Some notes on mathematical statistics

Lorenz Richter
BTU Cottbus-Senftenberg

September 5, 2017

Contents

1 The statistical model

We consider the sample space $\mathcal{X}$, a σ-algebra $\mathcal{F}$ and a family of probability measures $(\mathbb{P}_\vartheta)_{\vartheta\in\Theta}$. Usually we observe i.i.d. data $X_1,\dots,X_n \sim \mathbb{P}_\vartheta$. The goal of statistics is to infer the true parameter ϑ from the observed data, i.e. to find a good estimator $\hat{\vartheta}$.

Definition 1.1 (Decision rule)**.** *A decision rule is a measurable map $\rho : \mathcal{X} \to A$, where A is called action space.*

Definition 1.2 (Loss function, risk)**.** *A function $l : \Theta \times A \to \mathbb{R}^+$ that is measurable in the second argument is called loss function. The "risk" (or "error") of a decision rule ρ under the true parameter ϑ is given by*

$$R(\vartheta,\rho) = \mathbb{E}_\vartheta[l(\vartheta,\rho)] = \int_{\mathcal{X}} l(\vartheta,\rho(x))\mathbb{P}_\vartheta(\mathrm{d}x).$$

Example 1.3 (James-Stein estimator)**.** *Let $X_1,\dots,X_n \sim \mathcal{N}(\mu,\mathbb{1})$ i.i.d. with unknown $\mu \in \mathbb{R}^d$ for $d \geq 3$. Then the estimator*

$$\hat{\mu}_{JS} := \left(1 - \frac{d-2}{n|\bar{X}|^2}\right)\bar{X}, \qquad \bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$$

satisfies

$$\mathbb{E}_\mu\left[|\hat{\mu}_{JS} - \mu|^2\right] = \frac{d}{n} - \mathbb{E}_\mu\left[\frac{(d-2)^2}{n^2|\bar{X}|^2}\right] < \frac{d}{n} = \mathbb{E}_\mu\left[|\bar{X} - \mu|^2\right]$$

for all $\mu \in \mathbb{R}^d$. In particular, $\bar{X}$ is not admissible for the quadratic loss in dimensions $d \geq 3$.

Definition 1.4 (Bayes risk)**.** *Let $\vartheta \in \Theta$ be distributed according to the (prior) distribution π, then the Bayes risk of a decision rule is defined as*

$$R_\pi(\rho) = \mathbb{E}_\pi[R(\vartheta,\rho)] = \int_\Theta \int_{\mathcal{X}} l(\vartheta,\rho(x))\mathbb{P}_\vartheta(\mathrm{d}x)\pi(\mathrm{d}\vartheta).$$

ρ is called Bayes-optimal if $R_\pi(\rho) = \inf_{\rho'} R_\pi(\rho')$ for any decision rule ρ'.

Definition 1.5 (Maximum likelihood estimator)**.** *Let $\mathbb{P}_\vartheta$ have the (likelihood) density $\frac{\mathrm{d}\mathbb{P}_\vartheta}{\mathrm{d}\nu}(x) = f^{X|\vartheta}(x) =: L(\vartheta,x)$ for all $\vartheta \in \Theta$. The maximum likelihood estimator (MLE) is defined as*

$$\hat{\vartheta} := \operatorname*{arg\,sup}_{\vartheta\in\Theta} f^{X|\vartheta}(x).$$

Theorem 1.6 (Bayes theorem)**.** *Let π be the prior with density $f_\Theta(\vartheta)$ and let $\mathbb{P}_\vartheta$ have the (likelihood) density $f^{X|\vartheta}(x)$ for all $\vartheta \in \Theta$. Then the posterior density is defined as*

$$f^{\vartheta|X}(\vartheta) = \frac{f^{X|\vartheta}(x)f_\Theta(\vartheta)}{\int_\Theta f^{X|\vartheta'}(x)f_\Theta(\vartheta')\nu(\mathrm{d}\vartheta')}.$$

Definition 1.7 (Maximum a posteriori estimator)**.** *The maximum a posteriori (MAP) estimator is defined as*

$$\hat{\vartheta} := \operatorname*{arg\,sup}_{\vartheta\in\Theta} f^{\vartheta|X}(\vartheta).$$

Definition 1.8 (Exponential family)**.** *$(\mathbb{P}_\vartheta)_{\vartheta\in\Theta}$ is called exponential family if*

$$\frac{\mathrm{d}\mathbb{P}_\vartheta}{\mathrm{d}\nu}(x) = C(\vartheta)h(x)\exp(\langle\eta(\vartheta),T(x)\rangle)$$

with $x \in \mathcal{X}, \vartheta \in \Theta, \eta : \Theta \to \mathbb{R}^k, C : \Theta \to \mathbb{R}^+, T : \mathcal{X} \to \mathbb{R}^k, h : \mathcal{X} \to \mathbb{R}^k$. T is called natural sufficient statistics and $\eta(\vartheta)$ natural parameter.

2 Asymptotics

2.1 Convergence speed in the law of large numbers

The weak law of large numbers suggests that deviations of the sample mean from the expectation go to zero with a rate $\frac{1}{\sqrt{n}}$. However, for the almost sure convergence, this does not hold exactly and we rather have deviations

$$\left|\bar{X} - \mathbb{E}[X_1]\right| = \mathcal{O}\left(\sqrt{\frac{2\log\log n}{n}}\right).$$

Theorem 2.1 (Law of iterated logarithm)**.** *For every sequence of i.i.d. random variables $X_1, \dots, X_n$ with $\mathbb{E}[X_1] = 0$ and $\mathrm{Var}(X_1) = 1$ we have*

$$\mathbb{P}\left(\limsup_{n\to\infty} \frac{X_1 + \ldots + X_n}{\sqrt{2n\log\log n}} = 1\right) = 1 \qquad \text{or} \qquad \mathbb{P}\left(\liminf_{n\to\infty} \frac{X_1 + \ldots + X_n}{\sqrt{2n\log\log n}} = -1\right) = 1\,.$$

