The Mathematics of Machine Learning

Summer term 2019

Mathias Trabs*
Universität Hamburg

This is a working version of the manuscript for this lecture, which will be continuously updated during the next months. Any suggestions for improvements or comment on mistakes are welcome. This version is from 14th October 2019.

Contents

*Email: mathias.trabs@uni-hamburg.de

1 Introduction

The term *machine learning* refers to a toolbox of methods which are able to extract information from a possibly huge amount of data and which have been extremely successfully applied in the last (few) decades. Due to the large datasets that have to be handled, their efficient implementation and application comes with computational and algorithmic challenges. Hence, many important aspects of machine learning are in the heart of computer science. On the other hand the methodological foundation of machine learning builds on mathematical disciplines, in particular, statistics, optimisation and approximation theory. In this lecture we will give a mathematical answer to the question: How and why do *machine learning methods* work? More precisely, we aim for a rigorous and mathematical analysis of some of these celebrated methods.

1.1 Literature

There is a growing number of text books for machine learning, most of them lacking for a precise mathematical analysis. Here is a short selection:

- Hastie, Tibshirani and Friedman (2009): The Elements of Statistical Learning. Data Mining, Inference, and Prediction. Springer,
- Shalev-Shwartz and Ben-David (2014): Understanding Machine Learning: From Theory to Algorithms. Cambridge University Press,
- Mohri, Rostamizadeh and Talwalkar (2012): Foundations of Machine Learning. MIT University Press,
- Bühlmann and van de Geer (2011): Statistics for High-Dimensional Data. Methods, Theory and Applications. Springer,
- Devroye, Györfi and Lugosi (1997): A Probabilistic Theory of Pattern Recognition. Springer,
- Giraud (2014): Introduction to High-Dimensional Statistics. Chapman and Hall/CRC,
- Steinwart and Christmann (2008): Support vector machines. Springer.

1.2 Machine learning and statistics

A general setting for (supervised) learning is given by a domain $\mathcal{X}$, the feature space, and an output set $\mathcal{Y}$ which might be finite or infinite and contains all possible labels. We then have a so called training set

$$(x_1, y_1), \dots, (x_n, y_n) \subseteq \mathcal{X} \times \mathcal{Y}.$$

A learner or, depending on the task, predictor or classifier is a map

$$h \colon \mathcal{X} \to \mathcal{Y}$$

which should predict the label of a point $x \in \mathcal{X}$. Next, we need a data-generating model. In the simplest case $x_1, \dots, x_n$ are i.i.d. sampled according to a distribution $\mathcal{D}^x$ on some probability space $(\Omega, \mathcal{F})$ and the labels are given by a *labelling function* $f \colon \mathcal{X} \to \mathcal{Y}$ such that $f(x_i) = y_i$. Instead of this restrictive *realisable assumption*, we can more generally assume a joint distribution $\mathcal{D}$ for the i.i.d. sample (x_i, y_i) and $f(x) = \mathbb{E}_{(X,Y)\sim\mathcal{D}}[Y|X = x]$. The distribution $\mathcal{D}$ is completely unknown.

In *statistical words*, we consider i.i.d. observations $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$ in the *non-parametric regression* model with random design and aim for an *estimator* h for the regression function f.

The error of the learner is measured with respect to a loss function $\ell\colon \mathcal{H} \times \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$, where $\mathcal{H}$ is the space of all learners under consideration. Typical choices are the zero-one-loss $\ell(h,x,y) = \mathbb{1}_{\{h(x)\neq y\}}$ or the square loss $\ell(h,x,y) = (h(x)-y)^2$. The aim is to minimise the risk

$$R(h) := \mathbb{E}_{(X,Y)\sim\mathcal{D}}[\ell(h,X,Y)]$$

by the choice $h \in \mathcal{H}$. A common approach is to estimate $L(h)$ by the *empirical risk*

$$R_n(h) := \frac{1}{n}\sum_{i=1}^{n} \ell(h,x_i,y_i)$$

leading to the empirical risk minimisation, i.e., we would like to choose

$$\widehat{h}_n := \operatorname*{arg\,min}_{h\in\mathcal{H}} R_n(h).$$

Note however, that this minimiser might be difficult to calculate ($\mathcal{H}$ might be very large an $\ell(\cdot,x,y)$ does not have to be convex, cf. zero-one-loss).

Important aspects for the machine learning are:

- No probabilistic model is assumed (or, more precisely, only a very general one).
- There is no focus on a (interpretable) model as in usually in statistics. Instead, a good forecasting performance is the Holy Grail as we see from the choice of the loss functions.
- The dimension of the space $\mathcal{X}$ might be very large ("big data"), while $\mathcal{Y}$ could be quite small, for instance $\mathcal{Y} = \{0,1\}$ for binary classification.

In this lecture, we will not study learnability concepts like PAC (probably approximately correct) in the above general/simple model which you find in machine learning textbooks, see (Shalev-Shwartz & Ben-David, 2014; Mohri et al., 2018). Instead, we will investigate more specific problems and their solutions together with a rigorous analysis.

2 High-dimensional regression

We have already seen that regression is a typical problem from supervised learning. While we mentioned above nonparametric regression, let us start even simpler and consider the linear regression which is well known from basic statistic courses:

$$Y = X\beta + \varepsilon, \tag{1}$$

where $X \in \mathbb{R}^{n\times p}$ is a (possibly random) design matrix , $\beta \in \mathbb{R}^p$ is the unknown parameter vector and $\varepsilon \in \mathbb{R}^n$ is a centred random vector, independent from X and satisfying $\mathbb{E}[\varepsilon\varepsilon^\top] = \sigma^2 I_n$ for the identity matrix $I_n \in \mathbb{R}^{n\times n}$. In order to construct and estimator $\widehat{\beta}$, we apply the *least squares criterion*

$$|Y - X\widehat{\beta}^{LS}|^2 = \min_{b\in\mathbb{R}^p} |Y - Xb|^2$$

In the language of machine learning this is exactly *empirical risk minimisation*. Under the assumption of normally distributed i.i.d. errors $\widehat{\beta}^{LS}$ is the maximum likelihood estimator, too. It is well known/ easy to verify that the resulting *least squares estimator* admits the representation

$$\widehat{\beta}^{LS} = (X^\top X)^{-1} X^\top Y,$$

provided $X^\top X \in \mathbb{R}^{p\times p}$ is invertible (which especially requires $p = \operatorname{rank}(X^\top X) \leqslant n \wedge p$)! The Gauß-Markov theorem tells us that this estimator is the best linear unbiased estimator in the model (1). Let us calculate the prediction error:

Lemma 2.1. *Let $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$, X be deterministic and* $\operatorname{rank}(X) = p \leqslant n$. *Then we have $\sigma^{-2}|X\beta - X\widehat{\beta}^{LS}|^2 \sim \chi^2(p)$ and thus*

$$\frac{\mathbb{E}[|X\beta - X\widehat{\beta}^{LS}|^2]}{n} = \sigma^2 \frac{p}{n}.$$

Proof. We have

$$X\widehat{\beta} = X(X^\top X)^{-1}X^\top Y = \underbrace{X(X^\top X)^{-1}X^\top X\beta}_{=X\beta} + X(X^\top X)^{-1}X^\top \varepsilon.$$

The matrix $\Pi_X := X(X^\top X)^{-1}X^\top$ is a projection matrix (i.e. $\Pi_X = \Pi_X^\top$ and $\Pi_X \Pi_X = \Pi_X$) of rank p. Therefore, there is an orthogonal matrix T such that $\Pi_X = T D_p T^\top$ with $D_p = \operatorname{diag}(\underbrace{1, \dots, 1}_{p \text{ times}}, 0, \dots, 0)$. Since $T\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$, we obtain

$$\sigma^{-2}|X\beta - X\widehat{\beta}^{LS}|^2 = \sigma^{-2}|\Pi_X \varepsilon|^2 = \sigma^{-2}\varepsilon^\top \Pi_X \varepsilon = (\sigma^{-1}T\varepsilon)^\top D_p (\sigma^{-1}T\varepsilon) = \sum_{i=1}^{p} Z_i^2 \sim \chi^2(p)$$

for i.i.d. $Z_i \in \mathcal{N}(0,1)$. □

We conclude that if p is large and gets closer to the sample size n, the prediction error will deteriorate. This is okay for classical statistical problems, but might be quite limiting for machine learning applications:

- Since we observe and accumulate more and more data, numerous modern applications demand for high-dimensional covariables in a multiple regression context

  $$Y_i = X_i\beta + \varepsilon_i \quad \text{with covariable vector} \quad X_i \in \mathbb{R}^p, \quad i = 1, \dots, n$$

  (here X_i denotes the ith *row vector* of the design matrix) with $p \gg n$.

- Recalling the nonparametric regression problem

  $$Y_i = f(x_i) + \varepsilon_i, \quad i = 1, \dots, n,$$

  with x_i in some domain $\mathcal{X}$. We would like to approximate $f \approx \sum_{j=1}^{p} \beta_j \psi_j$ with a linear combination from a large flexible dictionary $(\psi_j)_{j=1,\dots,p}$ of functions $\psi_j \colon \mathcal{X} \to \mathbb{R}$. In this case the design matrix is given by $X = (\psi_j(x_i))_{i=1,\dots,n, j=1,\dots,p}$.

- ▶ Both cases lead to the high-dimensional linear model $Y = X\beta + \varepsilon$ with possibly $p \gg n$. In this case the classical least squares estimator does not work, and we require something else.

- ▶ In both cases, we may hope than only a few non-zero coefficients β_j are sufficient to describe the model quite well, i.e., to have a small prediction error $|X\widehat{\beta} - X\beta|$. In this case, we say there is a sparse representation**.**

A direct way to implement sparsity in the estimator is to modify the least squares criterion to

$$\min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{n}|Y - X\beta|^2 + \lambda |\beta|_0 \right\} \qquad \text{with} \qquad |\beta|_0 = \sum_j \mathbb{1}_{\{\beta_j \neq 0\}}.$$

However, this optimisation problem is not convex. We thus replace the ℓ_0-penalisation by an ℓ_1-penalisation, leading to:

Definition 2.2. We consider the general regression model $Y = f + \varepsilon$ for some unknown mean vector $f \in \mathbb{R}^n$, observation errors $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$ and some explanatory deterministic design matrix $X \in \mathbb{R}^{n\times p}$. The Least absolute shrinkage and selection operator, shortly the Lasso, is defined by

$$\widehat{\beta} := \operatorname*{arg\,min}_{\beta\in\mathbb{R}^p} \Big\{ \frac{1}{n}|Y - X\beta|^2 + \lambda|\beta|_1 \Big\}$$

with ℓ_1-Norm $|\beta|_1 := \sum_{j=1}^p |\beta_j|$. Hereby, $\lambda > 0$ is a penalisation parameter that have to be chosen.

The Lasso minimisation problem is convex such that there always exists a solution. However, for $p > n$ this solution might not be unique. Fortunately, the set of non-zero coefficients is invariant under all possible solutions as can be verified with a bit of subdifferential calculus (see Exercise). A detailed discussion of the uniqueness of the Lasso is given by Tibshirani (2013).

Remark 2.3. The linear regression model corresponds to $f = X\beta^*$ for some $\beta^* \in \mathbb{R}^p$. The general form in the above definition also allows for an approximation error, where the observations are not exactly determined by the linear model, as for instance in the nonparametric regression model. In that case we aim for a good approximation $f \approx X\beta^*$ ideally with only few non-zero coefficients in β^*. Since f is unknown, β^* is not known, too. Consequently, we understand β^* as an *oracle choice* which is fixed and deterministic.

Lemma 2.4 (Basic inequality)**.** *For each $\beta^* \in \mathbb{R}^p$ we have in the model $Y = f + \varepsilon$ for $f \in \mathbb{R}^n$*

$$\frac{1}{n}|X\widehat{\beta} - f|^2 + \lambda|\widehat{\beta}|_1 \leqslant \frac{1}{n}|X\beta^* - f|^2 + \frac{2}{n}\langle \varepsilon, X(\widehat{\beta} - \beta^*)\rangle + \lambda|\beta^*|_1.$$

Proof. By definition we have

$$\frac{1}{n}|Y - X\widehat{\beta}|^2 + \lambda|\widehat{\beta}|_1 \leqslant \frac{1}{n}|Y - X\beta^*|^2 + \lambda|\beta^*|_1$$

implying

$$\frac{1}{n}\Big(|f|^2 - 2\langle Y, X\widehat{\beta}\rangle + |X\widehat{\beta}|^2\Big) + \lambda|\widehat{\beta}|_1 \leqslant \frac{1}{n}\Big(|f|^2 - 2\langle Y, X\beta^*\rangle + |X\beta^*|^2\Big) + \lambda|\beta^*|_1.$$

The assertion the follows from $Y = f + \varepsilon$. □

Without the stochastic term in the above inequality, we could immediately conclude that the Lasso is at least as good as the oracle with respect to the ℓ_1-penalised prediction error. The stochastic error can be roughly bounded by

$$\frac{1}{n}|\langle \varepsilon, X(\widehat{\beta} - \beta^*)\rangle| = \frac{1}{n}|\langle X^\top\varepsilon, \widehat{\beta} - \beta^*\rangle| \leqslant \Big(\sum_{i=1}^p |\widehat{\beta}_i - \beta_i^*|\Big) \max_{j=1,\dots,p} |(X^\top\varepsilon)_j|/n,$$

where $X^\top\varepsilon \sim \mathcal{N}(0, \sigma^2 X^\top X)$. We thus expect that $\max_{j=1,\dots,p} |(X^\top\varepsilon)_j|/n$ concentrates around the origin. Let us first consider the event, where the stochastic error behaves nicely:

Lemma 2.5. *For $\beta^* \in \mathbb{R}^p$ let $S_* = \{j|\beta_j^* \neq 0\}$ and set $|b|_S = \sum_{j\in S} |b_j|$ for any $b \in \mathbb{R}^p$, $S \subseteq \{1, \dots, p\}$. On the event*

$$\mathcal{G} := \Big\{ \max_{j=1,\dots,p} |(X^\top\varepsilon)_j|/n \leqslant \frac{\lambda}{8} \Big\}$$

we have

$$\frac{4}{n}|X\widehat{\beta} - f|^2 + 3\lambda|\widehat{\beta}|_{S_*^c} \leqslant 5\lambda|\widehat{\beta} - \beta^*|_{S_*} + \frac{4}{n}|X\beta^* - f|^2.$$

Proof. We have on $\mathcal{G}$ that

$$\frac{4}{n}|X\widehat{\beta} - f|^2 + 4\lambda|\widehat{\beta}|_1 \leqslant \lambda|\widehat{\beta} - \beta^*|_1 + \frac{4}{n}|X\beta^* - f|^2 + 4\lambda|\beta^*|_1.$$

The identities $|\beta|_1 = |\beta|_{S_*^c} + |\beta|_{S_*}$ and $|\beta^*|_{S_*^c} = 0$ imply

$$\frac{4}{n}|X\widehat{\beta} - f|^2 + 4\lambda|\widehat{\beta}|_{S_*^c} \leqslant -4\lambda|\widehat{\beta}|_{S_*} + \lambda|\widehat{\beta} - \beta^*|_{S_*} + \lambda|\widehat{\beta}|_{S_*^c} + \frac{4}{n}|X\beta^* - f|^2 + 4\lambda|\beta^*|_{S_*}.$$

It remains to note $|\beta^*|_{S_*} \leqslant |\widehat{\beta} - \beta^*|_{S_*} + |\widehat{\beta}|_{S_*}$. □

To proceed, we would like to bound the estimation error $|\widehat{\beta} - \beta^*|_{S_*}$ with the prediction error $|X(\widehat{\beta} - \beta^*)|$. An eigenvalue condition would give

$$|\widehat{\beta} - \beta^*|_{S_*}^2 \leqslant S_*|\widehat{\beta} - \beta^*|^2 \leqslant \frac{|S_*|}{\varphi_n}\langle\Sigma_n(\widehat{\beta} - \beta^*), \widehat{\beta} - \beta^*\rangle = \frac{|S_*|}{n\varphi_n}|X(\widehat{\beta} - \beta^*)|^2$$

for the smallest eigenvalue φ_n of the Gram matrix $\Sigma_n := \frac{1}{n}X^\top X$. Such a condition is to restrictive for $p \gg n$, but can be relaxed:

Definition 2.6. An index set $S \subseteq \{1, \dots, p\}$ satisfies the compatibility condition, if there exists some $\varphi_n(S) > 0$ such that

$$\forall \beta \in \mathbb{R}^p \text{ with } |\beta|_{S^c} \leqslant 3|\beta|_S : \quad \langle\Sigma_n\beta, \beta\rangle \geqslant \frac{\varphi_n^2(S)}{|S|}|\beta|_S^2.$$

Otherwise we set $\varphi_n(S) = 0$.

Assuming the above condition for all $\beta \in \mathbb{R}^p$ without the restriction $|\beta|_{S^C} \leqslant 3|\beta|_S$, the above condition would imply that the smallest eigenvalue of Σ_n is bounded from below by $\frac{\varphi_n^2(S)}{|S|}$. However, this assumption would be very limiting, noting that if $p > n$ the matrix Σ_n cannot have full rank such that the smallest eigenvalue is always 0. Hence, the restriction to only those β which satisfy $|\beta|_{S^C} \leqslant 3|\beta|_S$ is a considerable relaxation and the compatibility condition can be understood as a *restricted eigenvalue condition*. It also has to be noted that the compatibility condition is difficult to check in practice.

Proposition 2.7. *On $\mathcal{G}$ we have for all $\beta^* \in \mathbb{R}^p$ such that the active set S_* satisfies the compatibility condition:*

$$\frac{2}{n}|X\widehat{\beta} - f|^2 + \lambda|\widehat{\beta} - \beta^*|_1 \leqslant \frac{6}{n}|X\beta^* - f|^2 + 24\lambda^2\frac{|S_*|}{\varphi_n^2(S_*)}.$$

Proof. Let us first consider the case $\lambda|\widehat{\beta} - \beta^*|_{S_*} \geqslant \frac{1}{n}|X\beta^* - f|^2$. In this case the previous lemma yields

$$\frac{4}{n}|X\widehat{\beta} - f|^2 + 3\lambda|\widehat{\beta}|_{S_*^c} \leqslant 9\lambda|\widehat{\beta} - \beta^*|_{S_*}.$$

In particular, $|\widehat{\beta} - \beta^*|_{S_*^c} \leqslant 3|\widehat{\beta} - \beta^*|_{S_*}$ such that the compatibility condition for S_* and $\widehat{\beta} - \beta^*$ implies

$$\begin{aligned}
\frac{4}{n}|X\widehat{\beta} - f|^2 + 3\lambda|\widehat{\beta} - \beta^*|_1 &\leqslant 9\lambda|\widehat{\beta} - \beta^*|_{S_*} + 3\lambda|\widehat{\beta} - \beta^*|_{S_*} \\
&\leqslant 12\lambda\frac{\sqrt{|S_*|}}{\varphi_n(S_*)}\frac{1}{\sqrt{n}}|X\widehat{\beta} - X\beta^*| \\
&\leqslant 12\lambda\frac{\sqrt{|S_*|}}{\varphi_n(S_*)}\Big(\frac{1}{\sqrt{n}}|X\widehat{\beta} - f| + \frac{1}{\sqrt{n}}|X\beta^* - f|\Big) \\
&\leqslant 24\lambda^2\frac{|S_*|}{\varphi_n^2(S_*)} + \frac{2}{n}|X\widehat{\beta} - f|^2 + \frac{6}{n}|X\beta^* - f|^2.
\end{aligned}$$

where we have used $12AB \leqslant 18A^2 + 2B^2$ and $12AB \leqslant 6A^2 + 6B^2$ in the last estimate.

