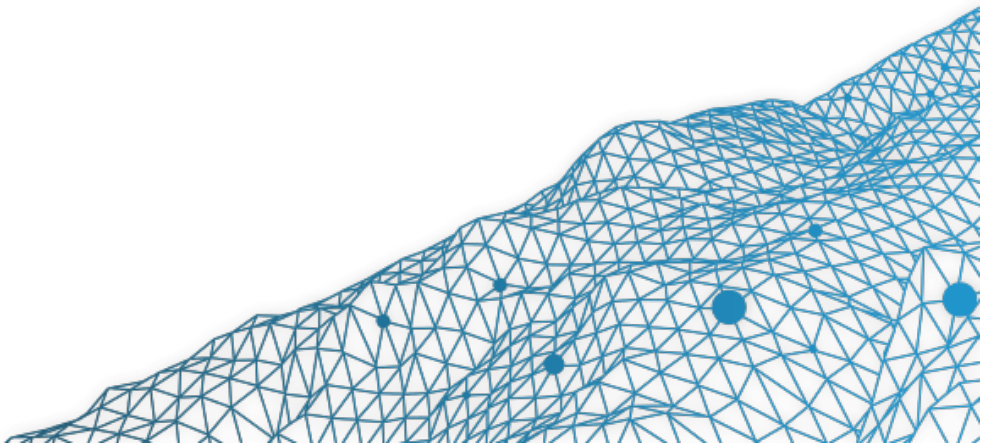
1 Introduction & Geometry of Least Squares

Arthur Charpentier (Université du Québec à Montréal)

Machine Learning & Econometrics

SIDE Summer School - July 2019



Preamble : complex does not mean better...

Econometrics

Machine Learning

Econometrics Courses in a Nutshell

Haavelno (1944, The Probabilistic Approach in Econometrics)

THE PROBABILITY APPROACH
IN ECONOMETRICS

By

Trygve Haavelmo

RESEARCH ASSOCIATE
COWLES COMMISSION FOR
RESEARCH IN ECONOMICS

Supplement to Econometrica, Volume 12, July, 1944

THE ECONOMETRIC SOCIETY
The University of Chicago
Chicago 37, Illinois

CHAPTER III

Stochastical Schemes as a Basis for Econometrics

As far as is known, the scheme of probability and random variables is, at least for the time being, the only scheme suitable for formulating such theories. We may have objections to using this scheme, but among these objections there is at least one that can be safely dismissed, viz., the objection that the scheme of probability and random variables is not general enough for application to economic data. Since, however, this is apparently not commonly accepted by economists we find ourselves justified in starting our discussion in this chapter with a brief outline of the modern theory of stochastical variables, with particular emphasis on certain points that seem relevant to economics.

The more recent developments in statistical theory are based upon the so-called modernized classical theory of probability. Here "probability" is defined as an absolutely additive and nonnegative *set-function*,[1] satisfying certain formal properties.[2]

Let us first take an example to illustrate this probability concept.

Data $(y_i, \boldsymbol{x}_i)$ are seen as realizations of (iid) random varibles $(Y, \boldsymbol{X})$ on some probabilistic space $(\Omega, \mathcal{F}, \mathbb{P})$.

The Linear Regression

Consider some sample $\mathcal{S}_n = \{(y_i, \boldsymbol{x}_i)\}$ where $y_i \in \mathbb{R}$ and $\boldsymbol{x}_i \in \mathbb{R}^p$.

$\mathcal{S}_n$ is the realization of n i.i.d. random vectors $(Y_i, X_{1,i}, \cdots, X_{p,i})$ with unknown distribution $\mathbb{P}$

Assume that $Y_i = \beta_0 + \beta_1 x_{1,i} + \cdots + \beta_p x_{p,i} + \varepsilon_i$, or $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$

where ε_i's satisfy $\mathbb{E}[\varepsilon_i] = 0$, $\text{Var}[\varepsilon_i] = \sigma^2$, $\text{Var}[\varepsilon_i, \varepsilon_j] = 0$

Least square estimator of (unknown) $\boldsymbol{\beta}$ is

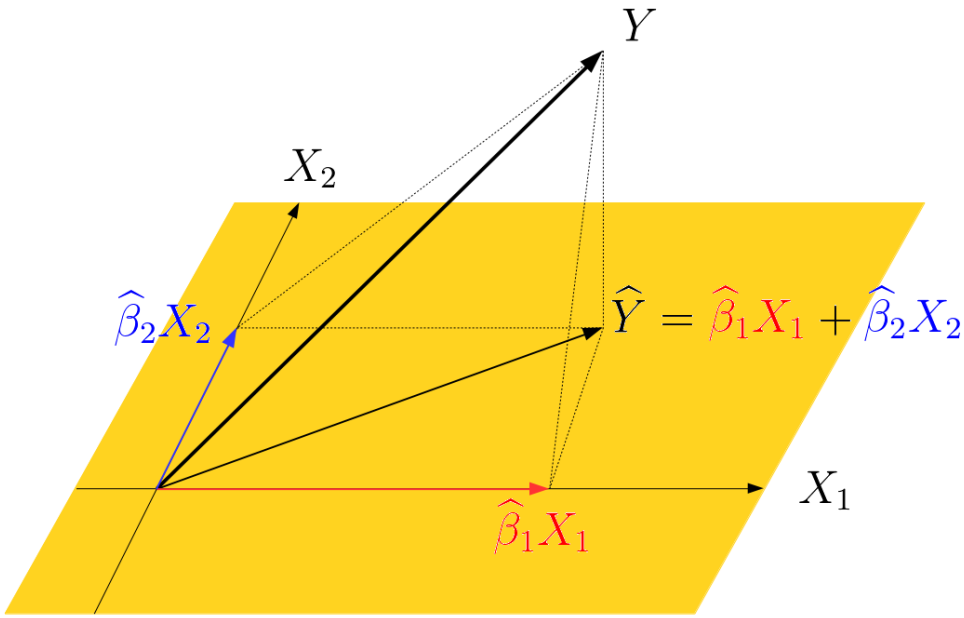
$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^{p+1}}{\text{argmin}} \left\{ \left\| \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} \right\|^2 \right\} = \left(\boldsymbol{X}^\top \boldsymbol{X} \right)^{-1} \boldsymbol{X}^\top \boldsymbol{y}$$