2.2 Chapman-Robbins bound

Theorem 2.2. *Let $\hat{g}$ be an unbiased estimator of the derived parameter $g(\vartheta_0)$. Then for all $\vartheta, \vartheta_0 \in \Theta$ with $\mathbb{P}_\vartheta \ll \mathbb{P}_{\vartheta_0}$ and $\frac{\mathrm{d}\mathbb{P}_\vartheta}{\mathrm{d}\mathbb{P}_{\vartheta_0}} \in L^2(\mathbb{P}_{\vartheta_0})$*

$$\mathrm{Var}_{\vartheta_0}(\hat{g}) = \mathbb{E}_{\vartheta_0}[(\hat{g} - g(\vartheta_0))^2] \geq \frac{(g(\vartheta) - g(\vartheta_0))^2}{\mathrm{Var}_{\vartheta_0}\left(\frac{\mathrm{d}\mathbb{P}_\vartheta}{\mathrm{d}\mathbb{P}_{\vartheta_0}}\right)}.$$

Example 2.3. *We observe $X \sim Exp(\vartheta)$ and want to estimate ϑ. Note that $\frac{\mathrm{d}\mathbb{P}_\vartheta}{\mathrm{d}\mathbb{P}_{\vartheta_0}} = \frac{\vartheta}{\vartheta_0} e^{(\vartheta_0 - \vartheta)x}$ is in $L^2(\mathbb{P}_{\vartheta_0})$ only if $\vartheta > \frac{\vartheta_0}{2}$. In that case $\mathrm{Var}_{\vartheta_0}\left(\frac{\mathrm{d}\mathbb{P}_\vartheta}{\mathrm{d}\mathbb{P}_{\vartheta_0}}\right) = \frac{(\vartheta - \vartheta_0)^2}{\vartheta_0(2\vartheta - \vartheta_0)}$. The Chapman-Robbins bound then gives $\mathbb{E}_{\vartheta_0}[(\hat{\vartheta} - \vartheta_0)^2] \geq \sup_{\vartheta > \frac{\vartheta_0}{2}}(2\vartheta - \vartheta_0) = \infty$. Therefore there is no unbiased estimator of ϑ_0 with finite variance. One can show that it works for $g(\vartheta) = \frac{1}{\vartheta}$, however.*

2.3 Cramér-Rao bound

Theorem 2.4. *Let $\hat{g}$ be an unbiased estimator of $g(\vartheta)$, $L(\vartheta, x)$ the L^2-differentiable likelihood function, $l(\vartheta, x) = \log L(\vartheta, x)$ and $I(\vartheta) = \mathbb{E}_\vartheta[\nabla l(\vartheta) \nabla l(\vartheta)^\top] = -\mathbb{E}_\vartheta\left[\left(\frac{\partial^2}{\partial\vartheta_i \partial\vartheta_j} l(\vartheta)\right)_{ij}\right]$ the Fisher information matrix. Then*

$$\mathrm{Var}_\vartheta(\hat{g}) = \mathbb{E}_\vartheta[(\hat{g} - g(\vartheta))^2] \geq \langle I(\vartheta)^{-1} \nabla g(\vartheta), \nabla g(\vartheta)\rangle.$$

Remark 2.5. *The Cramér-Rao bound can be seen as a special case of the Chapman-Robbins bound. Note that even if it is applicable, this does not necessarily imply that the lower bound is reached by any estimator. However, there are conditions that guarantee the existence of an estimator reaching the bound. When considering the exponential family for instance, $T(x)$ is and unbiased estimator of $\mathbb{E}[T(x)]$ that attains the Cramer-Rao bound if $I(\vartheta)$ is positive definite.*

2.4 Asymptotics of MLE

Theorem 2.6. *Let $\hat{\vartheta}_n$ be the MLE and let certain technical assumptions be valid. Then*

$$\sqrt{n}(\hat{\vartheta}_n - \vartheta) \to \mathcal{N}(0, I(\vartheta)^{-1})$$

in distribution for $n \to \infty$.

3 Statistical tests

3.1 Test theory

We consider the statistical model $(\mathcal{X}, \mathcal{F}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ and split $\Theta = \Theta_1 \dot\cup \Theta_2$. With the test φ we want to find out which Θ_i the true ϑ comes from (more formally, a (randomized) test $\varphi : \mathcal{X} \to [0,1]$ is a measurable function). We usually consider the null hypothesis $H_0 : \vartheta \in \Theta_0$ versus the alternative hypothesis $H_1 : \vartheta \in \Theta_1$.

Definition 3.1. *A (non-randomized) test is a measurable function*

$$\varphi : \mathcal{X} \to \{0,1\}.$$

We consider the convention: $\varphi(x) = 1 \Leftrightarrow H_0$ is rejected, $\varphi(x) = 0 \Leftrightarrow H_0$ is not rejected. This yields a rejection domain $B := \{x : \varphi(x) = 1\}$ and a domain where we keep H_0, namely $A := \{x : \varphi(x) = 0\}$. A test has level α if $\mathbb{E}_\vartheta[\varphi] \le \alpha$ for all $\vartheta \in \Theta_0$.

The following situations are possible:

	H_0 true	H_1 true
choose H_0	w.p. $1-\alpha$ (specificity, true negative rate)	type 2 error, w.p. β
choose H_1	type 1 error, w.p. α	w.p. $1-\beta$ (power, sensitivity, true positive rate)

Remark 3.2. *Usually it is not possible to minimize α and β at the same time. Practically we often set an α a priori and then try to find a minimal β. Note that $1-\beta = \mathbb{E}_\vartheta[\varphi]$ for $\vartheta \in \Theta_1$. A test φ is called "uniformly most powerful" at level α if it has level α and if any other test φ' at level α satisfies $\mathbb{E}_\vartheta[\varphi] \ge \mathbb{E}_\vartheta[\varphi']$ for all $\vartheta \in \Theta_1$. If we have $\mathbb{E}_\vartheta[\varphi] \ge \alpha$ for all $\vartheta \in \Theta_1$ we call it "unbiased".*

Definition 3.3. *A (one-sided) test φ is of Neyman-Pearson (NP) type if for all $\alpha \in (0,1)$ there is a constant C_α s.t.*

$$\varphi(x) = \begin{cases} 1, & T(x) > C_\alpha, \\ \gamma(x), & T(x) = C_\alpha, \\ 0, & T(x) < C_\alpha \end{cases}$$

for some randomization $\gamma(x) \in [0,1]$ and a test statistic T.

