PHYSICAL HUMAN-ROBOT INTERACTION

Linear Methods for Regression

Riccardo Muradore

UNIVERSITÀ di VERONA
Dipartimento di INFORMATICA

Linear Methods for Regression

Gray-box Identification: the class of the model is known, the specific parameters are unknown

Identification of the parameters of the DC motor

Transfer function

$$P(s) = \frac{\hat{\omega}(s)}{\hat{V}(s)} = \frac{K_m}{(Js+b)(Ls+R)+K_mK_e} \overset{L=0}{\simeq} \frac{K_m}{JRs+bR+K_mK_e} = \frac{k}{\tau s+1}$$

Differential equation

$$\tau\dot{\omega}(t) + \omega(t) = kV(t)$$

How can we determine τ and k?

Re-writing the differential equation as

$$\frac{\tau}{k}\dot{\omega}(t) + \frac{1}{k}\omega(t) = V(t)$$

and

$$\begin{bmatrix} \dot{\omega}(t) & \omega(t) \end{bmatrix} \begin{bmatrix} \frac{\tau}{k} \\ \frac{1}{k} \end{bmatrix} = V(t).$$

Let

$$\begin{aligned} x &\triangleq \begin{bmatrix} \dot{\omega} & \omega \end{bmatrix} \\ y &\triangleq V \\ \theta &\triangleq \begin{bmatrix} \frac{\tau}{k} \\ \frac{1}{k} \end{bmatrix} \end{aligned}$$

then

$$y = x\theta$$

with θ vector of unknowns.

If N samples are available for x and y

$$\begin{bmatrix} y(1) \\ y(2) \\ \vdots \\ y(N) \end{bmatrix} = \begin{bmatrix} x(1) \\ x(2) \\ \vdots \\ x(N) \end{bmatrix} \theta$$

we end up with

$$Y = X\theta$$

In the DC motor case

$$Y = \begin{bmatrix} V(t_1) \\ V(t_2) \\ \vdots \\ V(t_N) \end{bmatrix}, \qquad X = \begin{bmatrix} \dot{\omega}(t_1) & \omega(t_1) \\ \dot{\omega}(t_2) & \omega(t_2) \\ \vdots & \\ \dot{\omega}(t_N) & \omega(t_N) \end{bmatrix}$$

Notation

Notation:

- $\mathbf{x} \in \mathbb{R}^m$ random variable ($\mathbf{x}_i \in \mathbb{R}$ is its i-th component)
- $x \in \mathbb{R}^m$ an observation of the random variable $\mathbf{x} \in \mathbb{R}^m$
- $X \in \mathbb{R}^{N \times m}$ a collection of N observations ($x_i \in \mathbb{R}^m$ is its i-th row)

Linear Model: (from now on $p = 1$)

Input: $\mathbf{x} \in \mathbb{R}^m, x \in \mathbb{R}^m, X \in \mathbb{R}^{N \times m}$
Output: $\mathbf{y} \in \mathbb{R}^1, y \in \mathbb{R}^1, Y \in \mathbb{R}^{N \times 1}$
Prediction: $\hat{\mathbf{y}} \in \mathbb{R}^1, \hat{y} \in \mathbb{R}^1, \hat{Y} \in \mathbb{R}^{N \times 1}$

$$y = f(x) = x\beta$$

where $\beta \in \mathbb{R}^m$

Prediction

$$\hat{y} = x\hat{\beta}$$

where $\hat{\beta} \in \mathbb{R}^m$ is the matrix of coefficients that we have to determine

Remark. If $p = 1$, the gradient $f'(x) = \nabla_x f(x) = \beta$ is a vector pointing in the steepest uphill direction

Least Squares

Let $X \in \mathbb{R}^{N\times m}$ and $Y \in \mathbb{R}^N$ a training set of data (collection of N pairs (x, y))
How to choice β?
First of all we have to introduce an index as a function of β.

