Sylvain Le Corff

Introduction to statistical learning

Contents

Chapter 1

Supervised learning

Contents

Keywords 1.1 *Bayes classifier, empirical risk, oracle inequality, linear discriminant analysis, logistic regression.*

In a supervised learning framework, a set $\{(X_i,Y_i)\}_{1\leqslant i\leqslant n}$ of input data (also referred to as *features*) $X_i \in \mathscr{X}$ and output data $Y_i \in \mathscr{Y}$ (also referred to as *observations*), for $1 \leqslant i \leqslant n$, is available, where $\mathscr{X}$ is a general feture space and $\mathscr{Y}$ is a general observation space. In a supervised classification setting, the problem is to learn wether an individual from a given state space $\mathscr{X}$ belongs to some class, so that $\mathscr{Y} = \{1,\dots,M\}$ for some $M \geqslant 1$. In a regression framework, the observation set $\mathscr{Y}$ is usually a subset of $\mathbb{R}^m$. The state space $\mathscr{X}$ is usually a subset of $\mathbb{R}^d$ and an element of $\mathscr{X}$ contains all the features the observation prediction is based on.

One of the main goals of supervised learning is to design an automatic procedure, based on the *training dataset* $\{(X_i,Y_i)\}_{1\leqslant i\leqslant n}$, to predict the observation $y \in \mathscr{Y}$ associated with an input $x \in \mathscr{X}$ which is not in the training dataset.

The simulations presented in these notes can be found at `https://sylvainlc.github.io/`. Most elementary numerical solutions are based on scikit-learn, the website `https://scikit-learn.org/stable/supervised_learning.html` provides many helpful comments and advices.

1.1 Losses and risks

In these notes, we consider that $\{(X_i,Y_i)\}_{1\leqslant i\leqslant n}$ are independent and identically distributed (i.i.d.) with the same distribution as a couple of random variables (X,Y) defined on a measured space $(\Omega,\mathscr{F},\mathbb{P})$ and taking values in $\mathscr{X}\times\mathscr{Y}$. The joint distribution of (X,Y) is unknown. A loss function is used to evaluate the prediction of the observations: $\ell : \mathscr{Y}\times\mathscr{Y} \to \mathbb{R}$.

- In a classification setting where $\mathscr{Y} = \{1,\dots,M\}$ for some $M \geqslant 1$, a common loss function is $\ell : (y,y') \mapsto \mathbb{1}_{y\neq y'}$. This 0-1 loss outputs 1 if the prediction $y' \in \mathscr{Y}$ is different from the true class $y \in \mathscr{Y}$.
- In a regression setting, common loss functions are $\ell : (y,y') \mapsto \|y-y'\|_2^2$ and $\ell : (y,y') \mapsto \|y-y'\|_1$.

Once the loss function is chosen, the expected risk allows to evaluate all predictors $f : \mathscr{X} \to \mathscr{Y}$. It is defined as the expected loss between the observation Y and the prediction $f(X)$:

$$\mathsf{R}(f) = \mathbb{E}\left[\ell(Y, f(X))\right] .$$

- In a classification setting using the 0-1 loss, the risk function is $\mathsf{R}(f) = \mathbb{E}\left[\mathbb{1}_{Y \neq h(X)}\right] = \mathbb{P}(Y \neq f(X))$. It is also known as the *misclassification loss*.
- In a regression setting using the square loss, the risk function is $\mathsf{R}(f) = \mathbb{E}[\|Y - f(X)\|_2^2]$.

This risk is typically unknown as the joint distribution of (X,Y) is unknown, i.e. the expectation cannot be computed explicitly. Therefore, a classical surrogate is given by the empirical risk:

$$\mathsf{R}_n(f) = \frac{1}{n}\sum_{i=1}^{n} \ell(Y_i, f(X_i)) .$$

In empirical risk minimization, we often define a parameterized family of predictors $\{f_\theta\}_{\theta\in\Theta}$ where $\Theta \in \mathbb{R}^q$ and for all $\theta \in \Theta$, $f_\theta : \mathscr{X} \to \mathscr{Y}$ and seek to minimize the empirical risk over this parameterized family:

$$\widehat{\theta}_n \in \underset{\theta\in\Theta}{\arg\min} \left\{ \mathsf{R}_n(f_\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell(Y_i, f_\theta(X_i)) \right\} .$$

Cross-validation

In practice, several parameterized families of predictors can be considered to minimize the empirical risk. Each parameterized family provides an empirical best predictor. Cross-validation (CV) provides appealing strategies for algorithm selection. Cross-validation proposes to split the data to estimate the risk of each algorithm: part of data is used to train each algorithm (or minimize the empirical risk over each parameterized family), and the remaining part is used to estimate the risk of the estimated predictor. In [Arlot and Celisse, 2010], the authors provide a survey of the most common cross-validation techniques and a few guidelines to choose the best technique depending on the statistical learning setting.

The most widespread technique is probably the k-fold cross-validation approach, see [Geisser, 1975]. In this case, the training dataset is first randomly partitioned into k subset $\mathscr{D}_i$, $1 \leqslant i \leqslant k$. Each subset $\mathscr{D}_i$, $1 \leqslant i \leqslant k$, is iteratively removed from the dataset before training the algorithm. Then, the empirical risk of the trained algorithm is computed over $\mathscr{D}_i$. The estimated risk of the algorithm is given by the empirical mean of the risks obtained when each $\mathscr{D}_i$, $1 \leqslant i \leqslant k$, is removed from the training dataset and used to evaluate the risk. Additional details and illustrations can be found for instance here: `https://scikit-learn.org/stable/modules/cross_validation.html`.

1.2 Bayes classifier

Let $(\Omega, \mathscr{F}, \mathbb{P})$ be a probability space. Assume that (X,Y) is a couple of random variables defined on $(\Omega, \mathscr{F}, \mathbb{P})$ and taking values in $\mathscr{X} \times \{-1,1\}$ where $\mathscr{X}$ is a given state space, which means that the focus is set on a two-class classification problem. One aim of supervised classification is to define a function $h : \mathscr{X} \to \{-1,1\}$, called *classifier*, such that $h(X)$ is the best prediction of Y in a given context. For instance, the risk of misclassification of h is

$$\mathsf{R}_{\text{miss}}(h) = \mathbb{E}\left[\mathbb{1}_{Y \neq h(X)}\right] = \mathbb{P}(Y \neq h(X)) .$$

Note that $\mathbb{E}[Y|X]$ is a random variable measurable with respect to the σ-algebra $\sigma(X)$. Therefore, there exists a function $\eta : \mathscr{X} \to [-1,1]$ so that $\mathbb{E}[Y|X] = \eta(X)$ almost surely.

Lemma 1.1 *The classifier $h_\star$, defined for all $x \in \mathscr{X}$, by*

$$h_\star(x) = \begin{cases} 1 & \text{if } \eta(x) > 0\,, \\ -1 & \text{otherwise}\,, \end{cases}$$

is such that

$$h_\star = \underset{h:\mathscr{X}\to\{-1,1\}}{\arg\min}\ \mathsf{R}_{\text{miss}}(h)\,.$$

PROOF. For all $u, v \in \{-1,1\}$, $\mathbb{1}\{u \neq v\} = \mathbb{1}\{uv = -1\} = (1-uv)/2$. Since Y and $h(X)$ take values in $\{-1,1\}$, this implies

$$\mathsf{R}_{\text{miss}}(h) = \mathbb{P}(Y \neq h(X)) = (1 - \mathbb{E}[Yh(X)])/2\,. \tag{1.1}$$

Now, using sucessively the tower property, the equality $|u| = u \times \operatorname{sgn}(u)$, and the tower property again,

$$\mathbb{E}[Yh(X)] = \mathbb{E}[\mathbb{E}[Y|X]h(X)] \leq \mathbb{E}[|\mathbb{E}[Y|X]|\underbrace{|h(X)|}_{=1}] = \mathbb{E}[\mathbb{E}[Y|X]\underbrace{\operatorname{sgn}(\mathbb{E}[Y|X])}_{h_\star(X)}] = \mathbb{E}[Yh_\star(X)]$$

Plugging this into (1.1) yields $\mathsf{R}_{\text{miss}}(h) \geq \mathsf{R}_{\text{miss}}(h_\star)$, which concludes the proof. ■

Note that

$$\mathbb{E}[Y|X] = \mathbb{P}(Y=1|X) - \mathbb{P}(Y=-1|X) = 2\mathbb{P}(Y=1|X) - 1\,,$$

which motivates this alternative definition of $h_\star$.

Definition 1.2. The classifier $h_\star$ is called the Bayes classifier. It may also be written as follows:

$$h_\star(X) = \begin{cases} 1 & \text{if } \mathbb{P}(Y=1|X) > 1/2 \text{ i.e. if } \mathbb{P}(Y=1|X) > \mathbb{P}(Y=-1|X)\,, \\ -1 & \text{otherwise}\,. \end{cases}$$

The Bayes classifier is the optimal choice to minimize the probability of misclassification R_{miss}. However, as the conditional distribution of Y given X is usually unknown, it cannot be computed explicitly. Supervised classification aims at designing an approximate classifier $\widehat{h}_n$ using independent observations $(X_i, Y_i)_{1\leqslant i\leqslant n}$ with the same distribution as (X,Y) so that the error $\mathsf{R}_{\text{miss}}(\widehat{h}_n) - \mathsf{R}_{\text{miss}}(h_\star)$ may be controlled.

1.3 Parametric and semiparametric classifiers

1.3.1 Mixture of Gaussian distributions

In this first example, we consider a *parametric model*, that is, we assume that the joint distribution of (X,Y) belongs to a family of distributions parametrized by a vector θ with real components. For $k \in \{-1,1\}$, write $\pi_k = \mathbb{P}(Y=k)$. Assume that $\mathscr{X} = \mathbb{R}^d$ and that, conditionally on the event $\{Y=k\}$, X has a Gaussian distribution with mean $\mu_k \in \mathbb{R}^d$ and covariance matrix $\Sigma \in \mathbb{R}^{d\times d}$, whose density is denoted g_k. In this case, the parameter $\theta = (\pi_1, \mu_1, \mu_{-1}, \Sigma)$ belongs to the set $\Theta = [0,1] \times \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^{d\times d}$. The parameter π_{-1} is not part of the components of θ since $\pi_{-1} = 1 - \pi_1$. In this case, the parameter $\theta = (\pi_1, \mu_1, \mu_{-1}, \Sigma)$ belongs to the set $\Theta = [0,1] \times \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^{d\times d}$. The parameter π_{-1} is not part of the components of θ since $\pi_{-1} = 1 - \pi_1$. The explicit computation of $\mathbb{P}(Y=1|X)$ writes

$$\mathbb{P}(Y=1|X) = \frac{\pi_1 g_1(X)}{\pi_1 g_1(X) + \pi_{-1} g_{-1}(X)} = \frac{1}{1 + \frac{\pi_{-1} g_{-1}(X)}{\pi_1 g_1(X)}} = \sigma\left(\log(\pi_1/\pi_{-1}) + \log(g_1(X)/g_{-1}(X)\right)\,,$$

where $\sigma : x \mapsto (1+\mathrm{e}^{-x})^{-1}$ is the sigmoid function. Then,

$$\mathbb{P}(Y=1|X) = \sigma\left(X^\top \omega + b\right) , \tag{1.2}$$

where

$$\omega = \Sigma^{-1}(\mu_1 - \mu_{-1}) , b = \log(\pi_1/\pi_{-1}) + \frac{1}{2}(\mu_1+\mu_{-1})^\top \Sigma^{-1}(\mu_{-1}-\mu_1) .$$

Since

$$\mathbb{E}[Y|X] = \mathbb{P}(Y=1|X) - \mathbb{P}(Y=-1|X) ,$$

the Bayes classifier is such that for all $x \in \mathscr{X}$,

$$\begin{aligned}
h_\star(x) = 1 &\Leftrightarrow \mathbb{P}(Y=1|X)|_{X=x} > \mathbb{P}(Y=-1|X)|_{X=x} , \\
&\Leftrightarrow \pi_1 \exp\left\{-\frac{1}{2}(x-\mu_1)^\top \Sigma^{-1}(x-\mu_1)\right\} > \pi_{-1}\exp\left\{-\frac{1}{2}(x-\mu_{-1})^\top \Sigma^{-1}(x-\mu_{-1})\right\} , \\
&\Leftrightarrow \log\left(\frac{\pi_1}{\pi_{-1}}\right) > -\frac{1}{2}(x-\mu_{-1})^\top \Sigma^{-1}(x-\mu_{-1}) + \frac{1}{2}(x-\mu_1)^\top \Sigma^{-1}(x-\mu_1) , \\
&\Leftrightarrow \log\left(\frac{\pi_1}{\pi_{-1}}\right) > x^\top \Sigma^{-1}(\mu_{-1}-\mu_1) + \frac{1}{2}(\mu_1+\mu_{-1})^\top \Sigma^{-1}(\mu_1-\mu_{-1}) .
\end{aligned}$$

In this case, the Bayes classifier is given by

$$h_\star : x \mapsto \begin{cases} 1 \quad \text{if } \left\langle \Sigma^{-1}(\mu_1-\mu_{-1}); x - \frac{\mu_1+\mu_{-1}}{2}\right\rangle + \log\left(\frac{\pi_1}{\pi_{-1}}\right) > 0 , \\ -1 \text{ otherwise} , \end{cases}$$

Additional numerical considerations can be found for instance here `https://scikit-learn.org/stable/modules/lda_qda.html#mathematical-formulation-of-the-lda-and-qda-classifiers`.

When Σ and μ_1 and μ_{-1} are unknown, this classifier cannot be computed explicitely. We will approximate it using the observations. Assume that $(X_i,Y_i)_{1\leqslant i\leqslant n}$ are independent observations with the same distribution as (X,Y). The loglikelihood of these observations is given by

$$\begin{aligned}
\log \mathbb{P}_\theta(X_{1:n}, Y_{1:n}) &= \sum_{i=1}^n \log \mathbb{P}_\theta(X_i,Y_i) , \\
&= -\frac{nd}{2}\log(2\pi) + \sum_{i=1}^n \sum_{k\in\{-1,1\}} \mathbb{1}_{Y_i=k}\left(\log\pi_k - \frac{\log\det\Sigma}{2} - \frac{1}{2}(X_i-\mu_k)^\top \Sigma^{-1}(X_i-\mu_k)\right) , \\
&= -\frac{nd}{2}\log(2\pi) - \frac{n}{2}\log\det\Sigma + \left(\sum_{i=1}^n \mathbb{1}_{Y_i=1}\right)\log\pi_1 + \left(\sum_{i=1}^n \mathbb{1}_{Y_i=-1}\right)\log(1-\pi_1) \\
&\qquad - \frac{1}{2}\sum_{i=1}^n \mathbb{1}_{Y_i=1}(X_i-\mu_1)^\top \Sigma^{-1}(X_i-\mu_1) - \frac{1}{2}\sum_{i=1}^n \mathbb{1}_{Y_i=-1}(X_i-\mu_{-1})^\top \Sigma^{-1}(X_i-\mu_{-1}) .
\end{aligned}$$

By Lemma 3.11, the gradient of $\log \mathbb{P}_\theta(X_{1:n},Y_{1:n})$ with respect to θ is therefore given by

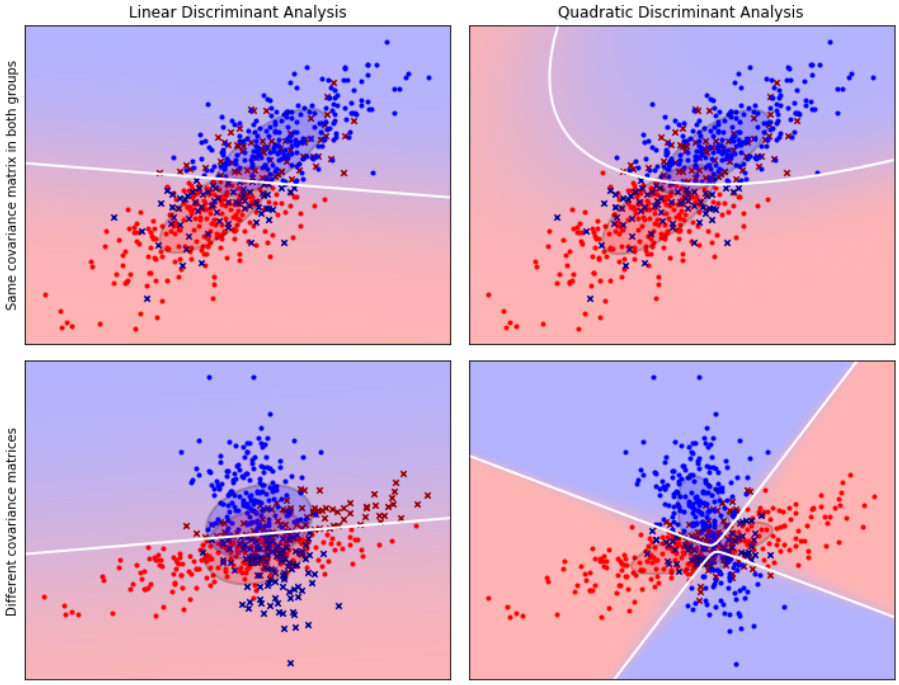

Fig. 1.1 (Top) Data are generated with the same covrariance matrix in each group. (Bottom) Data are generated with different covrariance matrices in the two groups. (Left) Classification boundary obtained with LDA, assuming the covariance matrix is the same in each group. (Right) Classification boundary obtained with QDA, assuming the covariance matrices are different in each group. Crosses are all false positives i.e. all data wrongly classified by the discriminant analysis. Simulations are inspired by `https://scikit-learn.org/stable/modules/lda_qda.html#mathematical-formulation-of-the-lda-and-qda-classifiers` and can be found here `https://sylvainlc.github.io/`.