Remark 3.4. *The two-sided version of the NP test is defined analogously. One can show that the NP test is uniformly most powerful for H_0 vs. H_1. The randomization γ is only necessary if the distribution is not continuous at the $(1-\alpha)$-quantile, as for instance in discrete distributions.*

3.2 Tests for parameters of the normal distribution

Yet we do not know how to compute the test statistics T or critical values C_α. For this we need to have adequate estimators for the quantities of interest and study their distributions. We for instance want to compare sample means $\bar X_n = \frac{1}{n}\sum_{i=1}^n X_i$ with one another. For this, we have to normalize the means and study how they are distributed.

Theorem 3.5. *If $X_1, \dots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ i.i.d. then $\sum_{i=1}^n X_i^2 \sim \chi_n^2$.*

Proof. Consider $X \sim \mathcal{N}(0,1)$, then $X^2 \sim \Gamma_{\frac12,\frac12}$, since for the cumulative distribution function (CDF) we have

$$F_{X^2}(x) = \mathbb{P}(X^2 \le x) = \mathbb{P}(|X| \le \sqrt{x}) = \Phi(\sqrt{x}) - \Phi(-\sqrt{x}),$$

where $\Phi(x)$ is the CDF of the standard normal distribution. To get the density we differentiate:

$$f_{X^2}(x) = F'_{X^2}(x) = \frac{1}{2\sqrt{x}}\Phi'(\sqrt{x}) + \frac{1}{2\sqrt{x}}\Phi'(-\sqrt{x}) = \frac{1}{\sqrt{x}}\Phi'(\sqrt{x}) = \frac{1}{\sqrt{2\pi x}} e^{-\frac{x}{2}}.$$

This is just the density of the gamma distribution with $\alpha = \beta = \frac12$ (cf. A.1.1).

We have

$$\sum_{i=1}^n X_i^2 \sim \Gamma_{\frac{n}{2},\frac12} = \chi_n^2$$

for the sum of gamma distributed random variables (cf. A.1.1). We say that we have n degrees of freedom. □

Theorem 3.6 (Fisher's F-distribution)**.** *Let $X_1, \dots, X_m, Y_1, \dots, Y_n \sim \mathcal{N}(0,1)$ be i.i.d. Then the quantity*

$$F_{m,n} := \frac{\frac{1}{m}\sum_{i=1}^{m} X_i^2}{\frac{1}{n}\sum_{i=1}^{n} Y_i^2}$$

has the density

$$f_{m,n}(x) = \frac{m^{\frac{m}{2}} n^{\frac{n}{2}} x^{\frac{m}{2}-1}}{B(\frac{m}{2}, \frac{n}{2})(n+mx)^{\frac{m+n}{2}}} \mathbb{1}_{(0,\infty)}(x).$$

Proof. `http://webpages.math.luc.edu/~jdg/w3teaching/stat_305/sp06/pdf/densityF.pdf` □

Theorem 3.7 (Student's t-distribution)**.** *Let $X, Y_1, \dots, Y_n \sim \mathcal{N}(0,1)$ be i.i.d. Then the quantity*

$$T = \frac{X}{\sqrt{\frac{1}{n}\sum_{i=1}^{n} Y_i^2}}$$

has the density

$$f(x;n) = \frac{\left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}}}{B\left(\frac{1}{2}, \frac{n}{2}\right)\sqrt{n}}.$$

Proof. According to theorem 3.6 we have $T^2 \sim F_{1,n}$. Therefore with the same trick as in the proof of theorem 3.5, $|T| = \sqrt{T^2}$ has the density

$$x \mapsto f_{1,n}(x^2)\, 2x$$

for $x > 0$. Since $\mathcal{N}(0,1)$ is symmetric, also T is symmetrically distributed around 0, so T and $-T$ have the same distributions. Therefore T has the density

$$x \mapsto f_{1,n}(x^2)|x| = f(x;n).$$

□

Remark 3.8. *For $n = 1$ we rediscover the Cauchy distribution and for $n \to \infty$ the t-distribution converges to $\mathcal{N}(0,1)$.*

Theorem 3.9. *Assume $X_1, \dots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ i.i.d. Then*

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad \text{and} \qquad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$$

are stochastically independent. We have $\bar{X}_n \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$, $\frac{n-1}{\sigma^2}S^2 \sim \chi^2_{n-1}$ and

$$T_n := \frac{\sqrt{n}(\bar{X}_n - \mu)}{S} \sim t_{n-1}.$$

Proof. We leave the proof of the stochastic independence to the reader. It remains to show that $\frac{n-1}{\sigma^2}S^2 \sim \chi^2_{n-1}$. We can write

$$\frac{n-1}{\sigma^2}S^2 = \sum_{i=1}^{n}\left(\frac{X_i - \mu}{\sigma}\right)^2 - \frac{n}{\sigma^2}(\bar{X}_n - \mu)^2,$$

where the first term is distributed according to χ^2_n and the second term according to χ^2_1. Computing the moment generating functions we get (using independence of $\bar{X}_n$ and S)

$$M_{\frac{n-1}{\sigma^2}S^2}(x) = (1-2x)^{-\frac{n}{2}}(1-2x)^{\frac{1}{2}} = (1-2x)^{-\frac{n-1}{2}},$$

which is the moment generating function of a χ^2_{n-1} distributed random variable. Our result follows by the uniqueness of moment generating functions. The distribution of T_n follows then with theorem 3.7. □

Remark 3.10. *For the two sample t-test with the null hypothesis $H_0 : \mu_X - \mu_Y = \omega_0$ we have the test statistic*

$$T = \frac{\bar{X} - \bar{Y} - \omega_0}{S\sqrt{\frac{1}{n} + \frac{1}{m}}}, \qquad S^2 = \frac{(n-1)S_X^2 + (m-1)S_Y^2}{n+m-2}.$$

Remark 3.11. *If the two samples are assumed to have to same variance, the t-test is "uniformly most powerful unbiased". If they are not the same an "optimal" test is not known (Behrens-Fisher problem).*

Remark 3.12. *Note that in our calculations we assumed normally distributed data X_i and that often in practice this is not guaranteed. However, the central limit theorem implies that for large enough sample size n the sample mean $\bar{X}_n$ is normally distributed. Similar versions of a central limit theorem exist for the sample variance S^2 so that for large enough n we can do the t-test even without normally distributed data [7].*

