

Springer Texts in Statistics

Series Editors:

G. Casella

R. DeVeaux

S.E. Fienberg

I. Olkin

For further volumes:

<http://www.springer.com/series/417>

Wolfgang Karl Härdle • Vladimir Spokoiny
Vladimir Panov • Weining Wang

Basics of Modern Mathematical Statistics

Exercises and Solutions

Wolfgang Karl Härdle
Weining Wang
L.v.Bortkiewicz Chair of Statistics, C.A.S.E.
Centre f. Appl. Stat. and Econ.
Humboldt-Universität zu Berlin
Berlin
Germany

Vladimir Spokoiny
Weierstrass Institute for Applied Analysis
and Stochastics (WIAS)
Berlin
Germany

Vladimir Panov
Universität Duisburg-Essen
Essen
Germany

The quantlets of this book may be downloaded from <http://extras.springer.com> directly or via a link on <http://springer.com/978-3-662-52386-5> and from the www.quantlet.de

ISSN 1431-875X

ISBN 978-3-662-52386-5

ISBN 978-3-642-36850-9 (eBook)

DOI 10.1007/978-3-642-36850-9

Springer Heidelberg New York Dordrecht London

Mathematics Subject Classification (2010): 62F10, 62F03, 62J05, 62P20

© Springer-Verlag Berlin Heidelberg 2014

Softcover reprint of the hardcover 1st edition 2014

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed. Exempted from this legal reservation are brief excerpts in connection with reviews or scholarly analysis or material supplied specifically for the purpose of being entered and executed on a computer system, for exclusive use by the purchaser of the work. Duplication of this publication or parts thereof is permitted only under the provisions of the Copyright Law of the Publisher's location, in its current version, and permission for use must always be obtained from Springer. Permissions for use may be obtained through RightsLink at the Copyright Clearance Center. Violations are liable to prosecution under the respective Copyright Law.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

While the advice and information in this book are believed to be true and accurate at the date of publication, neither the authors nor the editors nor the publisher can accept any legal responsibility for any errors or omissions that may be made. The publisher makes no warranty, express or implied, with respect to the material contained herein.

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

“Wir behalten von unseren Studien am Ende doch nur das, was wir praktisch anwenden.”

“In the end, we really only retain from our studies that which we apply in a practical way.”

J. W. Goethe, Gespräche mit Eckermann, 24. Feb. 1824.

The complexity of statistical data nowadays requires modern and numerically efficient mathematical methodologies that can cope with the vast availability of quantitative data. Risk analysis, calibration of financial models, medical statistics and biology make extensive use of mathematical and statistical modeling.

Practice makes perfect. The best method of mastering models is working with them. In this book we present a collection of exercises and solutions which can be helpful in the advanced comprehension of *Mathematical Statistics*. Our exercises are correlated to Spokoiny and Dickhaus (2014). The exercises illustrate the theory by discussing practical examples in detail. We provide computational solutions for the majority of the problems. All numerical solutions are calculated with R and Matlab. The corresponding quantlets – a name we give to these program codes – are indicated by  in the text of this book. They follow the name scheme MSExyz123 and can be downloaded from the Springer homepage of this book or from the authors' homepages.

Mathematical Statistics is a global science. We have therefore added, below each chapter title, the corresponding translation in one of the world languages. We also head each section with a proverb in one of those world languages. We start with a German proverb from Goethe (see above) on the importance of practice.

We have tried to achieve a good balance between theoretical illustration and practical challenges. We have also kept the presentation relatively smooth and, for more detailed discussion, refer to more advanced text books that are cited in the reference sections.

The book is divided into three main parts where we discuss the issues relating to option pricing, time series analysis and advanced quantitative statistical techniques.

The main motivation for writing this book came from our students of the course *Mathematical Statistics* which we teach at the Humboldt-Universität zu Berlin. The students expressed a strong demand for solving additional problems and assured us that (in line with Goethe) giving plenty of examples improves learning speed and quality. We are grateful for their highly motivating comments, commitment and positive feedback. Very special thanks go to our students Shih-Kang Chao, Ye Hua, Yuan Liao, Maria Osipenko, Ceren Önder and Dedy Dwi Prastyo for advise and ideas on solutions. We thank Niels Thomas from Springer Verlag for continuous support and for valuable suggestions on writing style and the content covered.

Berlin, Germany
Essen, Germany
Berlin, Germany
January 2013

Wolfgang Karl Härdle
Vladimir Panov
Vladimir Spokoiny
Weining Wang

Contents

1	Basics	1
2	Parameter Estimation for an i.i.d. Model	9
3	Parameter Estimation for a Regression Model	53
4	Estimation in Linear Models	73
5	Bayes Estimation	107
6	Testing a Statistical Hypothesis.....	129
7	Testing in Linear Models	159
8	Some Other Testing Methods	167
	Index	183

Language List

Arabic	اللغة العربية
Chinese	中文
Croatian	Hrvatski jezik
Czech	Čeština
Dutch	Nederlands
English	English
French	Français
German(Colognian)	Deutsch (Kölsch)
Greek	ελληνική γλώσσα
Hebrew	עברית
Indonesian	Indonesia
Italian	Italiano
Japanese	日本語
Korean	한국말
Latin	Lingua Latina

Polish	język polski
Romanian	român
Russian	русский язык
Slovak	Slovenské
Spanish	español
Turkish	Türkçe
Ukrainian	українська
Vietnamese	tiếng Việt

Symbols and Notation

Basics

X, Y	random variables or vectors
$X_1, X_2, \dots, X_p$	random variables
$\mathbf{X} = (X_1, \dots, X_p)^\top$	random vector
$X \sim F$	X has distribution F
Γ, Δ	matrices
Σ	covariance matrix
$\mathbf{1}_n$	vector of ones $\underbrace{(1, \dots, 1)}_{n\text{-times}}^\top$
$\mathbf{0}_n$	vector of zeros $\underbrace{(0, \dots, 0)}_{n\text{-times}}^\top$
I_p	identity matrix
$\mathbf{1}(\cdot)$	indicator function, for a set M is $\mathbf{1} = 1$ on M , $\mathbf{1} = 0$ otherwise
$\mathbf{i}$	$\sqrt{-1}$
$\approx$	approximately equal
$\otimes$	Kronecker product
<i>iff</i>	if and only if, equivalence
W_t	standard Wiener process
$\mathbb{C}$	complex number set
$\mathbb{R}$	real number set
$\mathbb{N}$	positive integer set
$\mathbb{Z}$	integer set
$(X)^+$	$ X * \mathbf{1}(X > 0)$
$[\lambda]$	largest integer smaller than λ
<i>a.s.</i>	almost sure
<i>rv</i>	random variable

cdf	cumulative distribution function
edf	empirical distribution function
pdf	probability density function
$\propto$	proportionally equal
$\mathcal{O}(\beta_n)$	$\alpha_n = \mathcal{O}(\beta_n)$ iff $ \alpha_n/\beta_n \leq \text{constant}$, as $n \longrightarrow \infty$
$\mathcal{O}(\beta_n)$	$\alpha_n = \mathcal{O}(\beta_n)$ iff $\alpha_n/\beta_n \longrightarrow 0$, as $n \longrightarrow \infty$
$\mathcal{O}_p(B_n)$	$A_n = \mathcal{O}_p(B_n)$ iff $\forall \varepsilon > 0 \exists M, \exists N$ such that $\mathbb{P}[A_n/B_n > M] < \varepsilon, \forall n > N$.
$\mathcal{O}_p(B_n)$	$A_n = \mathcal{O}_p(B_n)$ iff $\forall \varepsilon > 0 : \lim_{n \rightarrow \infty} \mathbb{P}[A_n/B_n > \varepsilon] = 0$

Characteristics of Distributions

$f(x)$	pdf or density of X
$f(x, y)$	joint density of X and Y
$f_X(x), f_Y(y)$	marginal densities of X and Y
$f_{X_1}(x_1), \dots, f_{X_p}(x_p)$	marginal densities of $X_1, \dots, X_p$
$\hat{f}_h(x)$	histogram or kernel estimator of $f(x)$
$F(x)$	cdf or distribution function of X
$F(x, y)$	joint distribution function of X and Y
$F_X(x), F_Y(y)$	marginal distribution functions of X and Y
$F_{X_1}(x_1), \dots, F_{X_p}(x_p)$	marginal distribution functions of $X_1, \dots, X_p$
$f_{Y X=x}(y)$	conditional density of Y given $X = x$
$\varphi_X(t)$	characteristic function of X
m_k	k th moment of X
κ_j	cumulants or semi-invariants of X

Moments

$\mathbb{E}(X), \mathbb{E}(Y)$	mean values of random variables or vectors X and Y
$\mathbb{E}(Y X = x)$	conditional expectation of random variable or vector Y given $X = x$
$\mu_{Y X}$	conditional expectation of Y given X
$\text{Var}(Y X = x)$	conditional variance of Y given $X = x$
$\sigma_{Y X}^2$	conditional variance of Y given X
$\sigma_{XY} = \text{Cov}(X, Y)$	covariance between random variables X and Y

$\sigma_{XX} = \text{Var}(X)$	variance of random variable X
$\rho_{XY} = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X) \text{Var}(Y)}}$	correlation between random variables X and Y
$\Sigma_{XY} = \text{Cov}(X, Y)$	covariance between random vectors X and Y , i.e., $\text{Cov}(X, Y) = \mathbb{E}(X - \mathbb{E}X)(Y - \mathbb{E}Y)^\top$
$\Sigma_{XX} = \text{Var}(X)$	covariance matrix of the random vector X

Samples

x, y	observations of X and Y
$x_1, \dots, x_n = \{x_i\}_{i=1}^n$	sample of n observations of X
$\mathcal{X} = \{x_{ij}\}_{i=1, \dots, n; j=1, \dots, p}$	$(n \times p)$ data matrix of observations of $X_1, \dots, X_p$ or of $X = (X_1, \dots, X_p)^\top$
$x_{(1)}, \dots, x_{(n)}$	the order statistics of $x_1, \dots, x_n$

Empirical Moments

$\bar{x} = n^{-1} \sum_{i=1}^n x_i$	average of X sampled by $\{x_i\}_{i=1, \dots, n}$
$s_{XY} = n^{-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$	empirical covariance of random variables X and Y sampled by $\{x_i\}_{i=1, \dots, n}$ and $\{y_i\}_{i=1, \dots, n}$
$s_{XX} = n^{-1} \sum_{i=1}^n (x_i - \bar{x})^2$	empirical variance of random variable X sampled by $\{x_i\}_{i=1, \dots, n}$
$r_{XY} = \frac{s_{XY}}{\sqrt{s_{XX}s_{YY}}}$	empirical correlation of X and Y
$\mathcal{S} = \{s_{X_i X_j}\}$	empirical covariance matrix of $X_1, \dots, X_p$ or of the random vector $\mathbf{X} = (X_1, \dots, X_p)^\top$
$\mathcal{R} = \{r_{X_i X_j}\}$	empirical correlation matrix of $X_1, \dots, X_p$ or of the random vector $\mathbf{X} = (X_1, \dots, X_p)^\top$

Mathematical Abbreviations

$\text{tr}(A)$	trace of matrix A
$\text{diag}(A)$	diagonal of matrix A
$\text{rank}(A)$	rank of matrix A

$\det(A)$ or $|A|$
 $\text{hull}(x_1, \dots, x_k)$
 $\text{span}(x_1, \dots, x_k)$

determinant of matrix A
convex hull of points $\{x_1, \dots, x_k\}$
linear space spanned by $\{x_1, \dots, x_k\}$

Distributions

$\mathcal{N}(\mu, \sigma^2)$

normal distribution with mean μ , variance σ^2

Φ

cdf of $\mathcal{N}(0, 1)$

φ

pdf of $\mathcal{N}(0, 1)$

$B(n, p)$

binomial distribution with parameters n and p

$LN(\mu, \sigma^2)$

lognormal distribution with mean μ and
variance σ^2

$\xrightarrow{\mathbb{P}}$

convergence in probability

$\xrightarrow{a.s.}$

almost sure convergence

$\xrightarrow{\mathcal{L}}$

convergence in distribution

CLT

Central Limit Theorem

LLN

Law of Large Numbers

χ_p^2

χ^2 distribution with p degrees of freedom

$\chi_{1-\alpha; p}^2$

$1 - \alpha$ quantile of the χ^2 distribution with p
degrees of freedom

t_n

t -distribution with n degrees of freedom

$t_{1-\alpha/2; n}$

$1 - \alpha/2$ quantile of the t -distribution with n
degrees of freedom

$F_{n, m}$

F -distribution with n and m degrees of
freedom

$F_{1-\alpha; n, m}$

$1 - \alpha$ quantile of the F -distribution with n
and m degrees of freedom

$PR(\alpha, \mu)$

Pareto distribution with parameters α and μ

$U(a, b)$

uniform distribution with parameters a and b

$Be(\alpha, \beta)$

beta distribution with parameters α and β

Maximum Likelihood Estimation

LPA	local parametric approximation
W	$\left\{w_i = K\left(\frac{x-x_i}{h}\right)\right\}$ – weighting scheme
$\tilde{\theta}, \tilde{\theta}(W)$	local estimate for W
c_r	$= \mathbb{E} \xi ^{2r}$ risk bound for Gaussian shift model
r_r	risk bound for EF
$\mathfrak{R}_r$	risk bound in a parametric model
$W^{(k)}$	k -th weighting scheme
$\tilde{\theta}_k$	estimate for $W^{(k)}$
$\mathfrak{z}_k$	k -th critical value
$\hat{\theta}_k$	adaptive estimate after k steps
$\hat{\theta}$	final adaptive estimate
$\hat{k}$	selected model
k^*	“oracle choice”
$\Delta(W, \theta)$	modeling bias
SMB	“small modeling bias” condition

Other Notation

$L(\theta)$	log-likelihood of $\mathbb{P}_\theta$
$L(\theta, \theta')$	$= L(\theta) - L(\theta')$, log-likelihood ratio of $\mathbb{P}_\theta$ with respect to $\mathbb{P}_{\theta'}$
$\mathcal{K}(\mathbb{P}, \mathbb{Q})$	Kullback-Leibler divergence between measures P and Q
$\mathcal{K}(\theta, \theta')$	Kullback-Leibler divergence between measures P_θ and $P_{\theta'}$
$I(\theta)$	Fisher information matrix at θ
θ^*	true parameter $f \equiv f_{\theta^*}$

Some Terminology

Odabrana terminologija

"Zakon varijacije": Kada na cesti prijeđete u drugu traku, ona u kojoj ste bili će se početi micati brže od one u kojoj se trenutno nalazite.

