Kernel Machines

Alain Celisse

SAMM

Paris 1-Panthéon Sorbonne University

`alain.celisse@univ-paris1.fr`

Lecture 3: Kernelizing classical strategies

—

Master 2 Data Science – Centrale Lille, Lille University
Fall 2022

Outline of the lectures

Successive topics of the coming lectures:

1. Introduction to Kernel methods
2. Support vector classifiers and Kernel methods
3. Extending classical strategies to high dimension (Today!)
   - KPCA
   - KRR
4. Duality gap and KKT conditions
5. Designing reproducing kernels
6. Maximum Mean Discrepancy (MMD)
7. Change-point detection, KCP

Outline of the lecture

- KPCA
- KRR

Introduction

Limitations of classical learning approaches

- Cannot easily deal with "complex structured" data (metric to be defined)
  - Combining real measures with histograms, categorical data, and graphs is a hard problem
  - Extracting relevant information requires particular metric
- Easy to understand (linear classifier), but not versatile tools (see the SVM lecture)
- Measuring the dependence is not an easy task
  - Covariance and correlation measure linear dependence between variables
    ($Cov(X, Y) = 0$ does not imply $X \sqcup Y$)
  - General measures of dependence require comparison between marginal and joint distributions

Dealing with structured data

Combining kernels

- k_1, k_2: reproducing kernels $\Rightarrow \alpha k_1 + \beta k_2$: reproducing kernel ($\alpha, \beta \geq 0$)
- $(k_1)^\alpha$ ($\alpha > 0$): reproducing kernel

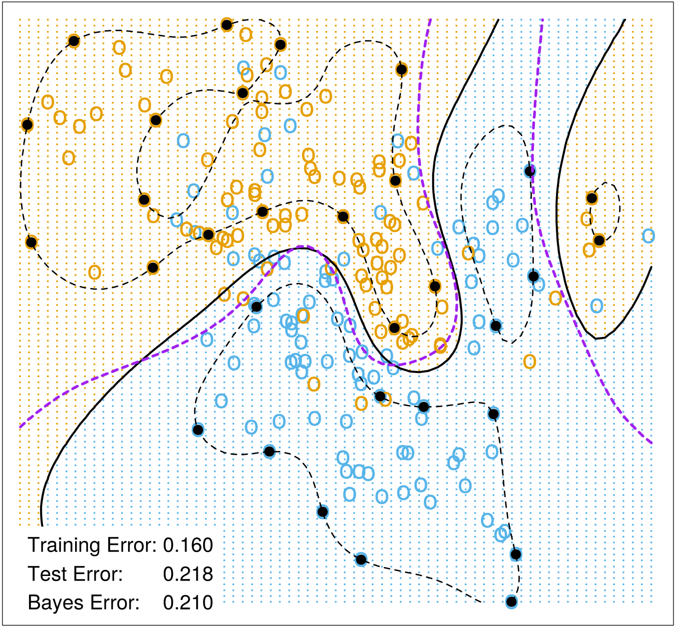
Kernels for structured objects

- The χ^2-kernel deals with histograms (G bins)

$$k(x, y) = \exp\left[-\sum_{g=1}^{G} \frac{(x_g - y_g)^2}{x_g + y_g}\right]$$

- Helpful in video streams analysis, dealing with texts (bag of words), ...

Versatile tool improving the performance



Measuring the dependence between variables with kernels

- $X_1, \ldots, X_n \in \mathbb{R}^d$
- Empirical covariance matrix:

$$\widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} X_i X_i^\top$$

- Empirical covariance operator (from $\mathcal{H}$ to $\mathcal{H}$):

$$\widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} (\phi(X_i) \otimes \phi(X_i))$$

  $\phi(X_i) \in \mathcal{H}$: Extended feature vector (centered)
  $\otimes$: tensor product defined by

$$(a \otimes b)g = \langle b, g\rangle_{\mathcal{H}}\, a \in \mathcal{H}, \qquad \forall g \in \mathcal{H}$$

- Captures non-linear dependencies between the d features of the X_is

Towards kernelized strategies

There are two classical lines of thought for extending a learning strategy to a kernelized version:

- Strategies relying on scalar products can be easily extended to the "kernel world"
  - SV classifier, Ridge regression, exponential families,...
- Strategies relying on similarity measures between "individuals" can be combined with any psd kernel
  - K-means, k-Nearest Neighbors, Spectral clustering,...

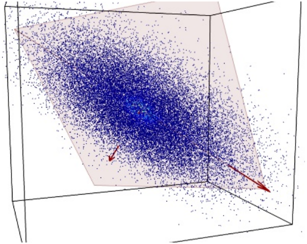
Learning procedures involving kernels

- Kernelized Tikhonov regularizarion (LS-SVM, or Kernel Ridge Regression)
- Kernel-principal component Aanalysis
- Kernel-Canonical Correlation Analysis
- MMD and HSIC criteria (coming lecture. . .)

- Kernelized Support Vector Classifier (SVM)
- Multi-task learning
- K-means, Kernelized HAC, Kernelized spectral clustering
- . . .