3.3 Confidence intervals

Definition 3.13. *We consider a statistical model $(\mathcal{X}, \mathcal{F}, \mathbb{P} = \{\mathbb{P}_\vartheta : \vartheta \in \Theta\})$. Then $\mathcal{C} = (C(x) : x \in \mathcal{X})$ with $C(x) \subset \Theta$ for all $x \in \mathcal{X}$ is a family of confidence intervals at level $1-\alpha$ for all $\vartheta \in \Theta$ if*

$$\mathbb{P}_\vartheta(\{x : \vartheta \in C(x)\}) \geq 1 - \alpha.$$

Remark 3.14. *The confidence interval is random. There is a one-to-one correspondence between confidence intervals and the rejection domains B from section 3.1. An estimator $\hat{\vartheta} \in B$ if and only if $\vartheta \notin C$.*

Example 3.15. *If we assume $X_i \sim \mathcal{N}(\mu, \sigma^2)$ i.i.d. with an unknown σ^2 then our confidence interval is given by $C(\omega) = \left[\bar{X} - \frac{\hat{\sigma}}{\sqrt{n}} t_{n-1,1-\frac{\alpha}{2}}, \bar{X} + \frac{\hat{\sigma}}{\sqrt{n}} t_{n-1,1-\frac{\alpha}{2}}\right]$.*

Definition 3.16 (p-value)**.** *We consider a statistical model $(\mathcal{X}, \mathcal{F}, \mathbb{P} = \{\mathbb{P}_\vartheta : \vartheta \in \Theta\})$ and a test φ for the hypothesis pair $H_0 \subset \Theta, H_1 = \Theta \setminus H_0$ based on a test statistics $T : \mathcal{X} \to \mathbb{R}$. Let B_α be the rejection domain for level $\alpha \in (0,1)$ s.t. $\varphi(x) = 1 \Leftrightarrow T(x) \in B_\alpha$ for $x \in \mathcal{X}$. Then the p-value is defined as*

$$p_\varphi(x) = \inf_{\alpha : T(x) \in B_\alpha} \mathbb{P}^*(T(X) \in B_\alpha),$$

where

$$\mathbb{P}^*(T(X) \in B_\alpha) = \sup_{\vartheta \in H_0} \mathbb{P}_\vartheta(T(X) \in B_\alpha)$$

if H_0 is not connected.

Remark 3.17. *If the hypothesis H_0 has one element we usually have $p_\varphi(x) = \inf\{\alpha : T(x) \in B_\alpha\}$. So the p-value is the probability that a random experiment according to the H_0 is at least as "extrem" as the observed one.*

Theorem 3.18. *If $X_i \sim \mathcal{N}(\mu, \sigma^2)$ with μ and σ^2 unknown, then under H_0 the p-value is uniformly distributed on $[0,1]$.*

Proof. Let us consider $H_0 : \vartheta = \vartheta_0$ and T a test statistics distributed according to the t-distribution – we call its cumulative distribution F_t. Let us investigate the cumulative distribution of the p-value, i.e.

$$\begin{aligned}
F(p \leq x) &= F\left(1 - \frac{p}{2} \geq 1 - \frac{x}{2}\right) \\
&= F\left(F_t(|T|) \geq 1 - \frac{x}{2}\right) \\
&= F\left(|T| \geq F_t^{-1}\left(1 - \frac{x}{2}\right)\right) \\
&= 2F\left(T \geq F_t^{-1}\left(1 - \frac{x}{2}\right)\right) \\
&= 2\left(1 - F_t\left(F_t^{-1}\left(1 - \frac{x}{2}\right)\right)\right) \\
&= x.
\end{aligned}$$

□

4 Generalized linear models

Definition 4.1 (Linear model)**.** *We consider observations $Y = (Y_1, \dots, Y_n)^\top \in \mathbb{R}^n$, a parameter vector $\beta \in \mathbb{R}^p, p \leq n$, a (design) matrix $X \in \mathbb{R}^{n \times p}$ with full rank p and a random vector $\varepsilon \in \mathbb{R}^n$ with $\mathbb{E}[\varepsilon_i] = 0$ and $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = \Sigma_{ij}$. Our model is*

$$Y = X\beta + \varepsilon.$$

and want to compute the least squares estimator that minimizes the weighted Euclidean distance

$$\hat{\beta} = \operatorname*{arg\,inf}_{b \in \mathbb{R}^p} |\Sigma^{-\frac{1}{2}}(Xb - Y)|^2.$$

We call $\Sigma = \sigma^2 \mathbb{1}$ the ordinary case.

Theorem 4.2 (Gauß-Markov)**.** *Set $X_\Sigma = \Sigma^{-\frac{1}{2}} X$ and Π_{X_Σ} shall be the orthogonal projection of $\mathbb{R}^n$ onto $\mathrm{ran}(X_\Sigma)$. Then we have $\Pi_{X_\Sigma} = X_\Sigma (X_\Sigma^\top X_\Sigma)^{-1} X_\Sigma^\top$ and the least squares estimator is*

$$\hat{\beta} = (X^\top \Sigma^{-1} X)^{-1} X^\top \Sigma^{-1} Y.$$

In the ordinary linear model we get

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

Proof. First note that $X_\Sigma^\top X_\Sigma = X^\top \Sigma^{-1} X$ is invertible since Σ is invertible. Since X has full rank we have for an arbitrary vector $v \in \mathbb{R}^p$

$$X^\top \Sigma^{-1} X v = 0 \Rightarrow v^\top X^\top \Sigma^{-1} X v = 0 \Rightarrow |\Sigma^{-\frac{1}{2}} X v| = 0 \Rightarrow |Xv| = 0 \Rightarrow v = 0.$$

We take Π_{X_Σ} as defined and consider $w = \Pi_{X_\Sigma} v$ for any $v \in \mathbb{R}^n$. Then $w \in \mathrm{ran}(X_\Sigma)$ and if $v = X_\Sigma u$ we have $w = \Pi_{X_\Sigma} X_\Sigma u = v$, so Π_{X_Σ} is indeed a projection. Π_{X_Σ} is self-adjoint since it is symmetric and therefore an orthogonal projection:

$$\forall u \in \mathbb{R}^n : \langle u - \Pi_{X_\Sigma} u, w \rangle = \langle u, w \rangle - \langle u, \Pi_{X_\Sigma} w \rangle = 0.$$

From the definition of $\hat{\beta}$ we know that it yields the best approximation of $\Sigma^{-\frac{1}{2}} Y$ in $\mathbb{R}^n$ via $X_\Sigma b$. It is defined by the orthogonal projection

$$\Pi_{X_\Sigma} \Sigma^{-\frac{1}{2}} Y = X_\Sigma \hat{\beta}.$$

Solving for $\hat{\beta}$ yields our estimator. □

Remark 4.3. *Alternatively we can get the result by direct minimization in β. We want to minimize*

$$|Y - X\beta|^2 = (Y - X\beta)^\top (Y - X\beta).$$

Differentiation w.r.t. β yields

$$-X^\top Y + X^\top X \beta = 0$$

and therefore the desired estimator.