Geometric Perspective

Define the orthogonal projection on $\mathcal{X}$,

$$\Pi_{\boldsymbol{X}} = \boldsymbol{X}[\boldsymbol{X}^\mathsf{T}\boldsymbol{X}]^{-1}\boldsymbol{X}^\mathsf{T}$$

$$\widehat{\boldsymbol{y}} = \underbrace{\boldsymbol{X}[\boldsymbol{X}^\mathsf{T}\boldsymbol{X}]^{-1}\boldsymbol{X}^\top}_{\Pi_{\boldsymbol{X}}}\boldsymbol{y} = \Pi_{\boldsymbol{X}}\boldsymbol{y}.$$



Pythagoras' theorem can be written

$$\|\boldsymbol{y}\|^2 = \|\Pi_{\boldsymbol{X}}\boldsymbol{y}\|^2 + \|\Pi_{\boldsymbol{X}^\perp}\boldsymbol{y}\|^2 = \|\Pi_{\boldsymbol{X}}\boldsymbol{y}\|^2 + \|\boldsymbol{y} - \Pi_{\boldsymbol{X}}\boldsymbol{y}\|^2$$

which can be expressed as

$$\underbrace{\sum_{i=1}^n y_i^2}_{n\times\text{total variance}} = \underbrace{\sum_{i=1}^n \widehat{y}_i^2}_{n\times\text{explained variance}} + \underbrace{\sum_{i=1}^n (y_i - \widehat{y}_i)^2}_{n\times\text{residual variance}}$$

Geometric Perspective

Define the angle θ between $\boldsymbol{y}$ and $\Pi_{\boldsymbol{X}}\boldsymbol{y}$,

$$R^2 = \frac{\|\Pi_{\boldsymbol{X}}\boldsymbol{y}\|^2}{\|\boldsymbol{y}\|^2} = 1 - \frac{\|\Pi_{\boldsymbol{X}^\perp}\boldsymbol{y}\|^2}{\|\boldsymbol{y}\|^2} = \cos^2(\theta)$$

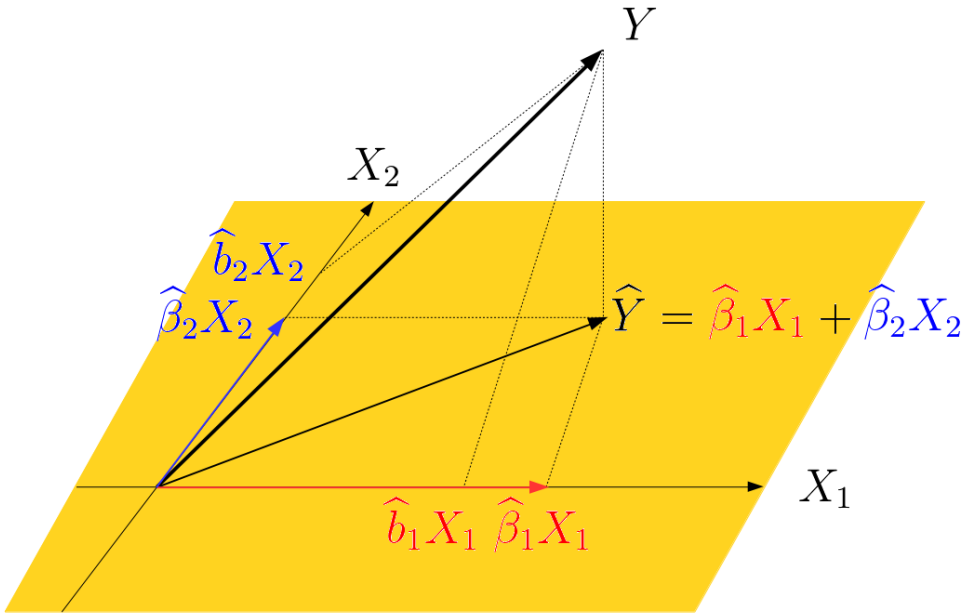
see Davidson & MacKinnon (2003)

$$\boldsymbol{y} = \beta_0 + \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \varepsilon$$

If $\boldsymbol{y}_2^\star = \Pi_{\boldsymbol{X}_1^\perp}\boldsymbol{y}$ and $\boldsymbol{X}_2^\star = \Pi_{\boldsymbol{X}_1^\perp}\boldsymbol{X}_2$, then

$$\widehat{\boldsymbol{\beta}}_2 = [\boldsymbol{X}_2^{\star\mathsf{T}}\boldsymbol{X}_2^\star]^{-1}\boldsymbol{X}_2^{\star\mathsf{T}}\boldsymbol{y}_2^\star$$

$\boldsymbol{X}_2^\star = \boldsymbol{X}_2$ if $\boldsymbol{X}_1 \perp \boldsymbol{X}_2$,
Frisch-Waugh theorem.



Least Squares ? The probabilistic interpretation

Recall that we assumed that $(Y_i, \boldsymbol{X}_i)$ were i.i.d. with unknown distribution $\mathbb{P}$.

In the (standard) linear model, we assume that $(Y|\boldsymbol{X}=\boldsymbol{x}) \sim \mathcal{N}(\boldsymbol{x}^\top\boldsymbol{\beta}, \sigma^2)$.

The maximum-likelihood estimator of $\boldsymbol{\beta}$ is then

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^{p+1}}{\text{argmax}} \left\{ -\sum_{i=1}^{n} (y_i - \boldsymbol{x}_i^\top\boldsymbol{\beta})^2 \right\}$$

For the logistic regression, $(Y|\boldsymbol{X}=\boldsymbol{x}) \sim \mathcal{B}(\text{logit}^{-1}(\boldsymbol{x}^\top\boldsymbol{\beta}))$.