Kernel PCA (KPCA)

Displaying the data in high dimension



- A large part of collected data is high-dimensional
- Displaying data in $\mathbb{R}^d$ is challenging if $d \geq 4$

Strategy

- Looking for directions along which the data exhibit the largest variability
- Use these directions as a new vector basis for drawing graphs
- These priviledged directions allow for visually identifying hidden structures (clusters)

PCA: Introduction

Definition (Covariance matrix)

- $X = (X^1, \ldots, X^d)^\top \in \mathbb{R}^d$: column vector
- $X^\top$: the transpose of X, row vector
- The covariance matrix is defined by

$$\mathbb{V}(X) = \mathbb{E}\left[(X - \mathbb{E}(X)) \cdot (X - \mathbb{E}(X))^\top \right]$$

Rks:

- "Covariance Matrix":

$$\begin{aligned}[\mathbb{V}(X)]_{i,j} &= \mathbb{E}\left[(X^i - \mathbb{E}(X^i)) \cdot (X^j - \mathbb{E}(X^j)) \right] \\ &= Cov(X^i, X^j)\end{aligned}$$

- The covariance matrix captures some dependence between the variables X^js
- If $X \in \mathbb{R}^d$ is Gaussian, $Cov(X^i, X^j) = 0 \quad \Leftrightarrow \quad X^i \perp X^j$

Estimating the covariance matrix

With $\mathbb{E}(X) = 0$,

$$\mathbb{V}(X) = \mathbb{E}\left[X \cdot X^\top\right] \in \mathcal{M}_d(\mathbb{R})$$

- $X \in \mathbb{R}^d$: column vector
- $\mathbb{V}(X)$: $d \times d$ matrix, postive semidefinite (PSD)

Definition (Empirical covariance matrix)

If $\mathbb{E}(X) = 0$,

$$\widehat{\mathbb{V}(X)} = \frac{1}{n}\sum_{i=1}^{n} X_i \cdot X_i^\top = \frac{1}{n}\mathrm{X}^\top \mathrm{X}$$

Rk:

- In practice, start by centering the X_is!
- If $d = 1$, $\widehat{\mathbb{V}(X)} = 1/n \sum_{i=1}^n X_i^2 \approx \mathbb{V}(X) = \mathbb{E}\left[X^2\right]$
- $O(nd^2)$ "elementary operations" to be computed
- With $d \gg 1$ (large), becomes computationally heavy

Classical PCA

PCA: Population version

Goal

- Find unit vector $v_1 \in \mathbb{R}^d$ such that

$$v_1 \in Arg \max_{\|u\|=1} \left\{ \mathrm{Var}(X^\top u) \right\} = Arg \max_{\|u\|=1} \left\{ \mathbb{E}\left[\left(X^\top u \right)^2 \right] \right\}$$
$$= Arg \max_{\|u\|=1} \left\{ u^\top \mathbb{V}(X) u \right\}$$

- Repeat with $v_{j+1} \in Vect(v_1, \ldots, v_j)^\perp$, for $1 \leq j \leq d$

Eigenvalues and eigenvectors

- Amounts to find the largest eigenvalue (and the corresponding eigenvector) of the PSD matrix $\mathbb{V}(X)$
- Amounts to find successive basis vectors (eigenvectors) maximizing the variance at each iteration

PCA: Eigenvalue decomposition

$\mathbb{V}(X)$: $d \times d$ psd matrix

Theorem

For any psd $d \times d$ matrix Σ, there exist

- O: *$d \times d$ orthogonal matrix* $\quad (O^\top \cdot O = O \cdot O^\top = I_d)$
- Λ: *$d \times d$ diagonal matrix* $\quad (\Lambda_{i,i} = \lambda_i,\ \Lambda_{i,j} = 0,\ i \neq j)$
- $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d \geq 0$

such that

$$\Sigma = O \cdot \Lambda \cdot O^\top = \sum_{j=1}^{d} \lambda_j O_{\cdot j} O_{\cdot j}^\top$$

Comments:

- O: rotation matrix from canonical to the new basis
- Each column of $O_{\cdot j}$ is an eigenvector of Σ
- Each diagonal element λ_i is an eigenvalue of Σ
- Each $O_{\cdot j}$ is the eigenvector of λ_j

PCA: Diagonalizing $\mathbb{V}(X)$

Application to $\mathbb{V}(X)$

With $\mathbb{V}(X) = V \Lambda V^\top$, ($V$: orthogonal)

$$w \in Arg \max_{\|u\|=1} \left\{ u^\top \mathbb{V}(X) u \right\}$$

$$\Leftrightarrow \quad w \in Arg \max_{\|u\|=1} \left\{ u^\top \sum_{j=1}^{d} \lambda_j V_{\cdot j} V_{\cdot j}^\top u \right\}$$

$$\Leftrightarrow \quad w \in Arg \max_{\|u\|=1} \left\{ \sum_{j=1}^{d} \lambda_j \left(V_{\cdot j}^\top u \right)^2 \right\}$$

$$\Leftrightarrow \quad w \in \left\{ u \in \mathbb{R}^d \mid \mathbb{V}(X) u = \lambda_1 u \right\} \cap B_{\|\cdot\|}(1)$$

Key ingredient

($v_j = V_{\cdot j}$)

Finding the λ_is and eigenvectors v_is such that

$$\mathbb{V}(X) = \sum_{j=1}^{d} \lambda_j v_j \cdot v_j^\top$$

PCA: Diagonalizing $\widehat{\mathbb{V}(X)}$

Application to $\widehat{\mathbb{V}(X)}$ (empirical covariance matrix)

With $\widehat{\mathbb{V}(X)} = \widehat{V}\widehat{\Lambda}\widehat{V}^{\top}$, ($\widehat{V}$: orthogonal)