Remark 4.4. *For the prediction error we get*

$$\mathbb{E}[|X\hat{\beta} - X\beta|^2] = \mathbb{E}[|\Pi_X \varepsilon|^2] = \sigma^2 p,$$

i.e. it grows with the number of parameters.

4.1 Linear regression

We observe realizations of the model

$$Y_i = aX_i + b + \varepsilon_i, \quad i = 1, \dots, n,$$

where we assume $\mathbb{E}[\varepsilon_i] = 0, \mathrm{Var}(\varepsilon_i) = \sigma^2$ for the residuals and $a, b, \sigma \in \mathbb{R}$ are unknown parameters. We want to find a linear function $y = ax + b$ that explains the data well and consider the least squares attempt:

$$(\hat{a}, \hat{b}) := \operatorname*{arg\,min}_{(a,b) \in \mathbb{R}^2} \sum_{i=1}^n (Y_i - aX_i - b)^2.$$

Compared to the definition 4.1 of the linear model we have $p = 2, \beta = (b, a)^\top, X = (X_{ij})$ with $X_{i,1} = 1, X_{i,2} = X_i$.

4.2 Multiple linear regression

We consider $X_i = (X_{i,1}, \dots, X_{i,d})^\top$ and observe

$$Y_i = a_0 + \langle a, X_i \rangle + \varepsilon_i,\ a = (a_1, \dots, a_d)^\top.$$

So we have $p = d + 1, \beta = (a_0, a_1, \dots, a_d)^\top$ and

$$X = \begin{pmatrix} 1 & X_{1,1} & \cdots & X_{1,d} \\ \vdots & \vdots & & \vdots \\ 1 & X_{n,1} & \cdots & X_{n,d} \end{pmatrix}.$$

4.3 Polynomial regression

We want to fit the data $Y_1, \dots, Y_n \in \mathbb{R}$ with a polynomial of order $m = p - 1$ and choose $\beta = (a_0, a_1, \dots a_{p-1})^\top$ and the Vandermonde matrix

$$X = \begin{pmatrix} 1 & X_1 & X_1^2 & \cdots & X_1^m \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & X_n & X_n^2 & \cdots & X_n^m \end{pmatrix}$$

$X_i \in \mathbb{R}$.

4.4 Logistic regression

Let us first generalize our definition of a linear model.

Definition 4.5 (Generalized linear model)**.** *In a generalized linear model (GLM) the linearity is related to the response variables $Y_i \in \mathbb{R}$ via a so called link function $g : \mathbb{R} \to \mathbb{R}$. We still consider variables $X_i \in \mathbb{R}^d$ and further assume that Y_i is in the exponential family. For the mean of Y_i we assume that*

$$\mu = \mathbb{E}[Y_i] = g^{-1}(\langle X_i, \beta \rangle),$$

where $\beta \in \mathbb{R}^p$ are the $p = d + 1$ unknown parameters, which are typically estimated with the maximum likelihood method.

Now we can look at the logistic regression, which is particularly useful for binary responses $Y_i \sim \mathrm{Ber}(p_i)$, as its outputs lie in the interval $[0, 1]$. Here, the link function is called logit function, defined by

$$g(\mu) = \log\left(\frac{\mu}{1-\mu}\right),\ \text{which brings } \mu = \frac{1}{1 + e^{-\langle X, \beta \rangle}}.$$

Note that the Bernoulli distribution is in the exponential family with sufficient statistics $T(p) = \log\left(\frac{p}{1-p}\right)$, since its likelihood for $p \in (0, 1)$ and $Y \in \{0, 1\}$ is

$$L(p, Y) = p^Y (1-p)^{1-Y} = (1-p)\left(\frac{p}{1-p}\right)^Y = (1-p)\exp(YT(p)).$$

We model the Bernoulli probabilities to be dependent on the data and obverse that $p_i := p(X_i) = \mathbb{E}_\beta[Y_i | X_i] = \frac{1}{1+e^{-\langle X_i, \beta \rangle}}$, again with the notation $X_i = (1, X_{i,1}, \dots, X_{i,d})^\top$. We want to identify a β s.t. our probabilities p_i are close to the observed data $X_1, \dots, X_n$ and therefore consider the maximum likelihood estimator $\hat{\beta} = \arg\max_{\beta \in \mathbb{R}^p} L(X, Y | \beta)$. We can equivalently minimize the negative loglikelihood $\mathcal{L}(p) = -\sum_{i=1}^n (Y_i \log(p_i) + (1 -$

$Y_i)\log(1-p_i))$, which implies $\nabla_\beta \mathcal{L} = 0$. Therefore we compute

$$\begin{aligned}
\frac{\partial}{\partial \beta_j}\mathcal{L} &= -\sum_{i=1}^{n}\left(Y_i \frac{\partial}{\partial \beta_j}\left(-\log(1+e^{-\beta^\top X_i})\right) + (1-Y_i)\frac{\partial}{\partial \beta_j}\left(-\beta^\top X_i - \log(1+e^{-\beta^\top X_i})\right)\right) \\
&= -\sum_{i=1}^{n}\left(Y_i \frac{e^{-\beta^\top X_i}}{1+e^{-\beta^\top X_i}} X_{ij} + (1-Y_i)\left(-X_{ij} + \frac{e^{-\beta^\top X_i}}{1+e^{-\beta^\top X_i}} X_{ij}\right)\right) \\
&= -\sum_{i=1}^{n}\left(Y_i(1-p)X_{ij} + (1-Y_i)(-pX_{ij})\right) \\
&= \sum_{i=1}^{n} X_{ij}(p_i - Y_i),
\end{aligned}$$

i.e.

$$\nabla_\beta \mathcal{L} = X^\top (p - Y) = 0.$$

One can show that the Hessian matrix $\left(\frac{\partial^2 \mathcal{L}}{\partial \beta_i \partial \beta_j}\right)$ is positive definite so that we indeed have a local minimum. Since this system of equations is usually not solvable in a closed form, we need iterative solvers to approach our minimization. The easiest attempt is a gradient descent for each data point X_i, i.e.

$$\beta^{(i+1)} := \beta^{(i)} - \eta_i X_i (p_i - Y_i)$$

with a suitable sequence of stepsizes $(\eta_i)_{i\in\mathbb{N}}$.