$$
\begin{aligned}
\frac{\partial \log \mathbb{P}_\theta\left(X_{1:n}, Y_{1:n}\right)}{\partial \pi_1} &= \left(\sum_{i=1}^n \mathbb{1}_{Y_i=1}\right)\frac{1}{\pi_1} - \left(\sum_{i=1}^n \mathbb{1}_{Y_i=-1}\right)\frac{1}{1-\pi_1}\,, \\
\frac{\partial \log \mathbb{P}_\theta\left(X_{1:n}, Y_{1:n}\right)}{\partial \mu_1} &= \sum_{i=1}^n \mathbb{1}_{Y_i=1}\left(2\Sigma^{-1}X_i - 2\Sigma^{-1}\mu_1\right)\,, \\
\frac{\partial \log \mathbb{P}_\theta\left(X_{1:n}, Y_{1:n}\right)}{\partial \mu_{-1}} &= \sum_{i=1}^n \mathbb{1}_{Y_i=-1}\left(2\Sigma^{-1}X_i - 2\Sigma^{-1}\mu_{-1}\right)\,, \\
\frac{\partial \log \mathbb{P}_\theta\left(X_{1:n}, Y_{1:n}\right)}{\partial \Sigma^{-1}} &= \frac{n}{2}\Sigma - \frac{1}{2}\sum_{i=1}^n \mathbb{1}_{Y_i=1}\left(X_i-\mu_1\right)\left(X_i-\mu_1\right)^\top - \frac{1}{2}\sum_{i=1}^n \mathbb{1}_{Y_i=-1}\left(X_i-\mu_{-1}\right)\left(X_i-\mu_{-1}\right)^\top\,.
\end{aligned}
$$

The maximum likelihood estimator is defined as the only parameter $\widehat{\theta}^n$ such that all these equations are set to 0. For $k \in \{-1, 1\}$, it is given by

$$
\begin{aligned}
\widehat{\pi}_k^n &= \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{Y_i=k}\,, \\
\widehat{\mu}_k^n &= \frac{1}{\sum_{i=1}^n \mathbb{1}_{Y_i=k}}\sum_{i=1}^n \mathbb{1}_{Y_i=k}X_i\,, \\
\widehat{\Sigma}^n &= \frac{1}{n}\sum_{i=1}^n \left(X_i - \widehat{\mu}_{Y_i}^n\right)\left(X_i - \widehat{\mu}_{Y_i}^n\right)^\top\,.
\end{aligned}
$$

Therefore, a natural surrogate for the bayes classifier is

$$\widehat{h}_n : x \mapsto \begin{cases} 1 & \text{if } \left\langle \widehat{\Omega}^n \left(\widehat{\mu}_1^n - \widehat{\mu}_{-1}^n\right); x - \frac{\widehat{\mu}_1^n + \widehat{\mu}_{-1}^n}{2} \right\rangle + \log\left(\frac{\widehat{\pi}_1^n}{\widehat{\pi}_{-1}^n}\right) > 0 \,, \\ -1 & \text{otherwise} \,, \end{cases}$$

where $\widehat{\Omega}^n = (\widehat{\Sigma}^n)^{-1}$. From the asymptotic properties of the Maximum Likelihood Estimator as n goes to infinity, this classifier converges almost surely to the Bayes classifier as the number of observations n tends to infinity.

1.3.2 Logistic regression

In some situations, it may be too restrictive to assume that the joint distribution of (X,Y) belongs to a parametric family. Instead, since the Bayes classifier defined in Lemma 1.1 only depends on the conditional distribution of Y given X, we only assume that this *conditional distribution* depends on a parameter. The model is said to be *semiparametric* instead of parametric. In the case where $\mathscr{X} = \mathbb{R}^d$, one of the most widely spread model for this conditional distribution is the *logistic regression* which is defined by

$$\mathbb{P}(Y = 1|X) = \sigma(\alpha + \beta^T X) \,, \tag{1.3}$$

where $\alpha \in \mathbb{R}$, $\beta \in \mathbb{R}^d$ and σ is the sigmoid function. The parameter θ is thus $\theta = (\alpha, \beta) \in \mathbb{R} \times \mathbb{R}^d$. Note that for all $x \in \mathscr{X}$,

$$\begin{aligned} \sigma(\alpha + \beta^T x) &= \frac{1}{1 + \mathrm{e}^{-\alpha - \langle \beta; x \rangle}} \,, \\ 1 - \sigma(\alpha + \beta^T x) &= \frac{1}{1 + \mathrm{e}^{\alpha + \langle \beta; x \rangle}} \,, \\ \log\left(\frac{\sigma(\alpha + \beta^T x)}{1 - \sigma(\alpha + \beta^T x)}\right) &= \alpha + \langle \beta; x \rangle \,. \end{aligned}$$

The Bayes classifier is then given by

$$h_\star : x \mapsto \begin{cases} 1 & \text{if } \alpha + \langle \beta; x \rangle > 0 \,, \\ -1 & \text{otherwise} \,. \end{cases}$$

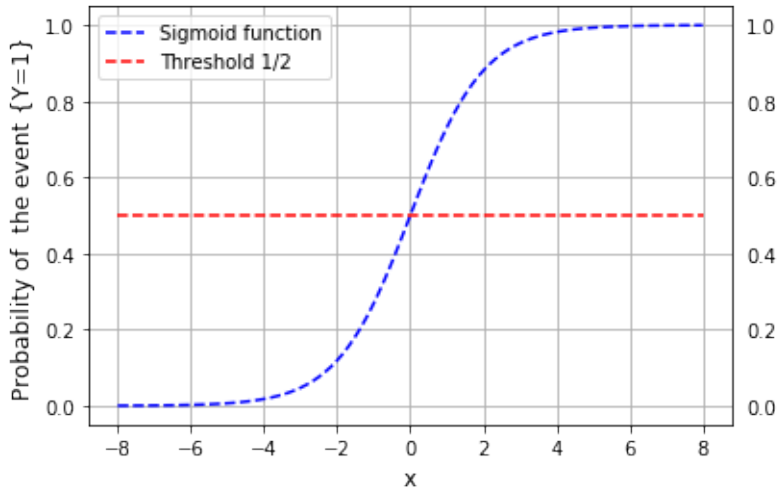

Fig. 1.2 Logistic function $\sigma : x \mapsto \mathrm{e}^x/(1+\mathrm{e}^x)$.

When α and β are unknown, this classifier cannot be computed explicitely and is approximated using the observations. Assume that $(X_i,Y_i)_{1\leqslant i\leqslant n}$ are independent observations with the same distribution as (X,Y). The conditional likelihood of the observations $Y_{1:n}$ given $X_{1:n}$ is:

$$\begin{aligned}\mathbb{P}_\theta\left(Y_{1:n}|X_{1:n}\right) &= \prod_{i=1}^n \mathbb{P}_\theta\left(Y_i|X_i\right) \ , \\ &= \prod_{i=1}^n \left(\sigma_{\alpha,\beta}\right)^{(1+Y_i)/2}\left(X_i\right)\left(1-\sigma_{\alpha,\beta}\left(X_i\right)\right)^{(1-Y_i)/2} \ , \\ &= \prod_{i=1}^n \left(\frac{\mathrm{e}^{\alpha+\langle\beta;X_i\rangle}}{1+\mathrm{e}^{\alpha+\langle\beta;X_i\rangle}}\right)^{(1+Y_i)/2}\left(\frac{1}{1+\mathrm{e}^{\alpha+\langle\beta;X_i\rangle}}\right)^{(1-Y_i)/2} \ .\end{aligned}$$

The associated conditional loglikelihood is therefore

$$\begin{aligned}\log\mathbb{P}_\theta\left(Y_{1:n}|X_{1:n}\right) &= \sum_{i=1}^n\left\{\frac{1+Y_i}{2}\log\left(\frac{\mathrm{e}^{\alpha+\langle\beta;X_i\rangle}}{1+\mathrm{e}^{\alpha+\langle\beta;X_i\rangle}}\right)+\frac{1-Y_i}{2}\log\left(\frac{1}{1+\mathrm{e}^{\alpha+\langle\beta;X_i\rangle}}\right)\right\} \ , \\ &= \sum_{i=1}^n\left\{\frac{1+Y_i}{2}\left(\alpha+\langle\beta;X_i\rangle\right)-\log\left(1+\mathrm{e}^{\alpha+\langle\beta;X_i\rangle}\right)\right\} \ .\end{aligned}$$

This conditional loglikelihood function cannot be maximized explictly yet numerous numerical optimization methods are available to maximize $(\alpha,\beta)\mapsto\log\mathbb{P}_\theta\left(Y_{1:n}|X_{1:n}\right)$. If $(\widehat{\alpha}_n,\widehat{\beta}_n)$ is an approximate solution to the optimization problem:

$$(\widehat{\alpha}_n,\widehat{\beta}_n)\in\underset{\theta=(\alpha,\beta)\in\mathbb{R}\times\mathbb{R}^d}{\arg\max}\ \log\mathbb{P}_\theta(Y_{1:n}|X_{1:n}) \ , \tag{1.4}$$

then the associated logistic regression classifier is given by

$$\widehat{h}_n : x\mapsto\begin{cases}1 & \text{if } \widehat{\alpha}_n+\langle\widehat{\beta}_n;x\rangle>0 \ , \\ -1 & \text{otherwise} \ ,\end{cases}$$

Even though, the model is semiparametric (and not parametric), it can be shown that, specifically for logistic regression model, the approximated classifier almost surely tends to the Bayes classifier as the number of observations n tends to infinity.

Additional numerical considerations can be found for instance here `https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression`.

1.4 Nonparametric Bayes classifier

In the case of *nonparametric* models, it is not assumed anymore that the joint law of (X,Y) belongs to any parametric or semiparametric family of models. The assumption on the distribution of (X,Y) is relaxed but instead, we will make some restrictions on the set of classifiers on which the optimisation occurs.

More precisely, we consider that the optimization of classifiers holds on a specific set $\mathscr{H}$ of classifiers (often called the *dictionary*), which may possibly not contain the Bayes classifier. Moreover, since in most cases, the classification risk R_{miss} cannot be computed nor minimized, it is instead estimated by the empirical classification risk defined as

$$\widehat{\mathsf{R}}^n_{\text{miss}}(h)=\frac{1}{n}\sum_{i=1}^n\mathbb{1}_{Y_i\neq h(X_i)} \ ,$$

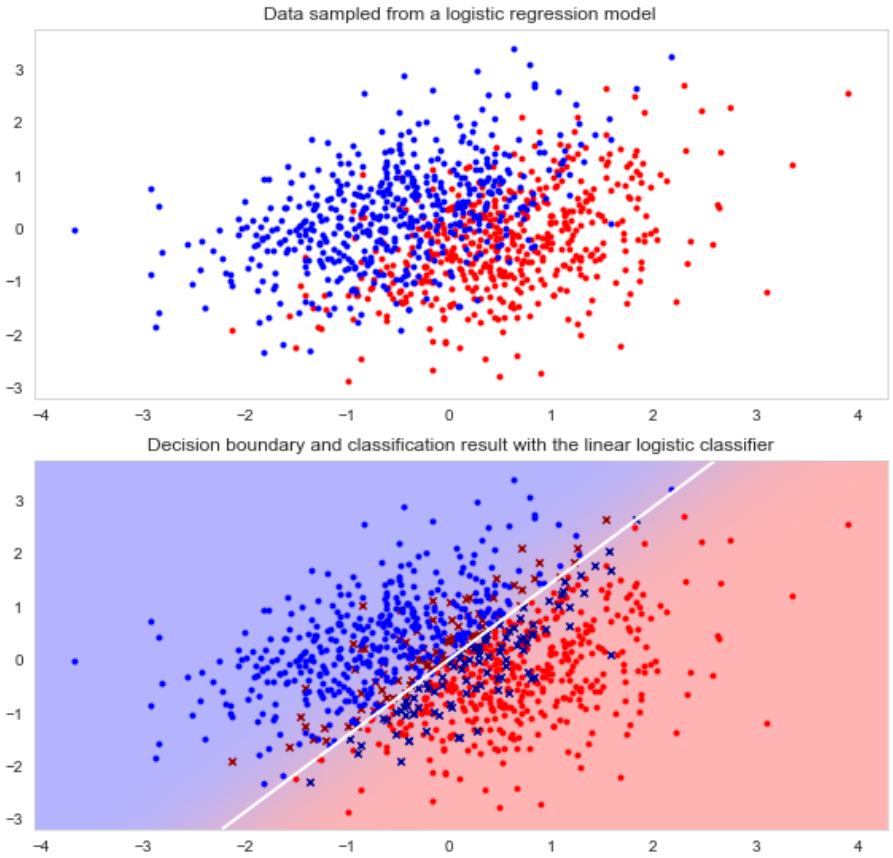

Fig. 1.3 (Top) Data are generated with a logistic regression model. The input data are Gaussian vectors in dimension $d = 2$ and the class of each data is chosen randomly according to (1.3) with a fixed weight w. (Bottom) Classification boundary obtained with the logistic classifier. Crosses are all false positives i.e. all data wrongly classified by the classifier (eventhough the weight w is known in this case). Simulations can be found here `https://sylvainlc.github.io/`.

where $(X_i, Y_i)_{1 \leqslant i \leqslant n}$ are independent observations with the same distribution as (X, Y). The classification problem then boilds down to solving

$$\widehat{h}^n_{\mathscr{H}} \in \arg\min_{h \in \mathscr{H}} \widehat{\mathsf{R}}^n_{\text{miss}}(h) \,. \tag{1.5}$$

In this context several practical and theoretical challenges arise from the minimization of the empirical classification risk. The choice of $\mathscr{H}$ is pivotal in designing an efficient classification procedure. Note that choosing $\mathscr{H}$ as all possible classifiers is meaningless, in this case, $\widehat{h}^n_{\mathscr{H}}$ is such that $\widehat{h}^n_{\mathscr{H}}(X_i) = Y_i$ for all $1 \leqslant i \leqslant n$ and $\widehat{h}^n_{\mathscr{H}}(x)$ is any element of $\{-1, 1\}$ for all $x \notin \{X_1, \ldots, X_n\}$. Although $\widehat{\mathsf{R}}^n_{\text{miss}}(h^n_{\mathscr{H}}) = 0$, is likely to be a poor approximation of $\mathsf{R}_{\text{miss}}(h^n_{\mathscr{H}})$. To understand this, the excess misclassification risk may be decomposed as follows

$$\mathsf{R}_{\text{miss}}\left(\widehat{h}^n_{\mathscr{H}}\right) - \mathsf{R}_{\text{miss}}\left(h_\star\right) = \mathsf{R}_{\text{miss}}\left(\widehat{h}^n_{\mathscr{H}}\right) - \min_{h \in \mathscr{H}} \mathsf{R}_{\text{miss}}\left(h\right) + \min_{h \in \mathscr{H}} \mathsf{R}_{\text{miss}}\left(h\right) - \mathsf{R}_{\text{miss}}\left(h_\star\right) \geqslant 0 \,.$$

The first term of the decomposition $\mathsf{R}_{\text{miss}}(\widehat{h}^n_{\mathscr{H}}) - \min_{h \in \mathscr{H}} \mathsf{R}_{\text{miss}}(h)$ is a **stochastic error** which is likely to grow when the size of $\mathscr{H}$ grows while $\min_{h \in \mathscr{H}} \mathsf{R}_{\text{miss}}(h) - \mathsf{R}_{\text{miss}}(h_\star)$ is **deterministic** and likely to decrease as the size of $\mathscr{H}$ grows.

Lemma 1.3 *For all set $\mathscr{H}$ of classifiers and all $n \geqslant 1$,*

$$\mathsf{R}_{\text{miss}}(\widehat{h}^n_{\mathscr{H}}) - \min_{h\in\mathscr{H}} \mathsf{R}_{\text{miss}}(h) \leqslant 2 \sup_{h\in\mathscr{H}} \left| \widehat{\mathsf{R}}^n_{\text{miss}}(h) - \mathsf{R}_{\text{miss}}(h) \right| . \tag{1.6}$$

PROOF. By definition of $\widehat{h}^n_{\mathscr{H}}$, for any $h \in \mathscr{H}$,

$$\begin{aligned}\mathsf{R}_{\text{miss}}(\widehat{h}^n_{\mathscr{H}}) - \min_{h\in\mathscr{H}} \mathsf{R}_{\text{miss}}(h) &= \mathsf{R}_{\text{miss}}(\widehat{h}^n_{\mathscr{H}}) - \widehat{\mathsf{R}}^n_{\text{miss}}(\widehat{h}^n_{\mathscr{H}}) + \widehat{\mathsf{R}}^n_{\text{miss}}(\widehat{h}^n_{\mathscr{H}}) - \min_{h\in\mathscr{H}} \mathsf{R}_{\text{miss}}(h) , \\ &\leqslant \mathsf{R}_{\text{miss}}(\widehat{h}^n_{\mathscr{H}}) - \widehat{\mathsf{R}}^n_{\text{miss}}(\widehat{h}^n_{\mathscr{H}}) + \widehat{\mathsf{R}}^n_{\text{miss}}(h) - \min_{h\in\mathscr{H}} \mathsf{R}_{\text{miss}}(h) .\end{aligned}$$

For all $\varepsilon > 0$ there exists $h_\varepsilon \in \mathscr{H}$ such that $\mathsf{R}_{\text{miss}}(h_\varepsilon) < \min_{h\in\mathscr{H}} \mathsf{R}_{\text{miss}}(h) + \varepsilon$ so that

$$\begin{aligned}\mathsf{R}_{\text{miss}}(\widehat{h}^n_{\mathscr{H}}) - \min_{h\in\mathscr{H}} \mathsf{R}_{\text{miss}}(h) &\leqslant \mathsf{R}_{\text{miss}}(\widehat{h}^n_{\mathscr{H}}) - \widehat{\mathsf{R}}^n_{\text{miss}}(\widehat{h}^n_{\mathscr{H}}) + \widehat{\mathsf{R}}^n_{\text{miss}}(h_\varepsilon) - \mathsf{R}_{\text{miss}}(h_\varepsilon) + \varepsilon , \\ &\leqslant 2 \sup_{h\in\mathscr{H}} \left| \widehat{\mathsf{R}}^n_{\text{miss}}(h) - \mathsf{R}_{\text{miss}}(h) \right| + \varepsilon ,\end{aligned}$$

which concludes the proof. ■

Oracle inequality when $\mathscr{H}$ is finite

This section considers the simple case where the dictionary is finite, i.e., $\mathscr{H} = \{h_1, \ldots, h_M\}$ where $M \geqslant 1$ and for all $1 \leqslant j \leqslant M$, $h_j : \mathscr{X} \to \{-1, 1\}$ is a given classifier.