- Find the $\widehat{\lambda}_j$s and $\widehat{v}_j$s such that

$$\widehat{\mathbb{V}(X)} = \sum_{j=1}^{d} \widehat{\lambda}_j \widehat{v}_j \cdot \widehat{v}_j^{\top} \qquad (\text{SVD of } \widehat{\mathbb{V}(X)})$$

- The "principal component" of individual i on the component j is the score

$$X_i^{\top} \cdot \widehat{v}_j \in \mathbb{R} \qquad (X_i \text{ centered})$$

- The "jth principal component" of all individuals is

$$\tilde{\mathrm{X}}_{\cdot j} = \mathrm{X} \cdot \widehat{v}_j \in \mathbb{R}^n \qquad (\text{projection})$$

- principal component decomposition:

$$\tilde{\mathrm{X}} = \mathrm{X} \cdot \widehat{V} \in \mathcal{M}_{n,d}(\mathbb{R}) \qquad (\text{Score matrix})$$

PCA and displaying the data

Summarizing the information

From

$$\tilde{\mathrm{X}} = \mathrm{X} \cdot \widehat{V} \in \mathcal{M}_{n,d}(\mathbb{R})$$

- The row i of $\tilde{\mathrm{X}}$ describes individual i in the new basis $\widehat{V}$
- Displaying component ℓ versus component k consists in displaying the point $(\tilde{\mathrm{X}}_{i,k}, \tilde{\mathrm{X}}_{i,\ell})$ for each individual i

Dimension reduction

- From $\widehat{\lambda}_1 \geq \cdots \geq \widehat{\lambda}_d \geq 0$, the relevant information about the cloud of points can be summarized by the r largest components
- To ease the storage of the data when $d \gg 1$, "build a smaller summary of the data" by means of

$$\tilde{\mathrm{X}}^{(r)} = \left[\tilde{\mathrm{X}}_{\cdot 1}, \tilde{\mathrm{X}}_{\cdot 2}, \ldots, \tilde{\mathrm{X}}_{\cdot r} \right]$$

K-PCA

New formulation

Projection of extended features

- For $i = 1, \ldots, n$, $\phi(x_i) = k_{x_i} \in \mathcal{H}$: Extended features
- Warning: The k_{x_i}s assumed to be centered in what follows
- Orthogonal projection of k_{x_i} onto $g \in \mathcal{H}$:

$$\left\langle k_{x_i}, \frac{g}{\|g\|_{\mathcal{H}}} \right\rangle_{\mathcal{H}} \cdot \frac{g}{\|g\|_{\mathcal{H}}}$$

Population variance of the projection

$$\mathrm{Var}\left(\frac{\langle k_X, g \rangle_{\mathcal{H}}}{\|g\|_{\mathcal{H}}} \right) = \mathbb{E}\left[\frac{(g(X))^2}{\|g\|_{\mathcal{H}}^2} \right]$$

Empirical variance of the projection

$$\widehat{\mathrm{Var}}\left(\frac{\langle k_X, g \rangle_{\mathcal{H}}}{\|g\|_{\mathcal{H}}} \right) = \widehat{\mathbb{E}}\left[\frac{(g(X))^2}{\|g\|_{\mathcal{H}}^2} \right] = \frac{1}{n} \sum_{i=1}^{n} \frac{(g(x_i))^2}{\|g\|_{\mathcal{H}}^2}$$

PCA formulated in the "Kernel World"

General Kernel PCA algorithm

- Find a vector $g_1 \in \mathcal{H}$ such that

$$g_1 \in Arg\max_{g \in \mathcal{H}} \left\{ \widehat{\mathrm{Var}}\left(\frac{\langle k_X, g \rangle_{\mathcal{H}}}{\|g\|_{\mathcal{H}}} \right) \right\}$$

$$= Arg\max_{g \in \mathcal{H}} \underbrace{\left\{ \frac{1}{n} \sum_{i=1}^{n} \frac{\left(g(x_i)\right)^2}{\|g\|_{\mathcal{H}}^2} \right\}}_{=-\Psi(g(x_1),\ldots,g(x_n),\|g\|_{\mathcal{H}})}$$

PCA formulated in the "Kernel World"

General Kernel PCA algorithm

- Find a vector $g_1 \in \mathcal{H}$ such that

$$g_1 \in \mathop{Arg\,max}_{g \in \mathcal{H}} \left\{ \widehat{\mathrm{Var}} \left(\frac{\langle k_X, g \rangle_{\mathcal{H}}}{\|g\|_{\mathcal{H}}} \right) \right\}$$

$$= \mathop{Arg\,max}_{g \in \mathcal{H}} \underbrace{\left\{ \frac{1}{n} \sum_{i=1}^{n} \frac{(g(x_i))^2}{\|g\|_{\mathcal{H}}^2} \right\}}_{=-\Psi(g(x_1),\ldots,g(x_n),\|g\|_{\mathcal{H}})}$$

(Representer theorem)

PCA formulated in the "Kernel World"