"Law of Variation": When you change lanes whilst driving, the lane you leave will always then move faster than the one you have joined.

This section contains an overview of some terminology that is used throughout the book. The notations are in part identical to those of Harville (2001). More detailed definitions and further explanations of the statistical terms can be found, e.g., in Breiman (1973), Feller (1966), Härdle and Simar (2011), Mardia et al. (1979), or Serfling (2002).

Asymptotic normality A sequence $X_1, X_2, \dots$ of random variables is *asymptotically normal* if there exist sequences of constants $\{\mu_i\}_{i=1}^{\infty}$ and $\{\sigma_i\}_{i=1}^{\infty}$ such that $\sigma_n^{-1}(X_n - \mu_n) \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1)$. The asymptotic normality means that for sufficiently large n , the random variable X_n has approximately $\mathcal{N}(\mu_n, \sigma_n^2)$ distribution.

Bias Consider a random variable X that is parametrized by $\theta \in \Theta$. Suppose that there is an estimator $\hat{\theta}$ of θ . The *bias* is defined as the systematic difference between $\hat{\theta}$ and θ , $\mathbb{E}\{\hat{\theta} - \theta\}$. The estimator is unbiased if $\mathbb{E}\hat{\theta} = \theta$.

Characteristic function Consider a random vector $\mathbf{X} \in \mathbb{R}^p$ with pdf f . The *characteristic function* (cf) is defined for $t \in \mathbb{R}^p$:

$$\varphi_{\mathbf{X}}(t) = \mathbb{E}[\exp(\mathbf{i}t^{\top} \mathbf{X})] = \int \exp(\mathbf{i}t^{\top} \mathbf{x}) f(\mathbf{x}) d\mathbf{x}.$$

The cf fulfills $\varphi_X(0) = 1$, $|\varphi_X(t)| \leq 1$. The pdf (density) f may be recovered from the cf: $f(x) = (2\pi)^{-p} \int \exp(-it^\top X) \varphi_X(t) dt$.

Characteristic polynomial (and equation) Corresponding to any $n \times n$ matrix A is its characteristic polynomial, say $p(\cdot)$, defined (for $-\infty < \lambda < \infty$) by $p(\lambda) = |A - \lambda I|$, and its characteristic equation $p(\lambda) = 0$ obtained by setting its characteristic polynomial equal to 0; $p(\lambda)$ is a polynomial in λ of degree n and hence is of the form $p(\lambda) = c_0 + c_1\lambda + \cdots + c_{n-1}\lambda^{n-1} + c_n\lambda^n$, where the coefficients $c_0, c_1, \dots, c_{n-1}, c_n$ depend on the elements of A .

Conditional distribution Consider the joint distribution of two random vectors $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^q$ with pdf $f(x, y) : \mathbb{R}^{p+1} \rightarrow \mathbb{R}$. The marginal density of X is $f_X(x) = \int f(x, y) dy$ and similarly $f_Y(y) = \int f(x, y) dx$. The *conditional density* of X given Y is $f_{X|Y}(x|y) = f(x, y)/f_Y(y)$. Similarly, the conditional density of Y given X is $f_{Y|X}(y|x) = f(x, y)/f_X(x)$.

Conditional moments Consider two random vectors $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^q$ with joint pdf $f(x, y)$. The *conditional moments* of Y given X are defined as the moments of the conditional distribution.

Contingency table Suppose that two random variables X and Y are observed on discrete values. The two entry frequency table that reports the simultaneous occurrence of X and Y is called a *contingency table*.

Critical value Suppose one needs to test a hypothesis H_0 . Consider a test statistic T for which the distribution under the null hypothesis is given by P_0 . For a given significance level α , the *critical value* is c_α such that $P_0(T > c_\alpha) = \alpha$. The critical value corresponds to the threshold that a test statistic has to exceed in order to reject the null hypothesis.

Cumulative distribution function (cdf) Let X be a p -dimensional random vector. The *cumulative distribution function* (cdf) of X is defined by $F(x) = \mathbb{P}(X \leq x) = \mathbb{P}(X_1 \leq x_1, X_2 \leq x_2, \dots, X_p \leq x_p)$.

Eigenvalues and eigenvectors An *eigenvalue* of an $n \times n$ matrix A is (by definition) a scalar (real number), say λ , for which there exists an $n \times 1$ vector, say x , such that $Ax = \lambda x$, or equivalently such that $(A - \lambda I_n)x = \mathbf{0}$; any such vector x is referred to as an *eigenvector* (of A) and is said to belong to (or correspond to) the eigenvalue λ . Eigenvalues (and eigenvectors), as defined herein, are restricted to real numbers (and vectors of real numbers).

Eigenvalues (not necessarily distinct) The characteristic polynomial, say $p(\cdot)$, of an $n \times n$ matrix A is expressible as

$$p(\lambda) = (-1)^n (\lambda - d_1)(\lambda - d_2) \cdots (\lambda - d_m) q(\lambda) \quad (-\infty < \lambda < \infty),$$

where $d_1, d_2, \dots, d_m$ are not-necessarily-distinct scalars and $q(\cdot)$ is a polynomial (of degree $n - m$) that has no real roots; $d_1, d_2, \dots, d_m$ are referred to as the *not-necessarily-distinct eigenvalues* of A or (at the possible risk of confusion) simply as the eigenvalues of A . If the spectrum of A has k members, say $\lambda_1, \dots, \lambda_k$, with algebraic multiplicities of $\gamma_1, \dots, \gamma_k$, respectively, then $m = \sum_{i=1}^k \gamma_i$, and (for $i = 1, \dots, k$) γ_i of the m not-necessarily-distinct eigenvalues equal λ_i .

Empirical distribution function Assume that $X_1, \dots, X_n$ are iid observations of a p -dimensional random vector. The *empirical distribution function* (edf) is defined through $F_n(x) = n^{-1} \sum_{i=1}^n \mathbf{1}(X_i \leq x)$.

Empirical moments The moments of a random vector X are defined through $m_k = \mathbb{E}(X^k) = \int x^k dF(x) = \int x^k f(x) dx$. Similarly, the *empirical moments* are defined through the empirical distribution function $F_n(x) = n^{-1} \sum_{i=1}^n \mathbf{1}(X_i \leq x)$. This leads to $\hat{m}_k = n^{-1} \sum_{i=1}^n X_i^k = \int x^k dF_n(x)$.

Estimate An *estimate* is a function of the observations designed to approximate an unknown parameter value.

Estimator An *estimator* is the prescription (on the basis of a random sample) of how to approximate an unknown parameter.

Expected (or mean) value For a random vector X with pdf f the *mean* or *expected value* is $\mathbb{E}(X) = \int x f(x) dx$.

Hessian matrix The *Hessian matrix* of a function f , whose value is an m dimension real vector, is the $m \times m$ matrix whose ij -th element is the ij -th partial derivative $\partial^2 f / \partial x_i \partial x_j$ of f .

Kernel density estimator The *kernel density estimator* $\hat{f}$ of a pdf f , based on a random sample $X_1, X_2, \dots, X_n$ from f , is defined by

$$\hat{f}(x) = (nh)^{-1} \sum_{i=1}^n K\left(\frac{x - X_i}{h}\right).$$

The properties of the estimator $\hat{f}(x)$ depend on the choice of the kernel function $K(\cdot)$ and the bandwidth h . The kernel density estimator can be seen as a smoothed histogram; see also Härdle et al. (2004).

Likelihood function Suppose that $\{x_i\}_{i=1}^n$ is an iid sample from a population with pdf $f(x; \theta)$. The *likelihood function* is defined as the joint pdf of the observations $x_1, \dots, x_n$ considered as a function of the parameter θ , i.e., $L(x_1, \dots, x_n; \theta) = \prod_{i=1}^n f(x_i; \theta)$. The log-likelihood function, $\ell(x_1, \dots, x_n; \theta) = \log L(x_1, \dots, x_n; \theta) = \sum_{i=1}^n \log f(x_i; \theta)$, is often easier to handle.

Linear dependence or independence A nonempty (but finite) set of matrices (of the same dimensions ($n \times p$)), say $A_1, A_2, \dots, A_k$, is (by definition) *linearly dependent* if there exist scalars $x_1, x_2, \dots, x_k$, not all 0, such that $\sum_{i=1}^k x_i A_i = 0_n 0_p^T$; otherwise (if no such scalars exist), the set is linearly independent. By convention, the empty set is linearly independent.

Marginal distribution For two random vectors X and Y with the joint pdf $f(x, y)$, the *marginal pdfs* are defined as $f_X(x) = \int f(x, y) dy$ and $f_Y(y) = \int f(x, y) dx$.

Marginal moments The *marginal moments* are the moments of the marginal distribution.

Mean The *mean* is the first-order empirical moment $\bar{x} = \int x dF_n(x) = n^{-1} \sum_{i=1}^n x_i = \hat{m}_1$.

Mean squared error (MSE) The *mean squared error* (MSE) is defined as $\mathbb{E}(\hat{\theta} - \theta)^2$.

Median Suppose that X is a continuous random variable with pdf $f(x)$. The *median* $\tilde{x}$ lies in the center of the distribution. It is defined as $\int_{-\infty}^{\tilde{x}} f(x)dx = \int_{\tilde{x}}^{+\infty} f(x)dx - 0.5$.

Moments The *moments* of a random vector X with the distribution function $F(x)$ are defined through $m_k = \mathbb{E}(X^k) = \int x^k dF(x)$. For continuous random vectors with pdf $f(x)$, we have $m_k = \mathbb{E}(X^k) = \int x^k f(x)dx$.

Normal (or Gaussian) distribution A random vector X with the *multinormal distribution* $\mathcal{N}(\mu, \Sigma)$ with the mean vector μ and the variance matrix Σ is given by the pdf

$$f_X(x) = |2\pi\Sigma|^{-1/2} \exp \left\{ -\frac{1}{2}(x - \mu)^\top \Sigma^{-1}(x - \mu) \right\}.$$

Orthogonal matrix An $(n \times n)$ matrix A is *orthogonal* if $A^\top A = AA^\top = I_n$.

Pivotal quantity A pivotal quantity or pivot is a function of observations and unobservable parameters whose probability distribution does not depend on unknown parameters.

Probability density function (pdf) For a continuous random vector X with cdf F , the *probability density function* (pdf) is defined as $f(x) = \partial F(x)/\partial x$.

Quantile For a random variable X with pdf f the α *quantile* q_α is defined through: $\int_{-\infty}^{q_\alpha} f(x)dx = \alpha$.

p-value The critical value c_α gives the critical threshold of a test statistic T for rejection of a null hypothesis H_0 . The probability $P_0(T > c_\alpha) = p$ defines that *p-value*. If the *p-value* is smaller than the significance level α , the null hypothesis is rejected.

Random variable(rv) Random events occur in a probability space with a certain even structure. A *random variable* (rv) is a function from this probability space to $\mathbb{R}$ (or $\mathbb{R}^p$ for random vectors) also known as the state space. The concept of a random variable (vector) allows one to elegantly describe events that are happening in an abstract space.

Scatterplot A *scatterplot* is a graphical presentation of the joint empirical distribution of two random variables.

Singular value decomposition (SVD) An $m \times n$ matrix A of rank r is expressible as

$$A = P \begin{pmatrix} D_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} Q^\top = P_1 D_1 Q_1^\top = \sum_{i=1}^r s_i P_i Q_i^\top = \sum_{j=1}^k \alpha_j U_j,$$

where $Q = (Q_1, \dots, Q_n)$ is an $n \times n$ orthogonal matrix and $D_1 = \text{diag}(s_1, \dots, s_r)$ an $r \times r$ diagonal matrix such that $Q^\top A^\top A Q = \begin{pmatrix} D_1^2 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}$, where $s_1, \dots, s_r$ are (strictly) positive, where $Q_1 = (Q_1, \dots, Q_r)$, $P_1 = (P_1, \dots, P_r) = A Q_1 D_1^{-1}$, and, for any $m \times (m - r)$ matrix P_2 such that $P_1^\top P_2 = \mathbf{0}$,

$P = (P_1, P_2)$, where $\alpha_1, \dots, \alpha_k$ are the distinct values represented among $s_1, \dots, s_r$, and where (for $j = 1, \dots, k$) $U_j = \sum_{\{i: s_i = \alpha_j\}} P_i Q_i^\top$; any of these four representations may be referred to as the *singular value decomposition* of A , and $s_1, \dots, s_r$ are referred to as the singular values of A . In fact, $s_1, \dots, s_r$ are the positive square roots of the nonzero eigenvalues of $A^\top A$ (or equivalently $AA^\top$), $Q_1, \dots, Q_n$ are eigenvectors of $A^\top A$, and the columns of P are eigenvectors of $AA^\top$.

Spectral decomposition A $p \times p$ symmetric matrix A is expressible as

$$A = \Gamma \Lambda \Gamma^\top = \sum_{i=1}^p \lambda_i \gamma_i \gamma_i^\top$$

where $\lambda_1, \dots, \lambda_p$ are the not-necessarily-distinct eigenvalues of A , $\gamma_1, \dots, \gamma_p$ are orthonormal eigenvectors corresponding to $\lambda_1, \dots, \lambda_p$, respectively, $\Gamma = (\gamma_1, \dots, \gamma_p)$, $D = \text{diag}(\lambda_1, \dots, \lambda_p)$.