Proposition 1.4 *Assume that $\mathscr{H} = \{h_1, \ldots, h_M\}$, then, for all $\delta > 0$,*

$$\mathbb{P}\left(\mathsf{R}_{\text{miss}}(\widehat{h}^n_{\mathscr{H}}) \leqslant \min_{1\leqslant j\leqslant M} \mathsf{R}_{\text{miss}}(h_j) + \sqrt{\frac{2}{n} \log\left(\frac{2M}{\delta}\right)} \right) \geqslant 1 - \delta .$$

PROOF. By Lemma 1.3, for all $u > 0$,

$$\mathbb{P}\left(\mathsf{R}_{\text{miss}}(\widehat{h}^n_{\mathscr{H}}) > \min_{1\leqslant j\leqslant M} \mathsf{R}_{\text{miss}}(h_j) + u \right) \leqslant \mathbb{P}\left(\sup_{h\in\mathscr{H}} \left| \widehat{\mathsf{R}}^n_{\text{miss}}(h) - \mathsf{R}_{\text{miss}}(h) \right| > \frac{u}{2} \right) \leqslant \sum_{j=1}^{M} \mathbb{P}\left(\left| \widehat{\mathsf{R}}^n_{\text{miss}}(h_j) - \mathsf{R}_{\text{miss}}(h_j) \right| > \frac{u}{2} \right) .$$

By Hoeffding's inequality, see Theorem 3.1,

$$\mathbb{P}\left(\mathsf{R}_{\text{miss}}(\widehat{h}^n_{\mathscr{H}}) > \min_{1\leqslant j\leqslant M} \mathsf{R}_{\text{miss}}(h_j) + u \right) \leqslant 2M \mathrm{e}^{-nu^2/2} ,$$

which concludes the proof by choosing

$$u = \sqrt{\frac{2}{n} \log\left(\frac{2M}{\delta}\right)} .$$

■

Proposition 1.5 *Assume that $\mathscr{H} = \{h_1, \ldots, h_M\}$, then,*

$$\mathbb{E}\left[\mathsf{R}_{\mathrm{miss}}(\widehat{h}^n_{\mathscr{H}})\right] \leqslant \min_{1\leqslant j\leqslant M} \mathsf{R}_{\mathrm{miss}}(h_j) + \sqrt{\frac{2\log(2M)}{n}}\,.$$

PROOF. By Lemma 1.3,

$$\mathbb{E}\left[\mathsf{R}_{\mathrm{miss}}(\widehat{h}^n_{\mathscr{H}})\right] - \min_{1\leqslant j\leqslant M} \mathsf{R}_{\mathrm{miss}}(h_j) \leqslant 2\mathbb{E}\left[\sup_{h\in\mathscr{H}} \left|\widehat{\mathsf{R}}^n_{\mathrm{miss}}(h) - \mathsf{R}_{\mathrm{miss}}(h)\right|\right] = \frac{2}{n}\mathbb{E}\left[\max_{1\leqslant j\leqslant M}\left\{n\left|\widehat{\mathsf{R}}^n_{\mathrm{miss}}(h_j) - \mathsf{R}_{\mathrm{miss}}(h_j)\right|\right\}\right]\,.$$

Note that

$$n\left\{\widehat{\mathsf{R}}^n_{\mathrm{miss}}(h_j) - \mathsf{R}_{\mathrm{miss}}(h_j)\right\} = \sum_{i=1}^n \left\{\mathbb{1}_{Y_i\neq h_j(X_i)} - \mathsf{R}_{\mathrm{miss}}(h_j)\right\}\,,$$

where the random variables $(\mathbb{1}_{Y_i\neq h_j(X_i)})_{1\leqslant i\leqslant n}$ are independent Bernoulli random variables with mean $\mathsf{R}_{\mathrm{miss}}(h_j)$. By Lemma 3.2, for all $t>0$,

$$\mathbb{E}\left[\exp\left\{t\sum_{i=1}^n \left\{\mathbb{1}_{Y_i\neq h_j(X_i)} - \mathsf{R}_{\mathrm{miss}}(h_j)\right\}\right\}\right] = \prod_{i=1}^n \mathbb{E}\left[\exp\left\{t\left(\mathbb{1}_{Y_i\neq h_j(X_i)} - \mathsf{R}_{\mathrm{miss}}(h_j)\right)\right\}\right] \leqslant \mathrm{e}^{nt^2/8}$$

and similarly

$$\mathbb{E}\left[\exp\left\{-t\sum_{i=1}^n \left\{\mathbb{1}_{Y_i\neq h_j(X_i)} - \mathsf{R}_{\mathrm{miss}}(h_j)\right\}\right\}\right] \leqslant \mathrm{e}^{nt^2/8}\,.$$

Then, for all $t>0$, by Jensen's inequality,

$$\begin{aligned}\exp\left\{t\mathbb{E}\left[\max_{1\leqslant j\leqslant M}\left\{n\left|\widehat{\mathsf{R}}^n_{\mathrm{miss}}(h_j) - \mathsf{R}_{\mathrm{miss}}(h_j)\right|\right\}\right]\right\} &\leqslant \mathbb{E}\left[\exp\left\{t\max_{1\leqslant j\leqslant M}\left\{n\left|\widehat{\mathsf{R}}^n_{\mathrm{miss}}(h_j) - \mathsf{R}_{\mathrm{miss}}(h_j)\right|\right\}\right\}\right]\,,\\ &\leqslant 2M\mathrm{e}^{nt^2/8}\,,\end{aligned}$$

which yields

$$\mathbb{E}\left[\max_{1\leqslant j\leqslant M}\left\{n\left|\widehat{\mathsf{R}}^n_{\mathrm{miss}}(h_j) - \mathsf{R}_{\mathrm{miss}}(h_j)\right|\right\}\right] \leqslant \frac{\log(2M)}{t} + \frac{nt}{8}\,.$$

Choosing $t=\sqrt{8\log(2M)/n}$,

$$\mathbb{E}\left[\max_{1\leqslant j\leqslant M}\left\{n\left|\widehat{\mathsf{R}}^n_{\mathrm{miss}}(h_j) - \mathsf{R}_{\mathrm{miss}}(h_j)\right|\right\}\right] \leqslant \sqrt{n\log(2M)/2}\,,$$

which concludes the proof. ■

Chapter 2

Multivariate regression

Contents

Keywords 2.1

2.1 Gaussian vectors

Definition 2.1. A random variable $X \in \mathbb{R}^d$ is a Gaussian vector if and only if, for all $a \in \mathbb{R}^d$, the random variable $\langle a\,;\,X\rangle$ is a Gaussian random variable.

For all random variable $X \in \mathbb{R}^d$, $X \sim \mathcal{N}(\mu, \Sigma)$ means that X is a Gaussian vector with mean $\mathbb{E}[X] = \mu \in \mathbb{R}^n$ and covariance matrix $\mathbb{V}[X] = \Sigma \in \mathbb{R}^{n\times n}$. The characteristic function of X is given (see exercises), for all $t \in \mathbb{R}^n$, by

$$\mathbb{E}[e^{i\langle t\,;\,X\rangle}] = e^{i\langle t\,;\,\mu\rangle - t^\top \Sigma t/2}\,.$$

Therefore, the law of a Gaussian vector is uniquely defined by its mean vector and its covariance matrix. If the covariance matrix Σ is nonsingular, then the law of X has a probability density with respect to the Lebesgue measure on $\mathbb{R}^n$ given by :

$$x \mapsto \det(2\pi\Sigma)^{-1/2} \exp\left\{-(x-\mu)^\top \Sigma^{-1}(x-\mu)/2\right\}\,,$$

where $\mu = \mathbb{E}[X]$.

Proposition 2.2 *Let $X \in \mathbb{R}^d$ be a Gaussian vector. Let $\{i_1,\dots,i_p\}$ be a subset of $\{1,\dots,d\}$, $p \geqslant 1$. If for all $1 \leqslant k \neq j \leqslant p$, $\mathrm{Cov}(X_{i_k}, X_{i_j}) = 0$, then $(X_{i_1},\dots,X_{i_p})$ are independent.*

PROOF. The random vector $(X_{i_1},\dots,X_{i_p})^\top$ is a Gaussian vector with mean $(\mathbb{E}[X_{i_1}],\dots,\mathbb{E}[X_{i_p}])^\top$ and diagonal covariance matrix $\mathrm{diag}(\mathbb{V}[X_{i_1}],\dots,\mathbb{V}[X_{i_p}])$. Consider $(\xi_{i_1},\dots,\xi_{i_p})$ i.i.d. random variables with distribution $\mathcal{N}(0,1)$ and define, for all $1 \leqslant j \leqslant p$,

$$Z_{i_j} = \mathbb{E}[X_{i_j}] + \sqrt{\mathbb{V}[X_{i_j}]}\xi_{i_j} .$$

Then, the random vector $(Z_{i_1},\dots,Z_{i_p})^\top$ is a Gaussian vector with the same mean and the same covariance matrix as $(X_{i_1},\dots,X_{i_p})^\top$. The two vectors have therefore the same characteristic function and the same law and $(X_{i_1},\dots,X_{i_p})$ are independent as $(\xi_{i_1},\dots,\xi_{i_p})$ are independent. ■

Theorem 2.3 (Cochran). *Let $X \sim \mathcal{N}(0,I_d)$ be a Gaussian vector in $\mathbb{R}^d$, F be a vector subspace of $\mathbb{R}^d$ and $F^\perp$ its orthogonal. Denote by $\pi_F(X)$ (resp. $\pi_{F^\perp}(X)$) the orthogonal projection of X on F (resp. on $F^\perp$). Then, $\pi_F(X)$ and $\pi_{F^\perp}(X)$ are independent, $\|\pi_F(X)\|_2^2 \sim \chi^2(p)$ and $\|\pi_{F^\perp}(X)\|_2^2 \sim \chi^2(d-p)$, where p is the dimension of F.*

PROOF. Let $(u_1,\dots,u_d)$ be an orthonormal basis of $\mathbb{R}^d$ where $(u_1,\dots,u_p)$ is an orthonormal basis of F and $(u_{p+1},\dots,u_d)$ and orthonormal basis of $F^\perp$. Consider the matrix $U \in \mathbb{R}^{d\times d}$ such that for all $1 \leqslant i \leqslant d$, the i-th column of U is u_i and $U_{(p)}$ (reps. $U^\perp_{(d-p)}$) the matrix made of the first p (resp. last $d-p$) columns of U. Note that

$$\pi_F(X) = \sum_{i=1}^{p} \langle X\,;\,u_i\rangle u_i ,$$

which can be written $\pi_F(X) = U_{(p)}U_{(p)}^\top X$. Similarly, $\pi_{F^\perp}(X) = U^\perp_{(d-p)}(U^\perp_{(d-p)})^\top X$ Therefore,

$$\begin{pmatrix}\pi_F(X)\\ \pi_{F^\perp}(X)\end{pmatrix} = \begin{pmatrix}U_{(p)}U_{(p)}^\top \\ U^\perp_{(d-p)}(U^\perp_{(d-p)})^\top\end{pmatrix} X$$

is a centered Gaussian vector with covariance matrix given by

$$\begin{pmatrix}U_{(p)}U_{(p)}^\top & 0\\ 0 & U^\perp_{(d-p)}(U^\perp_{(d-p)})^\top\end{pmatrix} .$$

By Proposition 2.2, $\pi_F(X)$ and $\pi_{F^\perp}(X)$ are independent. On the other hand,

$$\|\pi_F(X)\|_2^2 = \sum_{i=1}^{p}\langle X\,;\,u_i\rangle^2 \quad \text{and} \quad \|\pi_{F^\perp}(X)\|_2^2 = \sum_{i=p+1}^{d}\langle X\,;\,u_i\rangle^2 .$$

The random vector $(\langle X\,;\,u_i\rangle)_{1\leqslant i\leqslant d}$ is given by U^TX: it is a Gaussian random vector with mean 0 and covariance matrix I_d. The random variables $(\langle X\,;\,u_i\rangle)_{1\leqslant i\leqslant d}$ are therefore i.i.d. with distribution $\mathcal{N}(0,1)$, which concludes the proof. ■

2.2 Full rank multivariate regression

2.2.1 Preliminaries

In a supervised learning framework, a set $\{(X_i,Y_i)\}_{1\leqslant i\leqslant n}$ of input data (also referred to as *features*) $X_i \in \mathscr{X}$ and output data $Y_i \in \mathscr{Y}$ (also referred to as *observations*), $1 \leqslant i \leqslant n$, is available, where $\mathscr{X}$ is a general feture space and $\mathscr{Y}$ is a general observation space. For instance, in a supervised classification framework, the problem is to learn wether an individual from a given state space $\mathscr{X}$ belongs to some class in $\mathscr{Y} = \{1,\dots,M\}$. The state space $\mathscr{X}$ is usually a subset of $\mathbb{R}^d$ and an element of $\mathscr{X}$ contains all the features ued to predict the associated observation. In a regression framework, the observation set $\mathscr{Y}$ is usually a subset of $\mathbb{R}^m$.

Our aim is to introduce a regression function using the training dataset $\{(X_i,Y_i)\}_{1\leqslant i\leqslant n}$ employed to predict the observations associated with new features in a test dataset. In these lecture notes, we focus on empirical risk minimization. We consider a parameter set Θ and a family of regression functions $\{f_\theta\}_{\theta\in\Theta}$ where for all $\theta\in\Theta$, $f_\theta : \mathscr{X}\to\mathscr{Y}$. Considering first that $\mathscr{Y}=\mathbb{R}$ and $\mathscr{X}=\mathbb{R}^d$ we focus on solving the following optimization problem:

$$\widehat{\theta}_n \in \operatorname*{arg\,min}_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^{n}(Y_i - f_\theta(X_i))^2 \ .$$

The components of the vector $\widehat{\theta}_n$ are often referred to as the *weights* or the *regression coefficients*. Each component $\widehat{\theta}_n(j)$, $1\leqslant j\leqslant d$, specifies the expected change in the output when the input $X(j)$ is changed by one unit.

2.2.2 Least squares estimator

In a linear regression setting, we assume that $\Theta=\mathbb{R}^d$ and that for all $\theta\in\Theta$, $f_\theta : x\mapsto\theta^\top x$. Let $Y\in\mathbb{R}^d$ be the random (column) vector such that for all $1\leqslant i\leqslant n$, the i-th component of Y is Y_i and $X\in\mathbb{R}^{n\times d}$ the matrix with line i equal to $X_i^\top$. The least squares estimate is defined as a solution to

$$\widehat{\theta}_n \in \operatorname*{arg\,min}_{\theta\in\mathbb{R}^d} \|Y-X\theta\|_2^2 \ .$$

In a *well-specified setting*, we assumed that for all $1\leqslant i\leqslant n$, $Y_i = X_i^\top\theta_\star+\varepsilon_i$ for some unknown $\theta_\star\in\mathbb{R}^d$ where the $(\varepsilon_i)_{1\leqslant i\leqslant n}$ are i.i.d. random variables in $\mathbb{R}$. Let $\varepsilon\in\mathbb{R}^n$ be the random vector such that for all $1\leqslant i\leqslant n$, the i-th component of ε is ε_i. The model is then written

$$Y = X\theta_\star+\varepsilon \ .$$

Remark 2.4 *If the matrix X has full rank, i.e. its columns are linearly independent, then $X^\top X$ is positive definite since for all $u\in\mathbb{R}^d$, $u^\top X^\top Xu = \|Xu\|_2^2$ and therefore $u^\top X^\top Xu\geqslant 0$ and $u^\top X^\top Xu = 0$ if and only if $Xu=0$ i.e. if $u=0$.*

Proposition 2.5 *If the matrix X has full rank, then, $\widehat{\theta}_n = (X^\top X)^{-1}X^\top Y$. In the well-specified setting, this is an unbiased estimator of $\theta_\star$ and it satisfies $\mathbb{V}[\widehat{\theta}_n]=\sigma_\star^2(X^\top X)^{-1}$.*

PROOF. For all $\theta\in\mathbb{R}^d$,

$$\|Y-X\theta\|_2^2 = \|Y\|_2^2+\theta^\top X^\top X\theta+2Y^\top X\theta \ .$$

The function $\ell:\theta\mapsto\|Y\|_2^2+\theta^\top X^\top X\theta+2Y^\top X\theta$ is convex and for all $\theta\in\mathbb{R}^d$,

$$\nabla\ell(\theta) = 2X^\top X\theta+2X^\top Y \ .$$

As the matrix X has full rank, $X^\top X$ is nonsingular and $\nabla\ell(\theta)=0$ has a unique solution given by

$$\widehat{\theta}_n = (X^\top X)^{-1}X^\top Y \ .$$

First, note that $\widehat{\theta}_n$ is unbiased as

$$\mathbb{E}[\widehat{\theta}_n] = (X^\top X)^{-1}X^\top\mathbb{E}[Y] = (X^\top X)^{-1}X^\top X\theta_\star = \theta_\star \ .$$

In addition,

$$\mathbb{V}[\widehat{\theta}_n] = (X^\top X)^{-1}X^\top\mathbb{V}[Y]X(X^\top X)^{-1} = \sigma_\star^2(X^\top X)^{-1}X^\top X(X^\top X)^{-1} = \sigma_\star^2(X^\top X)^{-1} \ .$$

■

Remark 2.6 *If we assume that for all $1 \leqslant i \leqslant n$, $Y_i = X_i^\top \theta_\star + \varepsilon_i$ for some unknown $\theta_\star \in \mathbb{R}^d$ where the $(\varepsilon_i)_{1\leqslant i\leqslant n}$ are i.i.d. random variables in $\mathbb{R}$ with distribution $\mathcal{N}(0,\sigma_\star^2)$, $\widehat{\theta}_n$ is the maximum likelihood estimator of $\theta_\star$. The loglikelihood of the observations writes, for all $\theta \in \Theta$,*

$$\log p_\theta(Y_{1:n}) = \sum_{i=1}^n \log p_\theta(Y_i) = \sum_{i=1}^n \left\{ -\frac{1}{2}\log(2\pi\sigma_\star^2) - \frac{1}{2}\left(Y_i - \theta^\top X_i\right)^2 \right\} .$$

Therefore, maximizing $\theta \mapsto \log p_\theta(Y_{1:n})$ amounts to minimizing $\theta \mapsto \sum_{i=1}^n (Y_i - \theta^\top X_i)^2 = \|Y - X\theta\|_2^2$.