In the case $\lambda|\widehat{\beta} - \beta^*|_{S_*} < \frac{1}{n}|X\beta^* - f|$ we obtain from Lemma 2.5

$$\frac{4}{n}|X\widehat{\beta} - f|^2 + 3\lambda|\widehat{\beta}|_{S_*^c} \leqslant \frac{9}{n}|X\beta^* - f|^2.$$

Hence,

$$\frac{4}{n}|X\widehat{\beta} - f|^2 + 3\lambda|\widehat{\beta} - \beta^*|_1 \leqslant \frac{9}{n}|X\beta^* - f|^2 + 3\lambda|\widehat{\beta} - \beta^*|_{S_*} \leqslant \frac{12}{n}|X\beta^* - f|^2. \qquad \square$$

The above result can be understood as counterpart the the bias-variance trade-off from non-parametric statistics: The term $|X\beta^* - f|^2$ is an approximation error which might decay with a growing set S_*. In contrast, the stochastic error term grows with the dimension of the active set S_*, but also decays if λ gets smaller (which will be the case for growing n). Finally, we choose λ such that the probability of the event $\mathcal{G}$ is large. To this end, we apply the Gaussian concentration of ε which will give us

$$\max_{j=1,\dots,p} |(X^\top \varepsilon)_j/n| = \mathcal{O}_p\Big(\sqrt{\frac{\sigma^2 \log p}{n}}\Big).$$

Lemma 2.8. *Define* $\lambda_{max}(\Sigma_n) := \max_{j=1,\dots,p} |(\Sigma_n)_{jj}|$. *We have for* $\lambda = 16\sqrt{\lambda_{max}(\Sigma_n)}\sqrt{\frac{\sigma^2(\tau+\log p)}{n}}$ *with* $\tau > 0$*:*

$$\mathbb{P}(\mathcal{G}^c) \leqslant \sqrt{2}e^{-\tau}.$$

Proof. Define $Z_j = (X^T\varepsilon)_j/n \sim \mathcal{N}(0, \sigma_j^2)$ with

$$\sigma_j^2 := \frac{\sigma^2}{n^2}(X^\top X)_{jj} = \frac{\sigma^2}{n}(\Sigma_n)_{jj} \leqslant \frac{\sigma^2}{n}\lambda_{max}(\Sigma_n) =: s^2, \qquad j = 1, \dots, p.$$

Owing to $\mathbb{E}[\exp(Z^2/4)] = \sqrt{2}$ for $Z \sim \mathcal{N}(0,1)$, Markov's inequality yields

$$\begin{aligned}
\mathbb{P}(\mathcal{G}^c) &= \mathbb{P}\Big(\max_{j=1,\dots,p} |Z_i| > \frac{\lambda}{8}\Big) \\
&\leqslant \mathbb{P}\Big(\exp\Big(\max_{j=1,\dots,p} \frac{Z_i^2}{4\sigma_i^2}\Big) > \exp\Big(\frac{(\frac{\lambda}{8})^2}{4s^2}\Big)\Big) \\
&\leqslant \exp\Big(-\frac{\lambda^2}{4\cdot 64 s^2}\Big)\mathbb{E}\Big[\exp\Big(\max_{j=1,\dots,p} \frac{Z_j^2}{4\sigma_j^2}\Big)\Big] \\
&\leqslant \exp\Big(-\frac{\lambda^2}{4\cdot 64 s^2}\Big)\sum_{i=1}^p \mathbb{E}\Big[\exp\Big(\frac{Z_i^2}{4\sigma_i^2}\Big)\Big] \\
&= \sqrt{2}\exp\Big(-\frac{\lambda^2}{256 s^2} + \log p\Big).
\end{aligned}$$

It remains to plug in the choice of λ. $\square$

Remark 2.9. Similar concentration results can be derived without the Gaußianity assumption, for instance, via Bernstein's inequality supposing finite moments of ε_i of some order $q > 2$.

Combining these results yields the following oracle inequality as a corollary.

Theorem 2.10 (Oracle inequality)**.** *In the regression model* $Y = f + \varepsilon$ *with* $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$*, let* $\widehat{\beta}$ *be the Lasso associated to the design matrix* $X \in \mathbb{R}^{n\times p}$ *and with penalty parameter* $\lambda = 16\sqrt{\max_{j=1,\dots,p}|(\Sigma_n)_{jj}|}\sqrt{\frac{\sigma^2(\tau+\log p)}{n}}$ *for some* $\tau > 0$ *and* $\Sigma_n := \frac{1}{n}X^\top X$*. Let the oracle be given by*

$$\beta^* := \operatorname*{arg\,min}_{\beta\in\mathbb{R}^p}\Big\{\frac{6}{n}|X\beta - f|^2 + \frac{24\lambda^2|S_\beta|}{\varphi_n(S)^2}\Big\} \qquad \big(S_\beta := \{j|\beta_j \neq 0\}\big)$$

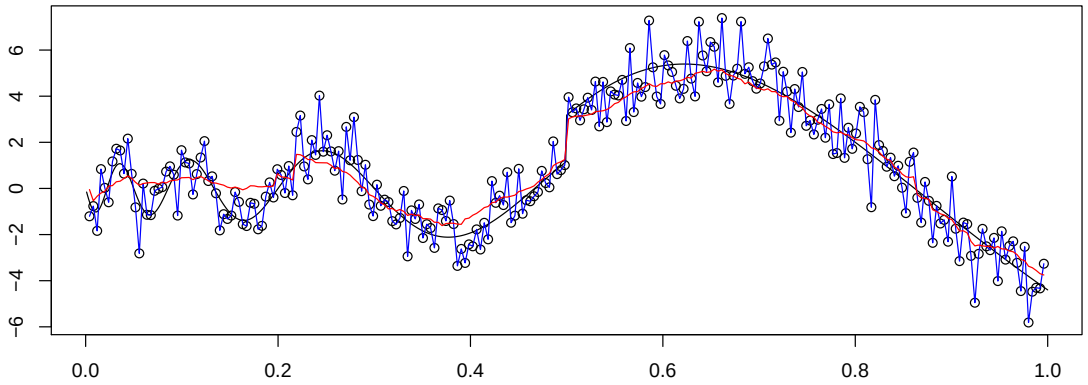

Figure 1: Least squares estimator (blue) and Lasso (red) applied to a non-parametric regression model with 250 equidistant observations and standard normal errors.

Then we have at least with probability $1 - \sqrt{2}e^{-\tau}$ *that*

$$\frac{2}{n}|X\widehat{\beta} - f|^2 + \lambda|\widehat{\beta} - \beta^*|_1 \leqslant \min_{\beta \in \mathbb{R}^p} \Big\{ \frac{6}{n}|X\beta - f|^2 + \frac{24\lambda^2|S_\beta|}{\varphi_n(S)^2} \Big\}$$
$$= \frac{6}{n}|X\beta^* - f|^2 + C(\Sigma_n, S_*)^2 \frac{\sigma^2(\tau + \log p)|S_*|}{n}$$

with $C(\Sigma_n, S) := \sqrt{6 \max_{j=1,\dots,p} |(\Sigma_n)_{jj}|} 2^5 / \varphi_n(S)$.

Roughly speaking, this oracle inequality tells us, that the Lasso $\widehat{\beta}$ is at least as good as the theoretically "best possible" choice β^*. Here "best possible" takes into account the approximation error $\frac{1}{n}|X\beta - f|^2$ for describing f via $X\beta$ and the stochastic error $\lambda^2|S_\beta| = \mathcal{O}(\frac{\sigma^2 \log p}{n}|S_\beta|)$. It is thus natural in include the penalty for non-zero coefficients also in the definition of the oracle.

For a simpler interpretation of this result let us consider the special case, where there is some $\beta \in \mathbb{R}^p$ such that $f = X\beta$. Then the oracle inequality implies

$$\frac{2}{n}|X(\widehat{\beta} - \beta)|^2 + \lambda|\widehat{\beta} - \beta|_1 \leqslant C(\Sigma_n, S_\beta)^2 \frac{\sigma^2(\tau + \log p)|S_\beta|}{n}.$$

If $|S_\beta| = o(\sqrt{n/\log p})$ and the compatibility constant $\varphi_n(S_\beta)$ behaves nicely, the right-hand side converges to zero for $n \to \infty$ and both the prediction error $\frac{1}{n}|X(\widehat{\beta} - \beta)|^2$ and the ℓ_1-estimation error $|\widehat{\beta} - \beta|_1$ converge to zero.

Example 2.11 (Dictionary learning)**.** We consider the regression model $Y_i = f(x_i) + \varepsilon_i$ with i.i.d. $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$, equidistant design points $x_i = \frac{i}{n+1}, i = 1, \dots, n$, and regression function

$$f(x) = \sin\Big(\frac{2\pi}{x + 0.2}\Big) e^{2x} + 2\mathbb{1}_{[0.5,1]}(x), \qquad x \in [0, 1].$$

Let $\sigma^2 = 1$ and $n = 250$. As dictionary we choose

$$\mathcal{D} := \big\{ \sin(2\pi j \cdot), \cos(2\pi j \cdot), (\cdot)^{j-1}, \mathbb{1}_{[\frac{j-1}{J}, \infty)} : j = 1, \dots, J \big\}.$$

For $J = n$ we obtain $p = 1000$. Figure 1 shows the least squares estimator as well as the Lasso with $\lambda = 1/8$ for a realisation of the observations errors. We clearly see that the least squares estimator only interpolates the observations. Contrarily, the Lasso has chosen only 48 non-zero coefficients.

A more detailed and comprehensive analysis of the Lasso can be found in Bühlmann & van de Geer (2011).

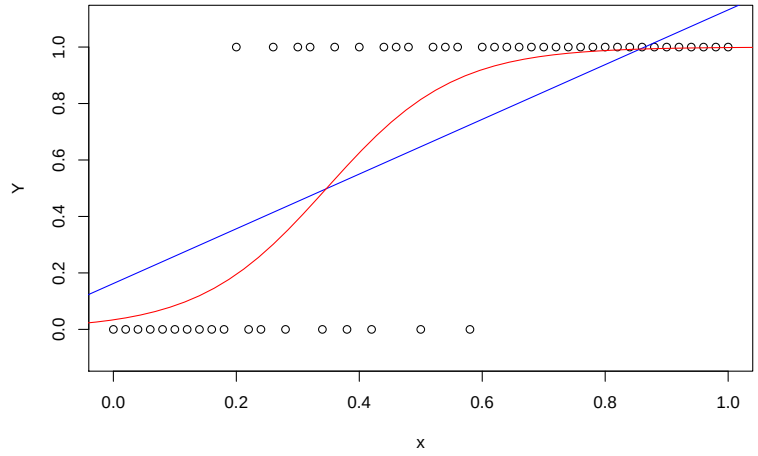

Figure 2: A sample of $n = 51$ simulated data points in a binary classification problem together with least square regression line

3 Classification

3.1 Regression approach

As in the regression setting, we observe $(x_1, Y_1), \dots, (x_n, Y_n) \in \mathcal{X} \times \mathcal{Y}$, but in classification the labels only have finitely many possible values, i.e., $|\mathcal{Y}| < \infty$. In the easiest case of a binary classification problem we have $\mathcal{Y} = \{0, 1\}$. A first naive classification method is to apply a regression estimator just as before. However, ignoring the fact that the domain of the responds variable has changed is not very convincing, cf. Figure 2. In particular, the model which is assumed in the classical regression analysis is violated.

In order to take the small output set into account, we could generalise the linear model. This leads to the following:

Definition 3.1. The independent observations $Y_1, \dots, Y_n$ on the measurable space $(\{0,1\}^n, \mathscr{P}(\{0,1\}^n))$ follow a logistic regression model if Y_i is Bernoulli $Ber(q_i)$ distributed for $i = 1, \dots, n$ and success probabilities $q_1, \dots, q_n \in (0, 1)$ satisfying

$$\left(\log \frac{q_1}{1-q_1}, \dots, \log \frac{q_n}{1-q_n} \right)^\top = X\beta$$

with design matrix $X \in \mathbb{R}^{n \times p}$ and unknown parameter vector $\beta \in \mathbb{R}^p$.

By definition the observations in a logistic regression model are in the correct space $\{0, 1\}^n$. The linear dependence on the covariables $x_i = (X_{i,1}, \dots, X_{i,p})^\top \in \mathbb{R}^p$ is moved from the expected value of the observations, as in the ordinary linear regression, to the success probabilities. Note that the Bernoulli distribution is an exponential family with natural parameter $\eta = \log(q/(1-q))$. The function $q \mapsto \log(q/(1-q))$ is called *logit* function. Therefore, the logistic regression model is a special case of generalised linear models with a canonical link function, cf. (Shao, 2003; Bühlmann & van de Geer, 2011).

To use the above model for classification, we first need to estimate the parameter vector β. To this end, we apply the maximum likelihood approach, noting that the success probabilities are given by

$$q_i = \frac{e^{x_i^\top \beta}}{1 + e^{x_i^\top \beta}} \in (0, 1), \qquad i = 1, \dots, n.$$

The likelihood and log-likelihood are thus given by

$$\begin{aligned} L(q,y) &:= \prod_{i=1}^n q_i^{y_i}(1-q_i)^{1-y_i}, \qquad \text{for } y \in \{0,1\}^n, \\ \ell(\beta,y) := \log L(q,y) &= \sum_{i=1}^n \Big(y_i \log q_i + (1-y_i)\log(1-q_i) \Big) \\ &= \sum_{i=1}^n \Big(y_i \log \frac{q_i}{1-q_i} + \log(1-q_i) \Big) \\ &= \sum_{i=1}^n \Big(y_i \cdot x_i^\top \beta - \log(1+e^{x_i^\top \beta}) \Big). \end{aligned}$$

The maximum likelihood estimator is

$$\widehat{\beta}^{MLE} := \operatorname*{arg\,max}_{\beta \in \mathbb{R}^p} \ell(\beta, Y).$$

Note that even for small dimensions p there is no closed formula for the solution of this estimation problem. In order to take large dimensions into account, we may add again an ℓ_1-penalty term:

$$\widehat{\beta} := \operatorname*{arg\,min}_{\beta \in \mathbb{R}^p} \big\{ -\frac{1}{n}\ell(\beta, Y) + \lambda |\beta|_1 \big\}. \tag{2}$$

Based on $\widehat{\beta}$, we can classify a new covariable vector with the following natural (but ad-hoc) approach.

Definition 3.2. Let $\widehat{\beta}$ be the Lasso from (2) in a logistic regression model based on observations $(x_1, Y_1), \ldots, (x_n, Y_n) \in \mathbb{R}^p \times \{0,1\}$ and $\tau \in (0,1)$ a threshold value. The logistic regression classifier is given by

$$h(x) := \mathbb{1}_{\{\widehat{q}(x) \geqslant \tau\}} \qquad \text{with} \qquad \widehat{q}(x) := \frac{e^{x^\top \widehat{\beta}}}{1 + e^{x^\top \widehat{\beta}}}, \qquad x \in \mathbb{R}^p.$$

The choice of the threshold depends on the application and the risk that should be minimised. Smaller values of τ reduce the miss-classifications for class 0, but increase the probability of a miss-classification for class 1.

Example 3.3. For $x_i = (1, i/50)^\top \in \mathbb{R}^2, i = 0, \ldots, 50$ and $\beta = (-2, 8)^\top \in \mathbb{R}^2$ we generate $Y_0, \ldots, Y_{50}$ according to the logistic regression model (the training set). The resulting MLE of the parameter vector is given by $\widehat{\beta} = (-3.35,\, 9.67)^\top$. Figure 2 shows the generated observations, the least squares regression line and the estimated success probability function $\widehat{q}$. Generating a test sample of 50 new observations with $x_i^{\text{test}} \overset{i.i.d.}{\sim} U([0,1])$ and applying the logistic regression classifier with $\tau = 0.5$ we obtain 6 miss-classifications (12%). The same approach applied to the ordinary least squares regression yields 8 miss-classificaitons (16%).

In order to analyse $\widehat{\beta}$, we have to adapt our findings from the previous section to another loss function. Set

$$\rho_f(x,y) := -yf(x) + \log(1 + e^{f(x)}) \qquad \text{and} \qquad \rho_\beta(x,y) := \rho_{f_\beta}(x,y) \quad \text{for } f_\beta(x) := \langle \beta, x \rangle.$$

Note that $f(x) \mapsto \rho_f(x,y)$ is convex for any $y \in \mathbb{R}$. We moreover introduce

$$P_n\rho_f := \underbrace{\frac{1}{n}\sum_{i=1}^n \rho_f(x_i, Y_i)}_{\text{empirical risk}} \qquad \text{and} \qquad P\rho_f := \underbrace{\frac{1}{n}\sum_{i=1}^n \widetilde{\mathbb{E}}\big[\rho_f(x_i, \widetilde{Y}_i)\big]}_{\text{theoretical risk}},$$

where $(\widetilde{Y}_1, \dots, \widetilde{Y}_n)$ is an independent sample from the same model and $\widetilde{\mathbb{E}}$ is the expectation only with respect to $\widetilde{Y}_i$. By the law of large numbers, $P_n\rho_f$ with be close to $P\rho_f$ for sufficiently large n and for some fixed f. However, we will plug in an estimator for f, namely the random variable $f_{\widehat{\beta}}$. Therefore, the distance $(P_n - P)\rho_{f_{\widehat{\beta}}}$ has to be analysed carefully. With this notation we can write

$$\widehat{\beta} := \arg\min_{\beta \in \mathbb{R}^p} \big\{ P_n\rho_\beta + \lambda|\beta|_1 \big\}.$$

We will measure the quality of $\widehat{\beta}$ via the excess risk

$$\mathcal{E}(\widehat{\beta}) := P\big[\rho_{\widehat{\beta}} - \rho_{f_0}\big] = P\rho_{\widehat{\beta}} - \min_f P\rho_f \qquad \text{for} \qquad f_0 := \arg\min_f P\rho_f$$

that is the distance of the risk of our estimator and the smallest possible risk achieved by the theoretical minimiser f_0. We take here the minimum over all functions $f\colon \mathbb{R}^p \to \mathbb{R}$, noting that for deterministic design, we can identify any such function with the vector $\vec{f} := (f(x_i), \dots, f(x_n))^\top \in \mathbb{R}^n$. It is important to notice that $\mathcal{E}(\widehat{\beta})$ is a random variable!

By definition we have $\mathcal{E}(\beta) \geqslant 0$ for any $\beta \in \mathbb{R}^p$. Moreover, $\mathcal{E}(\beta)$ is an upper bound for the prediction error in squared loss:

Lemma 3.4. *Suppose the observations Y_i are independently $Ber(q_i)$-distributed with $q_i \in (\delta, 1-\delta)$ for $i = 1, \dots, n$ and some $\delta \in (0, 1/2)$. For any $\eta > 0$ and the constant $C_{\delta,\eta} = \frac{1}{2}(1 + e^\eta/\delta)^{-2}$ we have for any β satisfying $|X\beta - \vec{f}_0|_\infty \leqslant \eta$:*

$$\mathcal{E}(\beta) \geqslant C_{\delta,\eta}\frac{1}{n}|X\beta - \vec{f}_0|^2.$$

Proof. For any $f\colon \mathbb{R}^p \to \mathbb{R}$ we have

$$\begin{aligned} P\rho_f = \frac{1}{n}\sum_{i=1}^n \mathbb{E}\big[\rho_f(x_i, \widetilde{Y}_i)\big] &= \frac{1}{n}\sum_{i=1}^n \Big(- \mathbb{E}[\widetilde{Y}_i] f(x_i) + \log(1 + e^{f(x_i)})\Big) \\ &= \frac{1}{n}\sum_{i=1}^n \Big(- q_i f(x_i) + \log(1 + e^{f(x_i)})\Big). \end{aligned}$$

The function $g_q(a) := -qa + \log(1 + e^a)$ fulfils

$$g_q'(a) = -q + \frac{e^a}{1 + e^a}, \qquad g_q''(a) = \frac{e^a}{(1 + e^a)^2}.$$

The global minimum of g_q' is obtained at $a_q = \log(\frac{q}{1-q})$ and a Taylor expansion yields for any $a \in \mathbb{R}$ and some intermediate point ξ

$$g_q(a) = g_q(a_q) + \frac{e^\xi}{2(1 + e^\xi)^2}(a - a_q)^2 \geqslant g_q(a_q) + \frac{e^{a_q - |a - a_q|}}{2(1 + e^{a_q + |a - a_q|})^2}(a - a_q)^2.$$

We conclude that $(\vec{f}_0)_i = f_0(x_i) = a_{q_i}$. Since $e^{a_{q_i}} = \frac{q_i}{1-q_i} \leqslant (1 - q_i)^{-1} < \delta^{-1}$ and $a^{a_{q_i}} \geqslant \delta$, for any β satisfying $|X\beta - \vec{f}_0|_\infty \leqslant \eta$ we obtain for the constant $C_{\eta,\delta} := \frac{1}{2}(2 + e^\eta/\delta)^{-2}$

$$\mathcal{E}(\beta) \geqslant C_{\delta,\eta}\frac{1}{n}\sum_{i=1}^n \big((X\beta)_i - f_0(x_i)\big)^2 = C_{\delta,\eta}\frac{1}{n}|X\beta - \vec{f}_0|^2. \qquad \square$$

Remark 3.5. The previous proof tells us that $f_0(x_i) = \log(\frac{q_i}{1-q_i})$ for $i = 1, \dots, n$. If the observations indeed follow an logistic regression model, there is some $\beta_0 \in \mathbb{R}^p$ such that $f_0(x_i) = \log(\frac{q_i}{1-q_i}) = (X\beta_0)_i$, $i = 1, \dots, n$. If there is no model miss-specification, the theoretical minimiser f_0 indeed coincides the true model and the excess risk is an upper bound for the prediction error:

$$\frac{1}{n}|X(\widehat{\beta} - \beta_0)|^2 \leqslant \mathcal{E}(\widehat{\beta})/C_{\delta,\eta} \qquad \text{provided that } |X(\widehat{\beta} - \beta_0)|_\infty \leqslant \eta.$$

By definition we have for any $\beta^* \in \mathbb{R}^p$

$$P_n \rho_{\widehat{\beta}} + \lambda |\widehat{\beta}|_1 \leqslant P_n \rho_{\beta^*} + \lambda |\beta^*|_1$$

implying

$$\mathcal{E}(\widehat{\beta}) + \underbrace{(P_n - P)\rho_{\widehat{\beta}}}_{=:v_n(\widehat{\beta})} + \lambda |\widehat{\beta}|_1 \leqslant \mathcal{E}(\beta^*) + \underbrace{(P_n - P)\rho_{\beta^*}}_{=:v_n(\beta^*)} + \lambda |\beta^*|_1.$$

Hence, we obtain the following basic inequality which is the analogous statement to Lemma 2.4:

$$\mathcal{E}(\widehat{\beta}) + \lambda |\widehat{\beta}|_1 \leqslant \mathcal{E}(\beta^*) + \lambda |\beta^*|_1 - \big(v_n(\widehat{\beta}) - v_n(\beta^*)\big). \tag{3}$$

The stochastic error term is thus $v_n(\widehat{\beta}) - v_n(\beta^*)$ and, in the analogy to the previous chapter, it is of the order $\lambda |\widehat{\beta} - \beta^*|_1$ with high probability.