Let $RSS(\beta)$ be the residual sum of squares

$$RSS(\beta) := \sum_{i=1}^{N} (y_i - x_i\beta)^T (y_i - x_i\beta) = (Y - X\beta)^T (Y - X\beta)$$

We search for

$$\hat{\beta} := \arg\min_{\beta} RSS(\beta)$$

Computing the first and second derivative we get the normal equations

$$\begin{aligned} \nabla_\beta RSS(\beta) &= -2X^T(Y - X\beta) \\ \nabla^2_{\beta\beta} RSS(\beta) &= 2X^T X \end{aligned}$$

Least Squares

If $X^T X$ is nonsingular (i.e. X has full column rank), the unique solution is given by the normal equations

$$\nabla_\beta RSS(\beta) = 0 \quad \Leftrightarrow \quad X^T(Y - X\beta) = 0$$

i.e.

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

and the prediction of y given a new value x is

$$\hat{y} = x\hat{\beta}$$

Observations:

- We assume that the underlying model is linear
- Statistics of x and y do not play any role (it seems ...)

Least Squares $p > 1$

Linear model

$$Y = XB + E$$

where $X \in \mathbb{R}^{N \times m}$, $Y \in \mathbb{R}^{N \times p}$, $E \in \mathbb{R}^{N \times p}$ and $B \in \mathbb{R}^{m \times p}$

The RSS takes the form

$$RSS(B) := \mathrm{trace}\{(Y - XB)^T (Y - XB)\}$$

and the least square estimation of B is written in the same way

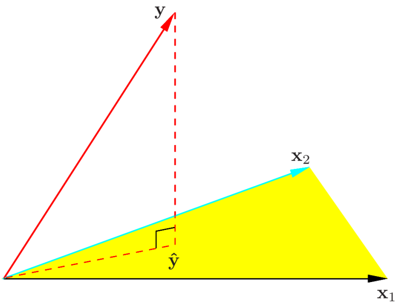
$$\hat{B} = (X^T X)^{-1} X^T Y$$

Multiple outputs do not affect one another's least squares estimates

If the component of the vector r.v $\mathbf{e}$ are correlated, i.e. $\mathbf{e} \sim \mathcal{N}(0, \Sigma)$, then we can define a weighted RSS

$$RSS(B, \Sigma) := \sum_{i=1}^{N} (Y_i - X_i B)^T \Sigma^{-1} (Y_i - X_i B)$$

Geometric interpretation



The normal equations

$$X^T(Y - X\beta) = 0$$

means the estimation $\hat{Y} = X\hat{\beta} = X(X^TX)^{-1}X^TY$ is the orthogonal projection of Y into the subspace X

Statistical interpretation

We now consider the r.v. $\mathbf{x}$ and $\mathbf{y}$ as input and output, respectively, and we seek a function $f(\mathbf{x})$ for predicting $\mathbf{y}$.

The criterion should be now deal with stochastic quantities: we introduce the expected squared prediction error EPE (strictly related with the mean squared error MSE)

$$\begin{aligned} EPE(f) &:= \mathbb{E}\left[(\mathbf{y}-f(\mathbf{x}))^T(\mathbf{y}-f(\mathbf{x}))\right] \\ &= \int_{S_x,S_y}(y-f(x))^T(y-f(x))p(x,y)dxdy \end{aligned}$$

where we implicitly assumed that $\mathbf{x}$ and $\mathbf{y}$ have a joint PDF. $EPE(f)$ is a $\mathcal{L}_2$ loss function

Conditioning on $\mathbf{x}$ we can re-write $EPE(f)$ as

$$EPE(f) := \mathbb{E}_x\left[\ \mathbb{E}_{y|x}\left[(\mathbf{y}-f(\mathbf{x}))^T(\mathbf{y}-f(\mathbf{x}))|\mathbf{x}\right]\ \right]$$

Statistical interpretation

We can determine $f(\cdot)$ pointwise

$$f(x) = \arg\min_{c} \mathbb{E}_{y|x}\left[(\mathbf{y}-c)^T(\mathbf{y}-c)|\mathbf{x}=x\right]$$

which means that

$$f(x) = \mathbb{E}\left[\mathbf{y}|\mathbf{x}=x\right]$$

i.e. the best $f(x)$ is the conditional mean (according to the *EPE* criterion).