General Kernel PCA algorithm

- Find a vector $g_1 \in \mathcal{H}$ such that

$$g_1 \in \underset{g\in\mathcal{H}}{Arg\max} \left\{ \widehat{\mathrm{Var}}\left(\frac{\langle k_X, g\rangle_{\mathcal{H}}}{\|g\|_{\mathcal{H}}} \right) \right\}$$

$$= \underset{g\in\mathcal{H}}{Arg\max} \underbrace{\left\{ \frac{1}{n}\sum_{i=1}^{n} \frac{(g(x_i))^2}{\|g\|_{\mathcal{H}}^2} \right\}}_{=-\Psi(g(x_1),\ldots,g(x_n),\|g\|_{\mathcal{H}})}$$

(Representer theorem)

- Repeat with $g_{j+1} \in Vect(g_1, \ldots, g_j)^{\perp}$, for $1 \leq j \leq n$

Simplified Kernel PCA algorithm

Representer theorem

- The solution $\widehat{g}_1 = \sum_{i=1}^n \widehat{\alpha}_i k_{x_i} \in \mathcal{H}$
- $\|\widehat{g}_1\|_{\mathcal{H}}^2 = \widehat{\alpha}^\top K \widehat{\alpha}$
- $\sum_{i=1}^n \widehat{g}_1(x_i)^2 = \widehat{\alpha}^\top K^2 \widehat{\alpha}$

Solving the problem amounts to compute:

Practical Kernel PCA algorithm

- Compute

$$\widehat{\alpha}_1 \in Arg \max_{\alpha \in \mathbb{R}^n} \left\{ \frac{\alpha^\top K^2 \alpha}{\alpha^\top K \alpha} \right\}$$

- Repeat with $\widehat{\alpha}_{j+1} \perp_K \{\widehat{\alpha}_1, \ldots, \widehat{\alpha}_j\}$ where

$$\alpha \perp_K \beta \qquad \Leftrightarrow \qquad \alpha^\top K \beta = 0$$

Conclusion

Solving KPCA means diagonalizing the centered Gram matrix

Interpretation

- The eigenvectors are now functions in $\mathcal{H}$

$$\widehat{g}_j = \sum_{i=1}^{n} \widehat{\alpha}_{j,i} k_{x_i}, \qquad \forall j \geq 1$$

- The "principal component of individual i on the component j is the score

$$\frac{\langle k_{x_i}, \widehat{g}_j \rangle_{\mathcal{H}}}{\|\widehat{g}_j\|_{\mathcal{H}}}$$

- The "jth principal component" of all individuals is

$$\left(\widehat{g}_j(x_1), \ldots, \widehat{g}_j(x_n)\right)^{\top} \times \frac{1}{\|\widehat{g}_j\|_{\mathcal{H}}}$$

- Any individual x can be displayed through its (i,j)th coordinates in the new basis by

$$\left(\frac{\langle k_x, \widehat{g}_i \rangle_{\mathcal{H}}}{\|\widehat{g}_i\|_{\mathcal{H}}}, \frac{\langle k_x, \widehat{g}_j \rangle_{\mathcal{H}}}{\|\widehat{g}_j\|_{\mathcal{H}}} \right)$$

Towards the covariance operator

For all $g \in H$, $(k_X\colon$ centered$)$

$$\mathrm{Var}\left(\left\langle k_X, \frac{g}{\|g\|_{\mathcal{H}}}\right\rangle_{\mathcal{H}}\right) = \mathbb{E}\left[\left\langle k_X, \frac{g}{\|g\|_{\mathcal{H}}}\right\rangle_{\mathcal{H}}^2\right]$$

Tensor product

For all $a, b \in \mathcal{H}$

$$(a \otimes b): \quad f \mapsto (a \otimes b)f = \langle b, f\rangle_{\mathcal{H}}\, a$$

Proposition

For all $f, g \in \mathcal{H}$

$$\langle (a \otimes b)f, g\rangle_{\mathcal{H}} = \langle a, f\rangle_{\mathcal{H}} \cdot \langle b, g\rangle_{\mathcal{H}}$$

Then

$$\langle (a \otimes a)f, f\rangle_{\mathcal{H}} = \langle a, f\rangle_{\mathcal{H}}^2$$

Covariance operator

For all $g \in H$, (k_X: centered)

$$\mathbb{E}\left[\left\langle k_X, \frac{g}{\|g\|_{\mathcal{H}}}\right\rangle_{\mathcal{H}}^2\right] = \mathbb{E}\left[\left\langle (k_X \otimes k_X)\frac{g}{\|g\|_{\mathcal{H}}}, \frac{g}{\|g\|_{\mathcal{H}}}\right\rangle_{\mathcal{H}}\right]$$
$$= \left\langle \mathbb{E}\left[k_X \otimes k_X\right]\frac{g}{\|g\|_{\mathcal{H}}}, \frac{g}{\|g\|_{\mathcal{H}}}\right\rangle_{\mathcal{H}}$$
$$= \left\langle \Sigma\frac{g}{\|g\|_{\mathcal{H}}}, \frac{g}{\|g\|_{\mathcal{H}}}\right\rangle_{\mathcal{H}}$$

Definition

Covariance operator: Unique operator Σ from $\mathcal{H}$ to $\mathcal{H}$ such that, for all $f, g \in \mathcal{H}$,

$$\langle \Sigma f, g\rangle_{\mathcal{H}} = \mathbb{E}\left[\langle k_X, f\rangle_{\mathcal{H}} \langle k_X, g\rangle_{\mathcal{H}}\right]$$
$$= \langle \mathbb{E}\left[k_X \otimes k_X\right] f, g\rangle_{\mathcal{H}}$$