Lemma 3.6. *Let Y_i be independently $Ber(q_i)$-distributed random variables with $q_i \in (\delta, 1-\delta)$, $i = 1, \dots, n$. On the event*

$$\mathcal{G} := \Big\{ \max_{j=1,\dots,p} |(X^\top (Y - q))_j / n| \leqslant \frac{\lambda}{4} \Big\}$$

we have $|v_n(\beta) - v_n(\beta^*)| \leqslant \frac{\lambda}{4} |\beta - \beta^*|_1$ for any $\beta \in \mathbb{R}^p$. Moreover, we have for $\lambda = 4\sqrt{2} \|X\|_{\max} (1 - \delta) \sqrt{\frac{\tau + \log p}{n}}$:

$$\mathbb{P}(\mathcal{G}) \geqslant 1 - 2e^{-\tau}.$$

Proof. Exercise ☑. □

We can now state and prove an oracle inequality for the Lasso in the logistic regression setting.

Theorem 3.7. *Suppose the observations Y_i are independently $Ber(q_i)$-distributed with $q_i \in (\delta, 1-\delta)$ for $i = 1, \dots, n$ and some $\delta \in (0, 1/2)$. Let $\widehat{\beta}$ be the Lasso from (2) associated to the design matrix $X \in \mathbb{R}^{n \times p}$ and with penalty parameter $\lambda = \sqrt{6} \|X\|_{\max} \delta \sqrt{\frac{\tau + \log p}{n}}$. For a sufficiently large $\eta > 0$ define the oracle as*

$$\beta^* := \operatorname*{arg\,min}_{\beta \in \mathbb{R}^p} \Big\{ \frac{3}{2} \mathcal{E}(\beta) + \frac{4\lambda^2 |S_\beta|}{C_{\delta,\eta} \varphi_n^2(S_\beta)} \Big\} \quad \text{and set} \quad \varepsilon^* := \frac{3}{2} \mathcal{E}(\beta^*) + \frac{4\lambda^2 |S_*|}{C_{\delta,\eta} \varphi_n^2(S_*)}.$$

Suppose that we have $|X\beta^* - \vec{f}_0|_\infty \leqslant \eta/2$ and $\varepsilon^* \|X\|_{\max} \leqslant \lambda\eta/12$. Then we have at least with probability $1 - 2e^{-\tau}$

$$\mathcal{E}(\widehat{\beta}) + \frac{3}{4} \lambda |\widehat{\beta} - \beta^*|_1 \leqslant 2\varepsilon^* = 3\mathcal{E}(\beta^*) + \frac{8}{C_{\delta,\eta}} \frac{\lambda^2 |S_*|}{\varphi_n^2(S_*)}.$$

Remark 3.8. The above oracle inequality is very similar to Theorem 3.7 in the Gaussian setting. In particular, the penalty term is again of the order $\frac{|S_*| \log p}{n}$. Assuming that $\mathcal{E}(\beta^*) = \mathcal{O}(\sqrt{(\log p)/n})$, i.e., the risk of the sparse oracle is close to the minimal risk, we conclude consistency of $\widehat{\beta}$. Under the latter condition, the assumption $\varepsilon^* \|X\|_{\max} \leqslant \lambda\eta/8$ is satisfied for some η if $(\log p)/n$ is small enough. It has to be noted that the assumption $|X\beta^* - \vec{f}_0|_\infty \leqslant \eta/2$ is a bit problematic, since large η also increase the weight of the penalty term in the definition of β^* (which is again compensated by small values of λ).

Proof. Define $M^* = 6\varepsilon^*/\lambda$ and

$$\widetilde{\beta} := t\widehat{\beta} + (1 - t)\beta^* \qquad \text{for} \qquad t := \frac{M^*}{M^* + |\widehat{\beta} - \beta^*|_1}.$$

Since $\beta \mapsto \rho_\beta$ and thus $\mathcal{E}(\beta)$ are convex, we obtain with (3) on $\mathcal{G}$ from 3.6

$$\mathcal{E}(\widetilde{\beta}) + \lambda|\widetilde{\beta}|_1 \leqslant t\big(\mathcal{E}(\widehat{\beta}) + \lambda|\widehat{\beta}|_1\big) + (1-t)\big(\mathcal{E}(\beta^*) + \lambda|\beta^*|_1\big) \leqslant \frac{\lambda}{4}t|\widehat{\beta} - \beta^*|_1 + \mathcal{E}(\beta^*) + \lambda|\beta^*|_1.$$

Owing to $t|\widehat{\beta} - \beta^*|_1 = |\widetilde{\beta} - \beta^*|_1$, we deduce

$$\mathcal{E}(\widetilde{\beta}) + \lambda|\widetilde{\beta}|_{S_*^c} \leqslant \frac{\lambda}{4}|\widetilde{\beta} - \beta^*|_1 + \mathcal{E}(\beta^*) + \lambda|\beta^*|_{S_*} - \lambda|\widetilde{\beta}|_{S_*} \leqslant \frac{\lambda}{4}|\widetilde{\beta}|_{S_*^c} + \mathcal{E}(\beta^*) + \frac{5}{4}\lambda|\widetilde{\beta} - \beta^*|_{S_*},$$

i.e.

$$\mathcal{E}(\widetilde{\beta}) + \frac{3}{4}\lambda|\widetilde{\beta}|_{S_*^c} \leqslant \mathcal{E}(\beta^*) + \frac{5}{4}\lambda|\widetilde{\beta} - \beta^*|_{S_*}. \tag{4}$$

We again distinguish two cases.

Case $\lambda|\widetilde{\beta} - \beta^*|_{S_*} \geqslant \mathcal{E}(\beta^*)$: We obtain $|\widetilde{\beta} - \beta^*|_{S_*^c} \leqslant 3|\widetilde{\beta} - \beta^*|_{S_*}$ such that the compatibility condition and $AB \leqslant \frac{1}{2}(A^2 + B^2)$ yield as in Proposition 2.7

$$\begin{aligned}
2\lambda|\widetilde{\beta} - \beta^*|_{S_*} &\leqslant 2\lambda\frac{\sqrt{|S_*|}}{\varphi_n(S_*)}\frac{1}{\sqrt{n}}|X\widetilde{\beta} - X\beta^*| \\
&\leqslant 2\lambda\frac{\sqrt{|S_*|}}{\varphi_n(S_*)}\Big(\frac{1}{\sqrt{n}}|X\widetilde{\beta} - f_0| + \frac{1}{\sqrt{n}}|X\beta^* - f_0|\Big) \\
&\leqslant 4\lambda^2\frac{|S_*|}{C_{\delta,\eta}\varphi_n^2(S_*)} + \frac{C_{\delta,\eta}}{2}\Big(\frac{1}{n}|X\widetilde{\beta} - f_0|^2 + \frac{1}{n}|X\beta^* - f_0|^2\Big).
\end{aligned}$$

Therefore, (4) yields

$$\begin{aligned}
\mathcal{E}(\widetilde{\beta}) + \frac{3}{4}\lambda|\widetilde{\beta} - \beta^*|_1 &\leqslant \mathcal{E}(\beta^*) + 2\lambda|\widetilde{\beta} - \beta^*|_{S_*} \\
&\leqslant \mathcal{E}(\beta^*) + \frac{C_{\delta,\eta}}{2}\Big(\frac{1}{n}|X\widetilde{\beta} - \vec{f}_0|^2 + \frac{1}{n}|X\beta^* - \vec{f}_0|^2\Big) + 4\lambda^2\frac{|S_*|}{C_{\delta,\eta}\varphi_n^2(S_*)}.
\end{aligned}$$

Lemma 3.4 and the assumption $|X\beta^* - \vec{f}_0|_\infty \leqslant \eta/2$ imply $\frac{C_{\delta,\eta}}{n}|X\beta^* - \vec{f}_0|^2 \leqslant \mathcal{E}(\beta^*)$. To apply the same result for $\widetilde{\beta}$, we note that

$$|X(\widetilde{\beta} - \beta^*)|_\infty = \max_{i=1,\dots,n}|\sum_{j=1}^p X_{ij}(\widetilde{\beta}_j - \beta_j^*)| \leqslant \underbrace{|\widetilde{\beta} - \beta^*|_1}_{=t|\widehat{\beta}-\beta^*|_1}\|X\|_{\max} \leqslant M^*\|X\|_{\max} \leqslant \frac{\eta}{2}.$$

Therefore, $|X\widetilde{\beta} - \vec{f}_0|_\infty \leqslant \eta$ and Lemma 3.4 yields $\frac{C_{\delta,\eta}}{n}|X\widetilde{\beta} - f_0|^2 \leqslant \mathcal{E}(\widetilde{\beta})$. We conclude

$$\frac{1}{2}\mathcal{E}(\widetilde{\beta}) + \frac{3}{4}\lambda|\widetilde{\beta} - \beta^*|_1 \leqslant \frac{3}{2}\mathcal{E}(\beta^*) + 4\lambda^2\frac{|S_*|}{C_{\delta,\eta}\varphi_n^2(S_*)} = \varepsilon^* = \frac{M^*\lambda}{6}.$$

We have deduced that $|\widetilde{\beta} - \beta^*|_1 \leqslant \frac{4}{18}M^* \leqslant M^*/2$.

Case $\lambda|\widetilde{\beta} - \beta^*|_{S_*} < \mathcal{E}(\beta^*)$: We obtain from (4)

$$\mathcal{E}(\widetilde{\beta}) + \frac{3}{4}\lambda|\widetilde{\beta}|_{S_*^c} \leqslant \frac{9}{4}\mathcal{E}(\beta^*)$$

and thus

$$\mathcal{E}(\widetilde{\beta}) + \frac{3}{4}\lambda|\widetilde{\beta} - \beta^*|_1 \leqslant \frac{9}{4}\mathcal{E}(\beta^*) + \frac{3}{4}\lambda|\widetilde{\beta} - \beta^*|_{S_*} \leqslant 3\mathcal{E}(\beta^*) \leqslant 2\varepsilon^* = \frac{M^*\lambda}{3}.$$

In *both cases* we have shown that $|\widetilde{\beta} - \beta^*|_1 \leqslant M^*/2$. Therefore,

$$\frac{M^*}{2} \geqslant |\widetilde{\beta} - \beta^*| = t|\widehat{\beta} - \beta^*|_1 = \frac{M^*|\widehat{\beta} - \beta^*|_1}{M^* + |\widehat{\beta} - \beta^*|_1}$$

implying $|\widehat{\beta} - \beta^*|_1 \leqslant M^*$. Consequently, the calculation from above can be applied to $\widehat{\beta}$ instead of $\widetilde{\beta}$ (noting that the ℓ_1-bound ensures the applicability of Lemma 3.4). We deduce that

$$\mathcal{E}(\widehat{\beta}) + \frac{3}{4}\lambda|\widehat{\beta} - \beta^*|_1 \leqslant 2\varepsilon^*. \qquad \square$$

The above theory for the Lasso can be extended to the board class of (high-dimensional) generalised linear models by modifying the considered loss function ρ_f, see Bühlmann & van de Geer (2011). If $f \mapsto \rho_f$ is convex, only an appropriate lower bound on the excess risk has to be verified (as in Lemma 3.4), the so called *margin condition*, and a concentration result for the stochastic error $v_n(\widehat{\beta}) - v_n(\beta^*)$ is required to apply the proof strategy from Theorem 3.7. For general loss functions, the stochastic error term can be handled with powerful theory for *empirical processes*.

3.2 Bayes classifier

The idea of the logistic regression was to describe the probability $\mathbb{P}(Y_i = 1)$ via a vector of covariables/ inputs $x_i \in \mathbb{R}^p$. If the covariables are random vectors, we have modelled the conditional probability $\mathbb{P}(Y = 1|X = x_i)$. Instead of the assuming a specific parametric structure together with a modification of the maximum likelihood method, we will now take a more general point of view.

Consider a random vector $(X, Y) \in \mathbb{R}^p \times \{0, \dots, K-1\}$ in classification problem with K different labels. For the zero-one-loss $\rho(h, x, y) = \mathbb{1}_{\{h(x)\neq y\}}$ the (theoretical) risk of a classifier $h\colon \mathbb{R}^p \to \{0, \dots, K-1\}$ is given by

$$R(h) := \mathbb{E}\big[\rho(h, X, Y)\big] = \mathbb{P}\big(h(X) \neq Y\big).$$

Note that $R(h) = P(h)$ in the notation of the Section 3.1. As before, the expectation in $R(h)$ is only taken with respect to X and Y, but not with respect to h if the latter is random.

The choice of the loss function emphasises that classification and statistical tests are different problems. There is neither a significance level nor hypotheses/alternatives. Usually, a symmetric risk is considered.

Lemma 3.9. *The theoretical risk minimiser $h_B\colon \mathbb{R}^p \to \{0, \dots, K-1\}$ satisfying $R(h_B) = \min_h R(h)$ is given by the so-called Bayes classifier*

$$h_B(x) := \operatorname*{arg\,max}_{k=0,\dots,K-1} \mathbb{P}(Y = k|X = x), \qquad x \in \mathbb{R}^p.$$

If $K = 2$, the Bayes classifier simplifies to $h_B(x) = \mathbb{1}_{\{\mathbb{P}(Y=1|X=x)\geqslant 1/2\}}$.

The Bayes classifier thus chooses the class with the largest conditional probability.

Proof. We have

$$R(h) = \mathbb{E}\big[\mathbb{E}[\mathbb{1}_{\{h(X)\neq y\}}|X]\big] = 1 - \mathbb{E}\big[\mathbb{P}(h(X) = Y|X)\big].$$

Hence, the risk of h is small, if the conditional probability $\mathbb{P}(h(X) = Y|X)$ is large. Moreover,

$$\begin{aligned} R(h) &= 1 - \mathbb{E}\Big[\mathbb{E}\Big[\sum_{k=0}^{K-1} \mathbb{1}_{\{Y=k\}}\mathbb{1}_{\{h(X)=k\}}\Big|X\Big]\Big] \\ &= 1 - \mathbb{E}\Big[\sum_{k=0}^{K-1} \mathbb{1}_{\{h(X)=k\}}\mathbb{E}\big[\mathbb{1}_{\{Y=k\}}|X\big]\Big] \\ &= 1 - \mathbb{E}\Big[\sum_{k=0}^{K-1} \mathbb{1}_{\{h(X)=k\}}\mathbb{P}(Y = k|X)\Big]. \end{aligned}$$

To minimise $R(h)$, it thus suffices to minimise

$$S(x) := \sum_{k=0}^{K-1} (1 - \alpha(x,k))\mathbb{P}(Y = k|X = x) \qquad \alpha(x,k) := \mathbb{1}_{\{h(x)=k\}}$$

for any $x \in \mathbb{R}^p$. Note that $S(x)$ can be read as posterior risk with respect to the posterior density $k \mapsto \mathbb{P}(Y = k|X = x)$ for the unknown parameter $Y \in \{0, \dots, K-1\}$ and given the observation $X = x$. Since $\alpha(x,k) \in \{0,1\}$ and $\sum_k \alpha(x,k) = 1$, the term $S(x)$ is minimal if most of the weight is on

$$\max_k \mathbb{P}(Y = k|X = x). \qquad \square$$

Example 3.10. Let $K = 2$ and $\mathbb{P}(Y = 0) = \mathbb{P}(Y = 1) = 1/2$ and $\mathbb{P}^{X|Y=i} = \mathcal{N}(\mu_i, 1)$, $i = 0, 1$, with $\mu_0 \neq \mu_1$. Denoting the probability density of $\mathcal{N}(\mu, 1)$ by φ_μ, the Bayes formula yields

$$\begin{aligned}\mathbb{P}(Y = 1|X = x) &= \frac{\varphi_{\mu_1}(x) \cdot \mathbb{P}(Y = 1)}{\varphi_{\mu_0}(x) \cdot \mathbb{P}(Y = 0) + \varphi_{\mu_1}(x) \cdot \mathbb{P}(Y = 1)} \\ &= \frac{\varphi_{\mu_1}(x)}{\varphi_{\mu_0}(x) + \varphi_{\mu_1}(x)} \overset{!}{\geqslant} \frac{1}{2}.\end{aligned}$$

The resulting Bayes classifier is thus given by

$$h(x) = \mathbb{1}_{\{\varphi_{\mu_1}(x) \geqslant \varphi_{\mu_0}(x)\}}.$$

Remark 3.11.

(i) While the Bayes classifier is optimal in theory, it is not admissible in practice because we do not know the conditional probabilities

$$\eta_k(x) = \mathbb{P}(Y = k|X = x). \tag{5}$$

Hence, the Bayes classifier can only be used as a benchmark or oracle.

(ii) The optimality of the Bayes classifier depends on the definition of the risk $R(h)$, where any miss-classification has the same weight. In some applications is might however be reasonable to penalise miss-classifications in some classes more severe than in other classes.

(iii) The minimal risk of the Bayes classifier is in general not zero. Any classifier might be wrong as we see in the previous example.

In practice many classification algorithms first estimate $\eta_k(x)$ and then classify according to the largest estimated conditional probability. We notice again that estimating $\eta_k(x)$ based on training data $(X_1, Y_1), \dots, (X_n, Y_n)$ is a regression problem, but in discrete responds variables/outputs and without an ordering on $\{0, \dots K-1\}$. The logistic regression was a first method for this aim. In the following subsections we will study a few more methods.

3.3 k-nearest-neighbours

For simplicity we again consider the binary classification based on observations $(X_1, Y_1), \dots, (X_n, Y_n) \in \mathbb{R}^p \times \{0,1\}$ being independent copies of (X, Y). Let $X_{(m)}(x)$ denote the mth nearest neighbour of $x \in \mathbb{R}^p$ and define the set of the k-nearest neighbours by

$$N_k(x) := \big\{X_{(1)}(x), \dots, X_{(k)}(x)\big\}$$

for some $k \in \{1, \dots, n\}$.