Beautiful but, given the data X, Y how can we compute the conditional expectation?!?

Statistical interpretation

Let us assume again

$$f(\mathbf{x}) = \mathbf{x}^T \beta$$

then

$$EPE(f) \quad := \quad \mathbb{E}\left[(\mathbf{y} - \mathbf{x}^T\beta)^T(\mathbf{y} - \mathbf{x}^T\beta)\right]$$

Differentiating w.r.t. β we end up with

$$\beta = \left(\mathbb{E}[\mathbf{x}\mathbf{x}^T]\right)^{-1} \mathbb{E}[\mathbf{x}^T\mathbf{y}]$$

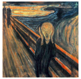
Computing the auto- and cross-correlation (i.e. using real numbers!)

$$\mathbb{E}[\mathbf{x}\mathbf{x}^T] \stackrel{N\to\infty}{\longrightarrow} S_{xx} := \frac{1}{N}\sum_{i=1}^{N} X_i^T X_i = \frac{1}{N} X^T X$$

$$\mathbb{E}[\mathbf{x}^T\mathbf{y}] \stackrel{N\to\infty}{\longrightarrow} S_{xy} := \frac{1}{N}\sum_{i=1}^{N} X_i Y_i^T = \frac{1}{N} X Y^T$$

Then we get

$$\hat{\beta} = \left(\frac{1}{N}X^T X\right)^{-1} \frac{1}{N} X Y^T$$
$$= (X^T X)^{-1} X Y^T$$



Again the normal equations !!!

But now we can provide a statistical interpretation of $\hat{\beta}$. Let $\mathbf{y} = \mathbf{x}^T\beta + \mathbf{e}$, $\mathbf{e} \sim \mathcal{N}(0, \sigma^2)$ be our model ($p = 1$), then $\hat{\beta}$ is a Gaussian variable

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

In fact, since $\hat{\beta} = (X^T X)^{-1} X\mathbf{y} - (X^T X)^{-1} X\mathbf{e}$

$$\hat{\mathbf{y}} = \mathbf{x}^T\hat{\beta} + \mathbf{e}$$

Given the linear model

$$y = x^T\beta, \qquad Y = X\beta$$

the least squares estimator $\hat{\phi}(x_0) = x_0^T\hat{\beta}$ of $\phi(x_0) = x_0^T\beta$ is unbiased because

$$\mathbb{E}[x_0^T\hat{\beta}] = x_0^T\beta$$

Theorem

If $\bar{\phi}(x_0)$ is any other unbiased estimation ($\mathbb{E}[\bar{\phi}(x_0)] = x_0^T\beta$) then

$$\mathrm{Var}(\hat{\phi}(x_0)) \leq \mathrm{Var}(\bar{\phi}(x_0))$$

Remark. Mean square error of a generic estimator $\bar{\phi}$ ($p = 1$)

$$MSE(\bar{\phi}) = \mathbb{E}[(\bar{\phi} - \phi)^2] \overset{(*)}{=} \underbrace{\mathrm{Var}(\bar{\phi})}_{\text{variance}} + \underbrace{(\mathbb{E}[\bar{\phi}] - \phi)^2}_{\text{bias}}$$

(*) = sum and subtract $\mathbb{E}[\bar{\phi}]$.

Gauss-Markov theorem

Given the stochastic linear model

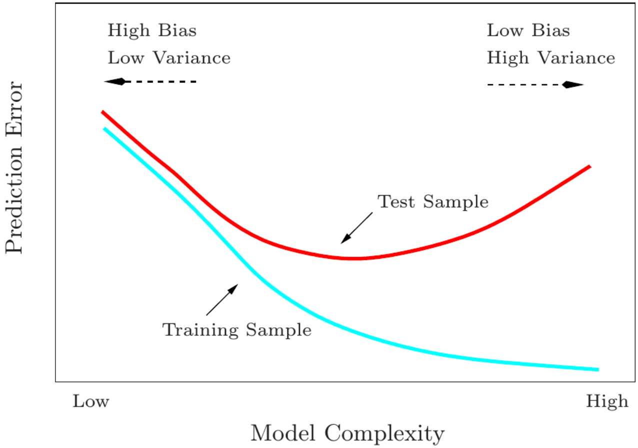
$$\mathbf{y} = \mathbf{x}^T \beta + \mathbf{e}, \qquad \mathbf{e} \sim \mathcal{N}(0, \sigma^2)$$