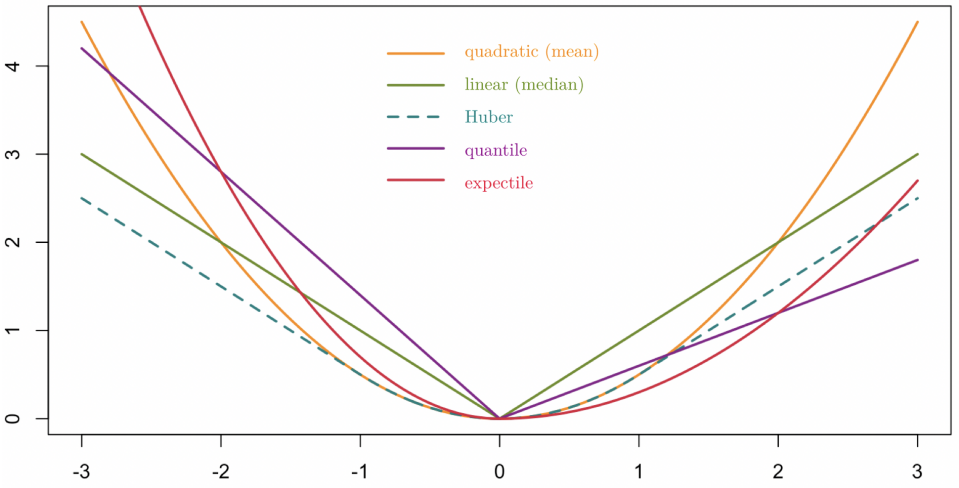
The maximum-likelihood estimator of $\boldsymbol{\beta}$ is then

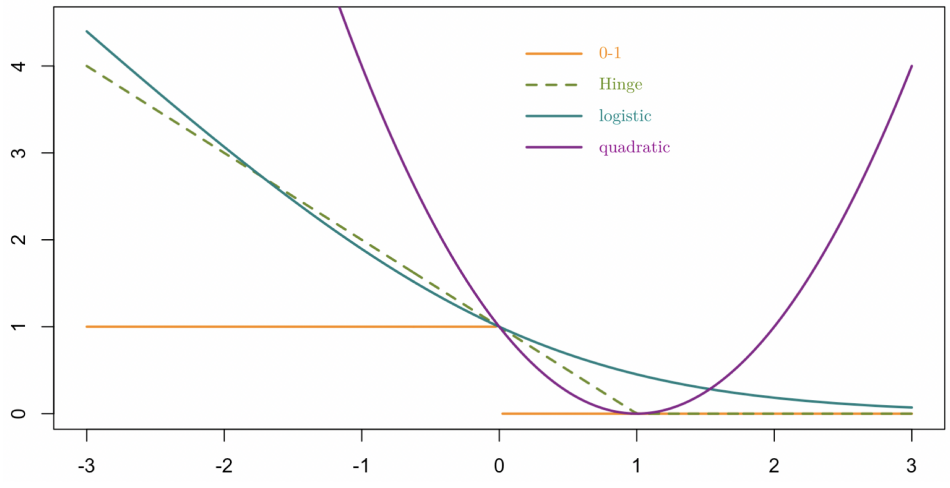
$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^{p+1}}{\text{argmin}} \left\{ \sum_{i=1}^{n} \left(y_i \cdot \log(\text{logit}^{-1}(\boldsymbol{x}_i^\top\boldsymbol{\beta})) + (1-y_i) \cdot \log(1-\text{logit}^{-1}(\boldsymbol{x}_i^\top\boldsymbol{\beta})) \right) \right\}$$

Least Squares ?

Consider a linear model, $y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \varepsilon_i$.

In a general setting, consider some loss function $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$, e.g. $\ell(y, m) = (y - m)^2$ (quadratic ℓ_2 norm)





For inference, we try to minimize the empirical risk associated with ℓ,

$$\widehat{\boldsymbol{\beta}} \in \underset{\boldsymbol{\beta} \in \mathbb{R}^{p+1}}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} \ell(y_i, \boldsymbol{x}_i^\top \boldsymbol{\beta}) \right\}$$

where $\boldsymbol{x}_i = (1, x_{1,i}, x_{2,i}, \cdots, x_{p,i})$. But cannot be used for validation.

Least Squares and Quadratic Risk

The risk associated with model m is $\mathcal{R}_{\mathbb{P}}(m) = \mathbb{E}_{\mathbb{P}}\big[\ell(Y, m(\boldsymbol{X}))\big]$. Hence

$$\mathbb{E}\big[Y|\boldsymbol{X}=\boldsymbol{x}\big] = \text{argmin}\left\{\mathcal{R}_{\mathbb{P}}(m) = \mathbb{E}_{\mathbb{P}}\big[\ell_2(Y, m(\boldsymbol{X}))\big]\right\}$$

where $\ell_2(y, m) = (y-m)^2$, so $m^\star$ depends on (unknown) $\mathbb{P}$.

Average Risks

The average risk associated with $\widehat{m}_n$ is $\mathcal{R}_{\mathbb{P}}(\widehat{m}_n) = \mathbb{E}_{\mathbb{P}}\big[\ell(Y, \widehat{m}_n(\boldsymbol{X}))\big]$

Empirical Risks

The empirical risk associated with $\widehat{m}_n$ is $\widehat{\mathcal{R}}_n(\widehat{m}_n) = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} \ell\big(y_i, \widehat{m}_n(\boldsymbol{x}_i)\big)$.

If we minimize the average risk, we overfit...