Definition 3.12. The k-nearest-neighbours classifier or knn classifier is defined by

$$\widehat{h}(x) := \mathbb{1}_{\{\widehat{\eta}(x)\geqslant 1/2\}} \qquad \text{for}$$
$$\widehat{\eta}(x) := \frac{1}{k}\sum_{i=1}^{n} \mathbb{1}_{\{X_i\in N_k(x)\}}\mathbb{1}_{\{Y_i=1\}} = \sum_{i=1}^{n} w_i(x)Y_i$$

with the (measurable) weights $w_i(x) = \frac{1}{k}\mathbb{1}_{\{X_i\in N_k(x)\}}$ satisfying $\sum_{i=1}^n w_i(x) = 1$.

The knn classifier thus performs a majority vote: it chooses the label of the majority of the trainingsdata in a neighbourhood of $x \in \mathbb{R}^p$. The excess risk of $\widehat{h}$ with respect to the zero-one loss can be bounded as follows

Lemma 3.13. *Denoting $\eta(x) = \mathbb{P}(Y = 1|X = x)$, we have*

$$\mathcal{E}(\widehat{h}) = R(\widehat{h}) - R(h_B) \leqslant 2\mathbb{E}_X[|\widehat{\eta}(X) - \eta(X)|].$$

Proof. We have for any classifier $h\colon \mathbb{R}^p \to \{0, 1\}$

$$\begin{aligned}\mathbb{P}(h(X) \neq Y|X = x) &= \mathbb{E}\big[(h(x) - Y)^2|X = x\big]\\ &= h(x)^2 + \mathbb{E}(Y|X = x) - 2h(x)\mathbb{P}(Y = 1|X = x)\\ &= h(x) + \eta(x) - 2h(x)\eta(x)\\ &= 1 - \eta(x) + \mathbb{1}_{\{h(x)=0\}}(2\eta(x) - 1).\end{aligned}$$

Therefore,

$$\begin{aligned}\big|\mathbb{P}_{(X,Y)}(\widehat{h}(X) \neq Y|X = x) - \mathbb{P}_{(X,Y)}(h_B(X) \neq Y|X = x)\big| &= |2\eta(x) - 1|\big|\mathbb{1}_{\{\widehat{h}(x)=0\}} - \mathbb{1}_{\{h_B(x)=0\}}\big|\\ &\leqslant 2|\widehat{\eta}(x) - \eta(x)|\mathbb{1}_{\{\widehat{h}(x)\neq h_B(x)\}},\end{aligned}$$

because $\widehat{h}(x) \neq h_B(x)$ implies that $|\widehat{\eta}(x) - \eta(x)| \geqslant |\eta(x) - \frac{1}{2}|$. We conclude

$$\begin{aligned}\mathcal{E}(\widehat{h}) &= |R(\widehat{h}) - R(h_B)|\\ &= \big|\mathbb{E}_{(X,Y)}\big[\mathbb{P}_{(X,Y)}(\widehat{h}(X) \neq Y|X) - \mathbb{P}(h_B(X) \neq Y|X)\big]\big|\\ &\leqslant 2\mathbb{E}_X\big[|\widehat{\eta}(X) - \eta(X)|\big].\end{aligned}$$

□

This lemma implies that for any L_1-consistent estimator $\widehat{\eta}$ of η, the resulting classifier $\widehat{h}(x) = \mathbb{1}_{\{\widehat{\eta}(x)\geqslant 1/2\}}$ is consistent in the sense that the expected excess risk converges to zero. For the knn rule we obtain the following:

Theorem 3.14. *Let $\eta\colon \mathbb{R}^p \to [0, 1], x \mapsto \mathbb{P}(Y = 1|X = x)$ be uniformly continuous. If $k = k_n \to \infty$ such that $\frac{k}{n} \to 0$, then*

$$\mathbb{E}[\mathcal{E}(\widehat{h})] = \mathbb{E}[R(\widehat{h})] - R(h_B) \to 0.$$

Proof. Due to Lemma 3.13 and $\sum_i w_i(x) = 1$, we have

$$\begin{aligned}\mathcal{E}(\widehat{h}) &\leqslant 2\mathbb{E}\Big[\Big|\sum_{i=1}^{n} w_i(X)\big(Y_i - \eta(X)\big)\Big|\Big]\\ &\leqslant 2\mathbb{E}\Big[\Big|\sum_{i=1}^{n} w_i(X)\big(Y_i - \eta(X_i)\big)\Big|\Big] + 2\mathbb{E}\Big[\Big|\sum_{i=1}^{n} w_i(X)\big(\eta(X_i) - \eta(X)\big)\Big|\Big]\\ &=: 2(S + D).\end{aligned} \tag{6}$$

While S is a stochastic error term given by the weighted distance of Y_i from this conditional mean $\eta(X_i) = \mathbb{E}[Y_i|X_i]$, D is an approximation error term which will be small if η is flat in a neighbourhood of X. We will bound both terms separately.

By independence of $X, (X_i, Y_i)_{i=1,\dots,n}$ we have for $i \neq j$:

$$\mathbb{E}\Big[\big(Y_i - \eta(X_i)\big)\big(Y_j - \eta(X_j)\big)\big|X, X_1, \dots, X_n\Big] = 0.$$

Together with the Cauchy-Schwarz inequality we have

$$\begin{aligned} S^2 &= \sum_{i,j=1}^n \mathbb{E}\Big[w_i(X)w_j(X)\big(Y_i - \eta(X_i)\big)\big(Y_j - \eta(X_j)\big)\Big] \\ &= \sum_{i=1}^n \mathbb{E}\Big[w_i(X)^2\big(\underbrace{Y_i - \eta(X_i)}_{\leqslant 1}\big)^2\Big] \\ &\leqslant \mathbb{E}\Big[\underbrace{\max_i w_i(X)}_{=1/k}\underbrace{\sum_{i=1}^n w_i(X)}_{=1}\Big] \leqslant \frac{1}{k} \to 0 \qquad \text{for } k \to \infty. \end{aligned}$$

To bound the second term, we note that for any $\varepsilon > 0$ there is some $\delta > 0$ such that

$$\forall y, z \in \mathbb{R}^p \text{ with } |y - z| \leqslant \delta : \qquad |\eta(y) - \eta(z)| \leqslant \varepsilon.$$

We conclude

$$\begin{aligned} D &\leqslant \mathbb{E}\Big[\Big|\sum_{i=1}^n w_i(X)\big(\eta(X_i) - \eta(X)\big)\mathbb{1}_{\{|X_i - X| > \delta\}}\Big|\Big] \\ &\quad + \mathbb{E}\Big[\Big|\sum_{i=1}^n w_i(X)\big(\eta(X_i) - \eta(X)\big)\mathbb{1}_{\{|X_i - X| \leqslant \delta\}}\Big|\Big] \\ &\leqslant \mathbb{E}\Big[\Big|\sum_{i=1}^n w_i(X)\mathbb{1}_{\{|X_i - X| > \delta\}}\Big|\Big] + \varepsilon\mathbb{E}\Big[\sum_{i=1}^n w_i(X)\Big] \\ &\leqslant \mathbb{E}\Big[\frac{1}{k}\sum_{i=1}^k \mathbb{1}_{\{|X_{(i)}(X) - X| > \delta\}}\Big] + \varepsilon \\ &\leqslant \mathbb{P}(|X_{(k)}(X) - X| > \delta) + \varepsilon. \end{aligned}$$

Since $|X_{(k)}(X) - X| > \delta$ implies that $\sum_{l=1}^n \mathbb{1}_{\{|X_l - X| \leqslant \delta\}} \leqslant k$, we obtain for $\frac{k}{n}$ sufficiently small

$$\begin{aligned} D &\leqslant \mathbb{P}\Big(\frac{1}{n}\sum_{l=1}^n \mathbb{1}_{\{|X_l - X| \leqslant \delta\}} \leqslant \frac{k}{n}\Big) + \varepsilon \\ &= \mathbb{E}\Big[\mathbb{P}\Big(\underbrace{\frac{1}{n}\sum_{l=1}^n \mathbb{1}_{\{|X_l - x| \leqslant \delta\}}}_{\to \mathbb{P}(|X_1 - x| \leqslant \delta), n \to \infty} \leqslant \underbrace{\frac{k}{n}}_{\to 0}\Big|X = x\Big)\Big] + \varepsilon. \end{aligned}$$

Since for each x in the support of $\mathbb{P}^X$ (i.e. the smallest closed set $C \in \mathcal{F}$ with $\mathbb{P}^X(C) = 1$), we have $\mathbb{P}(|X_1 - x| \leqslant \delta) > 0$, the conditional probability above converges $\mathbb{P}^X$-almost surely to zero. By dominated convergence, we conclude $D \leqslant 2\varepsilon$ for $n \to \infty$. Since $\varepsilon > 0$ was arbitrary, the assertion is proved. □

Under additional assumptions, we can quantify the convergence of the knn classfier:

Corollary 3.15. *If $p > 1$, η is globally Lipschitz continuous with Lipschitz constant $L \geqslant 1$ and $c\delta^p \leqslant \mathbb{P}(|X_1 - x| \leqslant \delta) \leqslant C\delta^p$ for constants $0 < c < C$, all $\delta \in (0, 1]$ and all x in the support of $\mathbb{P}^X$, then there is a constant $R = R(L, c, C, p) > 0$ such that*

$$\mathbb{E}[\mathcal{E}(\widehat{h})] \leqslant \frac{2}{\sqrt{k}} + R\Big(\big(\frac{k}{n}\big)^{1/p} + \frac{1}{n}\Big).$$

Proof. We again apply the error decomposition (6) and the above bound for S. Using Lipschitz continuity of η with Lipschitz constant $L \geqslant 1$, we obtain the upper bound

$$\begin{aligned}
D &= \mathbb{E}\Big[\Big|\sum_{i=1}^n w_i(X)\big(\eta(X_i) - \eta(X)\big)\Big|\Big] \\
&\leqslant \mathbb{E}\Big[\sum_{i=1}^n w_i(X)(L|X_i - X| \wedge 1)\Big] \\
&= \mathbb{E}\Big[\frac{1}{k}\sum_{i=1}^k (L|X_{(i)} - X| \wedge 1)\Big] \\
&\leqslant L\mathbb{E}\big[|X_{(k)} - X| \wedge 1\big] \\
&\leqslant L\mathbb{P}\big(|X_{(k)} - X| > 1\big) + L\mathbb{E}\Big[|X_{(k)} - X|\mathbb{1}_{\{|X_{(k)}-X|\leqslant 1\}}\Big]. \qquad (7)
\end{aligned}$$

For any $\delta \in (0,1]$, $\frac{k}{n}$ sufficiently small and $x \in \operatorname{supp}\mathbb{P}^X$ Chebyshev's inequality yields

$$\begin{aligned}
\mathbb{P}(|X_{(k)}(x) - x| > \delta) &\leqslant \mathbb{P}\Big(\frac{1}{n}\sum_{l=1}^n \mathbb{1}_{\{|X_l - x|\leqslant\delta\}} \leqslant \frac{k}{n}\Big) \\
&= \mathbb{P}\Big(\mathbb{E}[\mathbb{1}_{\{|X_1-x|\leqslant\delta\}}] - \frac{1}{n}\sum_{l=1}^n \mathbb{1}_{\{|X_l-x|\leqslant\delta\}} \geqslant \mathbb{E}[\mathbb{1}_{\{|X_1-x|\leqslant\delta\}}] - \frac{k}{n}\Big) \\
&\leqslant \frac{n^{-1}\operatorname{Var}(\mathbb{1}_{\{|X_1-x|\leqslant\delta\}})}{\big(\mathbb{E}[\mathbb{1}_{\{|X_1-x|\leqslant\delta\}}] - \frac{k}{n}\big)^2} \\
&\leqslant \frac{1}{n}\frac{\mathbb{P}(|X_1 - x| \leqslant \delta)}{\big(\mathbb{P}(|X_1 - x| \leqslant \delta) - \frac{k}{n}\big)^2} \leqslant \frac{C\delta^p}{(c\delta^p - k/n)^2 n}.
\end{aligned}$$

Therefore, the first term in (7) is bounded by

$$\mathbb{P}\big(|X_{(k)} - X| > 1\big) \leqslant \mathbb{E}\Big[\frac{\mathbb{P}(|X_1 - X| \leqslant 1|X)}{n(\mathbb{P}(|X_1 - X| \leqslant 1|X) - k/n)^2}\Big] \leqslant \frac{C}{(c - k/n)^2 n}.$$

For the second term in (7) we calculate

$$\begin{aligned}
\mathbb{E}\Big[|X_{(k)} - X|\mathbb{1}_{\{|X_{(k)}-X|\leqslant 1\}}\Big] &= \int_0^1 \mathbb{P}\big(|X_{(k)} - X| > \delta\big)d\delta \\
&\leqslant \big(\frac{2k}{cn}\big)^{1/p} + \int_{(2k/(cn))^{1/p}}^1 \mathbb{P}\big(|X_{(k)} - X| > \delta\big)d\delta \\
&\leqslant \big(\frac{2k}{cn}\big)^{1/p} + \frac{1}{n}\int_{(2k/(cn))^{1/p}}^1 \frac{C\delta^p}{(c\delta^p - k/n)^2}d\delta \\
&\leqslant \big(\frac{2k}{cn}\big)^{1/p} + \frac{1}{n}\frac{4C}{c^2}\int_{(2k/(cn))^{1/p}}^1 \delta^{-p}d\delta \\
&\leqslant \big(\frac{2k}{cn}\big)^{1/p} + \frac{1}{n}\frac{4C}{c^2(p-1)}\big(\frac{2k}{n}\big)^{(1-p)/p} \\
&\leqslant 2^{1/p}\Big(\frac{1}{c^{1/p}} + \frac{2C}{c^2(p-1)}\frac{1}{k}\Big)\big(\frac{k}{n}\big)^{1/p}.
\end{aligned}$$

Combining the above estimates yields,

$$D \leqslant L\Big(\frac{2}{c^{1/p}}\big(\frac{k}{n}\big)^{1/p} + \frac{4C}{c^2(p-1)}\frac{1}{k}\big(\frac{k}{n}\big)^{1/p} + \frac{4C}{c^2}\frac{1}{n}\Big) \leqslant R(L,c,C,p)\Big(\big(\frac{k}{n}\big)^{1/p} + \frac{1}{n}\Big)$$

for $n \to \infty$ and $\frac{k}{n} \to 0$. □

Remark 3.16.

(i) The knn method is non-parametric and its structure and analysis is related to the Nadaraya-Watson estimator with bandwidth $h = (k/n)^{1/p}$ (or equivalently $k = nh^p$) in the non-parametric regression setting, cf. Tsybakov (2009). Indeed, the Nadaraya-Watson estimator uses averages in a ball around x with radius h, where we expect approximately $h^p n$ observations under a regular distribution of the X_i's. The knn method uses local averages over the k nearest neighbours. A consequence of this non-parametric approach is, that the knn classifier also works for very flexible (and especially non-linear) decision boundaries.

(ii) Balancing deterministic and stochastic error term yields the upper bound $\mathbb{E}[\mathcal{E}(\widehat{h})] = \mathcal{O}(n^{-1/(2+p)})$ for $k^* = n^{-2/(2+p)}$ which coincides with the optimal non-parametric rate for a Lipschitz regular regression function. We observe the *curse of dimensionality*: The convergence rate deteriorates for large p. Note that this effect could be circumvented in the logistic regression approach due to the imposed parametric structure.

(iii) The assumption $c\delta^p \leqslant \mathbb{P}(|X_1 - x| \leqslant \delta) \leqslant C\delta^p$ is for instance satisfied if X_i have a compactly supported Lebesgue density which is bounded and strictly positive on its support. It ensures that the design is regular in the sense that the lower bound implies that the features are everywhere sufficiently dense distributed while the upper bound excludes a strong clustering of the input points.

Example 3.17. See for instance Figures 2.13, 2.15 and 2.16 by James et al. (2013).

3.4 Discriminant analysis

The above studied knn classification relies on a non-parametric approach. In contrast, the (linear) discriminant analysis is based on a parametric assumption on the feature distribution for each class. Suppose for each class $k \in \{0, \dots, K-1\}$ we have density

$$f_k(x)dx = \mathbb{P}(X \in dx | Y = k).$$

For prior probabilities $\pi_k = \mathbb{P}(Y = k) \in [0, 1]$ for each class, the Bayes formula implies that the posterior distribution of the probability mass function (*ger.: Zähldichte*) given the observation $x \in \mathbb{R}^p$ is

$$\eta_k(x) = \mathbb{P}(Y = k | X = x) = \frac{\pi_k f_k(x)}{\sum_{l=0}^{K-1} \pi_l f_l(x)}$$

for each $x \in \mathbb{R}^p$ with $\sum_{l=0}^{K-1} \pi_k f_l(x) > 0$. Fixing a parametric class for $(f_k)_k$, we will use the training set to estimate the unknown parameters and classify a new input according to the largest estimated posterior probability as suggested by the Bayes classifier.

In the linear discriminant analysis we impose a multivariate Gaussian density

Lemma 3.18. *Let $\pi_k = \mathbb{P}(Y = k) \in [0, 1]$ and assume $\mathbb{P}(X = dx | Y = k) = f_k(x)dx$ for all $k = 0, \dots, K-1$ with*

$$f_k(x) := \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\Big(-\frac{1}{2}(x - \mu_k)^\top \Sigma^{-1}(x - \mu_k)\Big), \qquad x \in \mathbb{R}^p, k \in \{0, \dots, K-1\},$$

for a common covariance matrix $\Sigma \in \mathbb{R}^{p \times p}$ and class dependent mean vectors $\mu_0, \dots, \mu_{K-1} \in \mathbb{R}^p$. Then the Bayes classifier is given by

$$h_B(x) = \underset{k=0,\dots,K-1}{\arg\max}\ \delta_k(x)$$

with the linear discriminant functions

$$\delta_k(x) := x^\top \Sigma^{-1}\mu_k - \frac{1}{2}\mu_k^\top \Sigma^{-1}\mu_k + \log \pi_k, \qquad x \in \mathbb{R}^p, k = 0, \dots, K-1.$$

Proof. Plugging the normal density into Bayes' formula yields

$$\begin{aligned}\eta_k(x) &= \frac{\pi_k e^{-\frac{1}{2}(x-\mu_k)^\top \Sigma^{-1}(x-\mu_k)}}{\sum_{l=0}^{K-1} \pi_l e^{-\frac{1}{2}(x-\mu_l)^\top \Sigma^{-1}(x-\mu_l)}} \\ &= \frac{\pi_k e^{x^\top \Sigma^{-1}\mu_k - \frac{1}{2}\mu_k^\top \Sigma^{-1}\mu_k}}{\sum_{l=0}^{K-1} \pi_l e^{x^\top \Sigma^{-1}\mu_l - \frac{1}{2}\mu_l^\top \Sigma^{-1}\mu_l}}.\end{aligned}$$

Since only the numerator depends on k, maximising $k \mapsto \eta_k(x)$ is equivalent to maximising $k \mapsto \delta_k(x)$. □

Due to the linearity of $x \mapsto \delta_k(x)$, the decision boundary

$$\{x \in \mathbb{R}^p : \delta_k(x) = \delta_l(x)\}$$

between two classes $k \neq l$ is linear, too. We now replace the unknown quantities π_k, μ_k and Σ by their canonical estimators

Definition 3.19. Based on an i.i.d. sample $(X_1, Y_1), \dots, (X_n, Y_n) \in \mathbb{R}^p \times \{0, \dots, K-1\}$ define $N_k := |\{j : Y_j = k\}|$ and

$$\widehat{\pi}_k := \frac{N_k}{n}, \quad \widehat{\mu}_k := \frac{1}{N_k} \sum_{j: y_j = k} X_j \quad \text{and} \quad \widehat{\Sigma} := \frac{1}{n-k} \sum_{k=1}^{K} \sum_{j: Y_j = k} (X_j - \widehat{\mu}_j)(X_j - \widehat{\mu}_j)^\top.$$

The classifier of the linear discriminant analysis is then given by

$$h_{\mathrm{LDA}}(x) := \operatorname*{arg\,max}_{k=0,\dots,K-1} \widehat{\delta}_k(x) \quad \text{with} \quad \widehat{\delta}_k(x) := x^\top \widehat{\Sigma}^{-1}\widehat{\mu}_k - \frac{1}{2}\widehat{\mu}_k^\top \widehat{\Sigma}^{-1}\widehat{\mu}_k + \log \widehat{\pi}_k.$$

Remark 3.20.