Remark 2.7 *The matrix $X(X^\top X)^{-1}X^\top \in \mathbb{R}^{n\times n}$ is the matrix of the orthogonal projection onto* Range(X)*, i.e., the vector space generated by the column vectors of X. First, let $v \in$ Range(X), then, there exists $u \in \mathbb{R}^d$ such that $v = Xu$ and $X(X^\top X)^{-1}X^\top v = X(X^\top X)^{-1}X^\top Xu = Xu = v$. Therefore, for all $v \in$ Range(X), $X(X^\top X)^{-1}X^\top v = v$. In addition, for all $v \in$ Range$(X)^\perp$, $X^\top v = 0$ so that $X(X^\top X)^{-1}X^\top v = 0$.*

Remark 2.8 *The projected value of Y is*

$$\widehat{Y} = X\widehat{\theta}_n = X(X^\top X)^{-1}X^\top Y .$$

In the special case where $d = 1$, $X \in \mathbb{R}^n$ and $\widehat{Y} = \{X^\top Y/(X^\top X)\}X = \{\langle X;Y\rangle/\langle X;X\rangle\}X$.

Proposition 2.9 *In the well-specified setting where $\varepsilon \sim \mathcal{N}(0,\sigma_\star^2 I_n)$, the random variable*

$$\widehat{\sigma}_n^2 = \frac{\|Y - X\widehat{\theta}_n\|_2^2}{n-d}$$

is an unbiased estimator of $\sigma_\star$. In addition, $(n-d)\widehat{\sigma}_n^2/\sigma_\star^2 \sim \chi^2(n-d)$, $\widehat{\theta}_n \sim \mathcal{N}(\theta_\star, \sigma_\star^2(X^\top X)^{-1})$ and $\widehat{\theta}_n$ and $\widehat{\sigma}_n^2$ are independent.

PROOF. By definition of $\widehat{\theta}_n$,

$$\widehat{\sigma}_n^2 = \frac{\|Y - X\widehat{\theta}_n\|_2^2}{n-d} = \frac{\|Y - X(X^\top X)^{-1}X^\top Y\|_2^2}{n-d} = \frac{\|(I_n - X(X^\top X)^{-1}X^\top)Y\|_2^2}{n-d}$$

By Remark 2.7, the matrix of the orthogonal projection on Range(X) is $X(X^\top X)^{-1}X^\top$ and therefore $(I_n - X(X^\top X)^{-1}X^\top$ is the matrix of the orthogonal projection on Range$(X)^\perp$. Then,

$$(I_n - X(X^\top X)^{-1}X^\top)Y = (I_n - X(X^\top X)^{-1}X^\top)(X\theta_\star + \varepsilon) = (I_n - X(X^\top X)^{-1}X^\top)\varepsilon .$$

By Theorem 2.3, $\|\sigma_\star^{-1}(I_n - X(X^\top X)^{-1}X^\top)\varepsilon\|_2^2$ has a χ^2 distribution with $n-d$ degrees of freedom which yields

$$\mathbb{E}[\|(I_n - X(X^\top X)^{-1}X^\top)Y\|_2^2] = \sigma_\star^2(n-d)$$

and $\mathbb{E}[\widehat{\sigma}_n^2] = \sigma_\star^2$. By Proposition 2.5, $\mathbb{E}[\widehat{\theta}_n] = \theta_\star$ and $\mathbb{V}[\widehat{\theta}_n] = \sigma_\star^2(X^\top X)^{-1}$ and $\widehat{\theta}_n$ is a Gaussian vector as an affine transformation of a Gaussian vector. Note that $(n-d)\widehat{\sigma}_n^2 = \|(I_n - X(X^\top X)^{-1}X^\top)\varepsilon\|_2^2$ and $\widehat{\theta}_n = (X^\top X)^{-1}X^\top X\theta_\star + (X^\top X)^{-1}X^\top\varepsilon$ and that $(X^\top X)^{-1}X^\top\varepsilon$ and $(I_n - X(X^\top X)^{-1}X^\top)\varepsilon$ are not correlated as

$$\mathbb{E}[(I_n - X(X^\top X)^{-1}X^\top)\varepsilon\varepsilon^\top X(X^\top X)^{-1}] = \sigma_\star^2\mathbb{E}[(I_n - X(X^\top X)^{-1}X^\top)X(X^\top X)^{-1}] = 0 .$$

The independence follows from Proposition 2.2. ■

2.2.3 Computational issues

Eventhough it is possible to compute the inverse of $X^\top X$ in a full rank setting, this matrix can be ill conditioned which may lead to numerical instability.

- In Scikit-learn, the fit function of `https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html` uses a SVD-based solver. By Proposition 3.7, if X has rank $r \geqslant 1$, there exist $\sigma_1 \geqslant \ldots \geqslant \sigma_r > 0$ such that

$$X = \sum_{k=1}^{r} \sigma_k u_k v_k^\top ,$$

where $\{u_1, \ldots, u_r\} \in (\mathbb{R}^n)^r$ and $\{v_1, \ldots, v_r\} \in (\mathbb{R}^d)^r$ are two orthonormal families. The vectors $\{\sigma_1, \ldots, \sigma_r\}$ are called singular values of A and $\{u_1, \ldots, u_r\}$ (resp. $\{v_1, \ldots, v_r\}$) are the left-singular (resp. right-singular) vectors of X. If U denotes the $\mathbb{R}^{n \times r}$ matrix with columns given by $\{u_1, \ldots, u_r\}$ and V denotes the $\mathbb{R}^{d \times r}$ matrix with columns given by $\{v_1, \ldots, v_r\}$, then the singular value decomposition of A may also be written as

$$X = U D_r V^\top ,$$

where $D_r = \mathrm{diag}(\sigma_1, \ldots, \sigma_r)$. Therefore

$$X^\top X = V D_r^2 V^\top \quad \text{and} \quad (X^\top X)^{-1} = V D_r^{-2} V^\top .$$

In this case, it is enough to compute V and D to compute $(X^\top X)^{-1}$.

- Using QR decomposition, we know that there exist an orthogonal matrix $Q \in \mathbb{R}^{n \times d}$, $Q^\top Q = I_d$, and an upper triangular matrix $R \in \mathbb{R}^{d \times d}$ such that $X = QR$. Then,

$$X^\top X \widehat{\theta}_n = X^\top Y \Leftrightarrow R \widehat{\theta}_n = Q^\top Y .$$

As the matrix R is upper triangular, the last equation can be solved using backsubstitution using for instance. The estimator $\widehat{\theta}_n$ can then de computed by i) computing the QR factorization of X, ii) computing $Q^\top Y$ and iii) solving $R\widehat{\theta}_n = Q^\top Y$.

2.3 Risk analysis of the full-rank multivariate regression

In our fixed-design setting, where the matrix X is deterministic, our aim is to minimize the fixed design risk:

$$\mathsf{R}(\theta) = \frac{1}{n} \mathbb{E}\left[\|Y - X\theta\|_2^2 \right] .$$

In the well-specified setting, note that

$$\mathsf{R}(\theta_\star) = \frac{1}{n} \mathbb{E}\left[\|Y - X\theta_\star\|_2^2 \right] = \frac{1}{n} \mathbb{E}\left[\|\varepsilon\|_2^2 \right] = \sigma_\star^2 .$$

Therefore, for all $\theta \in \Theta$,

$$\begin{aligned}
\mathsf{R}(\theta) - \mathsf{R}(\theta_\star) &= \frac{1}{n} \mathbb{E}\left[\|X\theta_\star + \varepsilon - X\theta\|_2^2 \right] - \sigma_\star^2 , \\
&= \frac{1}{n} \mathbb{E}\left[\|X(\theta_\star - \theta)\|_2^2 + \|\varepsilon\|_2^2 + 2(\theta_\star - \theta)^\top X^\top \varepsilon \right] - \sigma_\star^2 , \\
&= (\theta_\star - \theta)^\top \left(\frac{1}{n} X^\top X \right) (\theta_\star - \theta) ,
\end{aligned}$$

since $\mathbb{E}[\varepsilon] = 0$. On the other hand, a standard bias-variance decomposition yields

$$\begin{aligned}\mathbb{E}\left[\mathsf{R}(\widehat{\theta}_n)-\mathsf{R}(\theta_\star)\right] &= \mathbb{E}\left[\left(\theta_\star-\widehat{\theta}_n\right)^\top\left(\frac{1}{n}X^\top X\right)\left(\theta_\star-\widehat{\theta}_n\right)\right],\\ &= \mathbb{E}\left[\left(\theta_\star-\mathbb{E}\left[\widehat{\theta}_n\right]+\mathbb{E}\left[\widehat{\theta}_n\right]-\widehat{\theta}_n\right)^\top\left(\frac{1}{n}X^\top X\right)\left(\theta_\star-\mathbb{E}\left[\widehat{\theta}_n\right]+\mathbb{E}\left[\widehat{\theta}_n\right]-\widehat{\theta}_n\right)\right],\\ &= \left(\theta_\star-\mathbb{E}\left[\widehat{\theta}_n\right]\right)^\top\left(\frac{1}{n}X^\top X\right)\left(\theta_\star-\mathbb{E}\left[\widehat{\theta}_n\right]\right)+\mathbb{E}\left[\left(\mathbb{E}\left[\widehat{\theta}_n\right]-\widehat{\theta}_n\right)^\top\left(\frac{1}{n}X^\top X\right)\left(\mathbb{E}\left[\widehat{\theta}_n\right]-\widehat{\theta}_n\right)\right].\end{aligned}$$

Proposition 2.10 *In the well-specified setting,*

$$\mathbb{E}\left[\mathsf{R}(\widehat{\theta}_n)-\mathsf{R}(\theta_\star)\right]=\sigma_\star^2\frac{d}{n}.$$

PROOF. By Proposition 2.5, $\mathbb{E}[\widehat{\theta}_n]=\theta_\star$ so that

$$\mathbb{E}\left[\mathsf{R}(\widehat{\theta}_n)-\mathsf{R}(\theta_\star)\right]=\mathbb{E}\left[\left(\mathbb{E}\left[\widehat{\theta}_n\right]-\widehat{\theta}_n\right)^\top\left(\frac{1}{n}X^\top X\right)\left(\mathbb{E}\left[\widehat{\theta}_n\right]-\widehat{\theta}_n\right)\right].$$

In addition, by Proposition 2.5, $\mathbb{V}[\widehat{\theta}_n]=\sigma_\star^2(X^\top X)^{-1}$, hence

$$\mathbb{E}\left[\mathsf{R}(\widehat{\theta}_n)-\mathsf{R}(\theta_\star)\right]=\frac{\sigma_\star^2}{n}\mathbb{E}\left[\left(\mathbb{E}\left[\widehat{\theta}_n\right]-\widehat{\theta}_n\right)^\top\mathbb{V}[\widehat{\theta}_n]^{-1}\left(\mathbb{E}\left[\widehat{\theta}_n\right]-\widehat{\theta}_n\right)\right].$$

By Lemma 3.3,

$$\mathbb{E}\left[\mathsf{R}(\widehat{\theta}_n)-\mathsf{R}(\theta_\star)\right]=\frac{\sigma_\star^2}{n}\mathrm{Trace}(\mathbb{V}[\widehat{\theta}_n]^{-1}\mathbb{V}[\widehat{\theta}_n])=\sigma_\star^2\frac{d}{n}.$$

■

2.4 Confidence intervals and tests

Student's t-statistics

Proposition 2.11 *For all* $1\leqslant j\leqslant n$,

$$\frac{\widehat{\theta}_{n,j}-\theta_{\star,j}}{\widehat{\sigma}_n\sqrt{(X^TX)^{-1}_{j,j}}}\sim\mathscr{S}(n-d),$$

where $\mathscr{S}(n-d)$ *is the Student's t-distribution with* $n-p$ *degrees of freedom, i.e. the law of* $X/\sqrt{Y/(n-d)}$ *where* $X\sim\mathscr{N}(0,1)$ *is independent of* $Y\sim\chi^2(n-d)$.

PROOF. By definition, for all $1\leqslant j\leqslant d$,

$$\frac{\widehat{\theta}_{n,j}-\theta_{\star,j}}{\widehat{\sigma}_n\sqrt{(X^\top X)^{-1}_{j,j}}}=\frac{\sigma_\star^{-1}(\widehat{\theta}_{n,j}-\theta_{\star,j})}{\sigma_\star^{-1}\widehat{\sigma}_n\sqrt{(X^\top X)^{-1}_{j,j}}}=\frac{e_j^\top(\sigma_\star^{-1}(\widehat{\theta}_n-\theta_\star))}{\sigma_\star^{-1}\widehat{\sigma}_n\sqrt{(X^\top X)^{-1}_{j,j}}}.$$

Note that $\sigma_\star^{-1}(\widehat{\theta}_n-\theta_\star)\sim\mathscr{N}(0,(X^\top X)^{-1})$ so that $e_j^\top(\sigma_\star^{-1}(\widehat{\theta}_n-\theta_\star))\sim\mathscr{N}(0,e_j^\top(X^\top X)^{-1}e_j)$ and

$$\frac{e_j^\top(\sigma_\star^{-1}(\widehat{\theta}_n-\theta_\star))}{\sqrt{(X^\top X)^{-1}_{j,j}}}\sim\mathscr{N}(0,1).$$

In addition,

$$\sigma_\star^{-1}\widehat{\sigma}_n = \sqrt{\sigma_\star^{-2}\widehat{\sigma}_n^2} = \sqrt{\|\sigma_\star^{-1}(I_n - X(X^\top X)^{-1}X^\top)\varepsilon\|_2^2/(n-d)}\,,$$

where $\sigma_\star^{-2}\widehat{\sigma}_n^2 = \|\sigma_\star^{-1}(I_n - X(X^\top X)^{-1}X^\top)\varepsilon\|_2^2 \sim \chi^2(n-d)$. The proof is concluded by noting that $\widehat{\theta}_n$ and $\widehat{\sigma}_n^2$ are independent. ■

By Proposition 2.11, for $\alpha \in (0,1)$, if $s_{1-\alpha/2}^{n-d}$ denotes the quantile of order $1-\alpha/2$ of the law $\mathscr{S}(n-d)$, then

$$\mathbb{P}\left(\left|\frac{\widehat{\theta}_{n,j} - \theta_{\star,j}}{\widehat{\sigma}_n\sqrt{(X^TX)^{-1}_{j,j}}}\right| \leqslant s_{1-\alpha/2}^{n-d}\right) = 1-\alpha\,.$$

Therefore,

$$I_{n,j}^{n-p}(\theta_\star) = \left[\widehat{\theta}_{n,j} - \widehat{\sigma}_n s_{1-\alpha/2}^{n-d}\sqrt{(X^\top X)^{-1}_{j,j}}\,;\, \widehat{\theta}_{n,j} + \widehat{\sigma}_n s_{1-\alpha/2}^{n-d}\sqrt{(X^\top X)^{-1}_{j,j}}\right]$$

is a confidence interval for $\theta_{\star,j}$ with confidence level $1-\alpha$. The result of Proposition 2.11 may also be used to perform the test

$$\mathrm{H}_0:\ \theta_{\star,j} = 0 \quad \text{vs} \quad \mathrm{H}_1:\ \theta_{\star,j} \neq 0\,.$$

Under H_0, the random variable $T_{n,j}$ defined by

$$T_{n,j} = \frac{\widehat{\theta}_{n,j}}{\widehat{\sigma}_n\sqrt{(X^\top X)^{-1}_{j,j}}}$$

does not depend on $\theta_\star$ neither on $\sigma_\star$ and is distributed as a Student $\mathscr{S}(n-d)$ random variable. A statistical test with statistical signifiance $1-\alpha$ to decide wether $\theta_\star \neq 0$ is $T_{n,j} < s_{1-\alpha/2}^{n-d}$.

Fisher statistics

Proposition 2.12 *Let L be a $\mathbb{R}^{q\times d}$ matrix with rank $q \leqslant d$. Then,*

$$\frac{(\widehat{\theta}_n - \theta_\star)^\top L^\top (L(X^\top X)^{-1}L^\top)^{-1}L(\widehat{\theta}_n - \theta_\star)}{q\widehat{\sigma}_n^2} \sim \mathscr{F}(q, n-d)\,,$$

where $\mathscr{F}(q,n-d)$ is the Fisher distribution with q and $n-d$ degrees of freedom, i.e. the law of $(X/q)/(Y/(n-p))$ where $X \sim \chi^2(q)$ is independent of $Y \sim \chi^2(n-d)$.