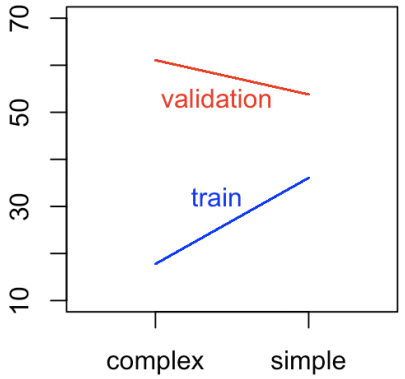
Cross Validation

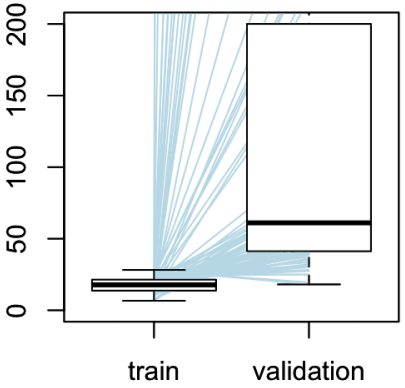
Classical Approach : split the sample $\mathcal{S}_n$ in two parts

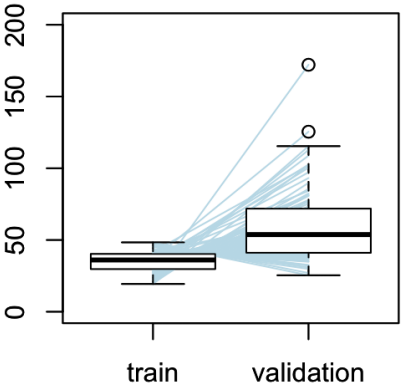
Hold-Out Cross Validation

1. Split $\{1, 2, \cdots, n\}$ in T (training) and V (validation)
2. Estimate $\widehat{m}$ on sample $(y_i, \boldsymbol{x}_i)$, $i \in T$: $\widehat{m}_T$
3. Compute $\dfrac{1}{|V|} \displaystyle\sum_{i \in V} \ell\big(y_i, \widehat{m}_T(x_{1,i}, \cdots, X_{p,i})\big)$

```
chicago <- read.table("http://freakonometrics.free.fr/
    chicago.txt", header=TRUE,sep=";")
idx <- sample(1:nrow(chicago),nrow(chicago)*.7)
train <- chicago[idx,]
valid <- chicago[-idx,]
```



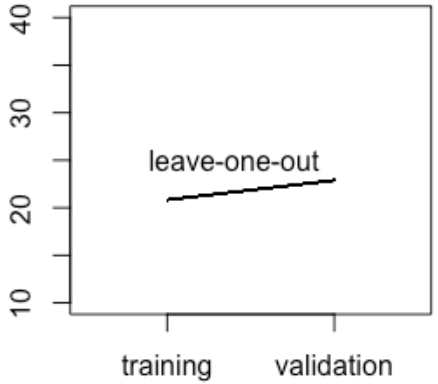




Cross Validation

Leave-one-Out Cross Validation

1. Estimate n models : $\widehat{m}_{-j}$ on sample $(y_i, \boldsymbol{x}_i)$, $i \in \{1, \cdots, n\} \backslash \{j\}$
2. Compute $\dfrac{1}{n}\sum_{i=1}^{n} \ell\big(y_i, \widehat{m}_{-i}(\boldsymbol{x}_i)\big)$



Can be computationally intensive...

In the case of a linear regression, there is a simple formula to compute $\widehat{\boldsymbol{\beta}}_{-j}$ when observation j is removed. Let $\boldsymbol{H} = \Pi_{\boldsymbol{X}} = \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top$

$$(\boldsymbol{X}_{(j)}^\top\boldsymbol{X}_{(j)})^{-1}\boldsymbol{X}_{(j)}^\top = \frac{1}{1-H_{j,j}}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}_{(j)}^\top$$

Cross Validation

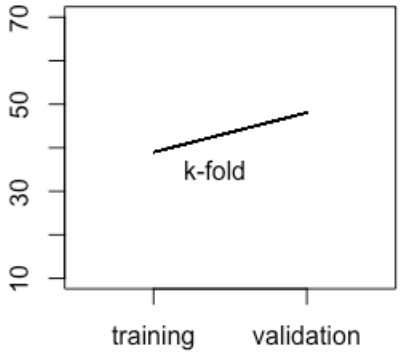
Instead of fitting n models, fit only K

K-Fold Cross Validation

1. Split $\{1, 2, \cdots, n\}$ in K groups $V_1, \cdots, V_K$
2. Estimate K models : $\widehat{m}_k$ on sample $(y_i, \boldsymbol{x}_i)$, $i \in \{1, \cdots, n\} \backslash V_k$
3. Compute $\dfrac{1}{K} \displaystyle\sum_{k=1}^{K} \frac{1}{|V_k|} \sum_{i \in V_k} \ell\big(y_i, \widehat{m}_k(x_{1,i}, \cdots, X_{p,i})\big)$

If $K = 10$ we fit on 90% of the observation, and validate on the remaining 10%

Here the K groups should be created randomly...



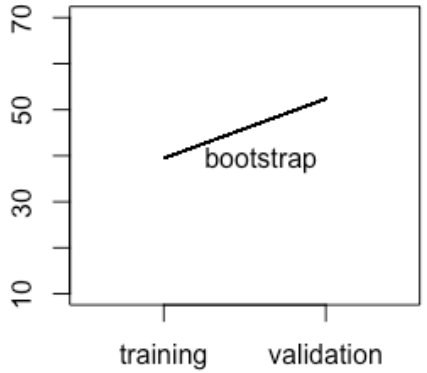
Cross Validation

Bootstrap Cross Validation

1. Generate B boostrap samples from $\{1, 2, \cdots, n\}$, $I_1, \cdots, I_B$
2. Estimate B models : $\widehat{m}_b$ on sample $(y_i, \boldsymbol{x}_i)$, $i \in I_b$
3. Compute $\dfrac{1}{B}\sum_{b=1}^{B}\dfrac{1}{n-|I_b|}\sum_{i\notin I_b}\ell\big(y_i, \widehat{m}_b(x_{1,i}, \cdots, X_{p,i})\big)$

The probability that $i \notin I_b$ is $\left(1-\dfrac{1}{n}\right)^n \sim e^{-1} (= 36.78\%)$

At stage b, we validate on $\sim 36.78\%$ of the dataset.