and let $\bar{\phi}(x_0)$ be the estimator for $y_0 = \phi(x_0) + e_0$, $\phi(x_0) = x_0^T \beta$.

The expected prediction error (EPE) of $\bar{\phi}(x_0)$ is

$$\begin{aligned} EPE(\bar{\phi}(x_0)) &= \mathbb{E}[(y_0 - \bar{\phi}(x_0))^2] \\ &= \sigma^2 + \mathbb{E}[(x_0^T \beta - \bar{\phi}(x_0))^2] \\ &= \sigma^2 + \underbrace{\mathrm{Var}(\bar{\phi}) + (\mathbb{E}[\bar{\phi}] - \phi)^2}_{MSE} \end{aligned}$$

underfitting VS overfitting



Recursive Least Square

From batch solution to recursive solution

If $X^T X$ is nonsingular (i.e. X has full column rank), the unique solution is given by the normal equations

$$\nabla_\beta RSS(\beta) = 0 \quad \Leftrightarrow \quad X^T(Y - X\beta) = 0$$

i.e.

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

where

$$X^T X = \sum_{i=1}^{N} x_i^T x_i \in \mathbb{R}^{m \times m}, \qquad X^T Y = \sum_{i=1}^{N} x_i^T y_i \in \mathbb{R}^{m \times 1}$$

Then

$$\hat{\beta}(N) = \left(\sum_{i=1}^{N} x_i^T x_i \right)^{-1} \sum_{i=1}^{N} x_i^T y_i$$

From batch solution to recursive solution

Let $P(k)$ be the sum till the k samples

$$P(k) = \left(\sum_{i=1}^{k} x_i^T x_i \right)^{-1}$$

We have

$$P^{-1}(k) = P^{-1}(k-1) + x_k^T x_k$$

The optimal estimation of β with k samples is

$$\begin{aligned}
\hat{\beta}(k) &= \left(\sum_{i=1}^{k} x_i^T x_i \right)^{-1} \sum_{i=1}^{k} x_i^T y_i \\
&= P(k) \sum_{i=1}^{k} x_i^T y_i \\
&= P(k) \left[\sum_{i=1}^{k-1} x_i^T y_i + x_k^T y_k \right] \qquad (1)
\end{aligned}$$

Form

$$\begin{aligned}\hat{\beta}(k-1) &= \left(\sum_{i=1}^{k-1} x_i^T x_i\right)^{-1} \sum_{i=1}^{k-1} x_i^T y_i \\ &= P(k-1)\sum_{i=1}^{k-1} x_i^T y_i\end{aligned}$$

it is possible to derive

$$\sum_{i=1}^{k-1} x_i^T y_i = P^{-1}(k-1)\hat{\beta}(k-1)$$

Substituting in (1)

$$\begin{aligned}\hat{\beta}(k) &= P(k)\left[\sum_{i=1}^{k-1} x_i^T y_i + x_k^T y_k\right] \\ &= P(k)\left[P^{-1}(k-1)\hat{\beta}(k-1) + x_k^T y_k\right] \qquad (2)\end{aligned}$$