PROOF. Note that $\mathrm{rank}(L(X^\top X)^{-1}L^\top) = \mathrm{rank}(LL^\top) = q$. The matrix $L(X^\top X)^{-1}L^\top$ is therefore positive definite. There exists a diagonal matrix $D \in \mathbb{R}^{q\times q}$ with positive diagonal terms and an orthogonal matrix $Q \in \mathbb{R}^{q\times q}$ such that $L(X^\top X)^{-1}L^\top = QDQ^{-1}$. The matrix $(L(X^\top X)^{-1}L^\top)^{-1/2}$ may be defined as $(L(X^\top X)^{-1}L^\top)^{-1/2} = QD^{-1/2}Q^{-1}$. It is then enough to note that $(L(X^\top X)^{-1}L^\top)^{-1/2}L(\widehat{\theta}_n - \theta_\star)/\sigma_\star \sim \mathscr{N}(0, I_q)$. Therefore,

$$\sigma_\star^{-2}\|(L(X^\top X)^{-1}L^\top)^{-1/2}L(\widehat{\theta}_n - \theta_\star)\|^2 = (\widehat{\theta}_n - \theta_\star)^\top L^\top (L(X^\top X)^{-1}L^\top)^{-1}L(\widehat{\theta}_n - \theta_\star)/\sigma_\star^2 \sim \chi^2(q)\,.$$

On the other hand, by Proposition 2.5,

$$(n-d)\sigma_\star^{-2}\widehat{\sigma}_n^2 \sim \chi^2(n-d)\,.$$

The proof is concluded by noting that $\widehat{\theta}_n$ and $\widehat{\sigma}_n^2$ are independent. ■

By Proposition 2.12, for $\alpha \in (0,1)$, if $f_{1-\alpha}^{q,n-d}$ denotes the quantile of order $1-\alpha$ of the law $\mathscr{F}(q,n-p)$, then

$$\mathbb{P}\left(\theta_\star \in \left\{\theta \in \mathbb{R}^d\,;\, (\widehat{\theta}_n - \theta)^\top L^\top (L(X^\top X)^{-1}L^\top)^{-1}L(\widehat{\theta}_n - \theta) \leqslant q\widehat{\sigma}_n^2 f_{1-\alpha}^{q,n-d}\right\}\right) = 1-\alpha\,.$$

Therefore,

$$I_n^{q,n-d}(\theta_\star) = \left\{ \theta \in \mathbb{R}^d \,;\, (\widehat{\theta}_n - \theta)^\top L^\top (L(X^\top X)^{-1} L^\top)^{-1} L(\widehat{\theta}_n - \theta) \leqslant q\widehat{\sigma}_n^2 f_{1-\alpha}^{q,n-d} \right\}$$

is a confidence region for $\theta_\star$ with confidence level $1-\alpha$. The result of Proposition 2.12 may also be used to perform the test

$$\mathrm{H}_0 : L\theta_\star = \bar{\theta} \quad \text{vs} \quad \mathrm{H}_1 : L\theta_\star \neq \bar{\theta} \,,$$

for a given $\bar{\theta} \in \mathbb{R}^d$.

Chapter 3

Penalized and sparse multivariate regression

Contents

3.1 Ridge regression

In the case where $X^\top X$ is singular (resp. has eigenvalues close to zero), the least squares estimate cannot be computed (resp. is not robust). A common approach to control the estimator variance is to solve the surrogate Ridge regression problem:

$$\widehat{\theta}^{\text{ridge}}_{n,\lambda} \in \operatorname*{arg\,min}_{\theta\in\mathbb{R}^d} \left\{ \frac{1}{n}\|Y - X\theta\|_2^2 + \lambda\|\theta\|_2^2 \right\} ,$$

where $\lambda > 0$.

Remark 3.1 *The matrix $n^{-1}X^\top X + \lambda I_n$ is definite positive for all $\lambda > 0$ as for all $u \in \mathbb{R}^d$,*

$$u^\top (n^{-1}X^\top X + \lambda I_n)u = n^{-1}\|Xu\|_2^2 + \lambda\|u\|_2^2 ,$$

which is positive for all $u \neq 0$. This remark allows to obtain the following result.

Proposition 3.2 *The unique solution to the Ridge regression problem is given by*

$$\widehat{\theta}^{\text{ridge}}_{n,\lambda} = \frac{1}{n}\left(\frac{1}{n}X^\top X + \lambda I_n\right)^{-1} X^\top Y .$$

In the well-specified setting, this estimator is biased and satisfies

$$\mathbb{E}[\widehat{\theta}^{\text{ridge}}_{n,\lambda}] - \theta_* = -\lambda \left(\frac{1}{n}X^\top X + \lambda I_n\right)^{-1} \theta_\star \,,$$
$$\mathbb{V}[\widehat{\theta}^{\text{ridge}}_{n,\lambda}] = \frac{\sigma_\star^2}{n}\left(\frac{1}{n}X^\top X + \lambda I_n\right)^{-2} \frac{1}{n}X^\top X \,.$$

PROOF. The unique expression of $\widehat{\theta}^{\text{ridge}}_{n,\lambda}$ is obtained similarly as in the proof of Proposition 2.5. Then,

$$\mathbb{E}[\widehat{\theta}^{\text{ridge}}_{n,\lambda}] = \frac{1}{n}\left(\frac{1}{n}X^\top X + \lambda I_n\right)^{-1} X^\top \mathbb{E}[Y] = \frac{1}{n}\left(\frac{1}{n}X^\top X + \lambda I_n\right)^{-1} X^\top X\theta_\star \,.$$

As the matrix $n^{-1}X^\top X$ is symmetric and real, $n^{-1}X^\top X$ is diagonalizable and $n^{-1}X^\top X$, $n^{-1}X^\top X + \lambda I_n$ and $(n^{-1}X^\top X + \lambda I_n)^{-1}$, are diagonalizable in the same orthonormal basis. Then, there exists nonnegative eigenvalues $\lambda_1 \geqslant \ldots \geqslant \lambda_d$ and orthonormal eigenvectors $u_1, \ldots u_d$ in $\mathbb{R}^d$ such that $n^{-1}X^\top X = \sum_{i=1}^d \lambda_i u_i u_i^\top$ and $(n^{-1}X^\top X + \lambda I_n)^{-1} = \sum_{i=1}^d (\lambda_i + \lambda)^{-1} u_i u_i^\top$. Therefore,

$$\mathbb{E}[\widehat{\theta}^{\text{ridge}}_{n,\lambda}] - \theta_\star = \sum_{i=1}^d \lambda_i(\lambda_i+\lambda)^{-1}u_i u_i^\top \theta_\star - \sum_{i=1}^d u_i u_i^\top \theta_\star = -\lambda \sum_{i=1}^d (\lambda_i+\lambda)^{-1} u_i u_i^\top \theta_\star = -\lambda\left(\frac{1}{n}X^\top X + \lambda I_n\right)^{-1}\theta_\star \,.$$

Similarly,

$$\mathbb{V}[\widehat{\theta}^{\text{ridge}}_{n,\lambda}] = \frac{1}{n^2}\left(\frac{1}{n}X^\top X + \lambda I_n\right)^{-1} X^\top \mathbb{V}[Y] X \left(\frac{1}{n}X^\top X + \lambda I_n\right)^{-1} = \frac{\sigma_\star^2}{n^2}\left(\frac{1}{n}X^\top X + \lambda I_n\right)^{-1} X^\top X \left(\frac{1}{n}X^\top X + \lambda I_n\right)^{-1} .$$

Therefore,

$$\mathbb{V}[\widehat{\theta}^{\text{ridge}}_{n,\lambda}] = \frac{\sigma_\star^2}{n}\left(\frac{1}{n}X^\top X + \lambda I_n\right)^{-2} \frac{1}{n}X^\top X \,.$$

■

Proposition 3.3 *In the well-specified setting, for all $\lambda > 0$,*

$$\mathbb{E}\left[\mathsf{R}(\widehat{\theta}^{\text{ridge}}_{n,\lambda}) - \mathsf{R}(\theta_\star)\right] = \lambda^2 \theta_\star^\top \left(\frac{1}{n}X^\top X + \lambda I_n\right)^{-2} \frac{1}{n}X^\top X\theta_\star + \frac{\sigma_\star^2}{n}\text{Trace}\left((n^{-1}X^\top X)^2 (n^{-1}X^\top X + \lambda I_n)^{-2}\right) \,.$$

PROOF. Following the full-rank risk analysis given in Section 2.3, we have

$$\begin{aligned}
&\mathbb{E}\left[\mathsf{R}(\widehat{\theta}^{\text{ridge}}_{n,\lambda}) - \mathsf{R}(\theta_\star)\right] \\
&\quad = \left(\theta_\star - \mathbb{E}\left[\widehat{\theta}^{\text{ridge}}_{n,\lambda}\right]\right)^\top \left(\frac{1}{n}X^\top X\right)\left(\theta_\star - \mathbb{E}\left[\widehat{\theta}^{\text{ridge}}_{n,\lambda}\right]\right) + \mathbb{E}\left[\left(\mathbb{E}\left[\widehat{\theta}^{\text{ridge}}_{n,\lambda}\right] - \widehat{\theta}^{\text{ridge}}_{n,\lambda}\right)^\top \left(\frac{1}{n}X^\top X\right)\left(\mathbb{E}\left[\widehat{\theta}^{\text{ridge}}_{n,\lambda}\right] - \widehat{\theta}^{\text{ridge}}_{n,\lambda}\right)\right] .
\end{aligned}$$

By Proposition 3.2, the bias term is given by

$$\begin{aligned}
\left(\theta_\star - \mathbb{E}\left[\widehat{\theta}^{\text{ridge}}_{n,\lambda}\right]\right)^\top \left(\frac{1}{n}X^\top X\right)\left(\theta_\star - \mathbb{E}\left[\widehat{\theta}^{\text{ridge}}_{n,\lambda}\right]\right) &= \frac{\lambda^2}{n}\left(\left(\frac{1}{n}X^\top X + \lambda I_n\right)^{-1}\theta_\star\right)^\top X^\top X \left(\left(\frac{1}{n}X^\top X + \lambda I_n\right)^{-1}\theta_\star\right) , \\
&= \frac{\lambda^2}{n}\theta_\star^\top \left(\frac{1}{n}X^\top X + \lambda I_n\right)^{-1} X^\top X \left(\frac{1}{n}X^\top X + \lambda I_n\right)^{-1}\theta_\star \,, \\
&= \lambda^2 \theta_\star^\top \left(\frac{1}{n}X^\top X + \lambda I_n\right)^{-2}\frac{1}{n}X^\top X\theta_\star \,.
\end{aligned}$$

By Lemma 3.3 and Proposition 3.2, the variance term is given by

$$\mathbb{E}\left[\left(\mathbb{E}\left[\widehat{\theta}^{\text{ridge}}_{n,\lambda}\right]-\widehat{\theta}^{\text{ridge}}_{n,\lambda}\right)^\top\left(\frac{1}{n}X^\top X\right)\left(\mathbb{E}\left[\widehat{\theta}^{\text{ridge}}_{n,\lambda}\right]-\widehat{\theta}^{\text{ridge}}_{n,\lambda}\right)\right] = \text{Trace}\left(\frac{1}{n}X^\top X\mathbb{V}[\widehat{\theta}^{\text{ridge}}_{n,\lambda}]\right),$$
$$= \frac{\sigma_\star^2}{n}\text{Trace}\left(\frac{1}{n}X^\top X\left(\frac{1}{n}X^\top X+\lambda I_n\right)^{-2}\frac{1}{n}X^\top X\right),$$
$$= \frac{\sigma_\star^2}{n}\text{Trace}\left((n^{-1}X^\top X)^2(n^{-1}X^\top X+\lambda I_n)^{-2}\right),$$

which concludes the proof. ■

Remark 3.4
- *The bias term increases with λ. It is 0 when $\lambda=0$ and it converges to $\theta_\star^\top X^\top X\theta_\star/n$ when $\lambda\to\infty$.*
- *The variance term decreases with λ. It is $\sigma_\star^2 d/n$ when $\lambda=0$ and it converges to 0 when $\lambda\to\infty$.*
- *The mean square error of the estimator is then given by*

$$\mathbb{E}\left[\left\|\widehat{\theta}^{\text{ridge}}_{n,\lambda}-\theta_*\right\|_2^2\right] = \text{Trace}\left(\mathbb{V}[\widehat{\theta}^{\text{ridge}}_{n,\lambda}]\right)+\left\|\mathbb{E}[\widehat{\theta}^{\text{ridge}}_{n,\lambda}]-\theta_*\right\|_2^2.$$

Let $(\vartheta_1,\dots,\vartheta_d)$ be an orthonormal basis of $\mathbb{R}^d$ of eigenvectors of $X^\top X$ associated with the eigenvalues $(\lambda_1,\dots,\lambda_d)\in\mathbb{R}^d$. Then,

$$\mathbb{E}\left[\left\|\widehat{\theta}^{\text{ridge}}_{n,\lambda}-\theta_*\right\|_2^2\right] = \sigma_\star^2\sum_{j=1}^d\frac{\lambda_j}{(\lambda_j+\lambda)^2}+\lambda^2\sum_{j=1}^d\frac{\langle\theta_*\,;\vartheta_j\,\rangle^2}{(\lambda_j+\lambda)^2}.$$

The mean square error is therefore a sum of two contributions, a bias related term which increases with λ and a variance related term which decreases with λ. In practice, the value of λ is chosen using cross-validation.

Using the risk analysis for the Ridge-based estimator, we can tune the regularization parameter λ to obtain a better bound than the $\sigma_\star^2 d/n$ bound of the case $\lambda=0$.

Proposition 3.5 *Choosing $\lambda=\lambda_\star$ where*

$$\lambda_\star = \frac{\sigma_\star\text{Trace}\left(X^\top X\right)^{1/2}}{\sqrt{n}\|\theta_\star\|_2}.$$

yields

$$\mathbb{E}\left[\mathsf{R}(\widehat{\theta}^{\text{ridge}}_{\lambda_\star})-\mathsf{R}(\theta_\star)\right]\leqslant\frac{\sigma_\star\|\theta_\star\|_2\text{Trace}\left(X^\top X\right)^{1/2}}{\sqrt{n}}.$$

PROOF. Let $(\vartheta_1,\dots,\vartheta_d)$ be an orthonormal basis of $\mathbb{R}^d$ of eigenvectors of $n^{-1}X^\top X$ associated with the eigenvalues $(\lambda_1,\dots,\lambda_d)\in\mathbb{R}^d$. Therefore,

$$\lambda^2\theta_\star^\top\left(\frac{1}{n}X^\top X+\lambda I_n\right)^{-2}\frac{1}{n}X^\top X\theta_\star = \lambda\sum_{i=1}^d\theta_\star^\top\frac{\lambda\lambda_i}{(\lambda_i+\lambda)^2}u_iu_i^\top\theta_\star\leqslant\frac{\lambda}{2}\|\theta_\star\|_2^2,$$

since for all $1\leqslant i\leqslant d$, $2\lambda\lambda_i\leqslant(\lambda+\lambda_i)^2$ implies $\lambda\lambda_i/(\lambda+\lambda_i)^2\leqslant 1/2$. On the other hand,

$$\frac{\sigma_\star^2}{n}\text{Trace}\left((n^{-1}X^\top X)^2(n^{-1}X^\top X+\lambda I_n)^{-2}\right)=\frac{\sigma_\star^2}{n}\text{Trace}\left(n^{-1}X^\top X\sum_{i=1}^d\frac{\lambda_i}{(\lambda+\lambda_i)^2}u_iu_i^\top\right)\leqslant\frac{\sigma_\star^2}{2n\lambda}\text{Trace}\left(n^{-1}X^\top X\right).$$

Therefore, by Proposition 3.3,

$$\mathbb{E}\left[\mathsf{R}(\widehat{\theta}^{\text{ridge}}_n)-\mathsf{R}(\theta_\star)\right]\leqslant\frac{\lambda}{2}\|\theta_\star\|_2^2+\frac{\sigma_\star^2}{2n\lambda}\text{Trace}\left(n^{-1}X^\top X\right).$$

The upper-bound is then minimized by choosing

$$\lambda_\star = \frac{\sigma_\star\text{Trace}\left(X^\top X\right)^{1/2}}{\sqrt{n}\|\theta_\star\|_2}.$$

■

3.2 Lasso regression

The Least Absolute Shrinkage and Selection Operator (Lasso) regression is a L_1 based regularized regression which aims at fostering sparsity. The objective is to solve the following minimization problem,

$$\widehat{\theta}^{\text{lasso}}_{\lambda,n} \in \underset{\theta\in\mathbb{R}^d}{\arg\min} \left\{ \frac{1}{n}\|Y - X\theta\|_2^2 + \lambda\|\theta\|_1 \right\} , \tag{3.1}$$

where $\lambda > 0$ and

$$\|\theta\|_1 = \sum_{j=1}^{d} |\theta_j| .$$

The function $\theta \mapsto n^{-1}\|Y - X\theta\|_2^2 + \lambda\|\theta\|_1$ is convex but not differentiable and the solution to this problem may not be unique.

3.2.1 Computational issues

A coordinate descent can be applied to solve the LASSO optimization problem. In this case, solving (3.1) amounts to producing iterative estimators, where at each iteration, a coordinate is selected to be updated. Then, the objective function is optimized explicitly with respect to the selected coordinate. For all $\theta \in \mathbb{R}^d$,

$$\nabla_\theta \|Y - X\theta\|_2^2 = -2X^\top (Y - X\theta) .$$

Then, for all $1 \leqslant j \leqslant d$, $(\nabla_\theta \|Y - X\theta\|_2^2)_j = -2\mathbf{X}_j^\top (Y - X\theta)$, where $\mathbf{X}_j$ is the j-th column of the matrix X. Define, for all $1 \leqslant j \leqslant d$,

$$\upsilon_j = \mathbf{X}_j^\top \left(Y - \sum_{\substack{i=1 \\ i\neq j}}^{d} \theta_i \mathbf{X}_i \right) .$$

Assuming that the columns of X are normalized, i.e. for all $1 \leqslant k \leqslant d$, $\mathbf{X}_k^\top \mathbf{X}_k = 1$, yields

$$(\nabla_\theta \|Y - X\theta\|_2^2)_j = -2(\upsilon_j - \theta_j) .$$

Consequently, for all $\theta_j \neq 0$,

$$(\nabla_\theta (n^{-1}\|Y - X\theta\|_2^2 + \lambda\|\theta\|_1))_j = \frac{2}{n}(\theta_j - \upsilon_j + \lambda n \operatorname{sign}(\theta_j)/2) .$$

For all $1 \leqslant j \leqslant d$, $\theta_j \mapsto n^{-1}\|Y - X\theta\|_2^2 + \lambda\|\theta\|_1$ is convex and grows to infinity when $|\theta_j| \to \infty$ and admits thus a minimum at some $\theta_j^\star \in \mathbb{R}$.