4.5 Analysis of variance (ANOVA)

The analysis of variance can be seen as a generalization of the two sample test. Assume we have data with a categorical factor. The model is

$$Y_{ij} = \mu_i + \varepsilon_{ij}$$

for factor $i = 1, \dots, p$ and samples $j = 1, \dots, n_i$ with $n = n_1 + \cdots + n_p$. We can therefore relate to definition 4.1 with

$$X = \begin{pmatrix} 1 & 0 & \cdots & \\ \vdots & \vdots & & \\ 1 & 0 & \cdots & \\ 0 & 1 & 0 & \cdots \\ \vdots & \vdots & \vdots & \\ 0 & 1 & 0 & \cdots \\ \vdots & \vdots & \vdots & \end{pmatrix} \in \{0,1\}^{n\times p}, \qquad \beta = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_p \end{pmatrix} \qquad \text{and} \qquad \varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2).$$

We want to test the hypothesis $H_0 : \mu_1 = \cdots = \mu_p$. The idea is to consider the one-dimensional subspace

$$U := \{x \in \mathbb{R}^p, x_i - x_j = 0, i, j = 1, \dots, p\}.$$

The estimator of each group is $\hat{\mu}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij}$ and the principle idea is to split the variance of the data in the variance coming from the linear model and the residual variance, namely

$$\|Y - \bar{Y}\|^2 = \|Y - \hat{Y}\|^2 + \|\hat{Y} - \bar{Y}\|^2$$

and we get for the one-way ANOVA

$$(n-1)V_{\mathrm{t}} = (n-p)V^* + (p-1)V_{\mathrm{b}}$$

with

$$V^* = \frac{1}{n-p}\sum_{i=1}^{p}\sum_{j=1}^{n_i}(Y_{ij} - \hat{\mu}_i)^2, \qquad V_{\mathrm{b}} = \frac{1}{p-1}\sum_{i=1}^{p} n_i(M - \hat{\mu}_i), \qquad M = \frac{1}{n}\sum_{i,j} Y_{ij}.$$

With those quantities we measure how well our model explains the data, namely we look at the ratio of the variance explained by the model and the total variance, i.e.

$$R^2 = \frac{\|\hat{Y} - \bar{Y}\|^2}{\|Y - \bar{Y}\|^2}.$$

If the variance between the groups is bigger than the variance in the groups, this suggests to reject the H_0. We have $\frac{n-p}{\sigma^2}V^* \sim \chi^2_{n-p}$ and $\frac{p-1}{\sigma^2}V_\mathrm{b} \sim \chi^2_{p-1}$ as well as our test statistics

$$F := \frac{V_b}{V^*} \sim F_{p-1,n-p}.$$

4.6 Kernel ridge regression

The idea of kernel ridge regression is to transfer the data $X_1, \dots, X_n$ to a higher dimensional space via $\Phi : \mathbb{R}^p \to \mathbb{R}^{\tilde{p}}$ and make computations there. It will turn out that those calculations do not have to be done explicitly, but can be done via a so called kernel.

In regression we want to predict $Y_i \in \mathbb{R}$ from $X_i \in \mathbb{R}^p$ via a function $f \in \mathcal{H}$ s.t. (quadratic) loss $l(Y_i, f(X_i)) = (Y_i - f(X_i))^2$ gets minimized. In order to not get all to complicated functions f (e.g. to avoid overfitting) we introduce a Tikhonov regularization with a parameter $\lambda \in \mathbb{R}$ and consider the so called ridge regression problem

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

For simplicity let us first consider the linear case again, i.e. $f(X_i) = \beta^\top X_i$ and formulate the optimization as a minimization with constraint:

$$\min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{2} \sum_{i=1}^{n} Z_i^2 + \frac{\lambda}{2} \|\beta\|^2 \right\}, \qquad Z_i = Y_i - \beta^\top X_i$$

With the Lagrange multipliers $\alpha \in \mathbb{R}^n$ we can formulate this as

$$\max_{\alpha} \min_{\beta, Z} L(\alpha, \beta, Z)$$

with

$$L(\alpha, \beta, Z) = \frac{1}{2} \sum_{i=1}^{n} Z_i^2 + \frac{\lambda}{2} \|\beta\|^2 + \sum_{i=1}^{n} \alpha_i (\beta^\top X_i - Y_i - Z_i).$$

We compute the minima in β and Z with

$$0 = \frac{\partial L}{\partial Z_i} = Z_i - \alpha_i$$
$$0 = \nabla_\beta L = \lambda \beta + \sum_{i=1}^{n} \alpha_i X_i$$

and get the dual problem

$$L(\alpha, \beta^*, Z^*) = -\frac{1}{2} \alpha^\top \alpha - \frac{1}{2\lambda} \alpha^\top K \alpha - \alpha^\top Y,$$

with $K_{ij} = X_i^\top X_j$, i.e. $K = XX^\top$. For this we find the optimal α via

$$0 = \nabla_\alpha L(\alpha, \beta^*, Z^*) = -\alpha - \frac{1}{\lambda} K \alpha - Y,$$

which yields

$$\alpha^* = -\lambda (K + \lambda \mathbb{1})^{-1} Y$$
$$\beta^* = -\frac{1}{\lambda} \sum_{i=1}^{n} \alpha_i^* X_i = -\frac{1}{\lambda} X^\top \alpha^* = X^\top (K + \lambda \mathbb{1})^{-1} Y.$$

So we get as a prediction for a new point X_k

$$f(X_k) = X_k^\top \beta^* = X_k^\top X^\top (K + \lambda \mathbb{1})^{-1} Y.$$

Note that for $\lambda = 0$ we get $\beta^* = X^\top (XX^\top)^{-1} Y$, which is the same as we have computed before, as can be seen by multiplying $(X^\top X)X^\top = X^\top (XX^\top)$ by $(X^\top X)^{-1}$ from the left and by $(XX^\top)^{-1} Y$ from the right.