From batch solution to recursive solution

By using

$$P^{-1}(k) = P^{-1}(k-1) + x_k^T x_k$$

we have

$$\begin{aligned}
\hat{\beta}(k) &= P(k)\left[P^{-1}(k-1)\hat{\beta}(k-1) + x_k^T y_k\right] \\
&= P(k)\left[\left(P^{-1}(k) - x_k^T x_k\right)\hat{\beta}(k-1) + x_k^T y_k\right] \\
&= P(k)\left[P^{-1}(k)\hat{\beta}(k-1) - x_k^T x_k \hat{\beta}(k-1) + x_k^T y_k\right] \\
&= \hat{\beta}(k-1) - P(k)x_k^T x_k \hat{\beta}(k-1) + P(k)x_k^T y_k \\
&= \hat{\beta}(k-1) + \underbrace{P(k)x_k^T}_{K(k)} \underbrace{\left(y_k - x_k\hat{\beta}(k-1)\right)}_{e(k)}
\end{aligned}$$

Finally

$$\hat{\beta}(k) = \hat{\beta}(k-1) + K(k)e(k)$$

with

$$\begin{aligned} K(k) &= P(k)x_k^T \\ e(k) &= y_k - x_k\hat{\beta}(k-1) \end{aligned}$$

and

$$P^{-1}(k) = P^{-1}(k-1) + x_k^T x_k$$

Problem: at each step a matrix inversion is needed!

Using the inversion matrix lemma

$$(A + BCD)^{-1} = A^{-1} - A^{-1}B(C^{-1} + DA^{-1}B)^{-1}DA^{-1}$$

it is possible to obtain

$$\begin{aligned} P(k) &= \left(P^{-1}(k-1) + x_k^T x_k\right)^{-1} \\ &= P(k-1) - \frac{P(k-1)x_k^T x_k P(k-1)}{1 + x_k P(k-1) x_k^T} \end{aligned}$$

with

$$\begin{aligned} A &\leftrightarrow P^{-1}(k-1) \\ B &\leftrightarrow x^T(k) \\ C &\leftrightarrow 1 \\ D &\leftrightarrow x(k) \end{aligned}$$

$$
\begin{aligned}
\hat{\beta}(k) &= \hat{\beta}(k-1) + K(k)e(k) \\
K(k) &= P(k)x_k^T \\
e(k) &= y_k - x_k\hat{\beta}(k-1) \\
P(k) &= P(k-1) - \frac{P(k-1)x_k^T x_k P(k-1)}{1 + x_k P(k-1)x_k^T}
\end{aligned}
$$

RLS with forgetting factor

What happens if we need to estimate time-varying parameters? We should weight differently the most recent measurements

$$RSS_\lambda(\beta(k)) := \sum_{i=1}^{k} \lambda^{k-i}(y_i - x_i\beta)^T(y_i - x_i\beta)$$

with $0 < \lambda \leq 1$.
Following the same kind of reasoning we have

$$\begin{aligned}
\hat{\beta}(k) &= \hat{\beta}(k-1) + K(k)e(k) \\
K(k) &= P(k)x_k^T \\
e(k) &= y_k - x_k\hat{\beta}(k-1) \\
P(k) &= \frac{1}{\lambda}\left[P(k-1) - \frac{P(k-1)x_k^T x_k P(k-1)}{\lambda + x_k P(k-1)x_k^T}\right]
\end{aligned}$$


Project

To do

- Identify the parameters k and τ (i.e. J and D) using the LS and the RLS on the DC motors data.

Adaptive Algorithm

N.B. continuous-time systems

Let

$$e(t) = y(t) - x(t)\beta$$

be the error.
To minimize the squared error

$$e^2(t) = (y(t) - x(t)\beta)^2$$

we can compute the gradient

$$\begin{aligned} \frac{\partial e^2(t)}{\partial \beta} &= -2x^T(t)\,(y(t) - x(t)\beta) \\ &= -2x^T(t)e(t) \end{aligned}$$

and move in the opposite direction

$$\dot{\beta} = gx^T(t)e(t)$$

with $g > 0$.

Regularization Methods

Statistical models

Statistical model:

$$\mathbf{y} = f(\mathbf{x}) + \mathbf{e}$$

where $\mathbf{y}$ is a random error with zero mean ($\mathbb{E}[\mathbf{e}] = 0$) and is independent of $\mathbf{x}$.