Empirical covariance operator

Definition

Empirical covariance operator: Unique operator $\widehat{\Sigma}$ from $\mathcal{H}$ to $\mathcal{H}$ such that, for all $f, g \in \mathcal{H}$,

$$\left\langle \widehat{\Sigma} f, g \right\rangle_{\mathcal{H}} = \frac{1}{n} \sum_{i=1}^{n} \langle k_{X_i}, f \rangle_{\mathcal{H}} \langle k_{X_i}, g \rangle_{\mathcal{H}} = \left\langle \frac{1}{n} \sum_{i=1}^{n} (k_{X_i} \otimes k_{X_i}) f, g \right\rangle_{\mathcal{H}}$$

Remark:

$$\left\langle \widehat{\Sigma} f, g \right\rangle_{\mathcal{H}} = \frac{1}{n} \sum_{i=1}^{n} f(X_i) g(X_i) \xrightarrow[n \to +\infty]{P} \langle \Sigma f, g \rangle_{\mathcal{H}} = \mathbb{E}\left[f(X) g(X) \right]$$

Kernel Ridge Regression (KRR)

Framework

Linear regression model

$$\mathrm{Y} = \mathrm{X} \cdot \beta^\star + \boldsymbol{\epsilon} \in \mathbb{R}^n$$

with $\beta^\star \in \mathbb{R}^d$, $\mathrm{Y} = (Y_1, \ldots, Y_n)^\top$, $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^\top$, and $\mathrm{X} = \left[\mathrm{X}^1, \ldots, \mathrm{X}^d\right]$: $n \times d$ matrix

Assumptions:

- $\mathbb{E}_\epsilon\left[\mathrm{Y}\right] = \mathrm{X} \cdot \beta^\star$
- $\mathrm{Var}_\epsilon(\boldsymbol{\epsilon}) = \sigma^2 I_n$, $\sigma^2 > 0$

Question

How to efficiently predict Y at a new $X \in \mathbb{R}^d$?

ℓ_0-norm and lack of convexity

Ideal optimization problem to solve

For every $k \geq 1$, solve

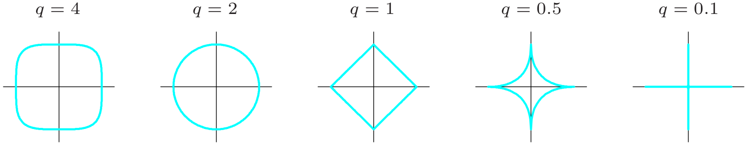
$$\widehat{\beta}_k = Arg \min_{\|\beta\|_0 \leq k} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - X_i^\top \beta \right)^2 \right\}$$

which is equivalent to:

For every $\lambda > 0$, solve $\qquad (\|\beta\|_0 = \sum_{j=1}^{d} \mathbb{1}_{\beta_j \neq 0})$

$$\widehat{\beta}_{NP,\lambda} = Arg \min_{\beta \in \mathbb{R}^d} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - X_i^\top \beta \right)^2 + \lambda \|\beta\|_0 \right\}$$

Remark: $\beta \mapsto \|\beta\|_0$ is non-convex! $\Rightarrow$ NP-hard to solve



Convex relaxation of the ℓ_0-norm

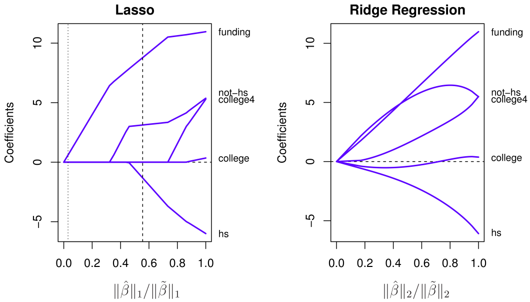
Convex relaxation with ℓ_2-norm: Ridge

For every $\lambda > 0$, solve $\qquad (\|\beta\|_2^2 = \sum_{j=1}^d |\beta_j|^2)$

$$\widehat{\beta}_\lambda = Arg \min_{\beta \in \mathbb{R}^d} \left\{ \frac{1}{n} \sum_{i=1}^n \left(Y_i - X_i^\top \beta \right)^2 + \lambda \|\beta\|_2^2 \right\}$$

$$= Arg \min_{\beta \in \mathbb{R}^d} \left\{ \|\mathrm{Y} - \mathrm{X}\beta\|_n^2 + \lambda \|\beta\|_2^2 \right\}$$

where $\|u\|_n^2 = \sum_{i=1}^n u_i^2 / n$



Computing the Ridge estimator

- Unlike LASSO, Ridge leads to a closed-form expression for the estimator
- The ℓ_2-norm differentiable everywhere (unlike the ℓ_1-norm)

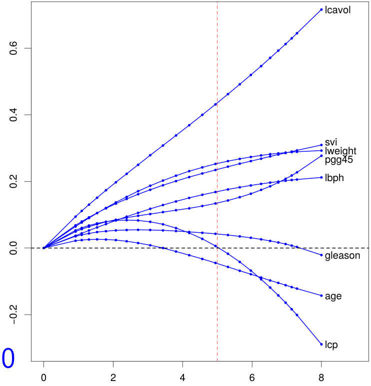
Ridge estimator: Closed-form expression

$$\widehat{\beta}_\lambda = \left(\frac{X^\top X}{n} + \lambda I_d \right)^{-1} \frac{X^\top Y}{n}, \qquad \forall \lambda > 0$$