Subspace A *subspace* of a linear space $\mathcal{V}$ is a subset of $\mathcal{V}$ that is itself a linear space.

Taylor expansion The *Taylor series* of a function $f(x)$ in a point a is the power series $\sum_{n=0}^{\infty} \frac{f^{(n)}(a)}{n!} (x - a)^n$. A truncated Taylor series is often used to approximate the function $f(x)$.

References

- Breiman, L. (1973). *Statistics: With a view towards application*. Boston: Houghton Mifflin Company.
- Feller, W. (1966). *An introduction to probability theory and its application* (Vol. 2). New York: Wiley.
- Härdle, W., Müller, M., Sperlich, S., & Werwatz, A. (2004). *Nonparametric and semiparametric models*. Berlin: Springer.
- Härdle, W., & Simar, L. (2011). *Applied multivariate statistical analysis* (3rd ed.). Berlin: Springer.
- Harville, D. A. (2001). *Matrix algebra: Exercises and solutions*. New York: Springer.
- Mardia, K. V., Kent, J. T. & Bibby, J. M. (1979). *Multivariate analysis*. Duluth/London: Academic.
- Serfling, R. J. (2002). *Approximation theorems of mathematical statistics*. New York: Wiley.
- Spokoiny, V., & Dickhaus, T. (2014). *Basics of modern parametric statistics*. Berlin: Springer.

List of Figures

Fig. 1.1	The shape of Jiao Bei	7
Fig. 2.1	The standard normal cdf (<i>thick line</i>) and the empirical distribution function (<i>thin line</i>) for $n = 100$.  MSEedfnormal....	12
Fig. 2.2	The standard normal cdf (<i>thick line</i>) and the empirical distribution function (<i>thin line</i>) for $n = 1,000$.  MSEedfnormal...	13
Fig. 2.3	The standard normal cdf (<i>thick line</i>) and the empirical distribution function (<i>thin line</i>) for $n = 1,000$. The maximal distance in this case occurs at $X_{i*} = 1.0646$ where $i* = 830$.  MSEGCthmnorm	13
Fig. 2.4	The exponential ($\lambda = 1$) cdf (<i>thick line</i>) and the empirical distribution function (<i>thin line</i>) for $n = 1,000$. The maximal distance in this case occurs at $X_{i*} = 0.9184$ where $i* = 577$.  MSEedfnormal	14
Fig. 2.5	Plots of density estimator and log-likelihood ratio function.  MSEloglikelihood	45
Fig. 3.1	The plot of scores with respect to response variable.  MSElogit.....	59
Fig. 3.2	Lorenz curve.  MSElorenz.....	59
Fig. 3.3	The kernel density estimator $\hat{f}_h(x)$ (<i>solid line</i>), $\hat{g}(x)$ with $f_0 = t(3)$ (<i>dashed line</i>), and $\hat{g}(x)$ with $f_0 = N(\hat{\mu}, \hat{\sigma}^2)$ (<i>dotted line</i>), for $n = 300$.  MSEnonpara1	61
Fig. 3.4	The kernel density estimator $\hat{f}_h(x)$ (<i>solid line</i>), $\hat{g}(x)$ with $f_0 = t(3)$ (<i>dashed line</i>), and $\hat{g}(x)$ with $f_0 = N(\hat{\mu}, \hat{\sigma}^2)$ (<i>dotted line</i>), for $n = 300$.  MSEnonpara2	62
Fig. 3.5	The linear regression line of Y_i on Z_i (<i>solid line</i>) and the linear regression line of Y_i on Ψ_i (<i>dashed line</i>), for $n = 300$.  MSEregression	63

Fig. 3.6	The kernel regression curve from the sample without measurement errors (<i>solid line</i>), the deconvoluted kernel regression curve (<i>dashed line</i>), and the kernel regression curve from the sample with measurement errors (<i>dotted line</i>), for $n = 3,000$.  MSEdecon	72
Fig. 4.1	Consider the model on a sample (i, Y_i) with $\theta^* = (1, 1)^\top$ and $\sigma = 1$. $\hat{\theta} = (1.012115, 1.099624)^\top$.  MSEExercise471	80
Fig. 4.2	Consider the model on a sample (X_i, Y_i) with $\theta^* = (1, 1)^\top$ and $\sigma = 1$. $\hat{\theta} = (1.012115, 1.099624)^\top$.  MSEExercise472	81
Fig. 5.1	A boy is trying to test the Robokeeper which is a machine more reliable than any human goalkeeper	124
Fig. 5.2	Germany goalkeeper Jens Lehmann's crumpled sheet that helped him save penalties against Argentina in the 2006 World Cup quarter-final shootout raised one million EUR (1.3 million USD) for charity	125
Fig. 5.3	The Jiao Bei pool	127
Fig. 6.1	The plot $y = f(x) = (1 + x^2)/(1 + (x - 1)^2)$.  MSEfcauchy	132
Fig. 6.2	The plot $y = f(\theta)$.  MSEklnatparam	139
Fig. 6.3	The plot $y = g(v)$.  MSEklcanparam	140
Fig. 6.4	The plot of $g(\theta) = 1 - G_{10}(10/\theta)$.  MSEEX0810	141
Fig. 6.5	The plot of $\hat{\theta} < \theta_0 - t_{\alpha}^-$.  MSEEX0711	143
Fig. 6.6	The plot of DAX returns from 20,000,103 to 20,111,227.  MSEDAXre	145
Fig. 6.7	Plot of S&P 500 index quarterly log-returns during the period Q2 1980–Q2 2012.  MSEspqlogret	157
Fig. 6.8	QQ-plot for S&P index quarterly log-returns during the period Q2 1980–Q2 2012.  MSEqlretqqplot	157
Fig. 7.1	A trajectory of y_t . $\theta_1 = 2$, $\theta_2 = 0.5$, $\omega_1 = 0.04$, $\omega_2 = 0.5$ and $\sigma = 0.8$.  MSESpectral	163
Fig. 8.1	The time series of DAX30.  MSENNormalityTests	172
Fig. 8.2	Example of population profiles  MSEprofil	179

List of Tables

Table 3.1	GLM results and overall model fit.  MSEglmest	58
Table 3.2	The goodness of the model.  MSEperformance.....	58
Table 5.1	The posterior probability when $z = 0, 1, 2, 3, 4, 5$	127

Chapter 1

Basics

기본

가랑비에 옷 젖는 줄 모른다

Constant sprinkle can make you wet

In this chapter on basics of mathematical statistics we present simple exercises that help to understand the notions of sample, observations and data modeling with parameterized distributions. We study the Bernoulli model, linear regression and discuss design questions for a variety of different applications.

Exercise 1.1. Let $Y = \{Y_1, \dots, Y_n\}$ be i.i.d. Bernoulli with the parameter θ^* .

1. Prove that the mean and the variance of the sum $S_n = Y_1 + \dots + Y_n$ satisfy

$$\begin{aligned}\mathbb{E}_{\theta^*} S_n &= n \theta^*, \\ \text{Var}_{\theta^*} S_n &\stackrel{\text{def}}{=} \mathbb{E}_{\theta^*} (S_n - \mathbb{E}_{\theta^*} S_n)^2 = n \theta^* (1 - \theta^*).\end{aligned}$$

2. Find θ^* that maximizes $\text{Var}_{\theta^*} S_n$.

1. Observe that the Y_i 's are i.i.d.

$$\begin{aligned}\mathbb{E}_{\theta^*} (Y_1 + Y_2 + Y_3 + \dots + Y_n) &= n \mathbb{E}_{\theta^*} (Y_1) \\ &= n \{\theta^* \times 1 + (1 - \theta^*) \times 0\} \\ &= n \theta^*\end{aligned}$$

Since the variance of a sum of i.i.d variables is the sum of the variances, we obtain:

$$\text{Var}_{\theta^*} S_n = n \text{Var} Y_1 = n\theta^*(1 - \theta^*)$$

2. Maximizing the function $u(1 - u)$ for u in $[0, 1]$ yields $u = 1/2$. The fair coin toss therefore has the maximum variance in this Bernoulli experiment.

Exercise 1.2. Consider the Bernoulli model with parameter θ^* and $\tilde{\theta} = n^{-1} \sum_{i=1}^n Y_i$ its estimator. Prove that $\tilde{\theta}(1 - \tilde{\theta})$ is estimating the population variance $\sigma^2 = \mathbb{E}_{\theta^*} (Y_1 - \theta^*)^2$

$\tilde{\theta}$ is a consistent estimator of θ^* . By the continuous mapping theorem, $\tilde{\theta}(1 - \tilde{\theta})$ estimates $\theta^*(1 - \theta^*)$. In fact, the empirical counterpart of σ^2 equals to $n^{-1} \sum_{i=1}^n Y_i^2 - (n^{-1} \sum_{i=1}^n Y_i)^2$. Since Y_i is either 0 or 1, this exactly equals to $n^{-1} \sum_{i=1}^n Y_i - (n^{-1} \sum_{i=1}^n Y_i)^2$, which is $\tilde{\theta}(1 - \tilde{\theta})$.

Exercise 1.3. Let $Y_i = \Psi_i^\top \theta^* + \varepsilon_i$ be a regression model with fixed design $\Psi_i = \{\psi_1(X_i), \dots, \psi_p(X_i)\}^\top \in \mathbb{R}^p$. Assume that the error ε_i are i.i.d. with mean 0 and $\text{Var}(\varepsilon) = \sigma^2$.

The LS estimator is:

$$\tilde{\theta} = (\Psi \Psi^\top)^{-1} \Psi Y.$$

Show that $\text{Var}(\tilde{\theta}) = \sigma^2 (\Psi \Psi^\top)^{-1}$.

$$\begin{aligned} \text{Var}(\tilde{\theta}) &= \text{Var} \left\{ (\Psi \Psi^\top)^{-1} \Psi Y \right\} \\ &= \text{Var} \left\{ (\Psi \Psi^\top)^{-1} \Psi (\Psi_i^\top \theta^* + \varepsilon) \right\} \\ &= \text{Var} \left\{ (\Psi \Psi^\top)^{-1} \Psi \varepsilon \right\} \\ &= (\Psi \Psi^\top)^{-1} \Psi \text{Var}(\varepsilon) \Psi^\top (\Psi \Psi^\top)^{-1} \sigma^2 I \\ &= \sigma^2 (\Psi \Psi^\top)^{-1}. \end{aligned}$$

Exercise 1.4. Consider a linear regression model $Y_i = \Psi_i^\top \theta^* + \varepsilon_i$ for $i = 1, \dots, n$ with uncorrelated ε_i satisfying $\mathbb{E} \varepsilon_i = 0$, $\mathbb{E} \varepsilon^2 = \sigma^2 < \infty$, $\Psi = (\psi_1, \psi_2, \dots, \psi_n)_{p \times n}$. Define a linear transformation of θ^* as $a^* \stackrel{\text{def}}{=} v^\top \theta^*$, $v \in \mathbb{R}$.

1. Show that $\Psi \phi = v_{p \times 1}$, where $\phi \in \mathbb{R}^n$, implies:

$$\text{Cov}(\phi^\top Y, \tilde{a}) \stackrel{\text{def}}{=} \mathbb{E}_{\theta^*} \{ (\phi^\top Y - a^*) (\tilde{a} - a^*) \} = \sigma^2 v^\top (\Psi \Psi^\top)^{-1} v$$

2. Check that $0 \leq \text{Var}(\phi^\top Y - \tilde{a}) = \text{Var}(\phi^\top Y) - \sigma^2 v^\top (\Psi \Psi^\top)^{-1} v$

1.

$$\begin{aligned} \text{Cov}(\phi^\top Y, v^\top \tilde{\theta}) &= \text{Cov}(\phi^\top Y, v^\top (\Psi \Psi^\top)^{-1} \Psi Y) \\ &= \text{Cov}(\phi^\top Y - a^*, v^\top (\Psi \Psi^\top)^{-1} \Psi Y - a^*) \\ &= \mathbb{E}\{(\phi^\top Y - a^*)(\tilde{a} - a^*)^\top\}. \end{aligned}$$

Since $\phi^\top Y - a^* = \phi^\top \varepsilon$ and $\tilde{a} - a^* = v^\top (\Psi \Psi^\top)^{-1} \Psi \varepsilon$, this yields:

$$\begin{aligned} \mathbb{E}[\phi^\top \varepsilon \{v^\top (\Psi \Psi^\top)^{-1} \Psi \varepsilon\}^\top] &= \sigma^2 \phi^\top \Psi^\top (\Psi \Psi^\top)^{-1} v \\ &= \sigma^2 v^\top (\Psi \Psi^\top)^{-1} \Psi \phi. \end{aligned}$$

2.

$$\begin{aligned} \text{Var}(\phi^\top Y - \tilde{a}) &= \text{Var}(\phi^\top Y) + \text{Var}(\tilde{a}) - 2 \text{Cov}(\phi^\top Y, \tilde{a}) \\ &= \text{Var}(\phi^\top Y) + \text{Var}\{v^\top (\Psi \Psi^\top)^{-1} \Psi Y\} - 2\sigma^2 v^\top (\Psi \Psi^\top)^{-1} \Psi \phi \\ &= \text{Var}(\phi^\top Y) + \sigma^2 v^\top (\Psi \Psi^\top)^{-1} v - 2\sigma^2 v^\top (\Psi \Psi^\top)^{-1} \Psi \phi \\ &= \text{Var}(\phi^\top Y) + \sigma^2 v^\top (\Psi \Psi^\top)^{-1} v - 2\sigma^2 v^\top (\Psi \Psi^\top)^{-1} v \\ &= \text{Var}(\phi^\top Y) - \sigma^2 v^\top (\Psi \Psi^\top)^{-1} v. \end{aligned}$$

Exercise 1.5. Let $Y_i = \Psi_i^\top \theta^* + \varepsilon_i$ for $i = 1, \dots, n$ with $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ and $\Psi_i, \theta^* \in \mathbb{R}^p$. Let $\text{rank}(\Psi) = p$ and let v be a given vector from $\mathbb{R}^p$. Denote the estimate $\tilde{a} = v^\top \tilde{\theta}$; denote the true value $a^* = v^\top \theta^*$. Prove that

1.

$$\tilde{a} - a^* \sim \mathcal{N}(0, s^2)$$

with $s^2 = \sigma^2 v^\top (\Psi \Psi^\top)^{-1} v$.

