymbol{\Theta}^{\star}$ solution of

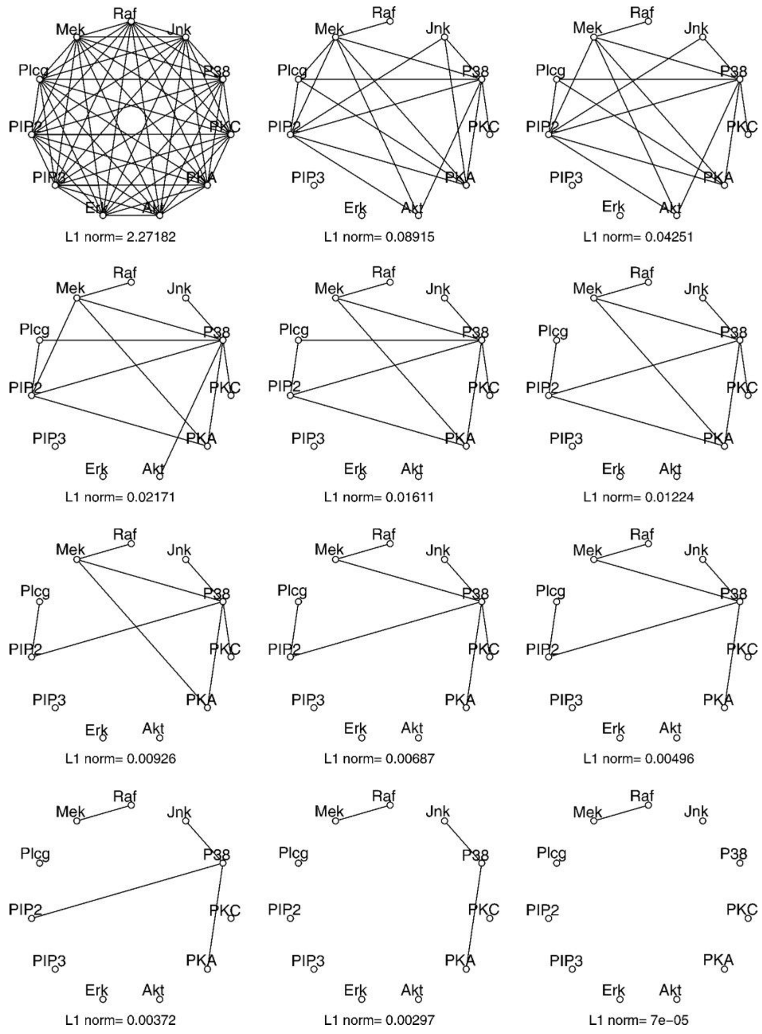
$$\boldsymbol{\Theta} \in \underset{\boldsymbol{\Theta}\in\mathcal{M}_{k\times k}}{\text{argmin}} \left\{-\log[\det(\boldsymbol{\Theta})] + \text{trace}[S\boldsymbol{\Theta}] + \lambda\|\boldsymbol{\Theta}\|_{\ell_1}\right\} \text{ where } S = \frac{\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X}}{n}$$

and where $\|\boldsymbol{\Theta}\|_{\ell_1} = \sum |\Theta_{i,j}|$.

See van Wieringen (2016) Undirected network reconstruction from high-dimensional data and https://github.com/kaizhang/glasso

Application to Network Simplification

Can be applied on networks, to spot 'significant' connexions...
Source: http://khughitt.github.io/graphical-lasso/



Extention of Penalization Techniques

In a more general context, we want to solve

$$\widehat{\boldsymbol{\theta}} \in \underset{\boldsymbol{\theta}\in\mathbb{R}^d}{\text{argmin}} \left\{ \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, m_{\boldsymbol{\theta}}(\boldsymbol{x}_i)) + \lambda \cdot \text{penalty}(\boldsymbol{\theta}) \right\}.$$