- If $\theta_j^\star \neq 0$, then

$$\theta_j^\star = \upsilon_j \left(1 - \frac{\lambda n \operatorname{sign}(\theta_j^\star)}{2\upsilon_j} \right) ,$$

  which yields, as $\operatorname{sign}(\theta_j^\star) = \operatorname{sign}(\upsilon_j)$,

$$\theta_j^\star = \upsilon_j \left(1 - \frac{\lambda n}{2|\upsilon_j|} \right)$$

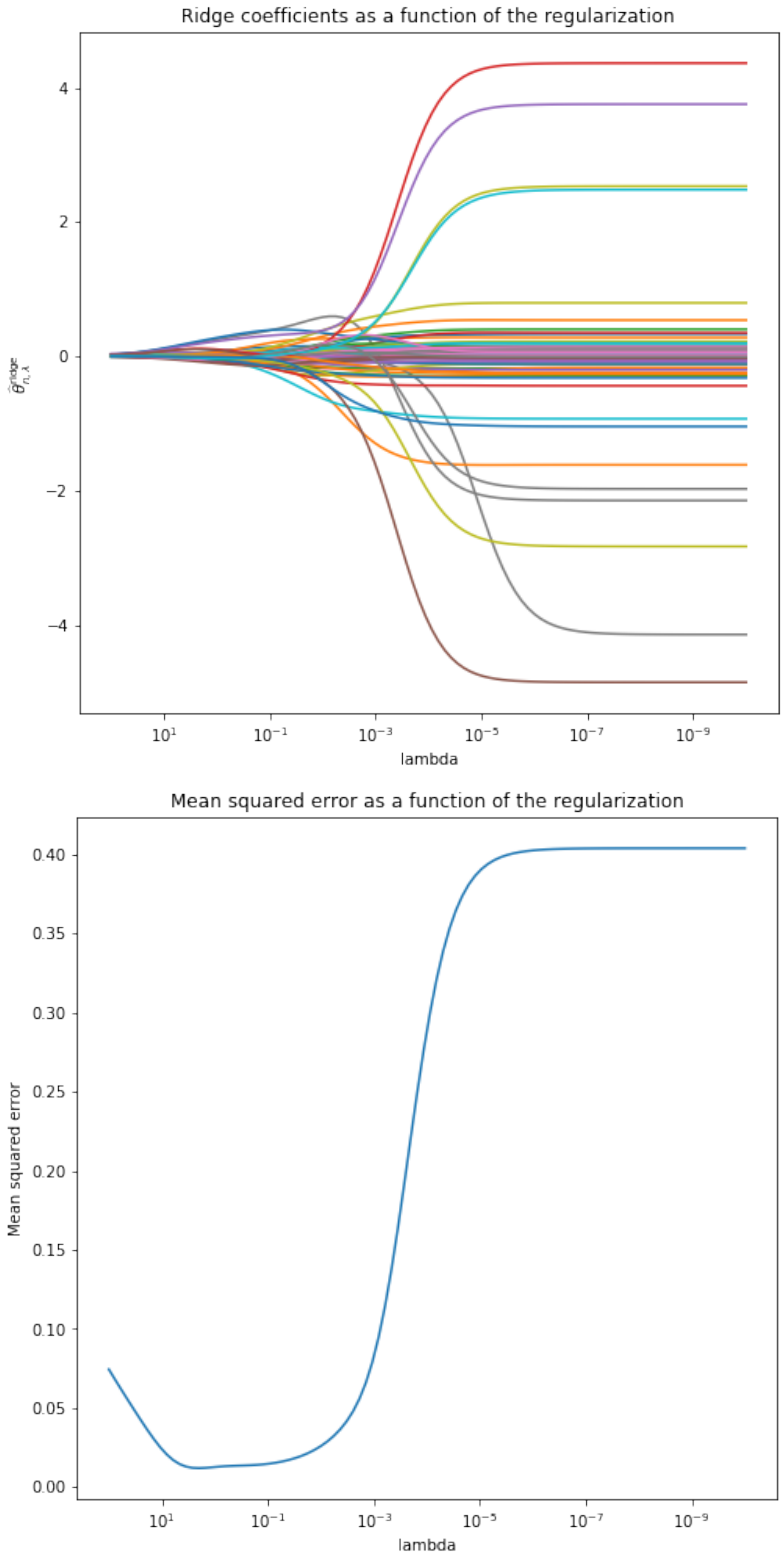

Fig. 3.1 Ridge regression is used to predict the Brazilian inflation based on many observed variables, see `https://github.com/gabrielrvsc/HDeconometrics/`. The model is trained using $n = 140$ data with for each $1 \leqslant i \leqslant n$, $X_i \in \mathbb{R}^{93}$, i.e. $d = 93$. The features are econometric data available each month. (Top) Estimated coefficient $\widehat{\theta}^{\text{ridge}}_{n,\lambda}$ as a function of λ. (Bottom) Mean squared error between the true observations and the predictions over the test set with 15 new data points.

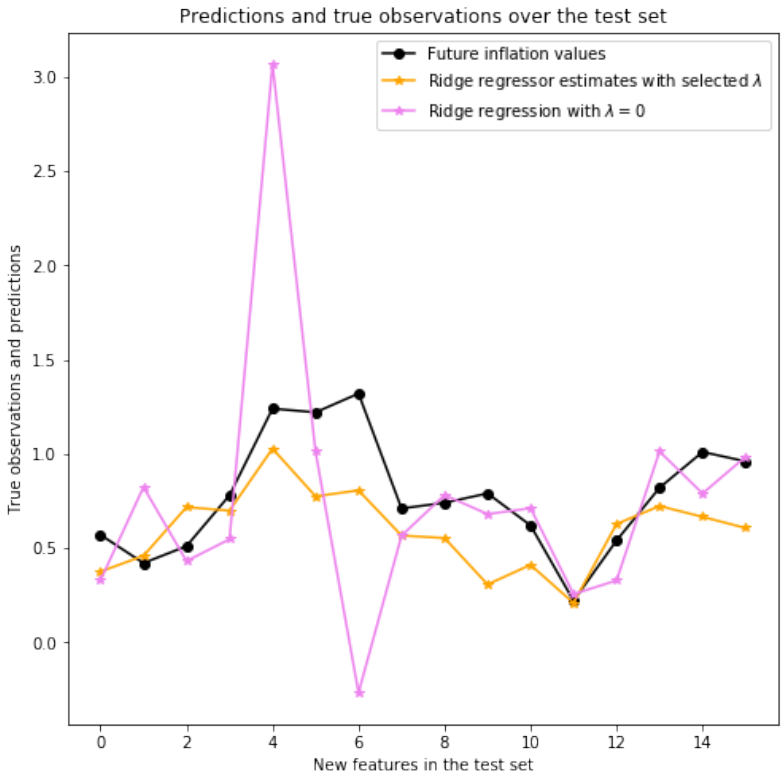

Fig. 3.2 Ridge regression is used to predict the Brazilian inflation based on many observed variables, see `https://github.com/gabrielrvsc/HDeconometrics/`. The model is trained using $n = 140$ data with for each $1 \leqslant i \leqslant n$, $X_i \in \mathbb{R}^{93}$, i.e. $d = 93$. The features are econometric data available each month. In this experiment, 15 new data points in a test set are used to evaluate the Ridge estimator. We present an ordinary least squares estimate (i.e with $\lambda = 0$) and the estimate obtained by selecting λ with a leave-one-out Cross-Validation. The MSE on the test set are 0.016 for the cross-validated λ and 0.398 for $\lambda = 0$.

and

$$1 - \frac{\lambda n}{2|\upsilon_j|} \geqslant 0\,.$$

- If $1 - \lambda n/(2|\upsilon_j|) < 0$, there is no solution to $(\nabla_\theta(n^{-1}\|Y - X\theta\|_2^2 + \lambda\|\theta\|_1))_j = 0$ for $\theta_j \neq 0$. Since $\theta_j \mapsto n^{-1}\|Y - X\theta\|_2^2 + \lambda\|\theta\|_1$ admits a minimum, $\theta_j^\star = 0$.

Therefore,

$$\theta_j^\star = \upsilon_j \left(1 - \frac{\lambda n}{2|\upsilon_j|}\right)_+ = \max\left(0; \upsilon_j\left(1 - \frac{\lambda n}{2|\upsilon_j|}\right)\right)\,.$$

An algorithm to approximatively solve the Lasso regression problem proceeds as described in Algorithm 1.

3.2.2 Risk analysis of LASSO regression problem

Proposition 3.6 *Consider a well specified model where $\varepsilon \sim \mathcal{N}(0, \sigma_\star^2 I_n)$. Then, choosing $n\lambda_\star^2/(16\sigma_\star^2\|\Sigma\|_\infty) = \log(dn)$ yields*

$$\frac{1}{n}\mathbb{E}\left[\|X(\widehat{\theta}^{\text{lasso}}_{\lambda_\star,n} - \theta_\star)\|_2^2\right] \leqslant 16\sigma_\star\sqrt{\frac{\log(dn)}{n}}\|\Sigma\|_\infty^{1/2}\|\theta_\star\|_1 + 3\sqrt{2}\frac{\sigma_\star^2}{n\sqrt{d}}\,.$$

Algorithm 1 Coordinate descent LASSO solver

Choose randomly an initial estimate $\widehat{\theta}_n^{(0)} \in \mathbb{R}^d$.
for $p = 1$ to $p = n_{\text{iter}}$ **do**
Choose randomly a coordinate $j \in \{1, \dots, d\}$.
Compute

$$\upsilon_j = \mathbf{X}_j^\top \left(Y - \sum_{\substack{i=1 \\ i \neq j}}^{d} \widehat{\theta}_{n,i}^{(p-1)} \mathbf{X}_i \right) .$$

If $1 - \lambda n/(2|\upsilon_j|) > 0$, set

$$\widehat{\theta}_{n,j}^{(p)} = \upsilon_j \left(1 - \frac{\lambda n}{2|\upsilon_j|} \right) .$$

If $1 - \lambda n/(2|\upsilon_j|) < 0$, set $\widehat{\theta}_{n,j}^{(p)} = 0$.
For all $1 \leqslant k \leqslant d$, $k \neq j$, set $\widehat{\theta}_{n,k}^{(p)} = \widehat{\theta}_{n,k}^{(p-1)}$.
end for

PROOF. By definition of $\widehat{\theta}_{\lambda_\star,n}^{\text{lasso}}$, for all $\theta \in \mathbb{R}^d$,

$$\frac{1}{n} \|Y - X\widehat{\theta}_{\lambda_\star,n}^{\text{lasso}}\|_2^2 + \lambda \|\widehat{\theta}_{\lambda_\star,n}^{\text{lasso}}\|_1 \leqslant \frac{1}{n} \|Y - X\theta_\star\|_2^2 + \lambda \|\theta_\star\|_1 .$$

As $Y = X\theta_\star + \varepsilon$, this yields

$$\frac{1}{n} \|\varepsilon - X(\widehat{\theta}_{\lambda_\star,n}^{\text{lasso}} - \theta_\star)\|_2^2 + \lambda \|\widehat{\theta}_{\lambda_\star,n}^{\text{lasso}}\|_1 \leqslant \frac{1}{n} \|\varepsilon\|_2^2 + \lambda \|\theta_\star\|_1 .$$

Therefore,

$$\frac{1}{n} \|X(\widehat{\theta}_{\lambda_\star,n}^{\text{lasso}} - \theta_\star)\|_2^2 + \lambda \|\widehat{\theta}_{\lambda_\star,n}^{\text{lasso}}\|_1 \leqslant \frac{2}{n} \varepsilon^\top X(\widehat{\theta}_{\lambda_\star,n}^{\text{lasso}} - \theta_\star) + \lambda \|\theta_\star\|_1 .$$

and

$$\begin{aligned} \|X(\widehat{\theta}_{\lambda_\star,n}^{\text{lasso}} - \theta_\star)\|_2^2 &\leqslant 2\varepsilon^\top X(\widehat{\theta}_{\lambda_\star,n}^{\text{lasso}} - \theta_\star) + \lambda n \|\theta_\star\|_1 - \lambda n \|\widehat{\theta}_{\lambda_\star,n}^{\text{lasso}}\|_1 , \\ &\leqslant 2\|X^\top \varepsilon\|_\infty \|\widehat{\theta}_{\lambda_\star,n}^{\text{lasso}} - \theta_\star\|_1 + \lambda n \|\theta_\star\|_1 - \lambda n \|\widehat{\theta}_{\lambda_\star,n}^{\text{lasso}}\|_1 . \end{aligned} \tag{3.2}$$

On the other hand, writing $\Sigma = n^{-1} X^\top X$,

$$\mathbb{P}\left(\|X^\top \varepsilon\|_\infty \geqslant \frac{n\lambda}{2} \right) \leqslant \sum_{j=1}^{d} \mathbb{P}\left(|X^\top \varepsilon|_j \geqslant \frac{n\lambda}{2} \right) \leqslant 2 \sum_{j=1}^{d} \exp\left\{ -n\lambda^2/(8\sigma_\star^2 \Sigma_{jj}) \right\} \leqslant 2d \exp\left\{ -n\lambda^2/(8\sigma_\star^2 \|\Sigma\|_\infty) \right\} ,$$

as for all $1 \leqslant j \leqslant d$, $X^\top \varepsilon \sim \mathcal{N}(0, \sigma_\star^2 X^\top X)$. Therefore, with probability at least $1 - 2d\exp\left\{ -n\lambda^2/(8\sigma_\star^2 \|\Sigma\|_\infty) \right\}$,

$$\|X(\widehat{\theta}_{\lambda_\star,n}^{\text{lasso}} - \theta_\star)\|_2^2 \leqslant \lambda n \|\widehat{\theta}_{\lambda_\star,n}^{\text{lasso}} - \theta_\star\|_1 + \lambda n \|\theta_\star\|_1 - \lambda n \|\theta_{\lambda_\star,n}^{\text{lasso}}\|_1$$

and

$$\frac{1}{n} \|X(\widehat{\theta}_{\lambda_\star,n}^{\text{lasso}} - \theta_\star)\|_2^2 \leqslant 2\lambda \|\theta_\star\|_1 .$$

Let $A = \{\|X^\top \varepsilon\|_\infty < n\lambda/2\}$. Then,

$$\mathbb{E}\left[\|X(\widehat{\theta}_{\lambda_\star,n}^{\text{lasso}} - \theta_\star)\|_2^2 \right] \leqslant 2n\lambda \|\theta_\star\|_1 + \mathbb{E}\left[\|X(\widehat{\theta}_{\lambda_\star,n}^{\text{lasso}} - \theta_\star)\|_2^2 \mathbb{1}_{A^c} \right] .$$

Note that, by (3.2), $\|X(\widehat{\theta}_{\lambda_\star,n}^{\text{lasso}} - \theta_\star)\|_2^2 \leqslant 2\|\varepsilon\|_2 \|X(\theta_{\lambda_\star,n}^{\text{lasso}} - \theta_\star)\|_2 + \lambda n \|\theta_\star\|_1 \leqslant \|\varepsilon\|_2^2/2 + \|X(\theta_{\lambda_\star,n}^{\text{lasso}} - \theta_\star)\|_2^2/2 + \lambda n \|\theta_\star\|_1$. Then,

$$\mathbb{E}\left[\|X(\widehat{\theta}_{\lambda_\star,n}^{\text{lasso}} - \theta_\star)\|_2^2 \right] \leqslant 2n\lambda \|\theta_\star\|_1 + \mathbb{E}\left[(\|\varepsilon\|_2^2 + 2\lambda n \|\theta_\star\|_1) \mathbb{1}_{A^c} \right]$$

and by Cauchy-Schwarz inequality,

$$\mathbb{E}\left[\|X(\widehat{\theta}_{\lambda_\star,n}^{\text{lasso}} - \theta_\star)\|_2^2 \right] \leqslant 4n\lambda \|\theta_\star\|_1 + \mathbb{E}\left[\|\varepsilon\|_2^4 \right]^{1/2} \mathbb{P}(A^c)^{1/2} .$$

Using that $\mathbb{E}[\|\varepsilon\|_2^4]^{1/2} \leqslant 3n\sigma_\star^2$ yields

$$\frac{1}{n}\mathbb{E}\left[\|X(\widehat{\theta}^{\text{lasso}}_{\lambda_\star,n}-\theta_\star)\|_2^2\right] \leqslant 4\lambda\|\theta_\star\|_1 + 3\sigma_\star^2\cdot\sqrt{2d}\exp\left\{-n\lambda^2/(16\sigma_\star^2\|\Sigma\|_\infty\right\} .$$

By choosing λ so that $n\lambda^2/(16\sigma_\star^2\|\Sigma\|_\infty) = \log(dn)$, we obtain

$$\frac{1}{n}\mathbb{E}\left[\|X(\widehat{\theta}^{\text{lasso}}_{\lambda_\star,n}-\theta_\star)\|_2^2\right] \leqslant 16\sigma_\star\sqrt{\frac{\log(dn)}{n}}\|\Sigma\|_\infty^{1/2}\|\theta_\star\|_1 + 3\sqrt{2}\frac{\sigma_\star^2}{n\sqrt{d}} .$$