2.

$$\mathbb{P}_{\theta^*}(|\tilde{a} - a^*| > z_\alpha s) = \alpha,$$

where $\Phi(z_\alpha) = 1 - \alpha/2$.

1. Note that

$$\begin{aligned} \tilde{a} - a^* &= v^\top (\tilde{\theta} - \theta^*) = v^\top \{(\Psi \Psi^\top)^{-1} \Psi Y - \theta^*\} \\ &= v^\top \{(\Psi \Psi^\top)^{-1} \Psi (\Psi^\top \theta^* + \varepsilon) - \theta^*\} \\ &= v^\top (\Psi \Psi^\top)^{-1} \Psi \varepsilon \end{aligned}$$

has a normal distribution, because it is a linear transformation of normally distributed vector $\boldsymbol{\varepsilon}$. So, it is sufficient to prove that

$$\mathbb{E}(\tilde{a} - a^*) = 0 \quad \text{and} \quad \text{Var}(\tilde{a} - a^*) = s^2.$$

First fact is exactly the Gauss-Markov theorem. Second fact can be checked via simple calculation:

$$\text{Var}(\tilde{a} - a^*) = \mathbb{E}(\tilde{a} - a^*)^2 = \mathbf{v}^\top (\Psi \Psi^\top)^{-1} \Psi \underbrace{\boldsymbol{\varepsilon} \boldsymbol{\varepsilon}^\top}_{=\sigma^2 I} \Psi^\top (\Psi \Psi^\top)^{-1} \mathbf{v} = s^2$$

2. The cdf of $Y \sim \mathcal{N}(0, s^2)$ is $\Phi(u/s)$, $u \in \mathbb{R}$. Hence

$$\begin{aligned} \mathbb{P}_{\theta^*}(|Y| > z_\alpha s) &= 2 \mathbb{P}_{\theta^*}(Y > z_\alpha s) = 2 \{1 - \Phi(z_\alpha)\} \\ &= 2 \{1 - (1 - \alpha/2)\} = \alpha. \end{aligned}$$

Exercise 1.6. Let $Y_1, \dots, Y_n$ be i.i.d. $U[0, \theta]$. For any integer k

$$\mathbb{E}_\theta(Y_1^k) = \theta^{-1} \int_0^\theta y^k dy = \theta^k / (k+1),$$

or $\theta = \{(k+1)\mathbb{E}_\theta(Y_1^k)\}^{1/k}$. For any k one defines

$$\tilde{\theta}_k = \left(\frac{k+1}{n} \sum_{i=1}^n Y_i^k \right)^{1/(k+1)}.$$

Prove that

$$\lim_{k \rightarrow \infty} \tilde{\theta}_k = \tilde{\theta}_\infty = \max\{Y_1, \dots, Y_n\}.$$

Define the order statistics $Y_{(1)} \leq Y_{(2)} \leq \dots \leq Y_{(n)}$. Since $Y_{(i)} \geq 0$ for all i we have

$$\begin{aligned} \left(\frac{k+1}{n} \right)^{1/(k+1)} Y_{(n)}^{\frac{1}{k+1}} &\leq \left(\frac{k+1}{n} \sum_{i=1}^n Y_i^k \right)^{1/(k+1)} = \left(\frac{k+1}{n} \sum_{i=1}^n Y_{(i)}^k \right)^{1/(k+1)} \\ &\leq (k+1)^{1/(k+1)} Y_{(n)}^{1/(k+1)}. \end{aligned}$$

The limit of $Y_{(n)}^{k/(k+1)}$ for $k \rightarrow \infty$ is $Y_{(n)} = \tilde{\theta}_\infty$. Both $\left(\frac{k+1}{n}\right)^{1/(k+1)}$ and $(k+1)^{1/(k+1)}$ tend to 1 as $k \rightarrow \infty$.

Exercise 1.7. A statistical decision problem is defined in terms of a decision space $\mathcal{D}$, a loss function $\wp(\cdot, \cdot)$ on $\mathcal{D} \times \Theta$ and the statistical decision $\rho = \rho(Y)$. Define the statistical decision problem for testing a simple hypothesis $\theta^* = \theta_0$ for a given point θ_0 .

Let $\Theta_0 = \{\theta_0\}$ and $\Theta_1 = \Theta \setminus \Theta_0$. The decision space $\mathcal{D}$ consists of two points $\{0, 1\}$, where $d = 0$ means that $H_0 : \theta^* = \theta_0$ is accepted, while $H_1 : \theta^* \neq \theta_0$ favors the alternative. The loss is defined as:

$$\wp(d, \theta) = \mathbf{1}(d = 1, \theta = \theta_0) + \mathbf{1}(d = 0, \theta \neq \theta_0).$$

A test is a binary valued function $\phi = \Phi(Y) \rightarrow \{0, 1\}$. The risk is calculated as:

$$\mathcal{R}(\phi, \theta^*) = \mathbb{E}_{\theta^*} \phi(Y),$$

i.e. the probability of selecting $\theta \neq \theta_0$.

Exercise 1.8. The risk of a statistical decision problem is denoted as $\mathcal{R}(\rho, \theta)$. The quality of a statistical decision can be measured by either the minimax or Bayes risk. The Bayes risk with prior π is given by $\mathcal{R}_\pi(\rho) = \int \mathcal{R}(\rho, \theta) \pi(d\theta)$, while the minimax risk is given by $\mathcal{R}(\rho^*) = \inf_\rho \mathcal{R}(\rho) = \inf_\rho \sup_{\theta \in \Theta} \mathcal{R}(\rho, \theta)$. Show that the minimax risk is greater than or equal to the Bayes risk whatever the prior measure π is.

Define $\forall \rho, \quad \mathcal{R}(\rho) \stackrel{\text{def}}{=} \sup_{\theta \in \Theta} \mathcal{R}(\rho, \theta)$

It is easy to see

$$\mathcal{R}(\rho) \geq \int \mathcal{R}(\rho, \theta) \pi(d\theta) \tag{1.1}$$

since

$$\int \mathcal{R}(\rho, \theta) \pi(d\theta) \leq \sup_{\theta} \mathcal{R}(\rho, \theta) \int \pi(d\theta) = \sup_{\theta} \mathcal{R}(\rho, \theta)$$

The relation in (1.1) will of course not change if we move to $\inf_\rho \mathcal{R}(\rho)$ leading to

$$\inf_{\rho} \mathcal{R}(\rho) \geq \inf_{\rho} \mathcal{R}_\pi(\rho)$$

which proves the claim.

Exercise 1.9. Consider the model in Exercise 1.9, where $\tilde{a} = \mathbf{v}^\top \tilde{\theta}$ and $\phi \in \mathbb{R}^p$. Check that the minimization of the quadratic form $\phi^\top \phi$ under the condition $\Psi \phi = \mathbf{v}$ leads to the equation $\phi^\top \phi = \mathbf{v}^\top (\Psi \Psi^\top)^{-1} \mathbf{v}$.

1. Define $\Pi = \Psi^\top (\Psi \Psi^\top)^{-1} \Psi$ and show that Π is a projector in $\mathbb{R}^n$ in the sense that $\Pi^2 = \Pi^\top = \Pi$.

2. Decompose $\phi^\top \phi = \phi^\top \Pi \phi + \phi^\top (I - \Pi) \phi$.
3. Check that $\sigma^2 \phi^\top \Pi \phi = \sigma^2 v^\top (\Psi \Psi^\top)^{-1} v = \text{Var}(\tilde{a})$ using $\psi \phi = v$.
4. Show that $\phi^\top (I - \Pi) \phi = 0$ iff $\Pi \phi = \phi$.

1. Define $\Pi = \Psi^\top (\Psi \Psi^\top)^{-1} \Psi$

We can prove that

$$\begin{aligned} \Pi^2 &= \Psi^\top (\Psi \Psi^\top)^{-1} \Psi \Psi^\top (\Psi \Psi^\top)^{-1} \Psi \\ &= \Psi^\top (\Psi \Psi^\top)^{-1} \Psi = \Pi \end{aligned}$$

and

$$\Pi^\top = (\Psi^\top (\Psi \Psi^\top)^{-1} \Psi)^\top = \Psi^\top (\Psi \Psi^\top)^{-1} \Psi = \Pi$$

so Π is a projector in $\mathbb{R}^n$ because

$$\Pi^2 = \Pi = \Pi^\top$$

2. Decompose $\phi^\top \phi = \phi^\top \Pi \phi + \phi^\top (I - \Pi) \phi$ where

$$\begin{aligned} \phi^\top \Pi \phi &= \phi^\top \Psi^\top (\Psi \Psi^\top)^{-1} \Psi \phi \\ &= v^\top (\Psi \Psi^\top)^{-1} v \end{aligned}$$

Therefore $\sigma^2 \phi^\top \Pi \phi = \sigma^2 v^\top (\Psi \Psi^\top)^{-1} v = \text{Var}(\tilde{a})$

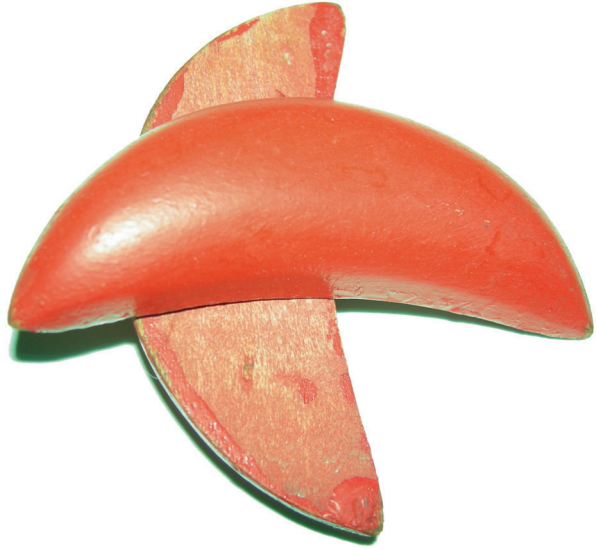
Recall that $(I - \Pi)$ is a projector matrix which just has eigenvalues 1 or 0. Thus it is non-negative definite and therefore $\phi^\top (I - \Pi) \phi \geq 0$ and $\phi^\top (I - \Pi) \phi = 0$ iff $\phi^\top \phi = \phi^\top \Pi \phi$.

$$\phi^\top (I - \Pi)^{1/2} (I - \Pi)^{1/2} \phi = 0$$

Set now $u \stackrel{\text{def}}{=} \phi^\top (I - \Pi)^{1/2}$, then we obtain:

$$\begin{aligned} u^\top u &= 0 \\ u &= 0 \\ \phi (I - \Pi)^{1/2} &= 0 \\ \phi (I - \Pi) \phi &= 0 \\ \phi &= \Pi \phi \end{aligned}$$

Fig. 1.1 The shape of Jiao Bei



then

$$\begin{aligned}\phi\phi &= \phi^\top \Pi \phi + \phi^\top (I - \Pi)\phi \\ &\geq \phi^\top \Pi \phi = \sigma^2 v^\top (\Psi \Psi^\top)^{-1} v\end{aligned}$$

if and only if $\phi = \Pi\phi$ for “=”.

Exercise 1.10. In Taiwanese culture, there is the “Jiao Bei” (茭杯, Fig. 1.1), which helps to know if the Gods agree with important matters such as marriage, home moving or dilemmas. This kind of divination—tossing “Jiao Bei”—is given by the outcome of the relative location of the two wooden pieces. Worshippers stand in front of the statue of the God they believe in, and speak the question in their mind. Finally they toss the Jiao Bei to see if the Gods agree or not.

As a pair of crescent moon-shaped wooden pieces, each Jiao Bei piece has a convex (C) and a flat side (F). When tossing Jiao Bei, there are four possible outcomes: (C,C), (F,F), (C,F), (F,C). The first two outcomes mean that the Gods disagree and one needs to restate or change the question. The last two outcomes mean that the Gods agree, and this outcome is called “Sheng Bei” (聖杯).

Suppose that each piece of Jiao Bei is fair and the probability to show C or F is equal. Sequential tossings of Jiao Bei can be viewed as sequence of i.i.d. Bernoulli trials.

1. What is the probability of the event of Sheng Bei?
2. If tossing Jiao Bei ten times, how many times of Sheng Bei would show up?
3. What is the probability that Sheng Bei finally shows up at the 5th tossing?

1. The probability for the event (C,C) is $1/4$, given the assumption that the events C and F have equal chances for each piece of the Jiao Bei. Similarly, the probabilities for the events (F,F), (C,F) and (F,C) are also $1/4$.
For the event of Sheng Bei, it would be either (C,F) or (F,C). Therefore the probability for the event Sheng Bei is $p = 1/4 + 1/4 = 1/2$.
2. Using the result of 1. in Exercise 1.1, the expected number of Sheng Bei if tossing ten times is $np = 10 * 1/2 = 5$.
3. We know that the probability for the event Sheng Bei is $1/2$. There are four failures before Sheng Bei shows up at the 5th tossing. So the probability for this event is

$$\left(\frac{1}{2}\right)^4 \frac{1}{2} = \left(\frac{1}{2}\right)^5.$$

Exercise 1.11. *The crucial assumption of Exercise 1.10 is the Jiao Bei fairness which is reflected in the probability $1/2$ of either C or F. A primary school student from Taiwan did a controlled experiments on a pair of Jiao Bei tossing 200 times, yielding the outcomes (C,C), (F,F), (F,C), (C,F). The outcomes (F,C), (C,F) are “Sheng Bei” and are denoted by 1, while the outcomes (C,C), (F,F) are not “Sheng Bei” and are denoted by 0. We have a sequence of experiment results:*

1	0	1	0	0	1	0	1	0	0	0	1	1	0	1	1	0	1	0	0
1	0	1	0	1	1	0	0	0	1	0	1	0	0	0	0	0	1	0	0
0	0	1	0	1	1	0	0	1	0	0	0	1	0	0	0	1	0	0	0
0	1	0	0	0	0	1	1	0	1	1	0	0	0	0	0	0	0	0	0
1	1	0	1	0	0	0	0	1	0	1	0	0	0	0	1	0	0	0	0
0	0	0	0	0	0	0	0	1	0	1	0	0	1	0	1	0	0	1	0
1	0	0	0	0	1	1	1	0	1	0	1	0	1	0	0	0	1	0	0
0	1	0	1	1	0	0	0	0	1	1	0	1	0	1	1	1	1	1	0
1	1	0	1	0	0	0	1	1	0	0	0	1	1	0	0	0	0	0	1
0	1	0	0	0	1	1	1	1	0	0	1	0	0	1	1	1	0	0	1

Can you conclude from this experiment that the Jiao Bei is fair?