Remark:

- $\frac{X^\top X}{n} + \lambda I_d$: always invertible
- $\frac{X^\top X}{n}$: $d \times d$ matrix $\Rightarrow$ Invertion is time consuming
- SVD of $XX^\top$ can be more efficient than that of $X^\top X$
- Ridge shrinks the coefficients (as LASSO does)

Ridge: No sparsity constraint



Continuous shrinking towards 0

$$\widehat{\beta}_\lambda = \left(\frac{X^\top X}{n} + \lambda I_d \right)^{-1} \frac{X^\top Y}{n}, \qquad \forall \lambda > 0$$

- Coefficient of $\widehat{\beta}_\lambda$ are shrunk continuously as λ grows (continuous function)
- Contrasts with the LASSO estimator

Estimated vector and predictor

Model (Reminder)

$$Y_i = \underbrace{X_i^\top \beta}_{=f_\beta(X_i)} + \epsilon_i$$

Ridge estimator (Reminder)

$$\widehat{\beta}_\lambda = \left(\frac{X^\top X}{n} + \lambda I_d \right)^{-1} \frac{X^\top Y}{n}, \qquad \forall \lambda > 0$$

Ridge predictor

$$f_{\widehat{\beta}_\lambda}(x) = x^\top \widehat{\beta}_\lambda = x^\top \left(\frac{X^\top X}{n} + \lambda I_d \right)^{-1} \frac{X^\top Y}{n}, \quad \forall x \in \mathbb{R}^d$$

Ridge predictor evaluated at the design points (the X_is)

$$\begin{aligned} F_{\widehat{\beta}_\lambda} &= X\widehat{\beta}_\lambda = X \left(\frac{X^\top X}{n} + \lambda I_d \right)^{-1} \frac{X^\top Y}{n} \\ &= \left(\frac{XX^\top}{n} + \lambda I_n \right)^{-1} \frac{XX^\top}{n} Y \end{aligned}$$

Kernel framework

From linear to Nonparametric regression model

- **Linear regression Model:** For all $1 \leq i \leq n$,

$$Y_i = \underbrace{X_i^\top \beta^\star}_{=f_{\beta^\star}(X_i)} + \epsilon_i = \langle X_i, \beta \rangle_{\mathbb{R}^d} + \epsilon_i$$

with $\mathbb{E}_{\epsilon_i}[\epsilon_i] = 0$, and $\mathrm{Var}_{\epsilon_i}(\epsilon_i) = \sigma^2 > 0$

- **Nonparametric regression model: For all** $1 \leq i \leq n$,

$$Y_i = f^\star(X_i) + \epsilon_i,$$

with $\mathbb{E}_{\epsilon_i}[\epsilon_i] = 0$, and $\mathrm{Var}_{\epsilon_i}(\epsilon_i) = \sigma^2 > 0$

Remark:

- No necessary Gaussian assumption!
- $f^\star$: not necessarily linear
- $f^\star(x) = \mathbb{E}[Y \mid X = x]$: regression function

Regression function and smoothness

$$Y_i = f^\star(X_i) + \epsilon_i,$$

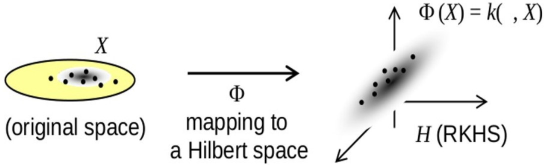
Estimating $f^\star$ and Smoothness

- Estimating $f^\star$ requires defining candidate functions (Statistical Model)
- Classical assumptions are that:
  - $f^\star$ bounded a.s. (because Y_i is so!)
  - $f^\star \in L^2(P_X)$, where P_X: Proba. distrib. of X
  - ...
- Here, we assume that there exists a reproducing kernel $k(\cdot,\cdot)$ and an RKHS $\mathcal{H}$ (uniquely defined from the kernel) such that $f^\star$ is "not too far" from $\mathcal{H}$

Kernel, Canonical feature map, RKHS

Mapping data from $\mathcal{X}$ to $\mathcal{H}$

- $k(\cdot,\cdot)$: reproducing kernel
- $x \in \mathbb{R}^d \mapsto k_x \in \mathcal{H}$: canonical feature map from $\mathbb{R}^d$ to $\mathcal{H}$
- $k_x = \phi(x) \in \mathcal{H}$: new "observation" in the RKHS



Reminder

Definition (Hilbert space)

Vector space endowed with a scalar product (pre-Hilbertian space), which is complete for the induced norm

Kernelized optimization problem

Classical Ridge

For every $\lambda > 0$, solve $\qquad (\|\beta\|_2^2 = \sum_{j=1}^d |\beta_j|^2)$

$$\widehat{\beta}_\lambda = Arg \min_{\beta \in \mathbb{R}^d} \left\{ \frac{1}{n} \sum_{i=1}^n \left(Y_i - f_\beta(X_i) \right)^2 + \lambda \|\beta\|_2^2 \right\}$$

where $f_\beta(x) = \langle x, \beta \rangle_{\mathbb{R}^d}$ and $\|u\|_n^2 = \sum_{i=1}^n u_i^2/n$

(Important normalization by $1/n$!)