(i) Allowing for different covariance matrices Σ_k for each class $k = 0, \dots, K-1$ leads to the quadratic discriminant functions

$$\delta_k'(x) := -\frac{1}{2}\log|\Sigma_k| - \frac{1}{2}(x-\mu_k)^\top \Sigma_k^{-1}(x-\mu_k) + \log \pi_k.$$

The decision boundary between two classes is then described by a quadratic equation. Replacing the unknown parameters in δ_k' by the relative frequencies for class k and the finite sample mean and covariance results in the quadratic discriminant analysis.

(ii) The parametric estimators of $\mu_k \in \mathbb{R}^{p\times p}$ and $\Sigma \in \mathbb{R}^{p\times p}$ that have been used above only work well if the dimension p is considerably smaller than n (see Exercise ☑). In order to apply the discriminant analysis in high-dimensional feature spaces, we may apply either dimension reduction methods or impose sparsity assumptions on μ_k and Σ such that Lasso ideas could be adapted (a resulting method is the so called *graphical Lasso* for estimating Σ^{-1}).

(iii) In the special case $K = 2$, we have $\eta_0(x) = 1 - \eta_1(x)$ and thus the log likelihood ratio in the linear discriminant analysis is of the form

$$\begin{aligned}\log\Big(\frac{\eta_1(x)}{1-\eta_1(x)}\Big) &= \log\Big(\frac{\eta_1(x)}{\eta_0(x)}\Big) \\ &= x^\top \underbrace{\Sigma^{-1}(\mu_1 - \mu_0)}_{=:b} \underbrace{-\frac{1}{2}(\mu_1+\mu_0)^\top \Sigma^{-1}(\mu_1-\mu_0) + \log\frac{\pi_1}{\pi_0}}_{=:a} \\ &= x^\top b + a,\end{aligned}$$

where $a \in \mathbb{R}$ and $b \in \mathbb{R}^p$ are unknown. The logistic regression with intercept, i.e., the features are given by $\binom{1}{x_i} \in \mathbb{R}^{p+1}$, was based on the model assumption

$$\log\Big(\frac{p_{Y|X=x}(1)}{1-p_{Y|X=x}(1)}\Big) = x_i^\top \beta_1 + \beta_0$$

with $\beta_1 \in \mathbb{R}^p$ and $\beta_0 \in \mathbb{R}$. Therefore both methods rely on the same structure of the likelihood, but the unknown parameters are estimated with different methods.

3.5 Support vector machines

In the previous sections, we have tried to mimic the Bayes classifier using parametric or non-parametric estimates for $\eta(x) = \mathbb{P}(Y = 1|X = x)$. A drawback of this strategy is that we have to solve the more complex regression problem instead of the potentially simpler classification problem $\{\eta(x) > 1/2\}$. Therefore, we now come back to the empirical risk minimisation (ERM) approach that we have already have touched in the first chapter: We will not estimate the theoretical risk minimiser, but estimate the risk and study the empirical risk minimiser.

We consider again the binary classification setting, i.e., an i.i.d. sequence $(X_1, Y_1), \dots, (X_n, Y_n) \in \mathcal{X} \times \{-1, 1\}$ (note the labels are denoted by ± 1 instead of $0, 1$) and we investigate the zero-one loss $\ell(h, x, y) = \mathbb{1}_{\{h(x)\neq y\}}$ for classifiers $h\colon \mathcal{X} \to \{-1, 1\}$. The resulting theoretical and empirical risks are given by

$$R(h) = \mathbb{P}(h(X) \neq Y) \qquad \text{and} \qquad R_n(h) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{h(X_i)\neq Y_i\}},$$

respectively.

Remark 3.21. The empirical risk minimiser $\widehat{h}_n$ in a given family $\mathcal{H}$ was defined via $R_n(\widehat{h}_n) = \min_{h\in\mathcal{H}} R_n(h)$ and it can be shown that

$$R(\widehat{h}_n) - \inf_{h\in\mathcal{H}} R(h) \leqslant 2 \sup_{h\in\mathcal{H}} \big|R_n(h) - R(h)\big|.$$

In particular, for a finite family $\mathcal{H} = \{h_1, \dots, h_M\}$ this upper bound can be estimated with Hoeffding's inequality and is of order $\mathcal{O}(\log(M)/n)$ (Exercise ☑). For general infinite families $\mathcal{H}$ the *empirical process theory* can be used to bound $\sup_{h\in\mathcal{H}} \big|R_n(h) - R(h)\big|$. In particular, the so called Vapnik-Chervonenkis dimension of $\mathcal{H}$ is an important measure of the complexity of $\mathcal{H}$, see e.g. Shalev-Shwartz & Ben-David (2014) or, for a comprehensive book on the theory of empirical processes, Giné & Nickl (2016).

Since $R_n(h)$ is not convex, the minimisation of the empirical risk is numerically difficult (a potentially NP-hard problem). Therefore, we will instead solve a related convex problem. To this end, we replace $\mathbb{1}_{\{y\neq h(x)\}} = \mathbb{1}_{\{-yh(x)>0\}}$ by $\varphi(-yh(x))$ for a convex function $\varphi\colon \mathbb{R} \to [0, \infty)$. In the sequel we will use the so-called *hinge loss* $\varphi(x) = (1 + x)_+$. We will moreover allow for real valued instead of $\{-1, +1\}$-valued functions in the optimisation.

Definition 3.22. For a convex function $\varphi\colon \mathbb{R} \to [0, \infty)$ and a family $\mathcal{F} \subseteq \{f\colon \mathcal{X} \to \mathbb{R}\}$ of real valued functions define the φ-Risk and the empirical φ-Risk for $f \in \mathcal{F}$:

$$R_\varphi(f) := \mathbb{E}\big[\varphi(-Yf(X))\big] \qquad \text{and} \qquad R_{n,\varphi}(f) := \frac{1}{n}\sum_{i=1}^n \varphi(-Y_i f(X_i)).$$

The classifier which is associated to f is defined via

$$h_f(x) := \begin{cases} +1, & \text{if } f(x) \geqslant 0, \\ -1, & \text{if } f(x) < 0. \end{cases}$$

The solution

$$\widehat{f}_{n,\varphi} := \arg\min_{f\in\mathcal{F}} R_{n,\varphi}(f)$$

is called the generalised φ-ERM classifier and $\widehat{h}_{n,\varphi} := h_{\widehat{f}_{n,\varphi}}$ is the φ-ERM classifier.

Lemma 3.23. *The Bayes classifier h^* and its Bayes risk R^* satisfy for the hinge loss $\varphi(x) = (1+x)_+$:*

$$\min_{f\colon \mathcal{X}\to\mathbb{R}, measureable} R_\varphi(f) = R_\varphi(h^*) = 2R^*.$$

Proof. We have

$$\begin{aligned}\mathbb{E}[\varphi(-Yf(X))|X=x] &= (1+f(x))_+(1-\eta(x)) + (1-f(x))_+\eta(x)\\ &\geqslant 2\big(\eta(x)\wedge(1-\eta(x))\big),\end{aligned}$$

where we used $(1+A)_+ + (1-A)_+ \geqslant (1+A)-(1+A) = 2$. Equality is achieved if and only if $f(x) = -1$ holds in the case $\eta(x) < 1/2$ and $f(x) = 1$ for $\eta(x) > 1/2$. Therefore, the Bayes classifier $h^*(x) = \mathbb{1}_{\{\eta(x)\geqslant 1/2\}} - \mathbb{1}_{\{\eta(x)<1/2\}}$ minimises the conditional expectation. Noting that the Bayes risk is given by

$$\begin{aligned}R(h^*) &= \mathbb{P}(Y=1, h^*(X)=-1) + \mathbb{P}(Y=-1, h^*(X)=1)\\ &= \mathbb{E}\Big[\eta(X)\mathbb{1}_{\{h^*(X)=-1\}} + (1-\eta(x))\mathbb{1}_{\{h^*(X)=1\}}\Big]\\ &= \mathbb{E}\big[\eta(X)\wedge(1-\eta(X))\big],\end{aligned}$$

we conclude

$$R_\varphi(f) \geqslant R_\varphi(h^*) = 2R^*. \qquad \square$$

The second idea behind SVMs is a general approach to choose the set $\mathcal{H}$ of classifiers. It relies on a transformation of the features (X_i) via kernels. To make this precise, we need some mathematical preliminaries.

Definition 3.24. A function $k\colon \mathcal{X}\times\mathcal{X}\to\mathbb{R}$ is called reproducing kernel, if a Hilbert space $(W, \langle\cdot,\cdot\rangle_w)$ of real valued functions on $\mathcal{X}$ exists such that

(a) $k(x,\cdot)\in W$ for all $x\in\mathcal{X}$,

(b) for all $x\in\mathcal{X}$ and $f\in W$ we have $f(x) = \langle f, k(x,\cdot)\rangle_W$ (reproduction).

In this case, we call W reproducing kernel Hilbert space (RKHS).

Lemma 3.25. *We have $\langle k(x,\cdot), k(y,\cdot)\rangle_W = k(x,y)$ and $\sup_{x\in\mathcal{X}}|f(x)| \leqslant \|f\|_W \sup_{x\in\mathcal{X}}\sqrt{k(x,x)}$.*

Proof. The first identity follows from the reproduction property $f(y) = \langle f, k(y,\cdot)\rangle_W$ applied to $f(y) = k(x,y)$. Together with the Cauchy-Schwarz inequality, we obtain

$$\begin{aligned}\sup_{x\in\mathcal{X}} f(x)^2 &= \sup_{x\in\mathcal{X}}\langle f, k(x,\cdot)\rangle_W^2\\ &\leqslant \|f\|_W^2 \sup_{x\in\mathcal{X}}\|k(x,\cdot)\|_W^2 = \|f\|_W^2 k(x,x).\end{aligned} \qquad \square$$

Example 3.26.

(i) The space of square (Lebesgue-)integrable functions $L^2(\mathcal{X})$ is not a RKHS because $f\mapsto f(x)$ is not continuous with respect to L^2 (also the values of are only a.e. determined). This contradicts the reproduction property where the point evaluation $\delta_x\colon W\ni f\mapsto f(x)$ is a bounded linear operator:

$$|\delta_x f| = |f(x)| = |\langle f, k(x,\cdot)\rangle_W| \leqslant \|k(x,\cdot)\|_W\|f\|_W.$$

(ii) Let $\varphi_1, \dots, \varphi_K$ be an orthonormal system in $L^2(\mathcal{X})$, then for any $(a_k)_{k=1,\dots,K} \subseteq (0,\infty)$

$$k(x,y) := \sum_{k=1}^{K} a_k \varphi_k(x)\varphi_k(y)$$

is a reproducing kernel of $W = \operatorname{span}(\varphi_1, \dots, \varphi_K)$ with respect to

$$\langle f, g\rangle_W := \sum_{k=1}^{K} \frac{1}{a_k} \langle f, \varphi_k\rangle_{L^2} \langle g, \varphi_k\rangle_{L^2}.$$

To see this, note that for every $f \in W$

$$\begin{aligned}\langle f, k(x,\cdot)\rangle_W &= \sum_{k=1}^{K} \frac{1}{a_k} \langle f, \varphi_k\rangle_{L^2} \langle k(x,\cdot), \varphi_k\rangle_{L^2} \\ &= \sum_{k=1}^{K} \frac{1}{a_k} \langle f, \varphi_k\rangle_{L^2} \sum_{l=1}^{K} a_l \varphi_l(x) \langle \varphi_l, \varphi_k\rangle_{L^2} \\ &= \sum_{k=1}^{K} \langle f, \varphi_k\rangle_{L^2} \varphi_k(x) = f(x).\end{aligned}$$

This example can be generalised to infinite series expansions ($K = \infty$), e.g. an orthonormal basis of a Hilbert space, under the condition that $k(x,y)$ is well defined. A particular example is the Fourier basis $\varphi_k(x) = e^{2\pi i k x}, x \in [0,1], k \in \mathbb{Z}$, with $(a_k) \in \ell^1$. Especially, $a_k = (1+k^2)^{-s}$ for $s > 1/2$ leads to the RKHS

$$W = H^s_{per}([0,1]) := \Big\{ f \in L^2([0,1]) \Big| \sum_{k\in\mathbb{Z}} (1+k^2)^s |\langle f, \varphi_k\rangle_{L^2}|^2 < \infty \Big\}$$

which is the periodic *Sobolev space* of order s.

(iii) Consider

$$W = \Big\{ f\colon [0,1] \to \mathbb{R} | f(0) = 0, \int_0^1 f'(x)^2 dx < \infty \Big\},$$

where the derivate shall exists in the weak sense (i.e., $\langle f', \varphi\rangle_{L^2} = -\langle f, \varphi'\rangle_{L^2}$ for all compactly supported C^∞-functions). W is a Hilbert space with respect to $\langle f, g\rangle_W := \langle f', g'\rangle_{L^2}$. With $k(x,y) = x \wedge y$ we have $k(x,\cdot) \in W$ and

$$\langle f, k(x,\cdot)\rangle_W = \int_0^1 f'(y) \mathbb{1}_{\{y \leqslant x\}} dy = f(x).$$

Hence, k is a reproducing kernel of W. Note that $k(x,y) = \mathbb{E}[B_x B_y]$ is the covariance function of the Brownian motion B and W is the *Cameron-Martin* space from stochastic analysis. In the sense of the stochastic integration, we have

$$\mathbb{E}\big[\langle f, B\rangle_W \langle g, B\rangle_W\big] := \mathbb{E}\Big[\int f' dB \int g' dB\Big] = \langle f, g\rangle_W.$$

More general, the covariance functions of centred Gaussian processes on $\mathcal{X}$ are kernels of RKHS which thus are also important for the theory of Gaussian processes.

(iv) Let a symmetric and positiv-definite function $k\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be given, i.e., $k(x,y) = k(y,x)$ and $\sum_{i,j=1}^n \alpha_i \alpha_j k(x_i, x_j) \geqslant 0$ with equality only if $\alpha_1 = \dots = \alpha_n = 0$ for all $x, y, x_i \in \mathcal{X} \subseteq \mathbb{R}^p$ and $\alpha_i \in \mathbb{R}, n \in \mathbb{N}$. Consider

$$W = \operatorname{span}\Big\{ k(x_1,\cdot), \dots, k(x_n,\cdot) \Big\}$$

for $x_1, \dots, x_n \in \mathcal{X}$, where we interpret $k(x_i, \cdot)$ as a function $\{x_1, \dots, x_n\} \to \mathbb{R}, x_j \mapsto k(x_i, x_j)$. We define for $f = \sum_{i=1}^n \alpha_i k(x_i, \cdot)$ and $g = \sum_{i=1}^n \beta_i k(x_i, \cdot)$

$$\langle f, g \rangle_W := \sum_{i,j=1}^n \alpha_i \beta_j k(x_i, x_j) = \alpha^\top K \beta \qquad \text{with } K := \big(k(x_i, x_j)\big)_{i,j} \in \mathbb{R}^{n \times n}.$$

We especially have that $f(x_j) = \sum_{i=1}^n \alpha_i k(x_i, x_j) = \langle f, k(x_j, \cdot) \rangle_W$. Given k and finitely many points $x_1, \dots, x_n$, we thus have constructed a RKHS, which is the usual starting point for SVMs in applications. A popular choice is the Gaussian radial kernel $k(x, y) = \exp(-\frac{|x-y|^2}{\gamma^2})$ for some $\gamma > 0$. In this case W consists of linear combinations of p-dimensional (discretised) Gaussians. An infinite dimensional version relies on the integral operator $T_k \colon L^2 \ni f \mapsto \int f(y) k(\cdot, y) dy$ for which Mercer's theorem (functional analysis) relates this example to (ii) via the eigenfunctions.

Theorem 3.27 (representation property)**.** *Let W be a RKHS with respect to $k \colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, $\Phi \colon \mathbb{R} \to \mathbb{R}$ strictly monotone increasing and $G \colon \mathbb{R}^n \to \mathbb{R}$ arbitrary. Then for any $x_1, \dots, x_n \in \mathcal{X}$ any solution to minimisation problem*

$$G(f(x_1), \dots, f(x_n)) + \Phi(\|f\|_W) \to \min_{f \in W}!$$

is of the form $f(x) = \sum_{i=1}^n \alpha_i k(x_i, x)$ with appropriate $\alpha_i \in \mathbb{R}$.

If G is convex and non-negative, then there exists for any $\lambda > 0$ a unique solution to the minimisation problem

$$G(f(x_1), \dots, f(x_n)) + \lambda \|f\|_W^2 \to \min_{f \in W}!$$

Proof. Consider $V := \operatorname{span}\{k(x_1, \cdot), \dots, k(x_n, \cdot)\}$ and $V^\perp := \{u \in W | \forall v \in V : \langle u, v \rangle_W = 0\}$. Obviously, we have for all $u \in V^\perp$ that $u(x_i) = \langle u, k(x_i, \cdot) \rangle_W = 0$. Hence, we obtain for any $f = u + v \in W$ with $u \in V^\perp$ and $v \in V$

$$\forall i = 1, \dots, n : \quad f(x_i) = v(x_i), \qquad \|f\|_W^2 = \|u\|_W^2 + \|v\|_W^2.$$

Consequently, $f = u + v$ can only be a solution to the minimisation problem, if $u = 0$ (otherwise the criterion for v will be smaller than for f). We thus have proved that $f \in V$.

If G is convex and non-negative, then

$$f \mapsto K(f) := G(f(x_1), \dots, f(x_n)) + \lambda \|f\|_W^2$$

is convex and non-negative, too. Furthermore, for any $f \in W$ with $\lambda \|f\|_W^2 > G(0, \dots, 0)$ we have $K(f) > K(0)$. Therefore, if $f_n \in W, n \in \mathbb{N}$, is a sequence such that

$$K(f_n) \to \inf_{f \in W} K(f),$$

then we can assume that for all $n \in \mathbb{N}$

$$\|f_n\|_W \leqslant \lambda^{-1/2} G(0, \dots, 0)^{1/2}$$

(all f_n not satisfying this assumption can be set to zero). Applying again the decomposition $f_n = v_n + u_n$ with $v_n \in V$, $u_n \in V^\perp$ shows that

$$K(v_n) \to \min_{f \in W} K(f).$$

Note that (v_n) is in the compact finite-dimensional ball $\{v \in V | \|v_n\|_W \leqslant \lambda^{-1/2} G(0, \dots, 0)^{1/2}\}$ and any accumulation point von (v_n) solves the minimisation problem.

We finally prove uniqueness. Suppose there are two solutions f_1, f_2, then $g = \frac{1}{2}(f_1 + f_2)$ satisfies

$$\|g\|_W^2 = \frac{1}{4}\big(2\|f_1\|_W^2 + 2\|f_2\|_W^2 - \|f_1 - f_2\|_W^2\big) < \frac{1}{2}\big(\|f_1\|_W^2 + \|f_2\|_W^2\big).$$

Convexity of G implies strict convexity of K and thus $K(g) < \frac{1}{2}(K(f_1) + K(f_2))$. This would contradict the optimality of f_1,f_2 such that the solution has to be unique. □

Definition 3.28. For a reproducing kernel $k\colon \mathcal{X}\times\mathcal{X}\to\mathbb{R}$, the corresponding RKHS W and some $\lambda>0$ set

$$\widehat{f}_n^{SVM} := \operatorname*{arg\,min}_{f\in W,\|f\|\leqslant\lambda}\Big(\frac{1}{n}\sum_{i=1}^n\big(1-Y_if(X_i)\big)_+\Big).$$

The resulting classifier $\widehat{h}_n^{SVM} := h_{\widehat{f}_n^{SVM}}$ is called the SVM classifier, support vector classifier or support vector machine.

Hence, the support vector machine is a φ-ERM classifier for which the set of possible (generalised) classifiers is given by a ball $\mathcal{F} := \{f\in W : \|f\|_W\leqslant\lambda\}$ in a RKHS on $\mathcal{X}$ with some radius $\lambda>0$. According to the Lagrange theory, there is some $\lambda'>0$ such that we obtain the representation (with hinge loss $\varphi(x)=(1+x)_+$)

$$\widehat{f}_n^{SVM} = \operatorname*{arg\,min}_{f\in W}\Big(R_{n,\varphi}(f)+\lambda'\|f\|_W^2\Big).$$

We thus recover the penalised empirical risk minimisation paradigm that already led to the Lasso.