This means that the relationship between $\mathbf{y}$ and $\mathbf{x}$ is not deterministic ($f(\cdot)$)

The additive r.v. $\mathbf{e}$ takes care of measurement noise, model uncertainty and non measured variables correlated with $\mathbf{y}$ as well

We often assume that the random variables $\mathbf{e}$ are independent and identically distributed (i.i.d.)

Statistical models

Assuming a linear basis expansion for $f_\theta(x)$ parametrized by the unknowns collected within the vector θ

$$f_\theta(x) = \sum_{1}^{K} h_k(x)\theta_k$$

where examples of $h_k(x)$ can be

$$\begin{aligned} h_k(x) &= x_k \\ h_k(x) &= (x_k)^2 \\ h_k(x) &= \sin(x_k) \\ h_k(x) &= \frac{1}{1+e^{-x^T\beta_k}} \end{aligned}$$

The optimization problem to solve is

$$\hat{\theta} = \arg\min_{\theta\in\Theta} RSS(\theta) = \sum_{1}^{N} (y_i - f_\theta(x_i))^2$$

where RSS stands for Residual Sum of Squares

Are there other kinds of criterion besides RSS, EPE?

YES, A more general principle for estimation is maximum likelihood estimation

Let $p_\theta(y)$ be the PDF of the samples $y_1, \ldots, y_N$

The log-probability (or log-likelihood) of the observed samples is

$$L(\theta) = \sum_1^N \log p_\theta(y_i)$$

Principle of maximum likelihood: the most reasonable values for θ are those for which the probability of the observed samples is largest

Statistical models

If the error $\mathbf{e}$ in the following statistical model

$$\mathbf{y} = f_\theta(\mathbf{x}) + \mathbf{e}$$

is Gaussian, $\mathbf{e} \sim \mathcal{N}(0, \sigma^2)$, then the conditional probability is

$$p(y|x, \theta) \sim \mathcal{N}(f_\theta(x), \sigma^2)$$

Then log-likelihood of the data is

$$\begin{aligned} L(\theta) &= \sum_{1}^{N} \log p(y_i | f_\theta(x_i), \theta) \\ &= -\frac{N}{2} \log(2\pi) - N \log \sigma - \frac{1}{2\sigma^2} \sum_{i=1}^{N} (y_i - f_\theta(x_i))^2 \end{aligned}$$

Least squares for the additive error model is equivalent to maximum likelihood using the conditional probability (The yellow is the $RSS(\theta)$)

Penalty function, or regularization methods, introduces our knowledge about the type of functions $f(x)$ we are looking for

$$PRSS(f, \lambda) := RSS(f) + \lambda g(f)$$

where the functional $g(f)$ will force our knowledge (or desiderata) on f

Example. One-dimension cubic smoothing spline is the solution of

$$PRSS(f, \lambda) := \sum_{i=1}^{N} (y_i - f(x_i))^2 + \lambda \int [f''(s)]^2 dx$$

Remark. Penalty function methods have a Bayesian interpretation:

- $g(f)$ is the log-prior distribution
- $PRSS(f, \lambda)$ is the log-posterior distribution
- the solution of $\arg\min_f PRSS(f, \lambda)$ is the posterior mode

If we want a local regression estimation of $f(x_0)$, we have to solve the problem

$$\hat{\theta} = \arg\min_{\theta} RSS(f_\theta, x_0) = \sum_{i=1}^{N} K_\lambda(x_0, x_i)(y_i - f_\theta(x_i))^2$$

where the kernel function $K_\lambda(x_0, x)$ weights the point x around x_0. The optimal estimation is $f_{\hat{\theta}}(x_0)$

An example of kernel function is the Gaussian kernel

$$K_\lambda(x_0, x) = \frac{1}{\lambda} \exp\left[-\frac{\|x - x_0\|^2}{2\lambda}\right]$$

Examples of $f_\theta(x)$ are

- $f_\theta(x) = \theta_0$, constant function
- $f_\theta(x) = \theta_0 + \theta_1 x$, linear regression