We can decide if this pair of Jiao Bei is fair by applying a test on the null hypothesis $H_0 : p_0 = 0.5$, where p is the probability that “Sheng Bei” shows up. Denote this set of data as $\{x_i\}_{i=1}^{200}$, and the event $x_i = 1$ is shown 75 times.

To compute the test statistics, first we have $\bar{x} = 75/200 = 0.375$. $\sqrt{\sigma^2/n} = \sqrt{0.5 * 0.5/200} = 0.0354$. The test statistics is $(\bar{x} - p_0)/\sqrt{\sigma^2/n} = -3.5311$. According to the asymptotic normality, the test statistics has p -value 0.0002. Thus, the null hypothesis is rejected by a significance level $\alpha = 0.001$.

Chapter 2

Parameter Estimation for an i.i.d. Model

Оценивание параметров в модели с независимыми одинаково
распределёнными наблюдениями

Кадры, овладевшие техникой, решают всё!

*Personnels that became proficient in technique decide
everything!*

Joseph Stalin

Exercise 2.1 (Glivenko-Cantelli theorem). Let F be the distribution function of a random variable X and let $\{X_i\}_{i=1}^n$ be an i.i.d. sample from F . Define the edf as

$$F_n(x) \stackrel{\text{def}}{=} n^{-1} \sum_{i=1}^n \mathbf{1}(X_i \leq x).$$

Prove that

$$\sup_x |F_n(x) - F(x)| \xrightarrow{a.s.} 0, \quad n \rightarrow \infty$$

1. If F is a continuous distribution function;
2. If F is a discrete distribution function.

1. Consider first the case when the function F is continuous in y . Fix any integer N and define with $\varepsilon = 1/N$ the points $t_1 < t_2 < \dots < t_N = +\infty$ such that

$$F(t_j) - F(t_{j-1}) = \varepsilon \text{ for } j = 2, \dots, N. \quad (2.1)$$

For every j , by the law of large numbers: $F_n(t_j) \xrightarrow{a.s.} F(t_j)$. This implies that for some $n(\varepsilon)$, it holds for all $n \geq n(\varepsilon)$

$$|F_n(t_j) - F(t_j)| \leq \varepsilon, \quad j = 1, \dots, N. \quad (2.2)$$

$F(t)$ and $F_n(t)$ are nondecreasing functions. This implies that for every $t \in [t_{j-1}, t_j]$ it holds

$$F(t_{j-1}) \leq F(t) \leq F(t_j), \quad F_n(t_{j-1}) \leq F_n(t) \leq F_n(t_j). \quad (2.3)$$

Let us subtract the first inequality (2.3) from the second:

$$F_n(t_{j-1}) - F(t_j) \leq F_n(t) - F(t) \leq F_n(t_j) - F(t_{j-1}), \quad (2.4)$$

Let us continue with the right hand side using (2.1) and (2.2):

$$\begin{aligned} F_n(t) - F(t) &\leq F_n(t_j) - F(t_{j-1}) \\ &= \underbrace{\{F_n(t_j) - F(t_j)\}}_{\leq \varepsilon} + \underbrace{\{F(t_j) - F(t_{j-1})\}}_{=\varepsilon} \leq 2\varepsilon, \end{aligned}$$

In the same way (considering the left part of (2.4)), one can prove that

$$F_n(t) - F(t) \geq -2\varepsilon$$

So,

$$|F_n(t) - F(t)| \leq 2\varepsilon. \quad (2.5)$$

Thus for all $\varepsilon > 0$ there exists constant $n(\varepsilon) > 0$ such that for every $n > n(\varepsilon)$ the inequality (2.5) holds for all $t \in \mathbb{R}$.

2. By $T = \{t_m\}_{m=1}^{+\infty}$ we denote points of discontinuity of function $F(x)$. Of course, these points are also points of discontinuity of function $F_n(t)$ (for any n).

Let us fix some $\varepsilon > 0$ and let us construct some finite set $S(\varepsilon)$. We include in $S(\varepsilon)$ the following points:

- (a) Points such that at least one inequality fulfills:

$$F(t_m) - F(t_{m-1}) > \varepsilon \quad \text{or} \quad F(t_{m+1}) - F(t_m) > \varepsilon$$

(b) Continuous set of points such that

$$F(t_m) - F(t_{m-1}) < \varepsilon$$

Denote amount of elements in $S(\varepsilon)$ by M .

We know that $F_n(t) \rightarrow F(t)$ almost sure. In particular

$$F_n(t_m) \xrightarrow{a.s.} F(t_m), \quad \forall m \in S(\varepsilon).$$

By definition

$$\exists n_m(\varepsilon) \in \mathbb{N} : \quad \forall n > n_m(\varepsilon) \quad |F_n(t_m) - F(t_m)| < \varepsilon$$

Define $n(\varepsilon) \stackrel{\text{def}}{=} \max\{n_1(\varepsilon), \dots, n_M(\varepsilon)\}$. Then for all $t_m \in S(\varepsilon)$

$$\forall n > n(\varepsilon) \quad |F_n(t_m) - F(t_m)| < \varepsilon.$$

Let us prove that the inequality

$$\forall n > n(\varepsilon) \quad |F_n(t_m) - F(t_m)| < 2\varepsilon. \quad (2.6)$$

is also true for all points $t_m \notin S(\varepsilon)$. Fix some $t_m \notin S(\varepsilon)$ and find index s such that

$$F(t_{s-1}) \leq F(t_m) \leq F(t_s), \quad F_n(t_{s-1}) \leq F_n(t_m) \leq F_n(t_s).$$

Consider

$$\begin{aligned} F_n(t_m) - F(t_m) &\leq F_n(t_s) - F(t_{s-1}) \\ &= \underbrace{\{F_n(t_s) - F(t_s)\}}_{< \varepsilon} + \underbrace{\{F(t_s) - F(t_{s-1})\}}_{\leq \varepsilon} \leq 2\varepsilon, \end{aligned}$$

Similarly, one can prove that

$$F_n(t_m) - F(t_m) \geq -2\varepsilon$$

This means that

$$|F_n(t_m) - F(t_m)| \leq 2\varepsilon$$

So, (2.6) is true for all $t_m \in T$.

For all t there exists some point $t_m \in T$ such that

$$F_n(t) = F_n(t_m) \quad \text{and} \quad F(t) = F(t_m).$$

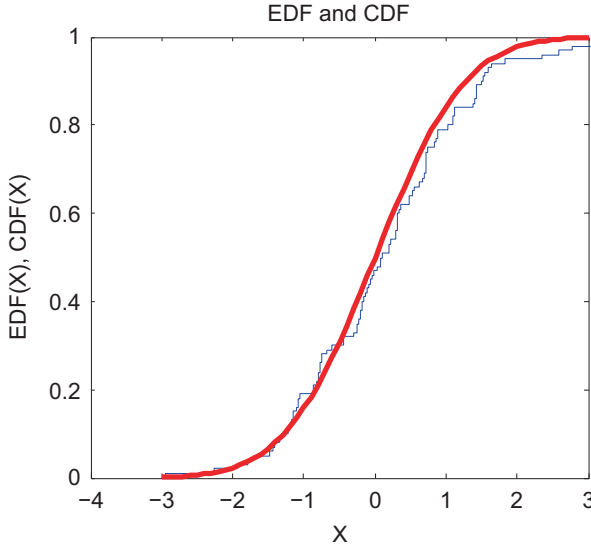


Fig. 2.1 The standard normal cdf (*thick line*) and the empirical distribution function (*thin line*) for $n = 100$.  MSEdfnormal

Thus

$$\forall n > n(\varepsilon) \quad |F_n(t) - F(t)| < \varepsilon.$$

This observation completes the proof.

For an illustration of the asymptotic property, we draw $\{X_i\}_{i=1}^n$ i.i.d. samples from the standard normal distribution. Figure 2.1 shows the case of $n = 100$ and Fig. 2.2 shows the case of $n = 1,000$. The empirical cdf and theoretical cdf are close in the limit as n becomes larger.

Exercise 2.2 (Illustration of the Glivenko-Cantelli theorem). Denote by F the cdf of

1. Standard normal law,
2. Exponential law with parameter $\lambda = 1$.

Consider the sample $\{X_i\}_{i=1}^n$. Draw the plot of the empirical distribution function F_n and cumulative distribution function F . Find the index $i^* \in \{1, \dots, n\}$ such that

$$|F_n(X_{i^*}) - F(X_{i^*})| = \sup_i |F_n(X_i) - F(X_i)|.$$

The examples for the code can be found in the Quantnet. The readers are suggested to change the sample size n to compare the results (Figs. 2.3 and 2.4).

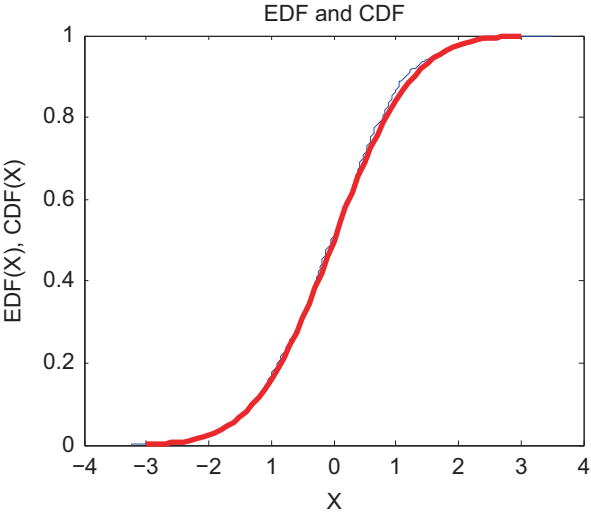


Fig. 2.2 The standard normal cdf (*thick line*) and the empirical distribution function (*thin line*) for $n = 1,000$.  MSEdnormal

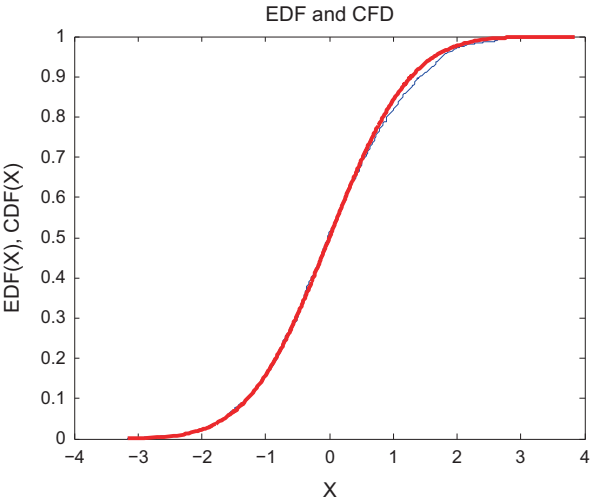


Fig. 2.3 The standard normal cdf (*thick line*) and the empirical distribution function (*thin line*) for $n = 1,000$. The maximal distance in this case occurs at $X_{i^*} = 1.0646$ where $i^* = 830$.  MSEGthmnorm

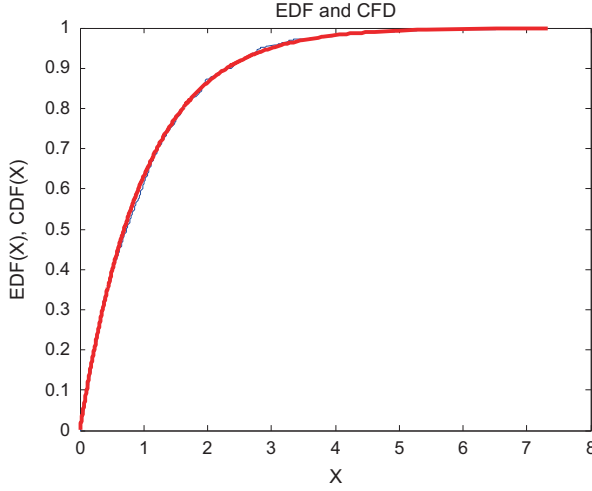


Fig. 2.4 The exponential ($\lambda = 1$) cdf (thick line) and the empirical distribution function (thin line) for $n = 1,000$. The maximal distance in this case occurs at $X_{i^*} = 0.9184$ where $i^* = 577$.
 MSEdnormal

Exercise 2.3. Compute the estimate of method of moments for the following parametric models:

1. Multinomial model:

$$\mathbb{P}_\theta(X = k) = \binom{m}{k} \theta^k (1 - \theta)^{m-k}, \quad k = 0, \dots, m.$$

2. Exponential model

$$\mathbb{P}_\theta(X > x) = e^{-x/\theta}.$$

In both cases one can follow the algorithm consisting of two steps:

- Calculate mathematical expectation $m(\theta) = \mathbb{E}_\theta X$;
- Solve the equation $m(\tilde{\theta}) = n^{-1} \sum_{i=1}^n X_i$; the solution is the required estimate.