■

3.3 Regression with infinite-dimensional models

3.3.1 Nonparametric regression

In a nonparametric regression framework, it is not assumed that the observations depend linearly on the covariates and a more general model is introduced. For all $1 \leqslant i \leqslant n$, the observation model is given by

$$Y_i = f^*(X_i) + \xi_i ,$$

where for all $1 \leqslant i \leqslant n$, $X_i \in \mathscr{X}$, and the $(\xi_i)_{1\leqslant i\leqslant n}$ are i.i.d. centered Gaussian random variables with variance σ^2. The function f^* is unknown and has to be estimated using the observations $(X_i,Y_i)_{1\leqslant i\leqslant n}$. A simple approach consists in defining an estimator of f^* as a linear combination of $M \geqslant 1$ known functions $(\varphi_1,\ldots,\varphi_M)$ defined on $\mathscr{X}$. Define $\mathscr{F}_\varphi$ as

$$\mathscr{F}_\varphi = \left\{ \sum_{j=1}^M \alpha_j\varphi_j \; ; \; (\alpha_1,\ldots,\alpha_M)\in\mathbb{R}^M \right\} .$$

Then, the least squares estimator of f^* on $\mathscr{F}_\varphi$ is defined as

$$\widehat{f}_n^\varphi \in \operatorname*{arg\,min}_{f\in\mathscr{F}_\varphi} \sum_{i=1}^n (Y_i - f(X_i))^2 .$$

Let Ψ be the $\mathbb{R}^{M\times n}$ matrix such as, for all $1 \leqslant i \leqslant n$ and $1 \leqslant j \leqslant M$, $\Psi_{i,j} = \varphi_j(X_i)$. Then, for all $f \in \mathscr{F}_\varphi$, there exists $\alpha = (\alpha_1,\ldots,\alpha_M) \in \mathbb{R}^M$ such that,

$$\sum_{i=1}^n (Y_i - f(X_i))^2 = \|Y - \Psi\alpha\|^2 .$$

Then, following the same steps as in Section 2.2, in the case where $\Psi^\top\Psi$ is nonsingular, the least squares estimate is

$$\widehat{f}_n^\varphi : x \mapsto \sum_{j=1}^M \widehat{\alpha}_{n,j}\varphi_j , \tag{3.3}$$

where

$$\widehat{\alpha}_n = (\Psi^\top\Psi)^{-1}\Psi^\top Y .$$

Introducing the function $\varphi : x \mapsto (\varphi_1(x)\ldots,\varphi_M(x))^\top$ yields the linear estimator

$$\widehat{f}_n^\varphi : x \mapsto \sum_{i=1}^n w_i(x)Y_i ,$$

where, for all $1 \leqslant i \leqslant n$,

$$w_i(x) = \left(\varphi(x)^\top(\Psi^\top\Psi)^{-1}\Psi^\top\right)_i .$$

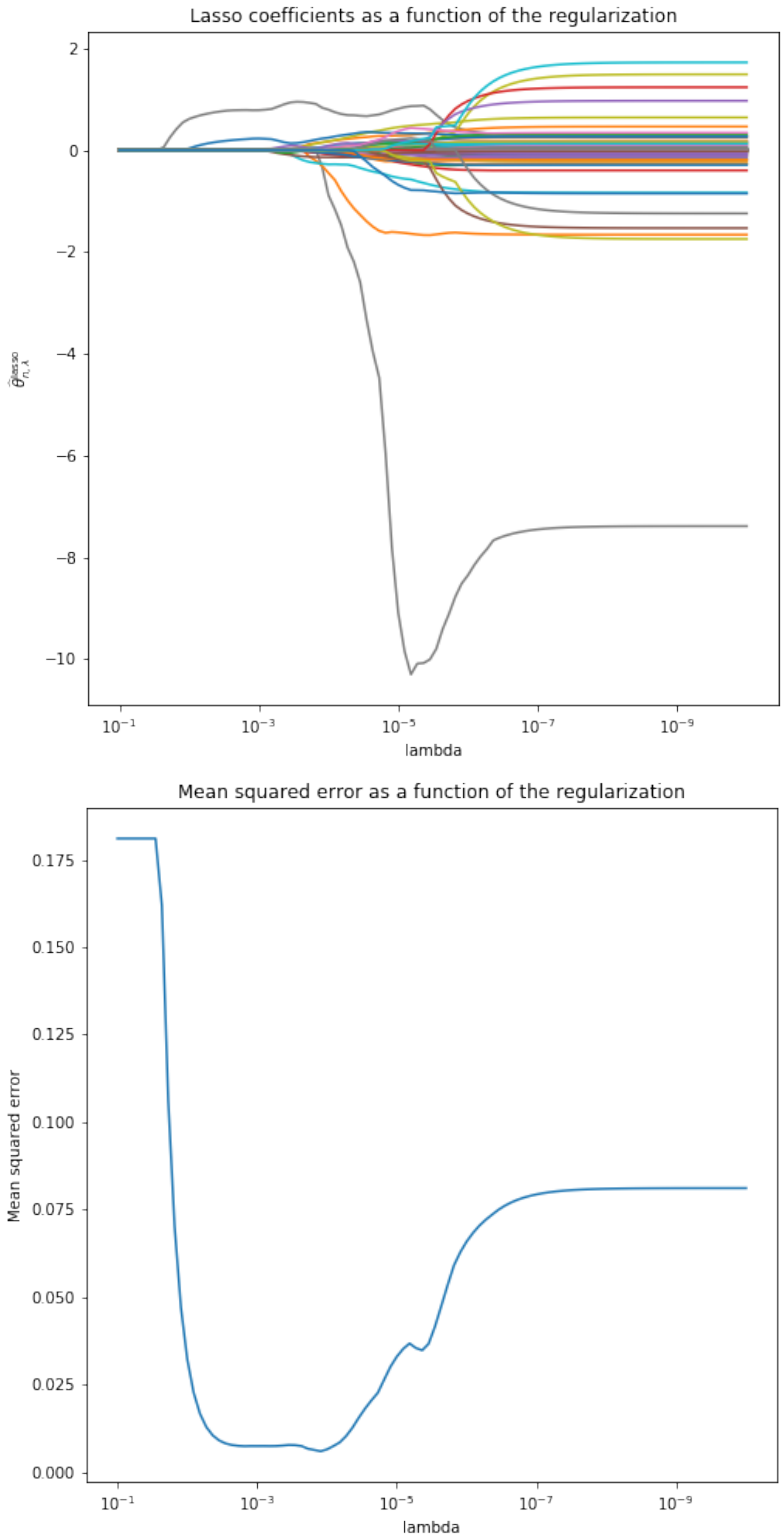

Fig. 3.3 Lasso regression is used to predict the Brazilian inflation based on many observed variables, see `https://github.com/gabrielrvsc/HDeconometrics/`. The model is trained using $n = 140$ data with for each $1 \leqslant i \leqslant n$, $X_i \in \mathbb{R}^{93}$, i.e. $d = 93$. The features are econometric data available each month. (Top) Estimated coefficient $\widehat{\theta}^{\text{lasso}}_{n,\lambda}$ as a function of λ. (Bottom) Mean squared error between the true observations and the predictions over the test set with 15 new data points.

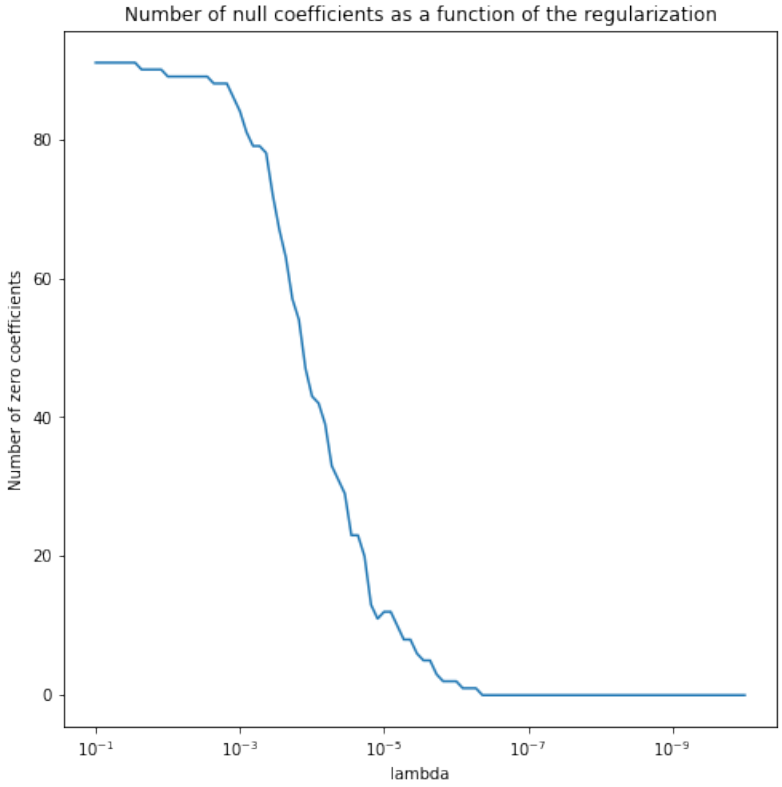

Fig. 3.4 Lasso regression is used to predict the Brazilian inflation based on many observed variables, see `https://github.com/gabrielrvsc/HDeconometrics/`. The model is trained using $n = 140$ data with for each $1 \leqslant i \leqslant n$, $X_i \in \mathbb{R}^{93}$, i.e. $d = 93$. The features are econometric data available each month. Number of null coefficient in $\widehat{\theta}^{\text{lasso}}_{n,\lambda}$ as a function of λ.

Proposition 3.7 *Let* $W = (w_i(X_j))_{1\leqslant i,j\leqslant n}$ *and* $\bar{f}^* = (f^*(X_1), \dots, f^*(X_n))^\top$. *Then,*

$$\frac{1}{n}\mathbb{E}\left[\sum_{i=1}^{n}(\widehat{f}_n^{\varphi}(X_i) - f^*(X_i))^2\right] = \frac{1}{n}\sum_{i=1}^{n}((W\bar{f}^*)_i - f^*(X_i))^2 + \frac{\sigma^2}{n}\text{Trace}(W^\top W)\,,$$

where $\widehat{f}_n^{\varphi}$ *is defined by* (3.3).

PROOF. See the exercises. ■

3.3.2 Introduction to kernel regression

Let $\mathscr{F}$ be a set of functions from $\mathscr{X} = \mathbb{R}^d$ to $\mathbb{R}$. The multivariate regression problem considered so far is part of the more general framework

$$\widehat{f}^n_{\mathscr{F}} \in \underset{f\in\mathscr{F}}{\arg\min}\ \frac{1}{n}\sum_{i=1}^{n}\ell(Y_i, f(X_i)) + \lambda\|f\|^2\,, \tag{3.4}$$

where $\lambda > 0$ and $\|\cdot\|$ is a norm on the space $\mathscr{F}$. In the case of the Ridge regression $\ell : (y,y') :\mapsto \|y-y'\|_2^2$, $\{f : \mathbb{R}^d \to \mathbb{R} \; ; \; \exists \theta \in \mathbb{R}^d \; \forall x \in \mathbb{R}^d \; , \; f(x) = f_\theta(x) = \theta^\top x\}$ and $\|f_\theta\|^2 = \|\theta\|_2^2$.

A useful case in practice consists in choosing $\mathscr{F}$ as a Reproducing Kernel Hilbert Space with positive definite reproducing kernel k on $\mathscr{X} \times \mathscr{X}$.

Definition 3.8. A function $k : \mathscr{X} \times \mathscr{X} :\to \mathbb{R}$ is said to be a positive semi-definite kernel if and only if it is symmetric and if for all $n \geqslant 1$, $(x_1,\ldots,x_n) \in \mathscr{X}^n$ and all $(a_1,\ldots,a_n) \in \mathbb{R}^n$,

$$\sum_{1 \leqslant i,j \leqslant n} a_i a_j k(x_i,x_j) \geqslant 0 \; .$$

Remark 3.9 *The following functions, defined on $\mathbb{R}^d \times \mathbb{R}^d$, are positive semi-definite kernels:*

$$k : (x,y) \mapsto x^T y \quad \text{and} \quad k : (x,y) \mapsto \exp\left(-\|x-y\|^2/(2\sigma^2)\right) \; , \; \sigma > 0 \; .$$

Definition 3.10. Let $\mathscr{F}$ be a Hilbert space of functions $f : \mathscr{X} \to \mathbb{R}$. A symmetric function $k : \mathscr{X} \times \mathscr{X} :\to \mathbb{R}$ is said to be a reproducing kernel of $\mathscr{F}$ if and only if for all $x \in \mathscr{X}$, $k(x,\cdot) \in \mathscr{F}$ and for all $x \in \mathsf{X}$ and all $f \in \mathscr{F}$, $\langle f; k(x,\cdot)\rangle = f(x)$. The space $\mathscr{F}$ is said to be a reproducing kernel Hilbert space with kernel k.

Remark 3.11 *For all $x \in \mathscr{X}$, the function $k(x,\cdot)$ is called afeature map and often written $\varphi(x)$. In this setting, all $x,x' \in \mathscr{X}$, $k(x,x') = \langle \varphi(x); \varphi(x')\rangle$. An important result given in [] states that a function $k : \mathscr{X} \times \mathscr{X} :\to \mathbb{R}$ is a positive semi-definite kernel if and only if there exist a Hilbert space $\mathscr{F}$ and a function $\varphi : \mathscr{X} \to \mathscr{F}$ such that $k(x,x') = \langle \varphi(x); \varphi(x')\rangle$.*

A reproducing kernel associated with a reproducing kernel Hilbert space is positive semi-definite since for all $n \geqslant 1$, $(x_1,\ldots,x_n) \in \mathscr{X}^n$ and all $(a_1,\ldots,a_n) \in \mathbb{R}^n$,

$$\sum_{1 \leqslant i,j \leqslant n} a_i a_j k(x_i,x_j) = \sum_{1 \leqslant i,j \leqslant n} a_i a_j \langle k(x_i,\cdot); k(x_j,\cdot)\rangle = \left\| \sum_{1 \leqslant i \leqslant n} a_i \langle k(x_i,\cdot) \right\|^2 \geqslant 0 \; .$$

Remark 3.12 *The positive semi-definite kernel $k : (x,y) \mapsto x^T y$ defined on $\mathbb{R}^d \times \mathbb{R}^d$ is a reproducing kernel of the space*

$$\mathscr{F} = \left\{ f : \mathbb{R}^d \to \mathbb{R} \; ; \; \exists \omega \in \mathbb{R}^d \; \forall x \in \mathbb{R}^d \; , \; f(x) = \omega^T x \right\} \; ,$$

equipped with the inner product defined, for all $(f,g) \in \mathscr{F} \times \mathscr{F}$, by

$$\langle f; g\rangle = \omega_f^T \omega_g \; ,$$

where $\omega_f, \omega_g \in \mathbb{R}^d$ and $f : x \mapsto \omega_f^T x$, $g : x \mapsto \omega_g^T x$.

Proposition 3.13 proves that the minimization of the penalized empirical loss amounts to solving a convex optimization problem on $\mathbb{R}^n$ for which many efficient numerical solution exist.