Norms, Inner Products and Kernels

Most properties are related to the geometry the Euclidean space of $\mathbb{R}^n$.

Inner Product

An inner product on a vector space $\mathcal{H}$ is the application $(f,g) \mapsto \langle f,g\rangle_{\mathcal{H}}$ (taking value in $\mathbb{R}$) bilinear, symmetric, definite positive:

- $\langle f,g\rangle_{\mathcal{H}} = \langle g,f\rangle_{\mathcal{H}}$
- $\langle \alpha f+\beta g,h\rangle_{\mathcal{H}} = \alpha\langle f,h\rangle_{\mathcal{H}} + \beta\langle g,h\rangle_{\mathcal{H}}$
- $\langle f,f\rangle_{\mathcal{H}} \geq 0$ and $\langle f,f\rangle_{\mathcal{H}} = 0$ if and only if $f=0$.

Example : $\mathcal{H} = \mathbb{R}^n$, $\langle \boldsymbol{x},\boldsymbol{y}\rangle = \boldsymbol{x}^\top \boldsymbol{y}$

Example : $\mathcal{H} = \ell_2 = \left\{ u : \sum_{i=1}^{\infty} u_i^2 < \infty \right\}$, $\langle u,v\rangle = \sum_{i=1}^{\infty} u_i v_i$

Example : $\mathcal{H} = L_2(\mu) = \left\{ f : \int f(x)^2 d\mu(x) < \infty \right\}$, $\langle f,g\rangle = \int f(x)g(x)d\mu(x)$

Norms, Inner Products and Kernels

If $\mathcal{H}$ is finite, $\mathcal{H} = \{h_1 \cdots, h_d\}$, $\langle x, y\rangle$ take value $K_{i,j}$ if $x = h_i$ and $y = h_j$. Let $\boldsymbol{K} = [K_{i,j}]$

$\boldsymbol{K}$ is a symmetric $d \times d$ matrix, $\boldsymbol{K} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^\top$ for some orthogonal matrix $\boldsymbol{V}$ where columns are eigenvectors, and $\boldsymbol{\Lambda} = \text{diag}[\lambda_i]$ (positive values). Let

$$\Phi(x) = \left(\sqrt{\lambda_1}V_{i,1}, \sqrt{\lambda_2}V_{2,i}, \cdots, \sqrt{\lambda_d}V_{d,i}\right) \text{ if } x = h_i$$

Note that

$$K_{i,j} = [\boldsymbol{K} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^\top]_{i,j} = \sum_{l=1}^{d} \lambda_l V_{i,l} V_{l,j} = \langle \Phi(h_i), \Phi(h_j)\rangle$$

Matrix K defines an inner product, it is called a kernel. It is symmetric, associated with a positive semi-definite matrix.

Then $K(u,u) \geq 0$ and $K(u,v) \leq \sqrt{K(u,u) \cdot K(v,v)}$.

Norms, Inner Products and Kernels

Let $\varphi : u \mapsto K(\cdot, u)$, then $K(x, y) = \langle \varphi(x), \varphi(y) \rangle$

In a general setting, let $\|f\|_{\mathcal{H}} = \langle f, f \rangle_{\mathcal{H}}^{1/2}$, and define the distance from f to $\mathcal{G} \subset \mathcal{H}$

$$d(f, \mathcal{G}) = \inf_{g \in \mathcal{G}} \{\|f - g\|_{\mathcal{H}}\} = d(f, g^\star) \text{ where } g^\star \in \mathcal{G}$$

Note that $\langle g, f - g^\star \rangle_{\mathcal{H}} = 0$, $\forall g \in \mathcal{G}$. And $\mathcal{H} = \mathcal{G} \oplus \mathcal{G}^\perp$.

Riesz representation theorem

For any continuous linear functionals L from $\mathcal{H}$ into the field $\mathbb{R}$, there exists a unique $g_L \in \mathcal{H}$ such that $\forall f \in \mathcal{H}$, $\langle g_L, f \rangle_{\mathcal{H}} = Lf$.

Consider the case where $\mathcal{H} = \mathbb{R}^n$. Let $\boldsymbol{\Sigma}$ denote some symmetric $n \times n$ positive definite matrix. Then

$$\langle \boldsymbol{f}, \boldsymbol{g} \rangle_{\boldsymbol{\Sigma}} = \boldsymbol{f}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{g} s \text{ is an inner product on } \mathbb{R}^n.$$

Norms, Inner Products and Kernels

Note that if $\boldsymbol{\sigma}_i$ denote columns of $\boldsymbol{\Sigma}$ $\langle \boldsymbol{\sigma}_i, \boldsymbol{\sigma}_j \rangle_{\boldsymbol{\Sigma}} = \boldsymbol{\sigma}_i^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\sigma}_{\boldsymbol{j}} = \Sigma_{i,j}$, and more generally, $\langle \boldsymbol{\sigma}_i, \boldsymbol{f} \rangle_{\boldsymbol{\Sigma}} = f_i$

The space $\mathcal{H}$ of functions $\mathbb{R}^p \to \mathbb{R}$ is a Reproducing Kernel Hilbert Space (RKHS) if there is an inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ such that $\mathcal{H}$ with $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ is an Hilbert space, and for all $\boldsymbol{x} \in \mathbb{R}^p$, linear functional $\delta_{\boldsymbol{x}} : \mathcal{H} \to \mathbb{R}$ defined as $\delta_{\boldsymbol{x}}(f) = f(\boldsymbol{x})$ is bounded.

Thus, $\mathcal{H}$ is a RKHS if and only if $\forall f \in \mathcal{H}$ and $\boldsymbol{x} \in \mathbb{R}^p$, there exists $M_{\boldsymbol{x}}$ such that $|f(\boldsymbol{x})| \leq M_{\boldsymbol{x}} \cdot \|f\|_{\mathcal{H}}$.