Let us apply this:

1. Multinomial model, we first calculate expectation:

$$\begin{aligned} m(\theta) &= n^{-1} \sum_{i=1}^n X_i = \sum_{k=0}^m k \binom{m}{k} \theta^k (1 - \theta)^{m-k} \\ &= m \sum_{k=1}^m \binom{m-1}{k-1} \theta^k (1 - \theta)^{m-k} = m\theta. \end{aligned}$$

Secondly we solve the equation

$$m(\tilde{\theta}) = \frac{1}{n} \sum_{i=1}^n X_i;$$

which gives the solution:

$$\tilde{\theta} = \frac{1}{nm} \sum_{i=1}^n X_i.$$

2. Exponential family. Both items are trivial: $m(\theta) = \frac{1}{\theta}$ and $\tilde{\theta} = n \left(\sum_{i=1}^n X_i \right)^{-1}$.

Exercise 2.4. Let $\{X_i\}_{i=1}^n$ be an i.i.d. sample from a distribution with Lebesgue density

$$f_{\theta}(x) = \frac{1}{2} (1 + \theta x) I_{[-1,1]}(x)$$

1. Find an estimator via the method of moments;
2. Find a consistent estimator.

Let us begin with calculation of the mathematical expectation:

$$\mathbb{E}_{\theta} X_1 = \frac{1}{2} \int_{-1}^1 (1 + \theta x) x \, dx = \frac{1}{3} \theta$$

Both items of the exercise follow immediately:

1. The estimator of method of moments is a solution of the equality

$$\mathbb{E}_{\tilde{\theta}} X_1 = n^{-1} \sum_{i=1}^n X_i$$

$$\text{So, } \tilde{\theta} = 3n^{-1} \sum_{i=1}^n X_i$$

2. By the law of large numbers,

$$n^{-1} \sum_{i=1}^n X_i \xrightarrow{a.s.} \mathbb{E} X_i = \frac{1}{3} \theta, \quad n \rightarrow \infty.$$

This means that

$$3n^{-1} \sum_{i=1}^n X_i \xrightarrow{a.s.} \theta, \quad n \rightarrow \infty,$$

hence the estimator $\hat{\theta} = 3n^{-1} \sum_{i=1}^n X_i$ is consistent.

Exercise 2.5. Consider the model

$$X_i = \theta^* + \varepsilon_i,$$

where θ^* is the parameter of interest and ε_i are independent normal errors $\mathcal{N}(0, \sigma_i^2)$.

Compute the MLE $\tilde{\theta}$ of the parameter θ^* and prove that this estimate has the following properties:

- (a) The estimate $\tilde{\theta}$ is unbiased: $\mathbb{E}_{\theta^*} \tilde{\theta} = \theta^*$.
(b) The quadratic risk of $\tilde{\theta}$ is equal to

$$\mathcal{R}(\tilde{\theta}, \theta^*) \stackrel{\text{def}}{=} \mathbb{E}_{\theta^*} |\tilde{\theta} - \theta^*|^2 = \left(\sum_{i=1}^n \sigma_i^2 \right)^{-1}.$$

The corresponding log-likelihood reads

$$L(\theta) = -\frac{1}{2} \sum_{i=1}^n \left\{ \log(2\pi\sigma_i^2) + \frac{(X_i - \theta)^2}{\sigma_i^2} \right\}.$$

The first derivative is equal to

$$\frac{\partial L(\theta)}{\partial \theta} = \sum_{i=1}^n \frac{X_i - \theta}{\sigma_i^2} = \sum_{i=1}^n \frac{X_i}{\sigma_i^2} - \theta \sum_{i=1}^n \frac{1}{\sigma_i^2}.$$

Then the MLE $\tilde{\theta}$ equals

$$\tilde{\theta} \stackrel{\text{def}}{=} \operatorname{argmax}_{\theta} L(\theta) = \frac{1}{N} \sum \frac{X_i}{\sigma_i^2},$$

where $N = \sum \sigma_i^{-2}$.

(a)

$$\mathbb{E}_{\theta^*} \tilde{\theta} = \frac{1}{N} \sum \frac{\mathbb{E} X_i}{\sigma_i^2} = \frac{1}{N} \sum \frac{\theta^*}{\sigma_i^2} = \frac{\theta^*}{N} \sum \frac{1}{\sigma_i^2} = \theta^*.$$

(b) The quadratic risk of $\tilde{\theta}$ is equal to the variance $\operatorname{Var}(\tilde{\theta})$:

$$\begin{aligned} \mathcal{R}(\tilde{\theta}, \theta^*) &\stackrel{\text{def}}{=} \mathbb{E}_{\theta^*} |\tilde{\theta} - \theta^*|^2 = \mathbb{E}_{\theta^*} \left| \frac{1}{N} \sum \frac{X_i}{\sigma_i^2} - \theta^* \right|^2 \\ &= \mathbb{E}_{\theta^*} \left| \frac{1}{N} \sum \frac{X_i}{\sigma_i^2} - \theta^* \frac{1}{N} \sum \frac{1}{\sigma_i^2} \right|^2 \\ &= \mathbb{E}_{\theta^*} \left| \frac{1}{N} \sum \frac{X_i - \theta^*}{\sigma_i^2} \right|^2 = \frac{1}{N^2} \sum \mathbb{E}_{\theta^*} \left| \frac{X_i - \theta^*}{\sigma_i^2} \right|^2. \end{aligned}$$

Note that random value $X_i - \theta^*$ has a normal distribution with zero mean and variance σ_i^2 . Then $(X_i - \theta^*)/\sigma_i^2 \sim \mathcal{N}(0, \sigma_i^{-2})$ and

$$\mathcal{R}(\tilde{\theta}^\circ, \theta^*) = \frac{1}{N^2} \sum \sigma_i^{-2} = \frac{1}{N}.$$

Exercise 2.6. Let $\{X_i\}_{i=1}^n$ be an i.i.d. sample with distribution that depends on some parameter θ . Let $\hat{\theta}_n$ be an estimate of parameter θ .

Assume that this estimate is root- n normal, i.e. there exists a function $\sigma(\theta)$ such that

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} \mathcal{N}(0, \sigma(\theta)^2), \quad n \rightarrow \infty.$$

Prove that $\hat{\theta}_n$ is consistent,

$$\hat{\theta}_n \xrightarrow{\mathbb{P}} \theta$$

This fact can be briefly formulated as “root- n normality implies consistency”. We need Slutsky’s Theorem:

1. Let a_n (sequence of real numbers) be convergent in probability,

$$a_n \xrightarrow{\mathbb{P}} a, \quad n \rightarrow \infty$$

Let η_n (sequence of random variables) be convergent in distribution,

$$\eta_n \xrightarrow{\mathcal{L}} \text{Law}(\eta), \quad n \rightarrow \infty$$

Then

$$a_n \eta_n \xrightarrow{\mathcal{L}} \text{Law}(a\eta), \quad n \rightarrow \infty$$

2. Let ξ_n be a sequence of random variables that converges in law to the distribution that is degenerated in some point c (we denote this degenerated distribution by $\text{Law}(c)$). Then ξ_n also tends to c in probability.

Let us apply these observations to our situation. We use the first part of Slutsky’s Theorem with $a_n = \frac{1}{\sqrt{n}}$ and $\xi_n = \sqrt{n}(\hat{\theta}_n - \theta)$.

The sequence a_n tends to zero and the sequence ξ_n tends in probability to a normal distribution. So,

$$a_n \xi_n \xrightarrow{\mathcal{L}} \text{Law}(0)$$

According to the second part, this sequence also tends to zero in probability. Thus,

$$a_n \xi_n = \hat{\theta}_n - \theta \xrightarrow{\mathbb{P}} 0.$$

Remark 2.1. In fact our proof is true for any estimate that has an asymptotic distribution (not necessarily normal).

Exercise 2.7. Let F be the distribution function of a random variable X and let $\{X_i\}_{i=1}^n$ be an i.i.d. sample from F . Let $g : \mathbb{R} \rightarrow \mathbb{R}$ be a function such that

$$\sigma_g^2 \stackrel{\text{def}}{=} \text{Var}\{g(X)\} < \infty$$

Denote

$$s \stackrel{\text{def}}{=} \mathbb{E}g(X), \quad S_n \stackrel{\text{def}}{=} n^{-1} \sum_{i=1}^n g(X_i)$$

1. Prove that

- (a) $S_n \xrightarrow{\mathbb{P}} s, \quad n \rightarrow \infty$
- (b) $\sqrt{n}(S_n - s) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma_g^2), \quad n \rightarrow \infty.$

2. Let $h(z)$ be a twice continuously differentiable function on the real line such that $h'(s) \neq 0$ and $h''(s)$ is bounded in some neighborhood of s . Prove that

- (a) $h(S_n) \xrightarrow{\mathbb{P}} h(s)$
- (b) $\sqrt{n}\{h(S_n) - h(s)\} \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma_h^2), \quad n \rightarrow \infty,$
where $\sigma_h^2 \stackrel{\text{def}}{=} |h'(s)|^2 \sigma_g^2.$

1. (a) Note that $\{g(X_i)\}_{i=1}^n$ is a sample from the distribution with expectation equal to $\mathbb{E}g(X)$.

One can apply the law of large numbers for the sequence $\{g(X_i)\}_{i=1}^n$:

$$n^{-1} \sum_{i=1}^n g(X_i) \xrightarrow{\mathbb{P}} \mathbb{E}g(X) \quad n \rightarrow \infty.$$

(b) This statement directly follows by the CLT for i.i.d. random variables:

$$\frac{n^{-1} \sum_{i=1}^n g(X_i) - \mathbb{E}g(X)}{\sqrt{\frac{1}{n} \text{Var}\{g(X)\}}} \sim \mathcal{N}(0, 1)$$

In other words,

$$\sqrt{n} \left\{ n^{-1} \sum_{i=1}^n g(X_i) - \mathbb{E}g(X) \right\} \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma_g^2), \quad n \rightarrow \infty.$$

2. (a) We know:

$$S_n \xrightarrow{\mathbb{P}} s, \quad n \rightarrow \infty$$

Then for any continuous function g :

$$g(S_n) \xrightarrow{\mathbb{P}} g(s), \quad n \rightarrow \infty$$

(b) One can find a neighborhood U of the point s such that

- (i) S_n belongs with high probability to U ;
- (ii) $h''(s)$ is bounded in U .

Applying the Taylor expansion to h in this neighborhood U :

$$\sqrt{n} \{h(S_n) - h(s)\} = \sqrt{n}h'(s)(S_n - s) + \frac{\sqrt{n}}{2}h''(\tilde{s})(S_n - s)^2, \quad (2.7)$$

where $\tilde{s}$ is some point between s and S_n . The right hand side of (2.7) is a sum of two random variables. First random variable $\sqrt{n}h'(s)(S_n - s)$ tends to $\mathcal{N}(0, |h'(s)|^2\sigma_g^2)$ in distribution.

Let us show that the second component tends to zero in probability. Actually,

$$\left| \frac{\sqrt{n}}{2}h''(\tilde{s})(S_n - s)^2 \right| \leq \frac{U}{2} \frac{1}{\sqrt{n}} \{ \sqrt{n}(S_n - s) \}^2,$$

where U is an upper bound for $h''(s)$ in the considering neighborhood. Expression in the right hand side is a product of the sequence $\frac{1}{\sqrt{n}}$, which tends to zero, and sequence $\{ \sqrt{n}(S_n - s) \}^2$, which converges in distribution. Then

$$\left| \frac{\sqrt{n}}{2}h''(\tilde{s})(S_n - s)^2 \right| \xrightarrow{\mathbb{P}} 0$$

Thus, the right hand side in (2.7) (and left hand side also) tends to $\mathcal{N}(0, |h'(s)|^2\sigma_g^2)$ in distribution.

Exercise 2.8. (Analogue of the Exercise 2.7 for multi-dimensional case) Let $\mathbf{g}(\cdot) = (g_1(\cdot), \dots, g_m(\cdot))^\top : \mathbb{R} \rightarrow \mathbb{R}^m$ be a function such that

$$\Sigma_{jk} \stackrel{\text{def}}{=} \mathbb{E}[g_j(X)g_k(X)] < \infty, \text{ for } j, k \leq m.$$

Denote

$$\begin{aligned} \mathbf{s} &= \mathbb{E}\mathbf{g}(X) = (\mathbb{E}g_1(X), \dots, \mathbb{E}g_m(X))^\top, \\ \mathbf{S}_n &= \frac{1}{n} \sum_{i=1}^n \mathbf{g}(X_i) = \left(\frac{1}{n} \sum_{i=1}^n g_1(X_i), \dots, \frac{1}{n} \sum_{i=1}^n g_m(X_i) \right)^\top. \end{aligned}$$

1. Prove that

$$\begin{aligned} (a) \quad \mathbf{S}_n &\xrightarrow{\mathbb{P}} \mathbf{s}, \quad n \rightarrow \infty \\ (b) \quad \sqrt{n}(\mathbf{S}_n - \mathbf{s}) &\xrightarrow{\mathcal{L}} \mathcal{N}(0, \Sigma), \quad n \rightarrow \infty, \\ &\text{where } \Sigma = (\Sigma_{jk})_{j,k=1,\dots,m} \end{aligned}$$

2. Let $H(z) : \mathbb{R}^m \rightarrow \mathbb{R}$ be a twice continuously differentiable function such that $\nabla H(z)$ and $\|\nabla^2 H(z)\|$ is bounded in some neighborhood of $\mathbf{s}$. Prove that

$$\begin{aligned} (a) \quad H(\mathbf{S}_n) &\xrightarrow{\mathbb{P}} H(\mathbf{s}) \\ (b) \quad \sqrt{n}\{H(\mathbf{S}_n) - H(\mathbf{s})\} &\xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma_H^2), \quad n \rightarrow \infty, \\ &\text{where } \sigma_H^2 \stackrel{\text{def}}{=} \nabla H(\mathbf{s})^\top \Sigma \nabla H(\mathbf{s}). \end{aligned}$$

First note that items 1a and 2a follow from items 1b and 2b correspondingly. Let us check items 1b and 2b.