The function f can be approximated using a set of M basis functions h_m

$$f_\theta(x) = \sum_{m=1}^{M} \theta_m h_m(x)$$

where $\theta = [\theta_1 \quad \cdots \quad \theta_M]$

Examples of basis functions:

- Radial basis functions:

$$f_\theta(x) = \sum_{m=1}^{M} \theta_m K_{\lambda_m}(\mu_m, x), \qquad K_\lambda(\mu, x) = e^{-\|x-\mu\|^2/2\lambda}$$

- Single-layer feed-forward neural network

$$f_\theta(x) = \sum_{m=1}^{M} \theta_m \sigma(\alpha_m^T x + b_m), \qquad \sigma(x) = \frac{1}{1+e^{-x}}$$

Remark. Linear methods can then be used with nonlinear input-output transformation because the model is linear in the parameters θ

"The least squares estimates often have low bias but large variance. Prediction accuracy can sometimes be improved by shrinking or setting some coefficients to zero. By doing so we sacrifice a little bit of bias to reduce the variance of the predicted values, and hence may improve the overall prediction accuracy."

Ridge Regression

Ridge regression shrinks the regression coefficients by imposing a penalty on their size. The coefficients $\hat{\beta}^{ridge}$ are obtained solving the minimization problem

$$\hat{\beta}^{ridge} = \arg\min_{\beta}\{\underbrace{\sum_{i=1}^{N}(Y_i - X_i\beta)^T(Y_i - X_i\beta)}_{RSS(\beta)} + \lambda \underbrace{\sum_{i=1}^{m}\beta_i^2}_{g(\beta)=\beta^T\beta}\}$$

with $\lambda \geq 0$, or the equivalent constrained problem

$$\begin{aligned}\hat{\beta}^{ridge} = \quad & \arg\min_{\beta} \quad \sum_{i=1}^{N}(Y_i - X_i\beta)^T(Y_i - X_i\beta) \\ & \text{s. to} \quad \sum_{i=1}^{m}\beta_i^2 \leq t\end{aligned}$$

The solution is

$$\hat{\beta}^{ridge} = (X^TX + \lambda I)^{-1}X^TY$$

The coefficients $\hat{\beta}^{lasso}$ are obtained solving the minimization problem

$$\hat{\beta}^{lasso} = \arg\min_{\beta} \{ \underbrace{\sum_{i=1}^{N} (Y_i - X_i\beta)^T (Y_i - X_i\beta)}_{RSS(\beta)} + \lambda \underbrace{\sum_{i=1}^{m} |\beta_i| \}}_{g(\beta)}$$

with $\lambda \geq 0$, or the equivalent constrained problem

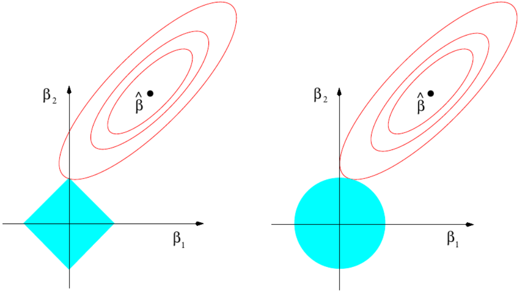
$$\begin{aligned} \hat{\beta}^{lasso} = \quad & \arg\min_{\beta} \quad \sum_{i=1}^{N} (Y_i - X_i\beta)^T (Y_i - X_i\beta) \\ & \text{s. to} \quad \sum_{i=1}^{m} |\beta_i| \leq t \end{aligned}$$

The are no closed form expression for $\hat{\beta}^{lasso}$

Remark 1. The Ridge Regression uses a $\mathcal{L}_2$ norm on β, whereas Lasso the $\mathcal{L}_1$ norm. This means that the solution is nonlinear in the data.

Remark 1. Decreasing t forces some of the coefficients to be set to zero (exactly).

Ridge regression VS Lasso



Lasso	Ridge
$|\beta_1| + |\beta_2| \leq t$	$\beta_1^2 + \beta_2^2 \leq t^2$

The red ellipses are the contours of the least squares error function