According to the representation theorem applied to $G(f(X_1),\dots,f(X_n)) = R_{n,\varphi}(f)$ the latter optimisation problem can be written as finite dimensional problem: The solution has to be of the form $\widehat{f}_n^{SVM}(x)=\sum_{i=1}^n\widehat{\alpha}_ik(X_i,x)$ for appropriate $\widehat{\alpha}=(\widehat{a}_i)_{i=1,\dots,n}\in\mathbb{R}^n$. Due to Lemma 3.25, we have

$$\Big\|\sum_{i=1}^n\alpha_ik(X_i,\cdot)\Big\|_W^2 = \sum_{i,j=1}^n\alpha_i\alpha_jk(X_i,X_j).$$

Therefore, the coefficients $\widehat{\alpha}$ of $\widehat{f}_n^{SVM}$ are given by

$$\widehat{\alpha} = \operatorname*{arg\,min}_{\alpha\in\mathbb{R}^n}\Big(\frac{1}{n}\sum_{i=1}^n\Big(1-Y_i\sum_{j=1}^n\alpha_jk(X_j,X_i)\Big)_+ + \lambda'\sum_{i,j=1}^n\alpha_i\alpha_jk(X_i,X_j)\Big).$$

It is worth noting that the RKHS W does not anymore appear in this representation. We only have to solve a finite dimensional optimisation problem depending on the kernel k.

Defining

$$I := \big\{i=1,\dots,n : Y_i\widehat{f}_n^{SVM}(X_i)\leqslant 1\big\},$$

$\widehat{f}_n^{SVM}$ is also the solution of the minimisation problem

$$\Big(\frac{1}{n}\sum_{i\in I}\big(1-Y_if(X_i)\big)_+ + \lambda'\|f\|_W^2\Big)\to\min_{f\in W}!$$

Since the representation theorem implies $\widehat{f}_n^{SVM}(x)=\sum_{i\in I}\widehat{\alpha}_ik(x_i,x)$, we conclude that $\widehat{\alpha}_i=0$ for all i for which $Y_i\widehat{f}_n^{SVM}(X_i)>1$, i.e. the corresponding observation (X_i,Y_i) is "significantly correctly specified". Hence, we often obtain a sparse representation of $\widehat{f}_n^{SVM}$. The features $(X_i)_{i\in I}$ are called support vectors.

To interpret this condition, we equivalently write the minimisation problem as

$$(\widehat{f}_n^{SVM},\widehat{\xi}) := \operatorname*{arg\,min}_{f\in W,\xi\in[0,\infty)^n,\forall i:Y_if(X_i)\geqslant 1-\xi_i}\Big(\lambda'\|f\|_W^2+\frac{1}{n}\sum_{i=1}^n\xi_i\Big). \tag{8}$$

The constraints on ξ ensure that $\xi_i\geqslant(1-Y_if(X_i))_+$ such that $\widehat{\xi}_i=(1-Y_if(X_i))_+$ and the equivalence of both minimisation problems follows.

Example 3.29. We again consider the setting from Example 3.26(ii) and assume the kernel is given as

$$k(x,y) := \sum_{k=1}^K a_k\varphi_k(x)\varphi_k(y),$$

where we allow for $K = \infty$ provided that the choice of (a_k) ensures that $k(x, y)$ is well defined. We have $W := \operatorname{span}\{\varphi_k, k = 1, \dots, K\}$ (with the obvious modification for $K = \infty$) equipped with

$$\langle f, g\rangle_W := \sum_{k=1}^{K} \frac{1}{a_k} \langle f, \varphi_k\rangle_{L^2} \langle g, \varphi_k\rangle_{L^2}.$$

Defining

$$\Psi \colon \mathcal{X} \to \mathbb{R}^K, x \mapsto (\sqrt{a_1}\varphi_1(x), \dots, \sqrt{a_K}\varphi_K(x))^\top,$$

we can write $k(x, y) = \Psi(x)^\top \Psi(y)$. For any $f \in W$, we find a representation $f(x) = \beta^\top \Psi(x)$, $\beta \in \mathbb{R}^K$. Due to the representation theorem, we know that the solution of (8) is of the form

$$f(x) = \sum_{i=1}^{n} \alpha_i k(x_i, x) = \sum_{i=1}^{n} \alpha_i \Psi(x_i)^\top \Psi(y) = \beta_f^\top \Psi(y) \quad \text{with } \beta_f := \sum_{i=1}^{n} \alpha_i \Psi(x_i) \in \mathbb{R}^K$$

for which

$$\|f\|_W^2 = \sum_{i,j=1}^{n} \alpha_i \alpha_j k(x_i, x_j) = \sum_{i,j=1}^{n} \alpha_i \alpha_j \Psi(x_i)^\top \Psi(y_j) = \Big(\sum_{i=1}^{n} \alpha_i \Psi(x_i)\Big)^\top \Big(\sum_{j=1}^{n} \alpha_j \Psi(x_j)\Big) = |\beta_f|^2.$$

Hence, (8) takes the form

$$\widehat{f}_n^{SVM}(x) := \widehat{\beta}_n^\top \Psi(x) \quad \text{for} \quad (\widehat{\beta}_n, \widehat{\xi}) := \operatorname*{arg\,min}_{\beta \in \mathbb{R}^K, \xi \in [0,\infty)^n, \forall i: Y_i(\beta^\top \Psi(X_i)) \geqslant 1 - \xi_i} \Big(|\beta|^2 + \frac{1}{n\lambda'} \sum_{i=1}^{n} \xi_i\Big).$$

If all (X_i, Y_i) can be significantly correctly specified in the sense of $Y_i \widehat{f}_n^{SVM}(X_i) = Y_i \widehat{\beta}_n^\top \Psi(X_i) \geqslant 1$, we have $\widehat{\xi}_i = 0$ for all $i = 1, \dots, n$ and $\widehat{f}_n^{SVM}$ minimises the norm $\|f\|_W = |\beta|$ under all such significantly correct classifiers. Thus in the augmented feature space $\Psi(X_i) \in \mathbb{R}^K$ (K might be much larger than the dimension of $\mathcal{X}$), we can separate both classes with the hyperplane $\{z \in \mathbb{R}^K : \widehat{\beta}_n^\top z = 0\}$. Note that $|\widehat{\beta}|_n^{-1} \widehat{\beta}_n^\top \Psi(X_i)$ measures the signed distance of $\Psi(X_i)$ from that hyperplane. Minimising $|\beta|^2$ thus results in finding the *separating hyperplane* with the largest possible distance to all $\Psi(X_i)$. The resulting distance between the two classes is called the *margin*. The points X_i where this maximal possible distance is attained must satisfy $Y_i \widehat{\beta}_n^\top \Psi(X_i) = 1$ such that these X_i are exactly the support vectors. In general $\widehat{f}_n^{SVM}$ will not classify all (training) observations correctly which is penalised by $\sum_{i=1}^{n} \xi_i$. This penalisation measures the total (aggregated) distance of all support vectors from their side margin.

From the discussion in this example we observe two main differences to the discriminant analysis: First, instead of relying on the global empirical means and covariances, the SVM classification is determined only by the critical support vectors near the decision boundary. Second, the *kernel trick* moreover allows for non-linear decision boundaries in $\mathcal{X}$ while the decision boundary is linear in the augmented feature space $\operatorname{span}\{k(X_1, \cdot), \dots, k(X_n, \cdot)\}$.

Next we will prove an oracle inequality for SVMs which provides control on the stochastic error. The approximation error severely depends on the considered problem and the choice of the kernel. We will not detail this out.

Theorem 3.30. *Let k be a kernel of the RKHS W satisfying $\sup_{x \in \mathcal{X}} k(x, x) < \infty$. The corresponding SVM classifier $\widehat{h}_n^{SVM}$ with radius $\lambda > 0$ fulfils*

$$\mathbb{E}\big[R(\widehat{h}_n^{SVM})\big] \leqslant \inf_{\|f\|_W \leqslant \lambda} R_\varphi(f) + \frac{8\lambda}{\sqrt{n}} \mathbb{E}\big[k(X, X)\big]^{1/2}.$$

Proof. Step 1: For $\varphi(x) = (1 + x)_+$ we have

$$R(\widehat{h}_n^{SVM}) = \mathbb{P}^{(X,Y)}\big(-Y \widehat{f}_n^{SVM}(X) > 0\big) \leqslant \mathbb{E}^{(X,Y)}\big[(1 - Y \widehat{f}_n^{SVM}(X))_+\big] = R_\varphi(\widehat{f}_n^{SVM}).$$

Therefore, we deduce from Remark 3.21 that

$$\begin{aligned} R(\widehat{h}_n^{SVM}) &\leqslant R_\varphi(\widehat{f}_n^{SVM}) - \inf_{\|f\|_W \leqslant \lambda} R_\varphi(f) + \inf_{\|f\|_W \leqslant \lambda} R_\varphi(f) \\ &\leqslant 2 \sup_{\|f\|_W \leqslant \lambda} |R_{n,\varphi}(f) - R_\varphi(f)| + \inf_{\|f\|_W \leqslant \lambda} R_\varphi(f). \end{aligned}$$

It remains to show

$$\mathbb{E}\Big[\sup_{\|f\|_W \leqslant \lambda} |R_{n,\varphi}(f) - R_\varphi(f)|\Big] \leqslant 4\lambda\sqrt{\mathbb{E}[k(x,x)]/n}.$$

Step 2: We use a symmetrisation argument (from empirical process theory) to bound the supremum. To this end, let $(X_i', Y_i')_{i=1,\dots,n}$ be an independent copy of $(X_i, Y_i)_{i=1,\dots,n}$ defined on the same probability space (a so called *ghost sample*). Jensen's inequality and $\sup_t \mathbb{E}[Z_t] \leqslant \mathbb{E}[\sup_t Z_t]$ imply

$$\begin{aligned} &\mathbb{E}\Big[\sup_{\|f\|_W \leqslant \lambda} \Big|\frac{1}{n}\sum_{i=1}^n \big(\varphi(-Y_i f(X_i)) - \mathbb{E}[\varphi(-Y_i' f(X_i'))]\big)\Big|\Big] \\ &\quad = \mathbb{E}\Big[\sup_{\|f\|_W \leqslant \lambda} \Big|\mathbb{E}\Big[\frac{1}{n}\sum_{i=1}^n \big(\varphi(-Y_i f(X_i)) - \varphi(-Y_i' f(X_i'))\big)\Big| X_1, \dots, X_n, Y_1, \dots, Y_n\Big]\Big|\Big] \\ &\quad \leqslant \mathbb{E}\Big[\sup_{\|f\|_W \leqslant \lambda} \Big|\frac{1}{n}\sum_{i=1}^n \big(\varphi(-Y_i f(X_i)) - \varphi(-Y_i' f(X_i'))\big)\Big|\Big]. \end{aligned}$$

Let moreover $(\varepsilon_i)_{i=1,\dots,n}$ be a Rademacher sequence, i.e., $\mathbb{P}(\varepsilon_i = 1) = \mathbb{P}(\varepsilon_i = -1) = \frac{1}{2}$, which is independent of $(X_i, Y_i)_{i=1,\dots,n}$ and $(X_i', Y_i')_{i=1,\dots,n}$. Since the distribution of

$$Z_i := \big(\varphi(-Y_i f(X_i)) - \varphi(-Y_i' f(X_i'))\big)$$

is symmetric, i.e., $Z_i \overset{d}{=} -Z_i$, we also have $\varepsilon_i Z_i \overset{d}{=} Z_i$:

$$\mathbb{P}(\varepsilon_i Z_i \in A) = \frac{1}{2}\mathbb{P}(Z_i \in A) + \frac{1}{2}\mathbb{P}(-Z_i \in A) = \mathbb{P}(Z_i \in A), \qquad \forall A \in \mathscr{B}_{\mathbb{R}}.$$

We conclude

$$\begin{aligned} &\mathbb{E}\Big[\sup_{\|f\|_W \leqslant \lambda} |R_{n,\varphi}(f) - R_\varphi(f)|\Big] \\ &\quad \leqslant \mathbb{E}\Big[\sup_{\|f\|_W \leqslant \lambda} \Big|\frac{1}{n}\sum_{i=1}^n \varepsilon_i\big(\varphi(-Y_i f(X_i)) - 1 + 1 - \varphi(-Y_i' f(X_i'))\big)\Big|\Big] \\ &\quad \leqslant 2\mathbb{E}\Big[\sup_{\|f\|_W \leqslant \lambda} \Big|\frac{1}{n}\sum_{i=1}^n \varepsilon_i\big(\varphi(-Y_i f(X_i)) - 1\big)\Big|\Big]. \end{aligned}$$

Next we use a contraction principle: If $\psi\colon [-1,1] \to \mathbb{R}$ is a contraction, i.e. $|\psi(x) - \psi(y)| \leqslant |x - y|$ and $\psi(0) = 0$, then for any family $\mathcal{G} \subseteq \{g\colon \mathcal{X} \times \{-1,+1\} \to [-1,1] \text{ measurable}\}$ we have (Ledoux & Talagrand, 2011, Thm. 4.12)

$$\mathbb{E}\Big[\sup_{g \in \mathcal{G}} \Big|\frac{1}{n}\sum_{i=1}^n \varepsilon_i \psi(g(X_i, Y_i))\Big|\Big] \leqslant 2\mathbb{E}\Big[\sup_{g \in \mathcal{G}} \Big|\frac{1}{n}\sum_{i=1}^n \varepsilon_i g(X_i, Y_i)\Big|\Big].$$

Using $\|f\| \leqslant \lambda$ and Lemma 3.25 we have $\|f\|_\infty \leqslant \lambda \sup_{x \in \mathcal{X}} \sqrt{k(x,x)} =: L < \infty$. Applying the above inequality to $\psi(u) := (\varphi(Lu) - 1)/L$ and $g(x,y) = -yf(x)/L \in [0,1]$, we obtain

$$\mathbb{E}\Big[\sup_{\|f\|_W \leqslant \lambda} \Big|\frac{1}{n}\sum_{i=1}^n \varepsilon_i \frac{\varphi(-Y_i f(X_i)) - 1}{L}\Big|\Big] \leqslant 2\mathbb{E}\Big[\sup_{\|f\|_W \leqslant \lambda} \Big|\frac{1}{n}\sum_{i=1}^n \varepsilon_i \frac{Y_i f(X_i)}{L}\Big|\Big].$$

Since $\varepsilon_i Y_i \stackrel{d}{=} \varepsilon_i$, we conclude

$$\mathbb{E}\Big[\sup_{\|f\|_W \leqslant \lambda} |R_{n,\varphi}(f) - R_\varphi(f)|\Big] \leqslant 4\mathbb{E}\Big[\sup_{\|f\|_W \leqslant \lambda} \Big|\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(X_i)\Big|\Big].$$

Step 3: Using the Hilbert space structure of the RKHS W, the Cauchy-Schwarz inequality yields

$$\begin{aligned}
\sup_{\|f\|_W \leqslant \lambda} \Big|\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(X_i)\Big|^2 &= \sup_{\|f\|_W \leqslant \lambda} \Big|\frac{1}{n}\sum_{i=1}^n \varepsilon_i \langle f, k(X_i,\cdot)\rangle_W\Big|^2 \\
&= \sup_{\|f\|_W \leqslant \lambda} \Big|\Big\langle f, \frac{1}{n}\sum_{i=1}^n \varepsilon_i k(X_i,\cdot)\Big\rangle_W\Big|^2 \\
&\leqslant \sup_{\|f\|_W \leqslant \lambda} \|f\|_W^2 \Big\|\frac{1}{n}\sum_{i=1}^n \varepsilon_i k(X_i,\cdot)\Big\|_W^2 \\
&= \lambda^2 \frac{1}{n^2}\sum_{i,j=1}^n \varepsilon_i\varepsilon_j k(X_i,X_j).
\end{aligned}$$

Due to $\mathbb{E}[\varepsilon_i\varepsilon_j] = \mathbb{1}_{\{i=j\}}$, we finally obtain

$$\begin{aligned}
\mathbb{E}\Big[\sup_{\|f\|_W \leqslant \lambda} \Big|\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(X_i)\Big|\Big] &\leqslant \lambda \mathbb{E}\Big[\frac{1}{n^2}\sum_{i,j=1}^n \varepsilon_i\varepsilon_j k(X_i,X_j)\Big]^{1/2} \\
&= \frac{\lambda}{n}\Big(\sum_{i=1}^n \mathbb{E}[k(X_i,X_i)]\Big)^{1/2} \\
&= \frac{\lambda}{\sqrt{n}}\sqrt{\mathbb{E}[k(X,X)]}.
\end{aligned}$$

In combination with the previous steps we deduce the claimed oracle inequality. □

Remark 3.31. The stochastic error bound is of the order $\mathcal{O}(1/\sqrt{n})$ in n. The complexity of the family of considered classifiers is bounded by $\lambda\mathbb{E}[k(X,X)]^{1/2}$. Clearly, λ should be chosen in a way that balances approximation error and stochastic error. In practice, cross validation is often used to find an adaptive choice of λ. A general theory of the approximation error and an analysis of the choice of λ can be found in Steinwart & Christmann (2008).

4 Principal component analysis

In the previous chapter we have studied several *supervised learning* methods, i.e., we observe features X_i together with corresponding labels Y_i where the latter live in a finite space (classification) or an infinite lable set (regression). In *unsupervised learning* we only observe the features $X_1, \dots, X_n \in \mathbb{R}^p$ and want to learn their structure. A particular aim that we will study in this chapter is dimension reduction, that is we would like to approximate/ describe X_i in a space $\mathbb{R}^q$ with much smaller dimension $q \ll p$. This is a fundamental and necessary task for any application due to the high-dimensionality of modern datasets. Another important problem in unsupervised learning is clustering where we would like to assign the X_i to different groups of "similar" features.

One of the first and most popular dimension reduction techniques is *principal component analysis*. In order to find a good low dimensional approximation of a given sample $X_1, \dots, X_n$ via an affine function

$$f(v) = \mu + Av \qquad \text{with} \qquad \mu \in \mathbb{R}^p, v \in \mathbb{R}^q \text{ and orthogonal matrix } A \in \mathbb{R}^{p\times q}$$

(i.e., $A^\top A = I_q$), we apply the least squares criterion:

$$\sum_{i=1}^n |X_i - \mu - A\nu_i|^2 \to \min_{\mu,(\nu_i),A} ! \tag{9}$$

For a fixed matrix A, we obtain the solutions (Exercise ☑)

$$\widehat{\mu} := \overline{X} := \frac{1}{n}\sum_{i=1}^n X_i \qquad \text{and} \qquad \widehat{\nu}_i := A^\top(X_i - \overline{X}).$$

For a fixed q, we thus have to solve the minimisation problem

$$\sum_{i=1}^n \left|X_i - \overline{X} - AA^\top(X_i - \overline{X})\right|^2 \to \min_A !$$

We set

$$X := \begin{pmatrix} X_1^\top \\ \vdots \\ X_n^\top \end{pmatrix} \in \mathbb{R}^{n\times p} \quad \text{and} \quad \widetilde{X} := \begin{pmatrix} X_1^\top - \overline{X}^\top \\ \vdots \\ X_n^\top - \overline{X}^\top \end{pmatrix} \in \mathbb{R}^{n\times p}$$

Theorem 4.1. *For $X \in \mathbb{R}^{n\times p}$ set $\overline{X} \in \mathbb{R}^p$ with $\overline{X}_j := \frac{1}{n}\sum_{i=1}^n X_{ij}$ for all $j = 1,\dots,p$ and $\widetilde{X} \in \mathbb{R}^{n\times p}$ as above. Let $\widehat{w}_1,\dots,\widehat{w}_p \in \mathbb{R}^p$ denote the normalised eigenvectors of $\widetilde{X}^\top\widetilde{X} \in \mathbb{R}^{p\times p}$ corresponding to the eigenvalues $\lambda_1^2 \geqslant \dots \geqslant \lambda_p^2 \geqslant 0$. For a given $q < p \wedge n$ the solution of the minimisation problem (9) is given by*

$$(\widehat{\mu}, (\widehat{\nu}_i)_{i=1,\dots,n}, \widehat{A}_q) = \operatorname*{arg\,min}_{\mu,(\nu_i),A} \sum_{i=1}^n |X_i - \mu - A\nu_i|^2$$

with

$$\widehat{\mu} := \overline{X}, \qquad \widehat{\nu}_i := \widehat{A}_q^\top(X_i - \overline{X}), \qquad \widehat{A}_q := (\widehat{w}_1 \cdots \widehat{w}_q).$$

Moreover,

$$\min_{\mu,(\nu_i),A} \sum_{i=1}^n |X_i - \mu - A\nu_i|^2 = \sum_{i=q+1}^p \lambda_i^2.$$

w_i is called i-th principal component.