Consider for every $\mathbf{v} = (v_1, \dots, v_m)^\top \in \mathbb{R}^m$ the scalar products $\mathbf{v}^\top \mathbf{g}(\cdot)$, $\mathbf{v}^\top \mathbf{s}$, $\mathbf{v}^\top \mathbf{S}_n$. For the statement 1b, it suffices to show that

$$\sqrt{n}\mathbf{v}^\top(\mathbf{S}_n - \mathbf{s}) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \mathbf{v}^\top \Sigma \mathbf{v}), \quad n \rightarrow \infty.$$

Actually

$$\begin{aligned} \sqrt{n}\mathbf{v}^\top(\mathbf{S}_n - \mathbf{s}) &= \sqrt{n} \sum_j v_j \left\{ \frac{1}{n} \sum_i g_j(X_i) - \mathbb{E}g_j(X) \right\} \\ &= \sqrt{n} \left[\frac{1}{n} \sum_i \left\{ \sum_j v_j g_j(X_i) \right\} - \mathbb{E} \left\{ \sum_j v_j g_j(X) \right\} \right] \\ &= \sqrt{n} \left\{ \frac{1}{n} \sum_i G(X_i) - \mathbb{E}G(X) \right\}, \end{aligned}$$

where $G(\cdot) = \sum_j v_j g_j(\cdot) = \mathbf{v}^\top \mathbf{g}(\cdot)$.

Now one can apply result of the Exercise 2.7 (item 1b) for the function $G(\cdot)$ and obtain the required statement.

For the statement 2b, consider the Taylor expansion

$$\sqrt{n} \{H(\mathbf{S}_n) - H(s)\} = \sqrt{n} \nabla H(s)^\top (\mathbf{S}_n - s) + \frac{\sqrt{n}}{2} (\mathbf{S}_n - s)^\top \nabla^2 H(\tilde{s}) (\mathbf{S}_n - s).$$

This formula is an analogue of (2.7). One can continue the line of reasoning in the same way as in the proof of (2.7) (item 2b).

In fact,

$$\sqrt{n} \nabla H(s)^\top (\mathbf{S}_n - s) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \nabla H(s)^\top \Sigma \nabla H(s)),$$

and

$$\left| \frac{\sqrt{n}}{2} (\mathbf{S}_n - s)^\top \nabla^2 H(\tilde{s}) (\mathbf{S}_n - s) \right| \leq \frac{1}{2\sqrt{n}} \|\sqrt{n} (\mathbf{S}_n - s)\|^2 \max_s \|\nabla^2 H(s)\|$$

$$\xrightarrow{\mathbb{P}} 0$$

These two observations conclude the proof.

Exercise 2.9.

1. Consider a sample $\{X_i\}_{i=1}^n$ from a distribution $P_{\theta^*} \in (P_\theta, \theta \in \Theta \in \mathbb{R})$. Let $\tilde{\theta}$ be an estimator of θ such that the bias

$$b(\tilde{\theta}, \theta^*) \stackrel{\text{def}}{=} \mathbb{E}_{\theta^*} \tilde{\theta} - \theta^*$$

and the variance $\text{Var}_{\theta^*}(\tilde{\theta})$ tend to zero as $n \rightarrow \infty$. Prove that $\tilde{\theta}$ is consistent.

2. Let $\{X_i\}_{i=1}^n$ be a sample from the uniform distribution on $[0, \theta]$. Using the first item of this exercise, prove that the estimator

$$\tilde{\theta}_1 = \max \{X_1, \dots, X_n\}$$

is consistent.

1. Applying the so called bias-variance decomposition, which is true for any estimate $\tilde{\theta}$:

$$\mathbb{E}_{\theta^*} (\tilde{\theta} - \theta^*)^2 = \text{Var}_{\theta^*}(\tilde{\theta}) + b^2(\tilde{\theta}, \theta^*). \quad (2.8)$$

Let us prove (2.8):

$$\begin{aligned}
 \mathbb{E}_{\theta^*} \left(\tilde{\theta} - \theta^* \right)^2 &= \mathbb{E}_{\theta^*} \left\{ \tilde{\theta} - \mathbb{E}(\tilde{\theta}) + \mathbb{E}(\tilde{\theta}) - \theta^* \right\}^2 \\
 &= \mathbb{E}_{\theta^*} \left\{ \tilde{\theta} - \mathbb{E}(\tilde{\theta}) + b(\tilde{\theta}, \theta^*) \right\}^2 \\
 &= \text{Var}_{\theta^*}(\tilde{\theta}) + 2b(\tilde{\theta}, \theta^*)\mathbb{E}_{\theta^*} \left(\tilde{\theta} - \mathbb{E}\tilde{\theta} \right) + b^2(\tilde{\theta}, \theta^*) \\
 &= \text{Var}_{\theta^*}(\tilde{\theta}) + b^2(\tilde{\theta}, \theta^*)
 \end{aligned}$$

If bias and variance tend to zero as $n \rightarrow \infty$, then

$$\mathbb{E}_{\theta^*} \left(\tilde{\theta} - \theta^* \right)^2 \rightarrow 0, \quad n \rightarrow \infty$$

This means that $\tilde{\theta}$ tends to θ^* in L_2 sense. Then $\tilde{\theta}$ also tends to θ^* in probability, i.e. $\tilde{\theta}$ is a consistent estimator.

2. First of all, let us calculate the cdf of $\tilde{\theta}_1$.

$$\begin{aligned}
 \mathbb{P}_{\theta^*} \left(\tilde{\theta}_1 \leq x \right) &= \mathbb{P}_{\theta^*} (X_1 \leq x, \dots, X_n \leq x) \\
 &= \{ \mathbb{P}_{\theta^*} (X_1 \leq x) \}^n = \left(\frac{x}{\theta^*} \right)^n, \quad x \in [0, \theta^*]
 \end{aligned}$$

Afterwards we can take the derivative and obtain the density function

$$p(x) = n(\theta^*)^{-n} x^{n-1} \mathbf{1}(0 \leq x \leq \theta^*)$$

For applying the first item, one has to calculate expectation and variance of $\tilde{\theta}_1$:

$$\mathbb{E}\tilde{\theta}_1 = \frac{n}{n+1}\theta^*, \quad \text{Var}(\tilde{\theta}_1) = \frac{n}{(n+1)^2(n+2)}\theta^{*2}$$

Now we are ready for applying the first item:

$$b(\tilde{\theta}_1, \theta^*) = \frac{n}{n+1}\theta^* - \theta^* = -\frac{1}{n+1}\theta^* \rightarrow 0, \quad n \rightarrow \infty.$$

$$\text{Var}_{\theta^*}(\tilde{\theta}) = \frac{n}{(n+1)^2(n+2)}\theta^{*2} \rightarrow 0, \quad n \rightarrow \infty.$$

So, assumptions are fulfilled. This concludes the proof.

Exercise 2.10. Check that the i.i.d. experiment from the uniform distribution on the interval $[0, \theta]$ with unknown θ is not regular.

First condition from the definition of the regular family is the following one: the sets $A(\theta) \stackrel{\text{def}}{=} \{y : p(y, \theta) = 0\}$ are the same for all $\theta \in \Theta$.

The uniform distribution on the interval $[0, \theta]$ doesn't satisfy this condition,

$$A(\theta) = (-\infty, 0) \cup (\theta, +\infty).$$

This exercise gives a local approximation of the Kullback-Leibler divergence.

Exercise 2.11. *Let (P_θ) be a regular family.*

1. *Show that the KL-divergence $\mathcal{K}(\theta, \theta')$ satisfies for any θ, θ' :*

(a)

$$\mathcal{K}(\theta, \theta') \Big|_{\theta'=\theta} = 0;$$

(b)

$$\frac{d}{d\theta'} \mathcal{K}(\theta, \theta') \Big|_{\theta'=\theta} = 0;$$

(c)

$$\frac{d^2}{d\theta'^2} \mathcal{K}(\theta, \theta') \Big|_{\theta'=\theta} = I(\theta).$$

2. *Show that in a small neighborhood of θ , the KL-divergence can be approximated by*

$$\mathcal{K}(\theta, \theta') \approx I(\theta) |\theta' - \theta|^2 / 2.$$

1. Note that

$$\mathcal{K}(\theta, \theta') = \mathbb{E}_\theta \log p(x, \theta) - \mathbb{E}_\theta \log p(x, \theta')$$

(a) First item is trivial.

(b)

$$\begin{aligned} \frac{d}{d\theta'} \mathcal{K}(\theta, \theta') &= -\frac{d}{d\theta'} \mathbb{E}_\theta \log p(x, \theta') \\ &= -\frac{d}{d\theta'} \int \log p(x, \theta') p(x, \theta) dx \\ &= -\int \frac{p'_{\theta'}(x, \theta')}{p(x, \theta')} p(x, \theta) dx, \end{aligned}$$

where $p'_{\theta'}(x, \theta') \stackrel{\text{def}}{=} \frac{d}{d\theta'} p(x, \theta')$. Substitution $\theta' = \theta$ gives

$$\begin{aligned} \frac{d}{d\theta'} \mathcal{K}(\theta, \theta') \Big|_{\theta'=\theta} &= - \int \frac{d}{d\theta'} \{p(x, \theta')\} dx \Big|_{\theta'=\theta} \\ &= - \frac{d}{d\theta'} \int p(x, \theta') dx \Big|_{\theta'=\theta} = 0. \end{aligned}$$

(c)

$$\begin{aligned} \frac{d^2}{d\theta'^2} \mathcal{K}(\theta, \theta') &= - \int \frac{d}{d\theta'} \left\{ \frac{p'_{\theta'}(x, \theta')}{p(x, \theta')} \right\} p(x, \theta) dx \\ &= - \int \left[\frac{p''_{\theta'}(x, \theta') p(x, \theta') - \{p'_{\theta'}(x, \theta')\}^2}{\{p(x, \theta')\}^2} \right] p(x, \theta) dx. \end{aligned}$$

Substitution $\theta' = \theta$ yields

$$\begin{aligned} \frac{d^2}{d\theta'^2} \mathcal{K}(\theta, \theta') \Big|_{\theta'=\theta} &= \underbrace{\int p''_{\theta'}(x, \theta') dx \Big|_{\theta'=\theta}}_{\frac{d^2}{d\theta'^2} \int p(x, \theta') dx \Big|_{\theta'=\theta} = 0} + \underbrace{\int \frac{\{p'_{\theta'}(x, \theta')\}^2}{p(x, \theta')} dx}_{=I(\theta)} = I(\theta). \end{aligned}$$

2. The required representation directly follows from the Taylor expansion at the point $\theta' = \theta$.

The following exercise

1. Illustrates two methods for checking the R-efficiency;
2. Shows that the Fisher information can depend on the parameter (for some parametric families), but can be a constant (for other parametric families).

Exercise 2.12. Consider two families:

- (a) the Gaussian shift (b) the Poisson family

1. Compute the Fisher Information for these families.
2. Check that the Cramér-Rao inequality for the empirical mean estimate $\tilde{\theta} = n^{-1} \sum_{i=1}^n X_i$ is in fact an equality, i.e.

$$\text{Var}_{\theta}(\tilde{\theta}) = n^{-1} I^{-1}(\theta).$$

3. Check R-efficiency of $\tilde{\theta}$

- (i) Using only the definition;
- (ii) Using the Theorem 2.6.3. of Spokoiny and Dickhaus (2014)

1. (a) Recall

$$p(x, \theta) = \frac{1}{\sqrt{2\pi}} \exp \left\{ -\frac{(x - \theta)^2}{2} \right\}.$$

Then

$$\begin{aligned} I(\theta) &= \mathbb{E}_\theta \left| \frac{\partial \log p(X, \theta)}{\partial \theta} \right|^2 = \mathbb{E}_\theta \left| \frac{\partial}{\partial \theta} \left\{ -\frac{(X - \theta)^2}{2} \right\} \right|^2 \\ &= \mathbb{E}_\theta |X - \theta|^2 \\ &= \mathbb{E}_\theta |X - \mathbb{E}_\theta X|^2 \\ &= \text{Var}(X) = 1. \end{aligned}$$

Therefore, the Fisher information is equal to 1 for any values of the parameter θ .

(b)

$$p(x, \theta) = \frac{\theta^x}{x!} e^{-\theta}, \quad x = 1, 2, \dots$$

$$\begin{aligned} I(\theta) &= \mathbb{E}_\theta \left| \frac{\partial \log p(X, \theta)}{\partial \theta} \right|^2 = \mathbb{E}_\theta \left| \frac{\partial}{\partial \theta} (X \log \theta - \log X! - \theta) \right|^2 \\ &= \mathbb{E}_\theta \left| \frac{X}{\theta} - 1 \right|^2 = \frac{1}{\theta^2} \mathbb{E}_\theta |X - \theta|^2 \\ &= \frac{1}{\theta^2} \mathbb{E}_\theta |X - \mathbb{E}_\theta X|^2 = \frac{1}{\theta^2} \text{Var}_\theta(X) = \frac{1}{\theta}. \end{aligned}$$

So, in the case of the Poisson family, the Fisher information depends on θ .

2. Estimator $\tilde{\theta}$ is unbiased for both cases. Then the Cramér-Rao inequality stands that

$$\text{Var}_\theta(\tilde{\theta}) = \text{Var}_\theta \left(\frac{1}{n} \sum_{i=1}^n X_i \right) = \frac{1}{n} \text{Var}_\theta(X_1) \geq n^{-1} I^{-1}(\theta).$$

So, the aim is to check that

$$\text{Var}_\theta(X_1) I(\theta) = 1. \quad (2.9)$$

- (a) For the Gaussian shift $\text{Var}_\theta(X_1) = 1$ and $I(\theta) = 1$. Hence, (2.9) is fulfilled.
 (b) For the Poisson family, $\text{Var}_\theta(X_1) = \theta$ and $I(\theta) = 1/\theta$. Hence, (2.9) is also fulfilled.