From Riesz theorem, there exists a unique $\zeta_{\boldsymbol{x}} \in \mathcal{H}$ associated with $\delta_{\boldsymbol{x}}$, i.e. $\langle \zeta_{\boldsymbol{x}}, f \rangle_{\mathcal{H}} = f(\boldsymbol{x})$

Reproducing Kernel of $\mathcal{H}$

Function $\boldsymbol{x} \mapsto \zeta_{\boldsymbol{x}}$ is called reproducing function in $\boldsymbol{x}$ and $K : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$ defined as $K(\boldsymbol{x}, \boldsymbol{y}) = \zeta_{\boldsymbol{x}}(\boldsymbol{t})$ is the reproducing kernel of $\mathcal{H}$.

Norms, Inner Products and Kernels

Observe that $\langle K(\boldsymbol{x},\cdot), K(\boldsymbol{y},\cdot)\rangle_{\mathcal{H}} = K(\boldsymbol{x},\boldsymbol{y})$.

The kernel is unique, and is (semi-)definite positive.

If $\mathcal{H}$ is a closed subspace of Hilbert space $\mathcal{X}$. For any function $f \in \mathcal{X}$, $\boldsymbol{x} \mapsto \langle f, K(\cdot,\boldsymbol{x})\rangle_{\mathcal{X}}$ is the projection of f on $\mathcal{H}$.

Note that conversely, Moore-Aronszajn's theorem allows to create a RKHS from a definite positive kernel K.

Norms, Inner Products and Kernels

Mercer's kernel

Let μ denote some measure on $\mathbb{R}^p$ and $\mathcal{H} = L^2(\mathbb{R}^p, \mu)$, define

$$(L_K f)(\boldsymbol{x}) = \int K(\boldsymbol{x}, \boldsymbol{y}) f(\boldsymbol{y}) d\mu(\boldsymbol{y})$$

which is a compact bounded linear operator, self-adjoint and positive. Let $\lambda_1 \geq \lambda_2 \geq \cdots$ denote eigenvalues of L_k, with (orthonormal) eigenvectors $\psi_1, \psi_2, \cdots$, then

$$K(\boldsymbol{x}, \boldsymbol{y}) = \sum_{k=1}^{p} \lambda_k \psi_k(\boldsymbol{x}) \psi_k(\boldsymbol{y}) = \Psi(\boldsymbol{x})^\top \Psi(\boldsymbol{y}) = \langle \Psi(\boldsymbol{x}), \Psi(\boldsymbol{y}) \rangle_{L^2}$$

where $\Psi(\boldsymbol{x}) = (\sqrt{\lambda_k} \psi_k(\boldsymbol{x}))$.

Norms, Inner Products and Kernels

Example : Consider the space $\mathcal{H}$ defined as

$$\mathcal{H}_1 = \left\{ f : [0,1] \to \mathbb{R} \text{ continuously differentiable, with } f' \in L^2([0,1]) \text{ and } f(0) = 0 \right\}$$

$\mathcal{H}_1$ is an Hilbert space on $[0,1]$ with inner product

$$\langle f, g \rangle_{\mathcal{H}_1} = \int_0^1 f'(t) g'(t) dt$$

with (definite positive) kernel $K_1(x,y) = \min\{x,y\}$:

$$\langle f, K(x,\cdot) \rangle_{\mathcal{H}_1} = \int_0^1 f'(t) \underbrace{\frac{\partial K_1(t,x)}{\partial x}}_{=\mathbf{1}_{[0,x]}(t)} = \int_0^x f'(t) dt = f(x)$$

Norms, Inner Products and Kernels

Example : Consider the Sobolev space $W^1([0,1])$ defined as

$$W^1([0,1]) = \left\{f : [0,1] \to \mathbb{R} \text{ continuously differentiable, with } f' \in L^2([0,1])\right\}$$

Observe that $W^1([0,1]) = \mathcal{H}_0 \oplus \mathcal{H}_1$ where

$$\mathcal{H}_0 = \left\{f : [0,1] \to \mathbb{R} \text{ continuously differentiable, with } f' = 0\right\}$$

The later is an Hilbert space with kernel $K_0(x,y) = 1$.

One can consider kernel $K(x,y) = K_0(x,y) + K_1(x,y)$ (related to linear splines). More generally, consider

$$\mathcal{H}_2 = \left\{f : [0,1] \to \mathbb{R} \text{ twice cont. diff., with } f'' \in L^2([0,1] \text{ with } f'(0) = 0)\right\}$$

Then $\langle f, g\rangle_{\mathcal{H}_2} = \displaystyle\int_0^1 f''(t)g''(t)dt$ is an inner product, with kernel

$$K_2(x,y) = \int_0^1 (x-t)_+(y-t)_+dt$$

Norms, Inner Products and Kernels

Consider some Hilbert space $\mathcal{H}$ with kernel K and some functional $\Psi : \mathbb{R}^{n+1} \to \mathbb{R}$, increasing in its last argument.

Representation theorem

Given $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n$, $\min_{f \in \mathcal{H}} \left\{ \Psi(f(\boldsymbol{x}_1), \cdots, f(\boldsymbol{x}_n), \|f\|_{\mathcal{H}}) \right\}$ admits solution

$$\forall \boldsymbol{x},\ f(\boldsymbol{x}) = \sum_{i=1}^{n} \alpha_i K(\boldsymbol{x}_i, \boldsymbol{x})$$

A classical expression for Ψ is, for some convex function ψ,

$$\Psi(f(\boldsymbol{x}_1), \cdots, f(\boldsymbol{x}_n), \|f\|_{\mathcal{H}}) = \psi(\boldsymbol{y}, f(\boldsymbol{x}_1), \cdots, f(\boldsymbol{x}_n)) + \lambda \|f\|_{\mathcal{H}}$$

$$\Psi(f(\boldsymbol{x}_1), \cdots, f(\boldsymbol{x}_n), \|f\|_{\mathcal{H}}) = \sum_{i=1}^{n} \ell(\boldsymbol{y}, f(\boldsymbol{x}_i)) + \lambda \|f\|_{\mathcal{H}}$$