Kernelized optimization problem

Classical Ridge

For every $\lambda > 0$, solve $\quad (\|\beta\|_2^2 = \sum_{j=1}^d |\beta_j|^2)$

$$\widehat{\beta}_\lambda = Arg \min_{\beta \in \mathbb{R}^d} \left\{ \frac{1}{n} \sum_{i=1}^n (Y_i - f_\beta(X_i))^2 + \lambda \|\beta\|_2^2 \right\}$$

where $f_\beta(x) = \langle x, \beta \rangle_{\mathbb{R}^d}$ and $\|u\|_n^2 = \sum_{i=1}^n u_i^2/n$

(Important normalization by $1/n$!)

Kernel Ridge Regression (KRR)

For every $\lambda > 0$, solve

$$\widehat{f}_\lambda = Arg \min_{f \in \mathcal{H}} \left\{ \frac{1}{n} \sum_{i=1}^n (Y_i - f(X_i))^2 + \lambda \|f\|_{\mathcal{H}}^2 \right\}$$

Remark:

Optimization over an inifinite dimensional space $\mathcal{H}$

Representer theorem

$$\widehat{f}_\lambda = Arg \min_{f \in \mathcal{H}} \left\{ \frac{1}{n} \sum_{i=1}^{n} (Y_i - f(X_i))^2 + \lambda \|f\|_{\mathcal{H}}^2 \right\}$$

Theorem (Representer theorem)

$\Psi : \mathbb{R}^n \times \mathbb{R}_+ \to \mathbb{R}$, *non-decreasing w.r.t. its* $n+1$*th argument*

$$Arg \min_{g \in \mathcal{H}} \{\Psi [g(x_1), \dots, g(x_n), \|g\|_{\mathcal{H}}]\}$$

Any solution $\widehat{g}$ *to the above optimization problem can be written as*

$$\widehat{g}(x) = \sum_{i=1}^{n} \widehat{\alpha}_i k(x_i, x), \quad \forall x \in \mathcal{X}$$

where $\widehat{\alpha}_i \in \mathbb{R}$, *for all* $1 \leq i \leq n$

Representer theorem (Proof)

- $S = \mathrm{Vect}(\{k_{x_1}, \ldots, k_{x_n}\}) \subset \mathcal{H}$
- For all $g \in \mathcal{H}$,

$$g = p_S^{\perp}(g) + \overbrace{\left(g - p_S^{\perp}(g)\right)}^{\in S^{\perp}}$$

Then, $\|g\|_{\mathcal{H}} \geq \left\|p_S^{\perp}(g)\right\|_{\mathcal{H}}$ (orthog. proj.)

- Besides,

$$\begin{aligned} g(x_i) &= \langle g, k_{x_i}\rangle_{\mathcal{H}} && \text{(reproducing property)} \\ &= \left\langle p_S^{\perp}(g), k_{x_i}\right\rangle_{\mathcal{H}} && \text{(since } k_{x_i} \in S\text{)} \end{aligned}$$

Then (Ψ non-decreasing)

$$\begin{aligned} &\Psi\left[g(x_1), \ldots, g(x_n), \|g\|_{\mathcal{H}}\right] \\ &= \Psi\left[\langle g, k_{x_1}\rangle_{\mathcal{H}}, \ldots, \langle g, k_{x_n}\rangle_{\mathcal{H}}, \|g\|_{\mathcal{H}}\right] \\ &= \Psi\left[\left\langle p_S^{\perp}(g), k_{x_1}\right\rangle_{\mathcal{H}}, \ldots, \left\langle p_S^{\perp}(g), k_{x_n}\right\rangle_{\mathcal{H}}, \|g\|_{\mathcal{H}}\right] \\ &\geq \Psi\left[\left\langle p_S^{\perp}(g), k_{x_1}\right\rangle_{\mathcal{H}}, \ldots, \left\langle p_S^{\perp}(g), k_{x_n}\right\rangle_{\mathcal{H}}, \left\|p_S^{\perp}(g)\right\|_{\mathcal{H}}\right] \end{aligned}$$

Hence, any minimizer belongs to S ∎

Closed-form expression

Conclusion for the KRR estimator

$$\widehat{f}_\lambda = \sum_{i=1}^n \widehat{\alpha}_i k_{x_i} \in \mathcal{H}, \qquad \forall \lambda > 0$$

where $\{\widehat{\alpha}_i\}_i \subset \mathbb{R}^n$ are to be calculated

Estimateur evaluated at the design points

$$\widehat{F}_\lambda = (\widehat{f}_\lambda(x_1), \ldots, \widehat{f}_\lambda(x_n))^\top = K\widehat{\alpha}$$

Then

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n \left(Y_i - \widehat{f}_\lambda(X_i)\right)^2 + \lambda \|f\|_{\mathcal{H}}^2 &= \left\|\mathrm{Y} - \widehat{F}_\lambda\right\|_n^2 + \lambda \|f\|_{\mathcal{H}}^2 \\ &= \|\mathrm{Y} - K\widehat{\alpha}\|_n^2 + \lambda \widehat{\alpha}^\top K \widehat{\alpha} \\ &= \varphi(\widehat{\alpha})\end{aligned}$$

Closed-form expression (Cont'd)

- $\partial\varphi(\widehat{\alpha}) = 0 \Leftrightarrow \frac{1}{n}K(\mathrm{Y} - K\widehat{\alpha}) = \lambda K\widehat{\alpha}$
  $\Leftrightarrow (K + n\lambda I_n)^{-1} K\mathrm{Y} = K\widehat{\alpha} = \widehat{F}_\lambda$

Closed-form expression (Cont'd)

- $\partial\varphi(\widehat{\alpha}) = 0 \Leftrightarrow \frac{1}{n}K(\mathrm{Y} - K\widehat{\alpha}) = \lambda K\widehat{\alpha}$
  $\Leftrightarrow (K + n\lambda I_n)^{-1} K\mathrm{Y} = K\widehat{\alpha} = \widehat{F}_\lambda$
- $\widehat{F}_\lambda = \left(\left\langle \widehat{f}_\lambda, k_{x_1} \right\rangle_{\mathcal{H}}, \ldots, \left\langle \widehat{f}_\lambda, k_{x_n} \right\rangle_{\mathcal{H}}\right)^\top$ (reprod. prop.)