Proof. Without loss of generality we can assume $\overline{X} = 0$ and thus $X = \widetilde{X}$. We thus have minimise

$$\begin{aligned}\sum_{i=1}^n |X_i - AA^\top X_i|^2 &= \sum_{i=1}^n\sum_{j=1}^p \left(X_{ij} - (AA^\top X^\top)_{j,i}\right)^2 \\ &= \left\|(I_p - AA^\top)X^\top\right\|_F^2 \to \min_A !\end{aligned}$$

with Frobenius norm $\|\cdot\|_F$. Note that for any matrix $B \in \mathbb{R}^{p\times n}$ and any square matrices $C, D \in \mathbb{R}^{p\times p}$

$$\|B\|_F^2 = \operatorname{tr}(B^\top B) = \operatorname{tr}(BB^\top) \qquad \text{and} \qquad \operatorname{tr}(CD) = \operatorname{tr}(DC)$$

Using orthogonality of A, we obtain

$$\begin{aligned}\left\|(I_p - AA^\top)X^\top\right\|_F^2 &= \operatorname{tr}\left((I_p - AA^\top)X^\top((I_p - AA^\top)X^\top)^\top\right) \\ &= \operatorname{tr}\left((I_p - AA^\top)X^\top X(I_p - AA^\top)\right) \\ &= \operatorname{tr}\left(X^\top X(I_p - AA^\top)^2\right) \\ &= \operatorname{tr}\left(X^\top X(I_p - AA^\top)\right). \end{aligned} \tag{10}$$

Since $X^\top X$ is symmetric and positive semi-definite with eigenvalues $\lambda_1^2 \geqslant \ldots \geqslant \lambda_p^2 \geqslant 0$, we find an orthogonal matrix $W = (\widehat{w}_1 \cdots \widehat{w}_p) \in \mathbb{R}^{p\times p}$ such that

$$X^\top X = W D^2 W^\top.$$

Therefore,

$$\begin{aligned}\big\|(I_p - AA^\top)X^\top\big\|_F^2 &= \operatorname{tr}\big(WD^2W^\top(I_p - AA^\top)\big)\\ &= \operatorname{tr}\big(D^2(W^\top W - W^\top AA^\top W)\big)\\ &= \operatorname{tr}\big(D^2(I_p - (W^\top A)(W^\top A)^\top\big).\end{aligned}$$

The matrix $\Pi_A := (W^\top A)(W^\top A)^\top$ is symmetric and idempotent ($\Pi_A\Pi_A = \Pi_A$), i.e., a orthogonal projection matrix, with rank q. Hence, the symmetric matrix $\Pi := I_p - \Pi_A$ is an orthogonal projection with rank $p-q$ and consequently, there is orthogonal matrix $U \in \mathbb{R}^{p\times p}$ and $E := \operatorname{diag}(\underbrace{1,\ldots,1}_{p-q \text{ times}},0,\ldots,0)$ such that $\Pi = U^\top EU$ and

$$\operatorname{tr}(\Pi) = \operatorname{tr}(U^\top EU) = \operatorname{tr}(E) = p - q.$$

Since $\Pi_{ii} = (\Pi^2)_{ii} = \sum_{j=1}^p \Pi_{ij}^2 \geqslant \Pi_{ii}^2$, we have $\Pi_{ii} \in [0,1]$. Owing to $\lambda_i \geqslant \lambda_{i+1}$, we conclude

$$\min_{\Pi} \operatorname{tr}(D^2\Pi) = \min_{\Pi} \sum_{i=1}^p \lambda_i^2 \Pi_{ii} \geqslant \sum_{i=q+1}^p \lambda_i^2.$$

The minimum is attained if Π projects onto the last $p-q$ coordinates that is for

$$\Pi_A = (W^\top A)(W^\top A)^\top = \operatorname{diag}(\underbrace{1,\ldots,1}_{q \text{ times}},\underbrace{0,\ldots,0}_{p-q \text{ times}}).$$

Therefore, we choose $\widehat{A}_q = (\widehat{w}_1 \cdots \widehat{w}_q)$. □

While reducing the dimension of the observations, PCA extracts significant features from the data by focussing on the subspace spanned by the eigenvectors corresponding to the largest eigenvalues. Extensions like "manifold learning" or "kernel PCA" allow for non-linear approximations, see e.g. Hastie et al. (2009).

In addition to the above (numerical) approach there is a statistical interpretation of PCA via empirical covariances: Suppose we have identically distributed observations $X_1,\ldots,X_n \in \mathbb{R}^p$ on some probability space $(\mathbb{R}^p, \mathscr{B}_{\mathbb{R}^p}, \mathbb{P})$ with $X_i \in L^2(\mathbb{P})$ and

$$\mu = \mathbb{E}[X_1] \in \mathbb{R}^p \qquad \text{and} \qquad \Sigma = \mathbb{E}\big[(X_1-\mu)(X_1-\mu)^\top\big] \in \mathbb{R}^{p\times p}.$$

The mean and the covariance matrix can be estimated (unbiased) via their empirical counterparts

$$\widehat{\mu}_n = \overline{X} = \frac{1}{n}\sum_{i=1}^n X_i \qquad \text{and} \qquad \widehat{\Sigma}_n = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X})(X_i-\overline{X})^\top = \frac{1}{n-1}\widetilde{X}^\top\widetilde{X},$$

with $\widetilde{X}$ from above. From the previous proof, we see that the objective function of our dimension reduction approach is

$$R_n(\mu, A) := \frac{1}{n-1}\sum_{i=1}^n |X_i - \mu - AA^\top(X_i-\mu)|^2 = \operatorname{tr}\big(\widehat{\Sigma}_n(I_p - AA^\top)\big).$$

Moreover, the eigenvectors $\widehat{w}_1,\ldots,\widehat{w}_p$ of $\widetilde{X}^\top\widetilde{X}$ coincide with the eigenvectors of $\widehat{\Sigma}_n$ and the corresponding eigenvalues of $\widehat{\Sigma}_n$ are given by

$$\frac{\lambda_1^2}{n-1} \geqslant \ldots \geqslant \frac{\lambda_p^2}{n-1} \geqslant 0.$$

For the matrix $W = (\widehat{w}_1 \cdots \widehat{w}_p)$ we have $\widetilde{X}^\top \widetilde{X} = WD^2W^\top$ and thus

$$W^\top \widehat{\Sigma}_n W = \frac{1}{n-1} W^\top W D^2 W^\top W = \frac{1}{n-1} D^2.$$

Hence, the empirical covariance matrix of the transformed data $(W^\top X_i)_{i=1,\dots,n}$ is a diagonal matrix, which can be understood as $(W^\top X_i)_{i=1,\dots,n}$ being empirically uncorrelated. This transformation is sometimes called "whitening tranformation". The PCA approximations

$$\big(\widehat{A}_q^\top \widetilde{X}_i\big)_{i=1,\dots,n}$$

from Theorem 4.1 especially have the empirical covariance matrix $\operatorname{diag}\big(\frac{\lambda_1^2}{n-1}, \cdots, \frac{\lambda_q^2}{n-1}\big)$. The following proposition shows that PCA chooses the q directions which explain the most variability in the data, that is the principal components are the directions with the maximal (empirical) variance.

Proposition 4.2. *Let $X_1, \dots, X_n \in \mathbb{R}^p$ be a sequence and let $\widehat{\Sigma}_n \in \mathbb{R}^{p\times p}$ and $\widehat{w}_1, \dots, \widehat{w}_p \in \mathbb{R}^p$ be defined as above. Denoting the unit sphere in $\mathbb{R}^p$ by $S_p = \{v \in \mathbb{R}^p : |v| = 1\}$, we have*

(i) $\widehat{w}_1 = \arg\max_{v\in S_p} \langle \widehat{\Sigma}_n v, v\rangle = \arg\max_{v\in S_p} \sum_{i=1}^n \langle \widetilde{X}_i, v\rangle^2$ *and*

(ii) $\widehat{w}_k = \arg\max_{v\in S_p : v\perp \widehat{w}_j, j<k} \langle \widehat{\Sigma}_n v, v\rangle = \arg\max_{v\in S_p : v\perp \widehat{w}_j, j<k} \sum_{i=1}^n \langle \widetilde{X}_i, v\rangle^2$ *for $k = 2, \dots, p$.*

Proof. The formulas for $\widehat{w}_k$ are the variational characterisations of eigenvectors of symmetric positive semi-definite matrices.

For $k = 1$ use $\widetilde{X}^\top \widetilde{X} = WD^2W^\top$ and thus for any $v \in S_p$

$$\begin{aligned}(n-1)\langle \widehat{\Sigma}_n v, v\rangle = \langle WD^2W^\top v, v\rangle =& \langle D^2W^\top v, W^\top v\rangle \\ \leqslant& \lambda_1^2 |W^\top v|^2 = \lambda_1^2 |v|^2 = \lambda_1^2.\end{aligned}$$

Equality is attained for $v = \widehat{w}_1$ because $(n-1)\langle \widehat{\Sigma}_n \widehat{w}_1, \widehat{w}_1\rangle = \langle D^2W^\top \widehat{w}_1, W^\top \widehat{w}_1\rangle = \lambda_1^2$.

For $k \geqslant 2$ we proceed inductively: Let $v \in S_p : v \perp \widehat{w}_j, j < k$. Then $v = (I_p - \sum_{i=1}^{k-1} \widehat{w}_i \widehat{w}_i^\top)v$ since

$$\widehat{w}_i \widehat{w}_i^\top v = \widehat{w}_i \langle \widehat{w}_i, v\rangle = 0 \quad \forall i = 1, \dots, k-1.$$

Using

$$\begin{aligned}(I_p - \sum_{i=1}^{k-1} \widehat{w}_i \widehat{w}_i^\top)^\top W &= (I_p - \sum_{i=1}^{k-1} \widehat{w}_i \widehat{w}_i^\top) W \\ &= W - \sum_{i=1}^{k-1} \widehat{w}_i \widehat{w}_i^\top (\widehat{w}_1 \ \cdots \ \widehat{w}_p) \\ &= W - \sum_{i=1}^{k-1} \widehat{w}_i (\widehat{w}_i^\top \widehat{w}_1, \cdots, \widehat{w}_i^\top \widehat{w}_p) = W \operatorname{diag}(\underbrace{0, \dots, 0}_{k-1 \text{ times}}, \underbrace{1, \dots, 1}_{p-k+1 \text{ times}}),\end{aligned}$$

we find

$$\begin{aligned}(n-1)\langle \widehat{\Sigma}_n v, v\rangle &= \Big((I_p - \sum_{i=1}^{k-1} \widehat{w}_i \widehat{w}_i^\top)v\Big)^\top \widetilde{X}^\top \widetilde{X} \Big((I_p - \sum_{i=1}^{k-1} \widehat{w}_i \widehat{w}_i^\top)v\Big) \\ &= v^\top \Big((I_p - \sum_{i=1}^{k-1} \widehat{w}_i \widehat{w}_i^\top)^\top W\Big) D^2 \Big((I_p - \sum_{i=1}^{k-1} \widehat{w}_i \widehat{w}_i^\top)^\top W\Big)^\top v \\ &= v^\top \operatorname{diag}(\underbrace{0, \dots, 0}_{k-1 \text{ times}}, \lambda_k^2, \dots, \lambda_p^2) v \\ &\leqslant \lambda_k^2 |v|^2 = \lambda_k^2.\end{aligned}$$

This upper bound is attained for $v = \widehat{w}_k$ since $(n-1)\langle\widehat{\Sigma}_n\widehat{w}_k, \widehat{w}_k\rangle = \langle D^2 W^\top \widehat{w}_k, W^\top \widehat{w}_k\rangle = \lambda_k^2$.

To conclude the proof note that

$$(n-1)\langle\widehat{\Sigma}_n v, v\rangle = v^\top \widetilde{X}^\top \widetilde{X} v = \sum_{i=1}^{p} \langle \widetilde{X}_i, v\rangle^2. \qquad \square$$

Remark 4.3.

(i) The dimension q is often chosen such that certain percent (e.g. 50% or 90%) of the overall empirical variance $\sum_{i=1}^p \lambda_i^2$ are explained by $\sum_{i=1}^q \lambda_i^2$.

(ii) If $p \gg n$ is very large, a diagonalisation of $\widetilde{X}^\top \widetilde{X} \in \mathbb{R}^{p\times p}$ might be too expensive. This can be circumvented by considering the smaller matrix $\widetilde{X}\widetilde{X}^\top \in \mathbb{R}^{n\times n}$ with eigenvalues $\mu_1^2 \geqslant \mu_2^2 \geqslant \ldots \geqslant \mu_n^2 \geqslant 0$ and eigenvectors $u_1, \ldots, u_n \in \mathbb{R}^n$. Then

$$\widetilde{X}^\top \widetilde{X}\widetilde{X}^\top u_i = \mu_i^2 \widetilde{X}^\top u_i.$$

Hence $\widehat{w}_i = \widetilde{X}^\top u_i \in \mathbb{R}^p$ is an eigenvector for the eigenvalue $\lambda_i^2 = \mu_i^2$. If $(u_i)_{i=1,\ldots,n}$ are chosen such that they build an orthonormal basis of $\mathbb{R}^n$, we have

$$|\widehat{w}_i|^2 = u_i^\top \widetilde{X}\widetilde{X}^\top u_i = \mu_i^2 |u_i|^2 = \lambda_i^2$$

and for all $v \in \mathbb{R}^p$ we obtain the *singular value decomposition* of $\widetilde{X} \in \mathbb{R}^{n\times p}$

$$\widetilde{X} u = \sum_{i=1}^{n} \langle \widetilde{X} v, u_i\rangle u_i = \sum_{i=1}^{n} \langle v, \widetilde{X}^\top u_i\rangle u_i = \sum_{i=1}^{n} \lambda_i \Big\langle v, \frac{\widehat{w}_i}{|\widehat{w}_i|}\Big\rangle u_i$$

In order to measure how well PCA approximates the observations in a low-dimensional subspace, we will now study the reconstruction error

$$R(\mu, A) := \mathbb{E}\big[|X_{n+1} - \mu - AA^\top (X_{n+1} - \mu)|^2\big].$$

of a new observation X_{n+1} independent of $X_1, \ldots, X_n$ and with the same distribution. We moreover assume that $(X_i)_{i=1,\ldots,n}$ are i.i.d. Replacing the expectation by its empirical counterpart, we obtain $R_n(\mu, A)$ and thus the PCA solution $\widehat{\mu}$ and $\widehat{A}_q$ is the corresponding empirical risk minimiser.

The smallest possible reconstruction error is determined in the following Lemma.

Lemma 4.4. *Let $X_{n+1} \in L^2(P)$ with $\mathbb{E}[X_{n+1}] = \mu^* \in \mathbb{R}$ and $\mathrm{Cov}(X_{n+1}) = \mathbb{E}[(X_{n+1} - \mu)(X_{n+1} - \mu)^\top] = \Sigma \in \mathbb{R}^{p\times p}$. Then the oracle reconstruction error is given by*

$$R^* := \min_{\mu\in\mathbb{R}^p, A\in\mathbb{R}^{p\times q} \text{ orthogonal}} R(\mu, A) = R(\mu^*, A_q^*) = \sum_{i=q+1}^{p} \lambda_i^2$$

for $A_q^* = (w_1 \cdots w_q) \in \mathbb{R}^{p\times q}$ where $w_1, \ldots, w_p \in \mathbb{R}^p$ denote the eigenvectors of Σ corresponding to eigenvalues $\lambda_1^2 \geqslant \ldots \geqslant \lambda_p^2 \geqslant 0$.

Proof. We abbreviate $X = X_{n+1}$. Differentiating with respect to μ reveals that the optimal choice is $\mu = \mu^*$. Further note that $\Pi = AA^\top$ is an orthogonal projection matrix (i.e. symmetric and $\Pi^2 = \Pi$) if and only if A is orthogonal ($A^\top A = I_q$). By the identity $a^\top a = \mathrm{tr}(a^\top a) = \mathrm{tr}(aa^\top)$ for any $a \in \mathbb{R}^p$ and linearity of the trace, we thus obtain

$$\begin{aligned}
R(\mu^*, A) = \mathbb{E}\big[|X - \mu^* - \Pi(X - \mu^*)|^2\big] &= \mathbb{E}\big[(X - \mu^*)^\top (I_p - \Pi)^\top (I_p - \Pi)(X - \mu^*)\big] \\
&= \mathbb{E}\big[\,\mathrm{tr}\,\big((I_p - \Pi)(X - \mu^*)(X - \mu^*)^\top (I_p - \Pi)^\top\big)\big] \\
&= \mathrm{tr}\,\big((I_p - \Pi)\mathbb{E}\big[(X - \mu^*)(X - \mu^*)^\top\big](I_p - \Pi)^\top\big) \\
&= \mathrm{tr}\,\big((I_p - \Pi)\Sigma(I_p - \Pi)^\top\big) \\
&= \mathrm{tr}\,\big(\Sigma(I_p - \Pi)^\top\big),
\end{aligned}$$

which has to be minimised with respect to Π. Noting that this is the same term as in (10) with $\widetilde{X}^\top \widetilde{X}$ replaced by Σ, the rest of the proof is analogous to the one of Theorem 4.1. □

Remark 4.5. In analogy to Proposition 4.2 we notice

$$\mathrm{Var}(\langle X_{n+1}, w_1\rangle) = \max_{v \in S_p} \mathrm{Var}(\langle X_{n+1}, v\rangle)$$

and

$$\mathrm{Var}(\langle X_{n+1}, w_k\rangle) = \max_{v \in S_p : v \perp w_j, j<k} \mathrm{Var}(\langle X_{n+1}, v\rangle)$$

such that the $\mathrm{span}(w_1, \dots, w_q)$ is the q-dimenional linear subspace with the largest variance of the projected observations.

In order to investigate how close the principal component analysis is to the oracle, we study the *excess risk*

$$\mathcal{E}(\widehat{\mu}, \widehat{A}_q) := R(\widehat{\mu}, \widehat{A}_q) - R^*.$$

Lemma 4.6. *Define* $\widehat{\Pi}_q := \widehat{A}_q \widehat{A}_q^\top$ *with PCA choice* $\widehat{A}_q$ *from Theorem 4.1 as well as* $\Pi_q^* = A_q^* (A_q^*)^\top$ *with* A_q^* *from Lemma 4.4. We have in terms of the Frobenius scalar product* $\langle A, B\rangle_F := \mathrm{tr}(A^\top B)$

$$\mathcal{E}(\widehat{\mu}, \widehat{A}_q) = \langle \Sigma, \Pi_q^* - \widehat{\Pi}_q\rangle_F + |(I_p - \widehat{\Pi}_q)(\widehat{\mu} - \mu^*)|^2.$$

Moreover,

$$\mathcal{E}(\widehat{\mu}, \widehat{A}_q) \leqslant \langle \Sigma - \widehat{\Sigma}_n, \Pi_q^* - \widehat{\Pi}_q\rangle_F + |\widehat{\mu} - \mu^*|^2$$

with the empirical covariance matrix $\widehat{\Sigma}_n \in \mathbb{R}^{p\times p}$.