Proposition 3.13 *Let $k : \mathscr{X} \times \mathscr{X} :\to \mathbb{R}$ be a positive definite kernel and $\mathscr{F}$ the RKHS with kernel k. Then,*

$$\widehat{f}_{\mathscr{F}}^n \in \operatorname*{arg\,min}_{f \in \mathscr{F}} \frac{1}{n} \sum_{i=1}^n \ell(Y_i, f(X_i)) + \lambda \|f\|_{\mathscr{F}}^2 \; ,$$

where $\|f\|_{\mathscr{F}}^2 = \langle f\,;\,f\rangle$, *is given by* $\widehat{f}_{\mathscr{F}}^n : x \mapsto \sum_{i=1}^n \widehat{\alpha}_i k(X_i,x)$, *where*

$$\widehat{\alpha} \in \operatorname*{arg\,min}_{\alpha\in\mathbb{R}^n} \left\{ \frac{1}{n}\sum_{i=1}^n \ell\left(Y_i, \sum_{j=1}^n \alpha_j k(X_j,X_i)\right) + \lambda \sum_{1\leqslant i,j\leqslant n} \alpha_i\alpha_j k(X_i,X_j) \right\} .$$

PROOF. Let V be the linear space spanned by $(k(X_i,\cdot))_{1\leqslant i\leqslant n}$. For all $f\in\mathscr{F}$, f can be written $f = f_V + f_{V^\perp}$ with $f_V \in V$ and $f_{V^\perp} \in V^\perp$. Since $\mathscr{F}$ is a RKHS with kernel k, for all $1\leqslant i\leqslant n$,

$$f_{V^\perp}(X_i) = 0 \quad \text{and} \quad f(X_i) = \langle f\,;\,k(X_i,\cdot)\rangle = f_V(X_i)\,.$$

Therefore,

$$\frac{1}{n}\sum_{i=1}^n \ell(Y_i,f(X_i)) + \lambda\|f\|^2 = \frac{1}{n}\sum_{i=1}^n \ell(Y_i,f_V(X_i)) + \lambda\|f_V\|^2 + \lambda\|f_{V^\perp}\|^2$$

and any minimizer of the target function is in V. There exist $(\alpha_1,\ldots,\alpha_n)\in\mathbb{R}^n$ such that

$$\widehat{f}_{\mathscr{F}}^n : x \mapsto \sum_{i=1}^n \alpha_i k(X_i,x)\,,$$

which concludes the proof. ■

Therefore, Proposition 3.13 establishes that solving

$$\widehat{f}_{\mathscr{F}}^n \in \operatorname*{arg\,min}_{f\in\mathscr{F}} \frac{1}{n}\sum_{i=1}^n \ell(Y_i,f(X_i)) + \lambda\|f\|_{\mathscr{F}}^2\,,$$

amounts to compute $\widehat{f}_{\mathscr{F}}^n : x \mapsto \sum_{i=1}^n \widehat{\alpha}_i k(X_i,x)$ where

$$\widehat{\alpha} \in \operatorname*{arg\,min}_{\alpha\in\mathbb{R}^n} \left\{ \frac{1}{n}\sum_{i=1}^n \ell\left(Y_i, K\alpha_i\right) + \lambda\,\alpha^\top K\alpha \right\} .$$

In a Ridge regression setting tuis yields:

$$\widehat{\alpha} \in \operatorname*{arg\,min}_{\alpha\in\mathbb{R}^n} \left\{ \frac{1}{n}\|Y - K\alpha\|_2^2 + \lambda\,\alpha^\top K\alpha \right\} .$$

Chapter 4

Technical results

Contents

4.1 Probabilistic inequalities

Theorem 4.1 (Hoeffding's inequality). *Let $(X_i)_{1\leqslant i\leqslant n}$ be n independent random variables such that for all $1 \leqslant i \leqslant n$, $\mathbb{P}(a_i \leqslant X_i \leqslant b_i) = 1$ where a_i, b_i are real numbers such that $a_i < b_i$. Then, for all $t > 0$,*

$$\mathbb{P}\left(\left|\sum_{i=1}^{n} X_i - \sum_{i=1}^{n} \mathbb{E}[X_i]\right| > t\right) \leqslant 2\exp\left(\frac{-2t^2}{\sum_{i=1}^{n}(b_i-a_i)^2}\right) .$$

PROOF. Without loss of generality, assume that $\mathbb{E}[X_i] = 0$ for all $1 \leqslant i \leqslant n$. It is enough to prove that, for all $t > 0$,

$$\mathbb{P}\left(\sum_{i=1}^{n} X_i > t\right) \leqslant \exp\left(\frac{-2t^2}{\sum_{i=1}^{n}(b_i-a_i)^2}\right) . \tag{4.1}$$

Equation (3.1) implies Hoeffding's inequality by noting that $\mathbb{P}(|\sum_{i=1}^n X_i| > t) \leq \mathbb{P}(\sum_{i=1}^n X_i > t) + \mathbb{P}(-\sum_{i=1}^n X_i > t)$ and by applying (3.1) to $(X_i)_{1\leqslant i\leqslant n}$ and $(-X_i)_{1\leqslant i\leqslant n}$. Write, for any $s,t > 0$,

$$\mathbb{P}\left(\sum_{i=1}^{n} X_i > t\right) = \mathbb{P}\left(\mathrm{e}^{s\sum_{i=1}^n X_i} > \mathrm{e}^{st}\right) < \mathrm{e}^{-st}\mathbb{E}\left[\mathrm{e}^{s\sum_{i=1}^n X_i}\right] = \mathrm{e}^{-st}\prod_{i=1}^{n}\mathbb{E}\left[\mathrm{e}^{sX_i}\right]$$

To bound the right hand side of this inequality, set, for all $1 \leqslant i \leqslant n$, $\phi_i : s \mapsto \log\left(\mathbb{E}\left[\mathrm{e}^{sX_i}\right]\right)$. Since X_i is almost surely bounded, ϕ_i is differentiable and for all $s > 0$, $\phi_i'(s) = \mathbb{E}\left[X_i\mathrm{e}^{sX_i}\right]/\mathbb{E}\left[\mathrm{e}^{sX_i}\right]$. Then, differentiating again,

$$\phi_i''(s) = \log''\left(\mathbb{E}\left[\mathrm{e}^{sX_i}\right]\right) = \frac{\mathbb{E}\left[X_i^2\mathrm{e}^{sX_i}\right]}{\mathbb{E}\left[\mathrm{e}^{sX_i}\right]} - \left(\frac{\mathbb{E}\left[X_i\mathrm{e}^{sX_i}\right]}{\mathbb{E}\left[\mathrm{e}^{sX_i}\right]}\right)^2 = \widetilde{\mathbb{E}}_i[X^2] - (\widetilde{\mathbb{E}}_i[X])^2 = \widetilde{\mathbb{E}}_i[(X - \widetilde{\mathbb{E}}_i[X])^2] ,$$

where

$$\widetilde{\mathbb{E}}_i[Z] = \frac{\mathbb{E}\left[Z\mathrm{e}^{sX_i}\right]}{\mathbb{E}\left[\mathrm{e}^{sX_i}\right]} .$$

Then,

$$\phi_i''(s) = \inf_{x\in[a_i,b_i]} \widetilde{\mathbb{E}}_i[(X-x)^2] \leqslant \widetilde{\mathbb{E}}_i\left[\left(X-\frac{a_i+b_i}{2}\right)^2\right] \leqslant \left(\frac{b_i-a_i}{2}\right)^2 .$$

Finally, using Taylor's expansion,

$$\phi_i(s) \leq \phi_i(0) + \phi_i'(0) + \frac{s^2}{2}\sup_{\alpha\in[0,1]}\phi_i''(\alpha s) \leq \frac{s^2(b_i-a_i)^2}{8} . \tag{4.2}$$

This implies

$$\mathbb{P}\left(\sum_{i=1}^n X_i > t\right) \leqslant \mathrm{e}^{-st}\mathrm{e}^{s^2\sum_{i=1}^n \frac{(b_i-a_i)^2}{8}} .$$

Choosing $s = 4t/(\sum_{i=1}^n (b_i-a_i)^2)$ minimizes the right hand side and yields (3.1).

■

Lemma 4.2 *Let X be a Bernoulli random variable. Then, for all $t > 0$,*

$$\Psi(t) = \mathbb{E}\left[\mathrm{e}^{t(X-\mathbb{E}[X])}\right] \leqslant \mathrm{e}^{t^2/8} .$$

PROOF. Let $p \in (0,1)$ be such that $p = \mathbb{P}(X=1)$ (cases $p=0$ and $p=1$ are straightforward). For all $t > 0$,

$$\varphi(t) = \log\Psi(t) = \log\left(1-p+p\mathrm{e}^t\right) - pt .$$

The proof then follows from proof of the Hoeffding inequality, i.e. (3.2) with $b_i = 1-p$ and $a_i = -p$. ■

4.2 Matrix calculus

Lemma 4.3 *Let X be a random vector in $\mathbb{R}^d$ with mean $\mu \in \mathbb{R}^d$ and covariance matrix $\Sigma \in \mathbb{R}^{d\times d}$ and A a symmetric matrix in $\mathbb{R}^{d\times d}$. Then,*

$$\mathbb{E}[X^\top AX] = \mu^\top A\mu + \mathrm{Trace}(A\Sigma) .$$

PROOF. As $X^\top AX$ is a real number, $\mathbb{E}[X^\top AX] = \mathbb{E}[\mathrm{Trace}(X^\top AX)] = \mathbb{E}[\mathrm{Trace}(AXX^\top)]$. By linearity, $\mathbb{E}[X^\top AX] = \mathrm{Trace}(A\mathbb{E}[XX^\top])$ which yields,

$$\mathbb{E}[X^\top AX] = \mathrm{Trace}(A\{\mathbb{V}[X] + \mathbb{E}[X]\mathbb{E}[X]^\top\}) = \mu^\top A\mu + \mathrm{Trace}(A\Sigma) .$$

■

Lemma 4.4 *Let A be a $n\times d$ matrix with real entries. Then,* $\mathrm{range}(A) = \mathrm{range}(AA^\top)$.

PROOF. First note that for all $x\in\mathbb{R}^n$, $AA^\top x = 0$ implies $\langle A^\top x; A^\top x\rangle = 0$ so that $A^\top x = 0$. The converse is obvious. Therefore, $\mathrm{Ker}(AA^\top) = \mathrm{Ker}(A^\top)$. Using that for any matrix B, $\mathrm{Ker}(B^\top) = (\mathrm{range}(B))^\perp$, yields $\mathrm{range}(AA^\top)^\perp = \mathrm{range}(A)^\perp$, which concludes the proof. ■

Lemma 4.5 *Let $\{U_k\}_{1\leqslant k\leqslant r}$ be a family of r orthonormal vectors of $\mathbb{R}^d$. Then, $\sum_{k=1}^r U_kU_k^\top$ is the matrix of the orthogonal projection onto*

$$\mathbf{H} = \left\{\sum_{k=1}^r \alpha_k U_k\,;\ \alpha_1,\ldots,\alpha_r \in \mathbb{R}\right\} .$$

Remark 4.6 *If A is a $n\times d$ matrix with real entries such that each column of A is in* H*, then,*

$$\left(\sum_{k=1}^{r} U_k U_k^\top\right) A = A \ .$$

PROOF. For all $X \in \mathbb{R}^d$, let $\pi_{\mathbf{H}}(X)$ be the orthogonal projection of X onto $\mathbf{H}$. Since $\{U_k\}_{1\leqslant k\leqslant r}$ is an orthonormal basis of $\mathbf{H}$,

$$\pi_{\mathbf{H}}(X) = \sum_{k=1}^{r} \langle X; U_k\rangle U_k = \left(\sum_{k=1}^{r} U_k U_k^\top\right) X \ .$$

This implies in particular that for each $X \in \mathbf{H}$, $X = \left(\sum_{k=1}^{r} U_k U_k^\top\right) X$. ■

Proposition 4.7 (Singular value decomposition) *For all $\mathbb{R}^{n\times d}$ matrix A with rank r, there exist $\sigma_1 \geqslant \ldots \geqslant \sigma_r > 0$ such that*

$$A = \sum_{k=1}^{r} \sigma_k u_k v_k^\top \ ,$$

where $\{u_1,\ldots,u_r\} \in (\mathbb{R}^n)^r$ and $\{v_1,\ldots,v_r\} \in (\mathbb{R}^d)^r$ are two orthonormal families. The vectors $\{\sigma_1,\ldots,\sigma_r\}$ are called singular values of A and $\{u_1,\ldots,u_r\}$ (resp. $\{v_1,\ldots,v_r\}$) are the left-singular (resp. right-singular) vectors of A.

Remark 4.8 *If U denotes the $\mathbb{R}^{n\times r}$ matrix with columns given by $\{u_1,\ldots,u_r\}$ and V denotes the $\mathbb{R}^{p\times r}$ matrix with columns given by $\{v_1,\ldots,v_r\}$, then the singular value decomposition of A may also be written as*

$$A = U D_r V^\top \ ,$$

where $D_r = \mathrm{diag}(\sigma_1,\ldots,\sigma_r)$.

Remark 4.9 *The singular value decomposition is closely related to the spectral theorem for symmetric semipositive definite matrices. In the framework of Proposition 3.7, $A^\top A$ and $AA^\top$ are positive semidefinite such that*

$$A^\top A = V D_r^2 V^\top \quad \text{and} \quad AA^\top = U D_r^2 U^\top \ .$$

PROOF. Since the matrix $AA^\top$ is positive semidefinite, its spectral decomposition is given by

$$AA^\top = \sum_{k=1}^{r} \lambda_k u_k u_k^\top \ ,$$

where $\lambda_1 \geqslant \ldots \geqslant \lambda_r > 0$ are the nonzero eigenvalues of $AA^\top$ and $\{u_1,\ldots,u_r\}$ is an orthonormal family of $\mathbb{R}^n$. For all $1 \leqslant k \leqslant r$, define $v_k = \lambda_k^{-1/2} A^\top u_k$ so that

$$\begin{aligned} \|v_k\|^2 &= \lambda_k^{-1} \langle A^\top u_k; A^\top u_k\rangle = \lambda_k^{-1} u_k^\top AA^\top u_k = 1 \ , \\ A^\top A v_k &= \lambda_k^{-1/2} A^\top AA^\top u_k = \lambda_k v_k \ . \end{aligned}$$

On the other hand, for all $1 \leqslant k \neq j \leqslant r$, $\langle v_k; v_j\rangle = \lambda_k^{-1/2}\lambda_j^{-1/2} u_k^\top AA^\top u_j = \lambda_k^{-1/2}\lambda_j^{1/2} u_k^\top u_j = 0$. Therefore, $\{v_1,\ldots,v_r\}$ is an orthonormal family of eigenvectors of $A^\top A$ associated with the eigenvalues $\lambda_1 \geqslant \ldots \geqslant \lambda_r > 0$. Define, for all $1 \leqslant k \leqslant r$, $\sigma_k = \lambda_k^{1/2}$ which yields

$$\sum_{k=1}^{r} \sigma_k u_k v_k^\top = \sum_{k=1}^{r} u_k u_k^\top A = \left(\sum_{k=1}^{r} u_k u_k^\top\right) A \ .$$

As $\{u_1,\ldots,u_r\}$ is an orthonormal family, by Lemma 3.5 $UU^\top = \sum_{k=1}^{r} u_k u_k^\top$ is the orthogonal projection onto the range$(AA^\top)$. And, by Lemma 3.4, range$(AA^\top)$ = range(A), which implies

$$\sum_{k=1}^{r} \sigma_k u_k v_k^\top = \left(\sum_{k=1}^{r} u_k u_k^\top\right) A = A \ .$$

■

Let M_d^+ the space of real-valued $d \times d$ symmetric positive matrices.

Lemma 4.10 *The function $\Sigma \mapsto \log\det\Sigma$ is concave on M_d^+.*

PROOF. Let $\Sigma, \Gamma \in M_d^+$ and $\lambda \in [0,1]$. Since $\Sigma^{-1/2}\Gamma\Sigma^{-1/2} \in M_d^+$, it is diagonalisable in some orthonormal basis and write $\mu_1, \ldots, \mu_d$ the (possibly repeated) entries of the diagonal. Note in particular that $\det\left(\Sigma^{-1/2}\Gamma\Sigma^{-1/2}\right) = \prod_{i=1}^d \mu_i$. Then,

$$\begin{aligned}
\log\det\left((1-\lambda)\Sigma + \lambda\Gamma\right) &= \log\det\left[\Sigma^{1/2}\left((1-\lambda)I + \lambda\Sigma^{-1/2}\Gamma\Sigma^{-1/2}\right)\Sigma^{1/2}\right] \\
&= \log\det\Sigma + \log\det\left((1-\lambda)I + \lambda\Sigma^{-1/2}\Gamma\Sigma^{-1/2}\right) \\
&= \log\det\Sigma + \sum_{i=1}^d \log(1-\lambda+\lambda\mu_i) \\
&\geq \log\det\Sigma + \sum_{i=1}^d (1-\lambda)\underbrace{\log(1)}_{=0} + \lambda\log(\mu_i) := D
\end{aligned}$$

where the last inequality follows from the concavity of the log. Now, rewrite the rhs D as:

$$\begin{aligned}
D &= (1-\lambda)\log\det\Sigma + \lambda\left(\log\det\Sigma^{1/2} + \log\det\Sigma^{-1/2}\Gamma\Sigma^{-1/2} + \log\det\Sigma^{1/2}\right) \\
&= (1-\lambda)\log\det\Sigma + \lambda\log\det\Gamma
\end{aligned}$$

This finishes the proof. ■

Lemma 4.11 *Let Σ be a symmetric and invertible matrix in $\mathbb{R}^{d\times d}$.*

(i) The derivative of the real valued function $\Sigma \mapsto \log\det(\Sigma)$ defined on $\mathbb{R}^{d\times d}$ is given by:

$$\partial_\Sigma\{\log\det(\Sigma)\} = \Sigma^{-1}\,,$$

where, for all real valued function f defined on $\mathbb{R}^{d\times d}$, $\partial_\Sigma f(\Sigma)$ denotes the $\mathbb{R}^{d\times d}$ matrix such that for all $1 \leqslant i,j \leqslant d$, $\{\partial_\Sigma f(\Sigma)\}_{i,j}$ is the partial derivative of f with respect to $\Sigma_{i,j}$.

(ii) The derivative of the real valued function $x \mapsto x^\top\Sigma x$ defined on $\mathbb{R}^d$ is given by:

$$\partial_x\{x^\top\Sigma x\} = 2\Sigma x\,.$$

PROOF.

(i) Recall that for all $i \in \{1,\ldots,d\}$ we have $\det(\Sigma) = \sum_{k=1}^d \Sigma_{i,k}\Delta_{i,k}$ where $\Delta_{i,j}$ is the (i,j)-cofactor associated to Σ. For any fixed i,j, the component $\Sigma_{i,j}$ does not appear in anywhere in the decomposition $\sum_{k=1}^d \Sigma_{i,k}\Delta_{i,k}$, except for the term $k=j$. This implies

$$\frac{\partial\log\det(\Sigma)}{\partial\Sigma_{i,j}} = \frac{1}{\det\Sigma}\frac{\partial\det(\Sigma)}{\partial\Sigma_{i,j}} = \frac{\Delta_{i,j}}{\det\Sigma}$$

Recalling the identity $\Sigma\,[\Delta_{j,i}]_{1\leq i,j\leq d} = (\det\Sigma)\,I_d$ so that $\Sigma^{-1} = \frac{[\Delta_{j,i}]_{1\leq i,j\leq d}^\top}{\det\Sigma}$, we finally get

$$\left[\frac{\partial\log\det(\Sigma)}{\partial\Sigma_{i,j}}\right]_{1\leq i,j\leq d} = (\Sigma^{-1})^\top = \Sigma^{-1}$$

where the last equality follows from the fact that Σ is symmetric.

(ii) Define $\varphi(x) = x^\top\Sigma x$. Then, by straightforward algebra, $\varphi(x+h) = \varphi(x) + 2h^\top\Sigma x + \varphi(h) = \varphi(x) + 2h^\top\Sigma x + o(\|h\|)$, which concludes the proof.

■

References

Arlot and Celisse, 2010. Arlot, S. and Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics surveys*, 4:40–79.

Geisser, 1975. Geisser, S. (1975). The predictive sample reuse method with applications. *Journal of the American statistical Association*, 70(350):320–328.