Closed-form expression (Cont'd)

- $\partial\varphi(\widehat{\alpha}) = 0 \Leftrightarrow \frac{1}{n}K(\mathrm{Y} - K\widehat{\alpha}) = \lambda K\widehat{\alpha}$
  $\Leftrightarrow (K + n\lambda I_n)^{-1} K\mathrm{Y} = K\widehat{\alpha} = \widehat{F}_\lambda$
- $\widehat{F}_\lambda = \left(\left\langle \widehat{f}_\lambda, k_{x_1} \right\rangle_{\mathcal{H}}, \ldots, \left\langle \widehat{f}_\lambda, k_{x_n} \right\rangle_{\mathcal{H}} \right)^\top$ (reprod. prop.)
- $$\begin{aligned}(K + n\lambda I_n)^{-1} K\mathrm{Y} &= K\left(K + n\lambda I_n\right)^{-1}\mathrm{Y} \\ &= \sum_{j=1}^n K_{\cdot,j}\left[\left(K + n\lambda I_n\right)^{-1}\mathrm{Y}\right]_j\end{aligned}$$

Closed-form expression (Cont'd)

- $\partial\varphi(\widehat{\alpha}) = 0 \Leftrightarrow \frac{1}{n}K(\mathrm{Y} - K\widehat{\alpha}) = \lambda K\widehat{\alpha}$
  $\Leftrightarrow (K + n\lambda I_n)^{-1} K\mathrm{Y} = K\widehat{\alpha} = \widehat{F}_\lambda$
- $\widehat{F}_\lambda = \left(\left\langle \widehat{f}_\lambda, k_{x_1}\right\rangle_{\mathcal{H}}, \ldots, \left\langle \widehat{f}_\lambda, k_{x_n}\right\rangle_{\mathcal{H}}\right)^\top$ (reprod. prop.)
- $$\begin{aligned}(K + n\lambda I_n)^{-1} K\mathrm{Y} &= K\left(K + n\lambda I_n\right)^{-1}\mathrm{Y} \\ &= \sum_{j=1}^n K_{\cdot,j}\left[\left(K + n\lambda I_n\right)^{-1}\mathrm{Y}\right]_j\end{aligned}$$

Since $K_{\cdot,j} = \left(\left\langle k_{x_1}, k_{x_j}\right\rangle_{\mathcal{H}}, \ldots, \left\langle k_{x_n}, k_{x_j}\right\rangle_{\mathcal{H}}\right)^\top$, we get

$$\widehat{f}_\lambda = \sum_{j=1}^n k_{x_j} \underbrace{\left[\left(K + n\lambda I_n\right)^{-1}\mathrm{Y}\right]_j}_{=\widehat{\alpha}_j} \in \mathcal{H}$$

No *a priori* unicity: one possible solution at this stage!

Unicity of the solution

- Assume $K\widehat{\alpha} = K\widehat{\beta} = \widehat{F}_\lambda \in \mathbb{R}^n$

Unicity of the solution

- Assume $K\widehat{\alpha} = K\widehat{\beta} = \widehat{F}_\lambda \in \mathbb{R}^n$
- $K(\widehat{\alpha} - \widehat{\beta}) = 0 \Leftrightarrow \widehat{\alpha} - \widehat{\beta} \in Null(K)$

Unicity of the solution

- Assume $K\widehat{\alpha} = K\widehat{\beta} = \widehat{F}_\lambda \in \mathbb{R}^n$
- $K(\widehat{\alpha} - \widehat{\beta}) = 0 \Leftrightarrow \widehat{\alpha} - \widehat{\beta} \in Null(K)$
- Moreover

$$
\begin{aligned}
& \left\| \sum_{i=1}^{n} \widehat{\alpha}_i k_{x_i} - \sum_{i=1}^{n} \widehat{\beta}_i k_{x_i} \right\|_{\mathcal{H}} \\
& = \left(\widehat{\alpha} - \widehat{\beta}\right)^\top K \left(\widehat{\alpha} - \widehat{\beta}\right) \\
& = 0
\end{aligned}
$$

Computing the whole solution path

Solution path

$$\lambda > 0 \mapsto \sum_{j=1}^{n} k_{x_j} \left[(K + n\lambda I_n)^{-1} \mathrm{Y} \right]_j$$

Efficient computations

- Computing the whole path is possible by means of only one (careful) SVD
- When $K = \mathrm{X}\mathrm{X}^\top$ (linear kernel $\leftrightarrow$ classical Ridge regression), both $\mathrm{X}\mathrm{X}^\top$ $(n \times n)$ and $\mathrm{X}^\top\mathrm{X}$ $(d \times d)$ share the same non-zero singular values