Norms, Inner Products and Kernels

Assume that $y_i = m(x_i) + \varepsilon_i$, where $m \in W_2([0,1])$, then polynomial splines of degree 2 is the solution of

$$\min_{m \in W_2} \left\{ \frac{1}{n} \sum_{i=1}^{n} \big(y_i - m(x_i)\big)^2 + \nu \int_0^1 [m''(t)]^2 dt \right\}$$

then $m^\star(x) = \beta_0 + \beta_1 x + \sum_{i=1}^{n} \gamma_i K_2(x_i, x)$ Note that one can use a matrix representation

$$\min \left\{ (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Q}\boldsymbol{\gamma})^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Q}\boldsymbol{\gamma}) + n\nu \boldsymbol{\gamma}^\top \boldsymbol{Q} \boldsymbol{\gamma} \right\}$$

where $\boldsymbol{Q} = [K_1(x_i, x_j)]$. If $\boldsymbol{M} = \boldsymbol{Q} + n\nu\mathbb{I}$,

$$\boldsymbol{\beta}^\star = \big(\boldsymbol{X}^\top \boldsymbol{M}^{-1} \boldsymbol{X}\big)^{-1} \boldsymbol{X}^\top \boldsymbol{M}^{-1} \boldsymbol{y} \text{ and } \boldsymbol{\gamma}^\star = \boldsymbol{M}^{-1} \big(\mathbb{I} - \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{M}^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{M}^{-1}\big) \boldsymbol{y}$$

Norms, Inner Products and Kernels

Kimeldorf & Wahba's representation theorem

Consider a kernel K and $\mathcal{H}_K$ the associated RKHS. For any (convex) loss function $\ell : \mathbb{R}^2 \to \mathbb{R}$, the solution

$$m^\star \in \underset{m \in \mathcal{H}_K}{\operatorname{argmin}} \sum_{i=1}^{n} \ell(y_i, m(\boldsymbol{x}_i)) + \|m\|^2_{\mathcal{H}_K}$$

can be expressed

$$m^\star(\cdot) = \sum_{i=1}^{n} \alpha_i K(\boldsymbol{x}_i, \cdot)$$

See Kimeldorf & Wahba (1971, Some results on Tchebycheffian spline functions), some slides and later on, on SVM.

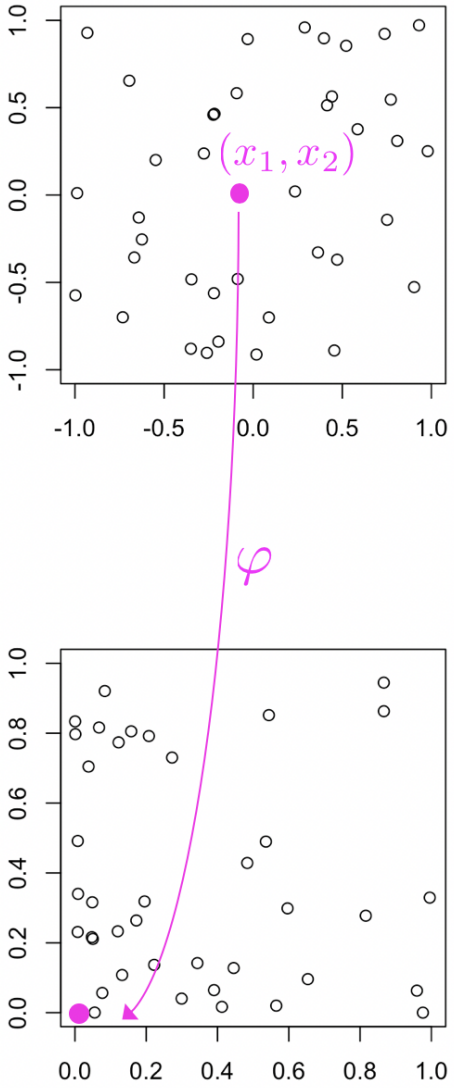
What's going on, here ?

For more technical (mathematical) results, see Wahba (1990, Spline Models for Observational Data)

We've seen that in many cases, $K(\boldsymbol{x}, \boldsymbol{y}) = \varphi(\boldsymbol{x})^\top \varphi(\boldsymbol{y})$ for some $\varphi : \mathbb{R}^p \to \mathbb{R}^q$ (here $q = p$)

Consider for example $\varphi : \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \mapsto \begin{pmatrix} x_1^2 \\ x_2^2 \end{pmatrix}$

so that $K(\boldsymbol{x}, \boldsymbol{y}) = x_1^2 y_1^2 + x_2^2 y_2^2$

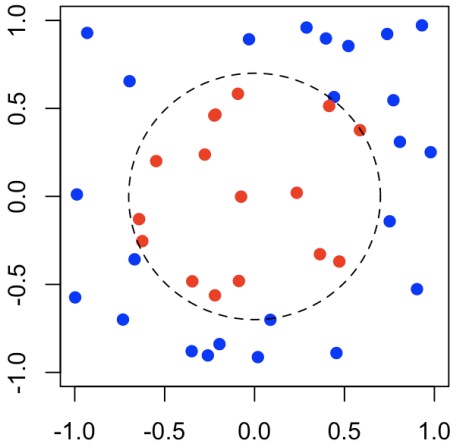


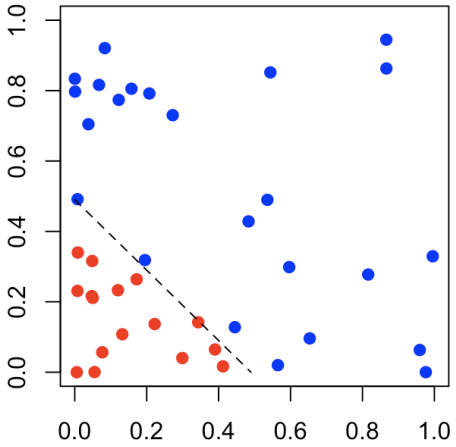
What's going on, here ?