Proof. In the previous proof, we have seen that for any orthogonal matrix $A \in \mathbb{R}^{p\times q}$ and $\Pi = AA^\top$ we have

$$R(\mu, A) = \mathrm{tr}\left(\mathbb{E}\big[(X-\mu)(X-\mu)^\top\big](I_p - \Pi)^\top\right).$$

The expectation equals (as in the bias-variance decomposition)

$$\mathbb{E}\big[(X-\mu)(X-\mu)^\top\big] = (\mu - \mu^*)(\mu - \mu^*)^\top + \Sigma.$$

Therefore,

$$\begin{aligned}
\mathcal{E}(\widehat{\mu}, \widehat{A}_n) &= \mathrm{tr}\left(((\widehat{\mu} - \mu^*)(\widehat{\mu} - \mu^*)^\top + \Sigma)(I_p - \widehat{\Pi})^\top\right) - \mathrm{tr}\left(\Sigma(I_p - \Pi^*)^\top\right) \\
&= \mathrm{tr}\left(\Sigma(\Pi_q^* - \widehat{\Pi}_q)^\top\right) + \mathrm{tr}\left((\widehat{\mu} - \mu^*)(\widehat{\mu} - \mu^*)^\top (I_p - \widehat{\Pi}_q)^\top\right) \\
&= \mathrm{tr}\left(\Sigma(\Pi_q^* - \widehat{\Pi}_q)^\top\right) + |(I_p - \widehat{\Pi}_q)(\widehat{\mu} - \mu^*)|^2,
\end{aligned}$$

using the auxiliary calculation for any $a \in \mathbb{R}^p$ and any orthogonal projection matrix Π:

$$\mathrm{tr}(aa^\top \Pi^\top) = \mathrm{tr}(aa^\top \Pi^\top \Pi) = \mathrm{tr}(\Pi a a^\top \Pi^\top) = \mathrm{tr}((\Pi a)(\Pi a)^\top) = (\Pi a)^\top (\Pi a).$$

We thus have proved the first equality. For the upper bound we use that $\widehat{\Pi}$ is the minimiser of $\min_\Pi \mathrm{tr}(\widehat{\Sigma}_n, I - \Pi) = \langle \widehat{\Sigma}_n, I_p - \widehat{\Pi}_q\rangle_F$, we deduce

$$\begin{aligned}
\langle \Sigma, \Pi_q^* - \widehat{\Pi}_q\rangle_F &= \langle \Sigma - \widehat{\Sigma}_n, \Pi_q^* - \widehat{\Pi}_q\rangle_F + \langle \widehat{\Sigma}_n, \Pi_q^* - \widehat{\Pi}_q\rangle_F \\
&= \langle \Sigma - \widehat{\Sigma}_n, \Pi_q^* - \widehat{\Pi}_q\rangle_F + \underbrace{\langle \widehat{\Sigma}_n, I_p - \widehat{\Pi}_q\rangle_F - \langle \widehat{\Sigma}_n, I_p - \Pi_q^*\rangle_F}_{\leqslant 0} \\
&\leqslant \langle \Sigma - \widehat{\Sigma}_n, \Pi_q^* - \widehat{\Pi}_q\rangle_F.
\end{aligned}$$

For the second term, we note that $I_p - \widehat{\Pi}_q$ is a projection matrix, such that all eigenvalues are in $\{0, 1\}$ and consequently

$$|(I_p - \widehat{\Pi}_q)(\widehat{\mu} - \mu^*)| \leqslant \|I_p - \widehat{\Pi}_q\|_{\mathrm{op}} |\widehat{\mu} - \mu^*| \leqslant |\widehat{\mu} - \mu^*|.$$ □

By the Cauchy-Schwarz inequality we obtain

$$\mathcal{E}(\widehat{\mu}, \widehat{A}_n) \leqslant \|\Sigma - \widehat{\Sigma}_n\|_F \|\Pi_q^* - \widehat{\Pi}_q\|_F + |\widehat{\mu} - \mu^*|^2.$$

Here, the term $\|\Sigma - \widehat{\Sigma}_n\|_F$ measures the error for estimating the covariance matrix Σ (cf. Exercise ☑), while $\|\Pi_q^* - \widehat{\Pi}_q\|_F$ quantifies the distance between the projection on the best possible or oracle q-dimensional subspace to the PCA projection. The second term $|\widehat{\mu} - \mu^*|^2$ is of the order p/n and will be negligible compared to the first term.

An easy upper bound for $\|\Pi_q^* - \widehat{\Pi}_q\|_F$ is given by

$$\begin{aligned}\|\Pi_q^* - \widehat{\Pi}_q\|_F^2 &= \|\Pi_q^*\|_F^2 + \|\widehat{\Pi}_q\|_F^2 - 2\langle \Pi_q^*, \widehat{\Pi}_q \rangle_F \\ &= 2(q - \langle \Pi_q^*, \widehat{\Pi}_q \rangle_F) \leqslant 2q,\end{aligned}$$

where we have used $\langle \Pi_q^*, \widehat{\Pi}_q \rangle_F = \operatorname{tr}((\Pi_q^*)^\top \widehat{\Pi}_q) = \operatorname{tr}((\Pi_q^* \widehat{\Pi}_q)^\top \Pi_q^* \widehat{\Pi}_q) = \|\Pi_q^* \widehat{\Pi}_q\|_F^2 \geqslant 0$. For a possibly sharper estimate we want to bound this projection error in terms of the estimation error $\widehat{\Sigma}_n - \Sigma$. A first reduction is provided by the following Lemma:

Lemma 4.7. *Denoting the eigenvectors of $\widehat{\Sigma}_n$ and Σ by $(\widehat{w}_i)_{i=1,\dots,p} \subseteq \mathbb{R}^p$ and $(w_i)_{i=1,\dots,p} \in \mathbb{R}^p$ ordered with respect to the size of the corresponding eigenvalues, respectively, we have*

$$\|\Pi_q^* - \widehat{\Pi}_q\|_F^2 \leqslant 2 \sum_{i=1}^{q} |\widehat{w}_i - w_i|^2.$$

Proof. We denote the orthogonal projection onto $\operatorname{span}(w_{q+1}, \dots, w_p) = \operatorname{span}(w_1, \dots, w_q)^\perp$ by $\Pi_{>q}^*$ and analogously define $\widehat{\Pi}_{>q}$. Using $I_p = \Pi_q^* + \Pi_{>q}^*$, we have

$$\begin{aligned}\|\Pi_q^* - \widehat{\Pi}_q\|_F{}^2 &= \|(\Pi_q^* + \Pi_{>q}^*)(\Pi_q^* - \widehat{\Pi}_q)\|_F^2 \\ &= \|\Pi_q^*(\Pi_p^* - \widehat{\Pi}_q)\|_F^2 + \|\Pi_{>q}^*(\Pi_q^* - \widehat{\Pi}_q)\|_F^2 + 2 \operatorname{tr}\big((\Pi_q^* - \widehat{\Pi}_q)^\top \underbrace{(\Pi_q^*)^\top \Pi_{>q}^*}_{=0} (\Pi_q^* - \widehat{\Pi}_q)\big) \\ &= \|\Pi_q^*(I_p - \widehat{\Pi}_p)\|_F^2 + \|\Pi_{>q}^* \widehat{\Pi}_q\|_F^2 \\ &= \|\Pi_q^* \widehat{\Pi}_{>q}\|_F^2 + \|\Pi_{>q}^* \widehat{\Pi}_q\|_F^2.\end{aligned}$$

Recall that for any operator $B \colon \mathbb{R}^p \to \mathbb{R}^p$ and any orthonormal basis $(e_i)_{i=1,\dots,p}$ of $\mathbb{R}^p$ we have

$$\|B\|_F^2 = \operatorname{tr}(B^\top B) = \sum_{i=1}^{p} |Be_i|^2.$$

Applied to $B = \Pi_q^* \widehat{\Pi}_q$ and $e_i = \widehat{w}_i$, we obtain with $\widehat{\Pi}_{>q} \widehat{w}_i = 1$ for $i > q$ and $\widehat{\Pi}_{>q} \widehat{w}_i = 0$ otherwise:

$$\begin{aligned}\|\Pi_q^* \widehat{\Pi}_{>q}\|_F^2 = \sum_{i=1}^{p} |\Pi_q^* \widehat{\Pi}_{>q} \widehat{w}_i|^2 &= \sum_{i=q+1}^{p} |\Pi_q^* \widehat{w}_i|^2 \\ &= \sum_{i=q+1}^{p} \sum_{j=1}^{q} \langle w_j, \widehat{w}_i \rangle^2 = \sum_{i=q+1}^{p} \sum_{j=1}^{q} \langle w_j - \widehat{w}_j, \widehat{w}_i \rangle^2 \leqslant \sum_{j=1}^{q} |\widehat{w}_j - w_j|^2\end{aligned}$$

using Parseval's identity in the last estimate. Analogously, we obtain $\|\Pi_{>q}^* \widehat{\Pi}_q\|_F^2 \leqslant \sum_{j=1}^{q} |\widehat{w}_j - w_j|^2$ which implies the asserted bound. □

The difference between estimated and true eigenvector can be bounded by the operator norm of the corresponding matrices as the following variant of the Davis & Kahan (1970) theorem verifies:

Proposition 4.8. *Let $A, B \in \mathbb{R}^{p\times p}$ be symmetric with eigenvalues*

$$\lambda_{1,C}^2 \geqslant \cdots \geqslant \lambda_{p,C}^2, \qquad C \in \{A, B\},$$

and normalised eigenvectors

$$w_{1,C}, \ldots, w_{p,C}, \qquad C \in \{A, B\},$$

where we suppose w.l.o.g. $\langle w_{i,A}, w_{i,B}\rangle \geqslant 0$ for all $i = 1, \ldots, p$. For $k \in \{1, \ldots, p\}$ define

$$\begin{aligned}\psi_{1,B} &:= \lambda_{1,B}^2 - \lambda_{2,B}^2, \qquad \psi_{p,B}^2 = \lambda_{p-1,B}^2 - \lambda_{p,B}^2 \quad \textit{and}\\ \psi_{k,B} &:= \min\{\lambda_{k,B}^2 - \lambda_{k+1,B}^2, \lambda_{k-1,B}^2 - \lambda_{k,B}^2\}, \quad k = 2, \ldots, p-1.\end{aligned}$$

If $\psi_{k,B} > 0$, then for all $k \in \{1, \ldots, p\}$:

$$|w_{k,A} - w_{k,B}| \leqslant \frac{2\sqrt{2}}{\psi_{k,B}} \|A - B\|_{\mathrm{op}}.$$

If A is perturbed version of the matrix B, the above inequality implies that the eigenvectors of A and B, respectively, are close together if the perturbation $A - B$ is small (in operator norm) and if the so-called spectral gap $\psi_{k,B}$, that is the distance of the k-th eigenvalue to the nearest other eigenvalue of B, is not too small.

Proof. We will use the following basic results(Exercise ☑):

$$\begin{aligned}|\lambda_{k,A}^2 - \lambda_{k,B}^2| &\leqslant \|A - B\|_{\mathrm{op}} \quad \text{for all } k = 1, \ldots, p,\\ \langle w_{k,A} - w_{k,B}, w_{k,B}\rangle &= -\frac{1}{2}|w_{k,A} - w_{k,B}|^2 \quad \text{for all } k = 1, \ldots, p.\end{aligned}$$

Based on the above equality and Parseval's identity, we have

$$\begin{aligned}|w_{k,A} - w_{k,B}|^2 &= \sum_{i=1}^{p} \langle w_{k,A} - w_{k,B}, w_{i,B}\rangle^2\\ &\leqslant \sum_{i=1,\ldots,p, i\neq k} \langle w_{k,A} - w_{k,B}, w_{i,B}\rangle^2 + \frac{1}{4}|w_{k,A} - w_{k,B}|^4.\end{aligned}$$

Due to $\langle w_{k,A}, w_{k,B}\rangle \geqslant 0$, we have

$$\begin{aligned}|w_{k,A} - w_{k,B}|^2 &= |w_{k,A}|^2 - 2\langle w_{k,A}, w_{k,B}\rangle + |w_{k,B}|^2\\ &\leqslant |w_{k,A}|^2 + |w_{k,B}|^2 = 2.\end{aligned}$$

Hence,

$$\frac{1}{2}|w_{k,A} - w_{k,B}|^2 \leqslant \sum_{i=1,\ldots,p, i\neq k} \langle w_{k,A} - w_{k,B}, w_{i,B}\rangle^2 = \sum_{i=1,\ldots,p, i\neq k} \langle w_{k,A}, w_{i,B}\rangle^2.$$

By symmetry of B and the ordering of $\lambda_{i,B}^2$, we can moreover estimate

$$\begin{aligned}|Bw_{k,A} - \lambda_{k,B}^2 w_{k,A}|^2 &= \sum_{i=1}^{p} \big(\langle Bw_{k,A}, w_{i,B}\rangle - \lambda_{k,B}^2\langle w_{k,A}, w_{i,B}\rangle\big)^2\\ &= \sum_{i=1}^{p} \big(\langle w_{k,A}, Bw_{i,B}\rangle - \lambda_{k,B}^2\langle w_{k,A}, w_{i,B}\rangle\big)^2\\ &= \sum_{i=1}^{p} (\lambda_{i,B}^2 - \lambda_{k,B}^2)^2\langle w_{k,A}, w_{i,B}\rangle^2\\ &\geqslant \psi_{k,B}^2 \sum_{i=1,\ldots,p, i\neq k} \langle w_{k,A}, w_{i,B}\rangle^2.\end{aligned}$$

Therefore,

$$\frac{1}{2}\psi_{k,B}^2|w_{k,A}-w_{k,B}|^2 \leqslant |Bw_{k,A}-\lambda_{k,B}^2 w_{k,A}|^2.$$

Applying the above uniform bound on $|\lambda_{k,A}^2-\lambda_{k,B}^2|$, we deduce

$$\begin{aligned}|Bw_{k,A}-\lambda_{k,B}^2 w_{k,A}| &\leqslant |(B-A)w_{k,A}|+|Aw_{k,A}-\lambda_{k,B}^2 w_{k,A}|\\ &\leqslant \|B-A\|_{\mathrm{op}}+|\lambda_{k,A}^2-\lambda_{k,B}^2|\\ &\leqslant 2\|B-A\|_{\mathrm{op}}.\end{aligned}$$

In combination with the previous estimate, we have proved

$$|w_{k,A}-w_{k,B}| \leqslant \frac{2\sqrt{2}}{\psi_{k,B}}\|A-B\|_{\mathrm{op}}. \qquad \square$$

As a corollary the combination of the previous results yields the following theorem.

Theorem 4.9. *Let the observations $X_1,\dots,X_n$ be identically distributed with $\mathbb{E}[X_i]=\mu^*$ and $\mathbb{E}[X_iX_i^\top]=\Sigma$ and set*

$$\widehat{\mu}=\frac{1}{n}\sum_{i=1}^n X_i \qquad \text{and} \qquad \widehat{\Sigma}_n=\frac{1}{n-1}\sum_{i=1}^n (X_i-\widehat{\mu})(X_i-\widehat{\mu})^\top$$

Define $\widehat{\Pi}_q := \widehat{A}_q\widehat{A}_q^\top$ with PCA choice $\widehat{A}_q$ from Theorem 4.1. Define for $k=2,\dots,p-1$

$$\psi_{k,\Sigma} := \min\{\lambda_k^2-\lambda_{k+1}^2, \lambda_{k-1,B}^2-\lambda_{k,B}^2\},$$

and $\psi_{1,\Sigma}=\lambda_1^2-\lambda_2^2, \psi_{p,\Sigma}=\lambda_{p-1}^2-\lambda_p^2$ where $\lambda_1^2\geqslant\ldots\geqslant\lambda_p^2$ denote the eigenvalues of Σ. Then

$$\mathcal{E}(\widehat{\mu},\widehat{A}_q) \leqslant \|\Sigma-\widehat{\Sigma}_n\|_F \min\Big(\sqrt{2q}, 4\|\widehat{\Sigma}_n-\Sigma\|_{\mathrm{op}}\Big(\sum_{i=1}^q \frac{1}{\psi_{i,\Sigma}^2}\Big)^{1/2}\Big)+|\widehat{\mu}-\mu^*|^2.$$

Proof. We have

$$\begin{aligned}\mathcal{E}(\widehat{\mu},\widehat{A}_q) &\leqslant \|\Sigma-\widehat{\Sigma}_n\|_F\|\Pi_q^*-\widehat{\Pi}_q\|_F+|\widehat{\mu}-\mu^*|^2\\ &\leqslant \|\Sigma-\widehat{\Sigma}_n\|_F \min\Big(\sqrt{2q}, \Big(2\sum_{i=1}^q |\widehat{w}_i-w_i|^2\Big)^{1/2}\Big)+|\widehat{\mu}-\mu^*|^2\\ &\leqslant \|\Sigma-\widehat{\Sigma}_n\|_F \min\Big(\sqrt{2q}, 4\|\widehat{\Sigma}_n-\Sigma\|_{\mathrm{op}}\Big(\sum_{i=1}^q \frac{1}{\psi_{i,\Sigma}^2}\Big)^{1/2}\Big)+|\widehat{\mu}-\mu^*|^2. \qquad \square\end{aligned}$$

The above theorem reveals that the excess risk of PCA determined by *(i)* the accuracy for estimating Σ and *(ii)* the so called *spectral gap* $\psi_{k,\Sigma}$. Since λ_k^2 are monotone decreasing, we will also have $\psi_{k,\Sigma}\to 0$. Hence, the additional price $\frac{1}{\psi_{q+1,\Sigma}}$ for taking into account one additional dimension grows in q. In terms of

$$\overline{\psi}_q = \min\{\lambda_k^2-\lambda_{k+1}^2, k=1,\dots,q-1\}$$

the theorem yields the upper bound

$$\mathcal{E}(\widehat{\mu},\widehat{A}_n) \leqslant \|\Sigma-\widehat{\Sigma}_n\|_F \min\Big(\sqrt{2q}, \frac{4}{\overline{\psi}_{q,\Sigma}}\sqrt{q}\|\widehat{\Sigma}_n-\Sigma\|_{\mathrm{op}}\Big)+|\widehat{\mu}-\mu^*|^2.$$

From the previous estimate we obtain the oracle inequality for the reconstruction error:

$$\begin{aligned}R(\widehat{\mu},\widehat{A}_n) &\leqslant \min_{\mu,A} R(\mu,A)+\frac{4}{\overline{\psi}_{q,\Sigma}}q\|\widehat{\Sigma}_n-\Sigma\|_{\mathrm{op}}\|\Sigma-\widehat{\Sigma}_n\|_F+|\widehat{\mu}-\mu^*|^2\\ &= \sum_{i=q+1}^p \lambda_i^2+\frac{4\sqrt{q}}{\overline{\psi}_{q,\Sigma}}\|\widehat{\Sigma}_n-\Sigma\|_{\mathrm{op}}\|\Sigma-\widehat{\Sigma}_n\|_F+|\widehat{\mu}-\mu^*|^2.\end{aligned}$$

The last line reveals the trade-off for choosing q: The approximation error decreases for larger choices of q, while the stochastic error term increases. These bounds can be improved in several ways, cf. Reiß & Wahl (2019). It is especially possible to replace $\sqrt{q}\|\widehat{\Sigma}_n - \Sigma\|_{\mathrm{op}}$ by $\|\widehat{\Sigma}_n - \Sigma\|_F$.

References

Bühlmann, P. & van de Geer, S. (2011). *Statistics for high-dimensional data*. Springer Series in Statistics. Springer, Heidelberg. Methods, theory and applications.

Davis, C. & Kahan, W. M. (1970). The rotation of eigenvectors by a perturbation. III. *SIAM J. Numer. Anal.*, 7, 1–46.

Giné, E. & Nickl, R. (2016). *Mathematical foundations of infinite-dimensional statistical models*. Cambridge Series in Statistical and Probabilistic Mathematics, [40]. Cambridge University Press, New York.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning*. Springer Series in Statistics. Springer, New York, second edition. Data mining, inference, and prediction.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning (with Applications in R)*. New York: Springer.

Ledoux, M. & Talagrand, M. (2011). *Probability in Banach spaces*. Classics in Mathematics. Springer-Verlag, Berlin. Isoperimetry and processes, Reprint of the 1991 edition.

Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2018). *Foundations of machine learning*. MIT press.

Reiß, M. & Wahl, M. (2019). Non-asymptotic upper bounds for the reconstruction error of pca. *Annals of Statistics*. To appear.

Shalev-Shwartz, S. & Ben-David, S. (2014). *Understanding machine learning: From theory to algorithms*. Cambridge university press.

Shao, J. (2003). *Mathematical Statistics*. New York: Springer.

Steinwart, I. & Christmann, A. (2008). *Support vector machines*. Information Science and Statistics. Springer, New York.

Tibshirani, R. J. (2013). The lasso problem and uniqueness. *Electron. J. Stat.*, 7, 1456–1490.

Tsybakov, A. B. (2009). *Introduction to nonparametric estimation*. Springer Series in Statistics. Springer, New York. Revised and extended from the 2004 French original, Translated by Vladimir Zaiats.