The idea of kernelization now is to replace $K_{ij} = X_i^\top X_j$ with the kernel $K_{ij} = k(x_i, x_j)$ and in $X_k^\top X^\top = (X_k^\top X_1, \ldots, X_k^\top X_n)$ replace each scalar product $X_k^\top X_i$ with $k(X_k, X_i)$. Then implicitly the minimization is solved in feature space.

5 Dimensionality reduction

5.1 Principal component analysis (PCA)

PCA tries to transfer possible correlated variables to uncorrelated variables through orthogonal projections. The first component shall have maximal variance and any succeeding component shall have maximal variance constrained to be orthogonal to all preceding variables, i.e. PCA can be thought of as revealing the internal structure of the data in a way that best explains the variance in the data.

Consider data $X_1, \dots, X_n \in \mathbb{R}^d$ w.r.t. to the canonical basis $e_1, \dots, e_d$. We define the mean and the covariance matrix

$$\bar{X} = \frac{1}{n}\sum_{k=1}^{n} X_k, \qquad C_{ij} = \frac{1}{n}\sum_{k=1}^{n}(X_k^i - \bar{X}^i)(X_k^j - \bar{X}^j).$$

and note that we can center the data to get $\bar{X} = 0$. Then the covariance matrix reads $C = \frac{1}{n}\sum_{k=1}^{n} X_k X_k^\top = \frac{1}{n} X^\top X$. The variance along an arbitrary direction $\tilde{e}$ is $\tilde{\sigma}^2 = \tilde{e}^\top C \tilde{e}$. We therefore want to solve the maximization

$$\max_{\tilde{e}} \tilde{e}^\top C \tilde{e}, \qquad \text{s.t. } |\tilde{e}| = 1,$$

which we can do with Lagrange multipliers, namely

$$\max_{\tilde{e}} f(\tilde{e})$$

with

$$f(\tilde{e}) = \tilde{e}^\top C \tilde{e} - \lambda(\tilde{e}^2 - 1) = \sum_{i,j=1}^{n} \tilde{e}_i C_{ij} \tilde{e}_j - \lambda\left(\sum_{i=1}^{n} \tilde{e}_i^2 - 1\right).$$

Therefore we set the first derivative to zero and get

$$\frac{\partial f}{\partial \tilde{e}_k} = \sum_{j=1}^{n} C_{kj}\tilde{e}_j + \sum_{i=1}^{n} C_{ik}\tilde{e}_i - 2\lambda\tilde{e}_k = 2\sum_{j=1}^{n} C_{kj}\tilde{e}_j - 2\lambda\tilde{e}_k = 0$$

for all components k, where we used the symmetry of C. This is just the eigenvalue equation

$$C\tilde{e} = \lambda\tilde{e},$$

i.e. the eigenvectors give us the directions of maximal variance and the eigenvalues are the corresponding variances (this can also be shown with the Rayleigh quotient; compare also to singular value decomposition).

Remark 5.1. *PCA is sensitive to the relative scaling of the original variables, the maximal variance criterion only makes sense if scales are comparable.*

5.2 Kernel PCA

The idea is to transform the data points $X_1, \dots, X_n \in \mathbb{R}^d$ to a higher dimensional space, since they can almost always be separated in $\tilde{d} > n$ dimensions. Therefore we define a mapping

$$\Phi : \mathbb{R}^d \to \mathbb{R}^{\tilde{d}}.$$

Then, Φ creates linearly independent vectors, so there is no covariance on which to perform eigendecomposition explicitly as we did in linear PCA. However, we never make computations explicitly in the higher dimensional space anyway, but rather consider the kernel

$$k(X, X') = \langle \Phi(X), \Phi(X') \rangle,$$

which represents the inner product space of the otherwise intractable feature space. So we actually do not compute the principal components themselves, but the projections of our data onto those components. In analogy to the situation in the linear case we have the covariance matrix

$$C = \frac{1}{n}\sum_{k=1}^{n} \Phi(X_k)\Phi(X_k)^\top$$

and we want to solve the eigenvalue equation

$$Cv = \lambda v.$$

Since all v lie in the span of $\Phi(X_1), \ldots, \Phi(X_n)$ we have coefficients $\alpha_1, \ldots, \alpha_n$ s.t. $v = \sum_{k=1}^{n} \alpha_k \Phi(X_k)$ and consider equivalently

$$\Phi(X_k) \cdot Cv = \lambda \Phi(X_k) \cdot v$$

for all $k = 1, \ldots, n$. This then leads to

$$n\lambda K\alpha = K^2\alpha \qquad \text{and therefore} \qquad n\lambda\alpha = K\alpha$$

with $K_{ij} = \langle \Phi(X_i), \Phi(X_j) \rangle$ and $\alpha = (\alpha_1, \ldots, \alpha_n)^\top$ (more precisely, the variables should have an additional index k). We then get for the projection of a test point $\Phi(X)$ onto an eigenvalue v

$$v \cdot \Phi(X) = \sum_{i=1}^{n} \alpha_i \Phi(X_i) \cdot \Phi(X)$$

and for $X = X_j$

$$\begin{aligned} v \cdot \Phi(X_j) &= \sum_{i=1}^{n} \alpha_i \Phi(X_i) \cdot \Phi(X_j) \\ &= \sum_{i=1}^{n} \alpha_i K_{ij} \\ &= n\lambda\alpha_j. \end{aligned}$$

5.3 K-means

K-means aims to partition observations $X_1, \ldots, X_n \in \mathbb{R}^d$ into K sets $S = \{S_1, \ldots, S_K\}$, yielding in a partitioning of data space into so called Voronoi cells. The objective is to minimize the variances, i.e. to find

$$S^* = \arg\min_{S} \sum_{i=1}^{K} \sum_{X \in S_i} d(X, \mu_i),$$

with $\mu_i = \frac{1}{|S_i|} \sum_{X \in S_i} X$.