Consider $\varphi : \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \mapsto \begin{pmatrix} x_1^2 \\ x_2^2 \end{pmatrix}$

$$K(\boldsymbol{x}, \boldsymbol{y}) = x_1^2 y_1^2 + x_2^2 y_2^2$$

From data $(y_i, \boldsymbol{x}_i)$, transform the covariates into $(y_i, \phi(\boldsymbol{x}_i))$, and use a (classical) linear model

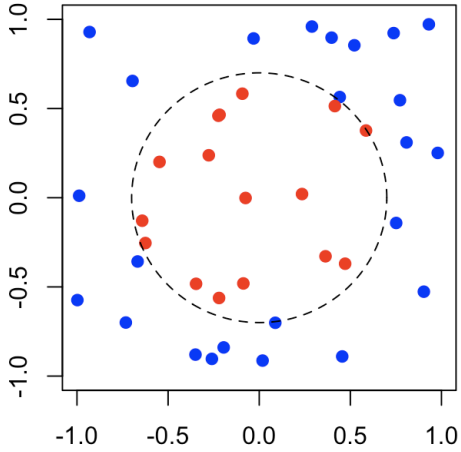


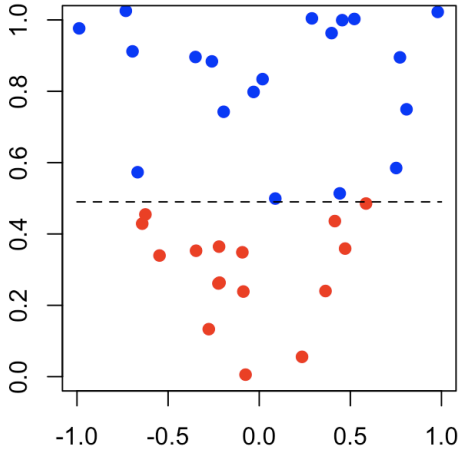


What's going on, here ?

Consider $\varphi : \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \mapsto \begin{pmatrix} x_1 \\ x_1^2 + x_2^2 \end{pmatrix}$

$$K(\boldsymbol{x}, \boldsymbol{y}) = x_1 y_1 + (x_1^2 + x_2^2)(y_1^2 + y_2^2)$$



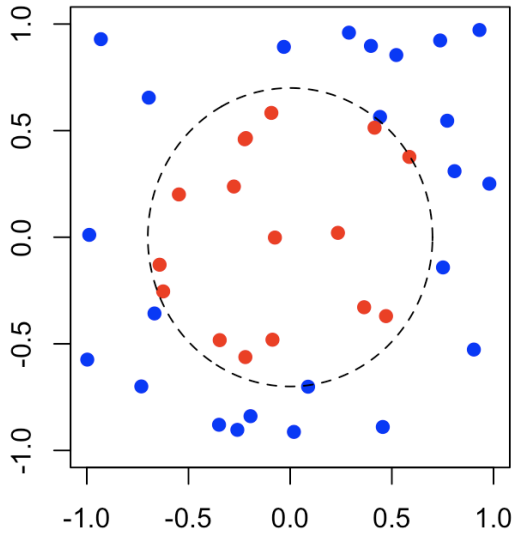


What's going on, here ?

But we can have $\varphi : \mathbb{R}^p \to \mathbb{R}^q$ (with $q \neq p$)

Consider $\varphi : \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \mapsto \begin{pmatrix} x_1 \\ x_2 \\ x_1 \cdot x_2 \end{pmatrix}$

A classical idea with SVM will be to consider $q > p$ to be able to find a linear separator of points red and blue



What's going on, here ?

$$\text{Consider } \varphi : \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \mapsto \begin{pmatrix} 1 \\ \sqrt{2}x_1 \\ \sqrt{2}x_2 \\ x_1^2 \\ x_2^2 \\ \sqrt{2}x_1x_2 \end{pmatrix}$$

$K(\boldsymbol{x}, \boldsymbol{y}) = \left(1 + \boldsymbol{x}^\top \boldsymbol{y}\right)^2$ is called polynomial kernel (of order 2). More generally

$K(\boldsymbol{x}, \boldsymbol{y}) = \left(1 + \boldsymbol{x}^\top \boldsymbol{y}\right)^d$, with $d \in \mathbb{N}$.

For any degree $d \geq 2$, $\varphi : \mathbb{R}^p \to \mathbb{R}^q$ with $q = \binom{p+d}{d}$ (default in R, $d = 3$)

How to get those kernels ?

If K_1 and K_2 are two kernels, so are

$$K(\boldsymbol{x},\boldsymbol{y}) = a_1K_1(\boldsymbol{x},\boldsymbol{y}) + a_2K_2(\boldsymbol{x},\boldsymbol{y}) \text{ and } K(\boldsymbol{x},\boldsymbol{y}) = K_1(\boldsymbol{x},\boldsymbol{y}) \cdot K_2(\boldsymbol{x},\boldsymbol{y})$$

If h is some $\mathbb{R}^n \to \mathbb{R}^n$ function, $K(\boldsymbol{x},\boldsymbol{y}) = K_1(h(\boldsymbol{x}),h(\boldsymbol{y}))$

If h is some $\mathbb{R}^n \to \mathbb{R}$ function, $K(\boldsymbol{x},\boldsymbol{y}) = h(\boldsymbol{x}) \cdot h(\boldsymbol{y})$

If P is a polynomial with positive coefficients, $k(\boldsymbol{x},\boldsymbol{y}) = P(K_1(\boldsymbol{x},\boldsymbol{y}))$ as well as $K(\boldsymbol{x},\boldsymbol{y}) = \exp[_1(\boldsymbol{x},\boldsymbol{y})]$