The standard algorithm goes as follows. Given an initial set of K means $m_1^{(1)}, \ldots, m_K^{(1)}$, the algorithm proceeds by alternating between two steps:

- Assign each data point to its nearest mean s.t. $S_i^{(m)} = \{X : d(X, m_i^{(m)}) \leq d(X, m_j^{(m)}), 1 \leq j \leq K\}$.
- Calculate the new means: $m_i^{(m+1)} = \frac{1}{|S_i^{(m)}|} \sum_{X \in S_i^{(m)}} X$.

Usually one chooses the Euclidean metric $d(x, y) = \|x - y\|^2$. There are various initialization strategies. In order to find a good global optimum one typically runs the algorithm several times with different randomly drawn initializations. Note that finding the optimum is not guaranteed by this algorithm.

5.4 t-SNE

In t-distributed stochastic neighbor embedding (t-SNE) the idea is to find a lower dimensional representation $\tilde{X}_1, \ldots, \tilde{X}_n \in \mathbb{R}^{\tilde{d}}$ of the original data $X_1, \ldots, X_n \in \mathbb{R}^d$, with $\tilde{d} \ll d$, s.t. "closeness" between points is similar in both spaces [5]. In the original space closeness of X_i and X_j is given by the probability distribution

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2n}, \qquad p_{i|j} = \frac{\exp\left(-\frac{\|X_i - X_j\|^2}{2\sigma^2}\right)}{\sum_{k \neq i} \exp\left(-\frac{\|X_i - X_k\|^2}{2\sigma^2}\right)}.$$

It can be interpreted as the conditional probability that X_i chooses X_j as its neighbor. In the smaller space $\mathbb{R}^{\tilde{d}}$ these probabilities are given by

$$q_{ij} = \frac{(1 + \|\tilde{X}_i - \tilde{X}_j\|^2)^{-1}}{\sum_{k \neq m}(1 + \|\tilde{X}_k - \tilde{X}_m\|^2)^{-1}}.$$

Note that this is just a t-distribution with one degree of freedom (which is equivalent to a Cauchy distribution), i.e. it has heavy tails, which allows for modelling dissimilar points as being far away from one another. The idea now is to learn a good representation $\tilde{X}_1, \ldots, \tilde{X}_n \in \mathbb{R}^{\tilde{d}}$ by minimizing the Kullback-Leibler divergence between the two distributions, namely

$$\mathrm{KL}(\mathbb{P}\|\mathbb{Q}) = \sum_{i \neq j} p_{ij} \log \frac{p_{ij}}{q_{ij}}.$$

Practically, this minimization is done with gradient descent and one computes

$$\frac{\partial \,\mathrm{KL}(\mathbb{P}\|\mathbb{Q})}{\partial \tilde{X}_i} = 4 \sum_j (p_{ij} - q_{ij})(\tilde{X}_i - \tilde{X}_j)(1 + \|\tilde{X}_i - \tilde{X}_j\|^2)^{-1}.$$

A Appendix

A.1 Distributions

A.1.1 Gamma distribution

$X \sim \Gamma(\alpha, \beta)$ with shape parameter α and an inverse scale parameter β. The density is

$$p(x; \alpha, \beta) = \frac{\beta^\alpha x^{\alpha-1} e^{-\beta x}}{\Gamma(\alpha)} \qquad \text{for } x > 0 \text{ and } \alpha, \beta > 0$$

with

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} \mathrm{d}x$$

being the gamma function, the generalization of the factorial, i.e. for any positive integer n we have $\Gamma(n) = (n-1)!$.

Note that $X_1 \sim \Gamma(\alpha_1, \beta)$, $X_2 \sim \Gamma(\alpha_2, \beta)$ implies that $X_1 + X_2 \sim \Gamma(\alpha_1 + \alpha_2, \beta)$ since

$$\begin{aligned}
f_{X_1+X_2}(z) &= \int_0^z f_{X_1}(x) f_{X_2}(z-x) \mathrm{d}x \\
&= \int_0^z \frac{\beta^{\alpha_1} x^{\alpha_1-1} e^{-\beta x} \beta^{\alpha_2} (z-x)^{\alpha_2-1} e^{-\beta(z-x)}}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \\
&= e^{-\beta z} \beta^{\alpha_1+\alpha_2} \int_0^z \frac{x^{\alpha_1-1}(z-x)^{\alpha_2-1}}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \\
&= e^{-\beta z} \beta^{\alpha_1+\alpha_2} z^{\alpha_1+\alpha_2-1} \int_0^1 \frac{t^{\alpha_1-1}(1-t)^{\alpha_2-1}}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \mathrm{d}t \\
&= \frac{e^{-\beta z} \beta^{\alpha_1+\alpha_2} z^{\alpha_1+\alpha_2-1}}{\Gamma(\alpha_1+\alpha_2)},
\end{aligned}$$

where in the last step we used the beta distribution.

A.1.2 Beta distribution

The density of the beta distribution is given by

$$f(x; \alpha, \beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha, \beta)}$$

with shape parameters α and β and beta function $B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$.

B Bibliography

[1] Dickhaus, T. (2012). Methoden der Statistik, lecture notes, Humboldt Universität Berlin.

[2] Hartmann, C. (2016). Wahrscheinlichkeitstheorie, lecture notes, BTU Cottbus-Senftenberg.

[3] Lehmann, E. L. and Casella, G. (2006). *Theory of point estimation*. Springer Science & Business Media.

[4] Lehmann, E. L. and Romano, J. P. (2006). *Testing statistical hypotheses*. Springer Science & Business Media.

[5] Maaten, L. v. d. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(Nov):2579–2605.

[6] Mathé, P. (2013). Computerorientierte Statistik, lecture notes, Freie Universität Berlin.

[7] Omey, E. and Van Gulck, S. (2008). Central limit theorems for variances and correlation coefficients.

[8] Reiß, M. (2015). Mathematische Statistik, lecture notes, Humboldt Universität Berlin.

[9] Shao, J. (2003). *Mathematical Statistics*. Springer Texts in Statistics. Springer.

[10] Van der Vaart, A. W. (1998). *Asymptotic statistics*, volume 3. Cambridge university press.