3. (i) The definition says that R-efficient estimators are exactly the estimators that give the equality in the Cramér-Rao inequality. So, this item is already proved.
(ii) The estimate $\tilde{\theta}$ can be represented as

$$\tilde{\theta} = n^{-1} \sum U(Y_i)$$

with $U(x) = x$. The aim is to show that the log-density $\ell(y, \theta)$ of P_θ can be represented as

$$\ell(x, \theta) = C(\theta)x - B(\theta) + \ell(x), \quad (2.10)$$

for some functions $C(\cdot)$ and $B(\cdot)$ on Θ and a function $\ell(\cdot)$ on $\mathbb{R}$.

(a)

$$\ell(x, \theta) = \theta x - \theta^2/2 + \left(-\frac{x^2}{2} + \log \frac{1}{\sqrt{2\pi}} \right),$$

and (2.10) follows with $C(\theta) = \theta$, $B(\theta) = \theta^2/2$, and $\ell(x) = -x^2/2 + \log 1/\sqrt{2\pi}$.

(b)

$$\ell(x, \theta) = \log(\theta)x - \theta + \log(x!), \quad (2.11)$$

and (2.11) follows with $C(\theta) = \log \theta$, $B(\theta) = \theta$, and $\ell(x) = \log(x!)$.

Exercise 2.13. Let X be a random variable with a distribution from $(P_\theta, \theta \in \Theta \subset \mathbb{R})$. Let also a function $\psi^\circ : \mathbb{R} \times \Theta \rightarrow \mathbb{R}$ be such that

$$\psi^\circ(x, \theta) = a(x - \theta)^2 + b(x - \theta) + c,$$

where $a, b, c \in \mathbb{R}$.

1. Find a condition on the constants a, b, c and the family (P_θ) such that the function $\psi^\circ(x, \theta)$ is a contrast.
2. Find a condition on the constants a, b, c such that the function $\psi^\circ(x, \theta)$ is a contrast for the model of the Gaussian shift $\mathcal{N}(\theta, 1)$.
1. By definition, the function ψ° is a contrast if and only if

$$\operatorname{argmin}_{\theta'} \mathbb{E}_\theta \psi^\circ(X, \theta') = \theta, \quad \forall \theta.$$

Introduce a function

$$\begin{aligned} f(\theta, \theta') &\stackrel{\text{def}}{=} \mathbb{E}_\theta [\psi^\circ(X, \theta')] \\ &= (a \mathbb{E}_\theta X^2 + b \mathbb{E}_\theta X + c) - (2a\mathbb{E}_\theta X + b) \theta' + a\theta'^2. \end{aligned} \quad (2.12)$$

The aim is to find a condition on the constants a, b, c and the family (P_θ) such that

$$\operatorname{argmin}_{\theta'} f(\theta, \theta') = \theta, \quad \forall \theta. \quad (2.13)$$

Take the derivative of the function $f(\theta, \theta')$ with respect to θ' and solve the equation $\partial f(\theta, \theta')/\partial \theta' = 0$:

$$\frac{\partial f(\theta, \theta')}{\partial \theta'} = -(2a\mathbb{E}_\theta X + b) + 2a\theta' = 0$$

This means that

$$\operatorname{argmin}_{\theta'} f(\theta, \theta') = \mathbb{E}_\theta X + \frac{b}{2a}.$$

Together with (2.13), this yields the required condition on the constants a, b and the family (P_θ) :

$$\theta = \mathbb{E}_\theta X + \frac{b}{2a}, \quad \forall \theta. \quad (2.14)$$

Constant c can be chosen arbitrary.

2. For the model of the Gaussian shift $\mathcal{N}(\theta, 1)$,

$$\mathbb{E}_\theta X = \theta, \quad \forall \theta.$$

Condition (2.14) in this case yields $b = 0$. This means, that any function $\psi^\circ(x, \theta)$ with $b = 0$ and any constants a and c is a contrast for the Gaussian shift.

Exercise 2.14. Let $\{X_i\}_{i=1}^n$ be an i.i.d. sample from a distribution $P_{\theta^*} \in (P_\theta, \theta \in \Theta \subset \mathbb{R})$.

1. Let also $g(x)$ satisfy $\int g(x)dP_{\theta^*}(x) = \theta^*$, leading to the moment estimate

$$\tilde{\theta} \stackrel{\text{def}}{=} n^{-1} \sum_{i=1}^n g(X_i).$$

Show that this estimate can be obtained as the M -estimate for a properly selected function $\psi(\cdot)$.

2. Let $\int g(x)dP_{\theta^*}(x) = m(\theta^*)$ for the given functions $g(\cdot)$ and strictly monotonic and continuously differentiable $m(\cdot)$. Show that the moment estimate $\tilde{\theta} = m^{-1}\{\sum g(X_i)/n\}$ can be obtained as the M -estimate for a properly selected function $\psi(\cdot)$.

1. It is the worth mentioning that

$$\tilde{\theta} \stackrel{\text{def}}{=} n^{-1} \sum_{i=1}^n g(X_i) = \underset{\theta}{\operatorname{argmin}} \sum_{i=1}^n \{g(X_i) - \theta\}^2. \quad (2.15)$$

This observation helps us to find an appropriate function ψ . Fix

$$\psi(x, \theta) \stackrel{\text{def}}{=} \{g(x) - \theta\}^2$$

and prove that

$$\theta^* = \underset{\theta}{\operatorname{argmin}} \mathbb{E}_{\theta^*} \psi(\mathbf{X}, \theta), \quad (2.16)$$

where $\mathbf{X}$ is a variable that has the distribution P_{θ^*} .

The proof of (2.16) is straightforward:

$$\mathbb{E}_{\theta^*} \psi(\mathbf{X}, \theta) = \mathbb{E}_{\theta^*} \{g(\mathbf{X}) - \theta\}^2 = \mathbb{E}_{\theta^*} g^2(\mathbf{X}) - 2\theta \mathbb{E}_{\theta^*} g(\mathbf{X}) + \theta^2$$

Minimizing the right hand side expression by θ yields

$$\theta_{\min} = \mathbb{E}_{\theta^*} g(\mathbf{X}) = \theta^*.$$

This concludes the proof.

2. The proof follows the same lines as the proof of the first statement. Note that

$$\tilde{\theta} \stackrel{\text{def}}{=} m^{-1} \left\{ \frac{1}{n} \sum_{i=1}^n g(X_i) \right\} = \underset{\theta}{\operatorname{argmin}} \sum_{i=1}^n \{g(X_i) - m(\theta)\}^2.$$

Function

$$\psi(x, \theta) \stackrel{\text{def}}{=} \{g(x) - m(\theta)\}^2$$

is appropriate because of

$$\theta^* = \underset{\theta}{\operatorname{argmin}} \mathbb{E}_{\theta^*} \{g(\mathbf{X}) - m(\theta)\}^2 \quad (2.17)$$

In fact, fix some $\theta \in \Theta$ and find a minimum value of the function

$$f(\theta) = \mathbb{E}_{\theta^*} \{g(\mathbf{X}) - m(\theta)\}^2$$

In order to minimize this function, solve the equation $f'(\theta) = 0$:

$$\frac{df(\theta)}{d\theta} = \frac{df(\theta)}{dm(\theta)} \frac{dm(\theta)}{d\theta} = 2\mathbb{E}_{\theta}^* \{g(X) - m(\theta)\} \frac{dm(\theta)}{d\theta} = 0$$

The first derivative of the function $m(\theta)$ doesn't change the sign because of monotonicity. This means that the minimum value of function f satisfies the following equation

$$m(\theta_{\min}) = \mathbb{E}_{\theta^*} g(X).$$

Then (2.17) fulfills. This completes the proof.

Exercise 2.15. Let $\{X_i\}_{i=1}^{n_1}$ be a sample from the distribution with the pdf

$$p(x, \theta) = \frac{2x}{\theta^2}, \quad x \in [0, \theta].$$

Find the MLE of the median of the distribution.

First let us find a relation between θ and the median m . By the definition of the median,

$$\int_{-\infty}^m p(x, \theta) dx = \int_0^m \frac{2x}{\theta^2} dx = 1/2,$$

i.e. $\theta = \sqrt{2}m$. Then the likelihood function

$$L(m) = \prod_{i=1}^n p(X_i, \sqrt{2}m) = \prod_{i=1}^n \frac{X_i}{m^2} \mathbf{1}(X_i \in [0, \sqrt{2}m])$$

has a maximum at the point $\hat{m} = \max_i X_i / \sqrt{2}$.

Exercise 2.16. Let $\{X_i^{(1)}\}_{i=1}^{n_1}$ and $\{X_i^{(2)}\}_{i=1}^{n_2}$ be two independent samples from the Poisson distributions with unknown parameters μ_1 and $\mu_2 = \mu_1 + \mu$ correspondingly. Find the maximum likelihood estimator for the parameter μ .
Hint: Is it possible to find separately $\hat{\mu}_1$ (the MLE for μ_1) from the first sample, $\hat{\mu}_2$ (the MLE for μ_2) from the second sample, and then obtain the MLE estimator for μ as the difference $\hat{\mu} = \hat{\mu}_1 - \hat{\mu}_2$?

Denote by $L_1(\mu_1)$ and $L_2(\mu_2)$ the log-likelihood functions for the first and the second samples correspondingly.

The MLE estimate for the parameter μ is determined as

$$(\hat{\mu}_1, \hat{\mu}) = \operatorname{argmax}_{\mu_1, \mu} \{L_1(\mu_1) + L_2(\mu_1 + \mu)\}.$$

Hence, $\hat{\mu} = \operatorname{argmax}_{\mu} L_2(\hat{\mu}_1 + \mu)$. The maximal value of the function L_2 is achieved at the point $\hat{\mu}_2$. This yields

$$\max_{\mu} L_2(\hat{\mu}_1 + \mu) = L_2(\hat{\mu}_2) = L_2\{\hat{\mu}_1 + (\hat{\mu}_2 - \hat{\mu}_1)\}.$$

So, $\hat{\mu} = \hat{\mu}_2 - \hat{\mu}_1$.

In the case of the Poisson distribution,

$$L_1(\mu_1) = \sum_{i=1}^{n_1} \log \left(e^{-\mu_1} \frac{\mu_1^{X_i^{(1)}}}{X_i^{(1)}!} \right) \quad \text{and} \quad L_2(\mu_2) = \sum_{i=1}^{n_2} \log \left(e^{-\mu_2} \frac{\mu_2^{X_i^{(2)}}}{X_i^{(2)}!} \right),$$

and the MLE of the parameter is the mean value, i.e. $\hat{\mu}_j = n_j^{-1} \sum_{i=1}^{n_j} X_i^{(j)} \stackrel{\text{def}}{=} \overline{X}^{(j)}$, $j = 1, 2$. Thus, we conclude that

$$\hat{\mu} = \overline{X}^{(2)} - \overline{X}^{(1)}.$$

Exercise 2.17. Let $\{X_i\}_{i=1}^n$ be an i.i.d. sample from a distribution with the Lebesgue density

$$p(x, \theta) = \frac{\alpha}{\beta^\alpha} x^{\alpha-1} I_{(0, \beta)}(x),$$

where $\alpha, \beta > 0$ and $\theta \stackrel{\text{def}}{=} (\alpha, \beta)$. Find estimators for the multivariate parameter θ using the following approaches:

1. Maximum likelihood approach;
2. Method of moments.

1. The likelihood function in this case

$$\begin{aligned} L(\theta) &= \prod_{i=1}^n p(X_i, \theta) = \frac{\alpha^n}{\beta^{\alpha n}} \prod_{i=1}^n X_i^{\alpha-1} I_{(0, \beta)}(X_i) \\ &= \frac{\alpha^n}{\beta^{\alpha n}} I_{(0, \beta)}(X_{(n)}) \prod_{i=1}^n X_i^{\alpha-1} \end{aligned}$$

is equal to zero if $\beta < X_{(n)}$ and decreases for $\beta \geq X_{(n)}$. Therefore the maximum likelihood estimator for the parameter β is $\tilde{\beta} = X_{(n)}$. In order to find MLE for the parameter α , one should maximize the function

$$f(\alpha) = C_1 \alpha^n C_2^{\alpha-1},$$

where $C_1 = I_{(0, \tilde{\beta})}(X_{(n)})\tilde{\beta}^{-n}$, $C_2 = \prod_{i=1}^n X_i / \tilde{\beta} = \prod_{i=1}^n X_i / X_{(n)}$. The equation $f'(\alpha) = 0$ gives the MLE of the parameter α :

$$\tilde{\alpha} = \frac{n}{-\log C_2} = \frac{n}{\sum_{i=1}^n \log \frac{X_{(n)}}{X_i}}.$$

So, the MLE is

$$\tilde{\theta} = (\tilde{\alpha}, \tilde{\beta}) = \left(\frac{n}{\sum_{i=1}^n \log \frac{X_{(n)}}{X_i}}, X_{(n)} \right).$$

2. Firstly we compute the first and the second moments:

$$\begin{aligned} m_1(\theta) &= \mathbb{E}_{\theta} X_1 = \int x p(x, \theta) dx = \int_0^{\beta} \frac{\alpha}{\beta^{\alpha}} x^{\alpha} dx = \frac{\alpha\beta}{\alpha+1} \\ m_2(\theta) &= \mathbb{E}_{\theta} X_1^2 = \int x^2 p(x, \theta) dx = \int_0^{\beta} \frac{\alpha}{\beta^{\alpha}} x^{\alpha+1} dx = \frac{\alpha\beta^2}{\alpha+2} \end{aligned}$$

The empirical counterparts are

$$M_1 = \frac{1}{n} \sum_{i=1}^n X_i, \quad M_2 = \frac{1}{n} \sum_{i=1}^n X_i^2. \quad (2.18)$$

The required estimators are the solutions of the system of equations

$$\begin{cases} M_1 = \alpha\beta/(\alpha+1) \\ M_2 = \alpha\beta^2/(\alpha+2) \end{cases} \quad (2.19)$$

Raise both parts of the first equation to the second power and divide it to the second equation:

$$\frac{M_1}{M_2} = \frac{\alpha(\alpha+2)}{(\alpha+1)^2}.$$

This yields the following quadratic equation w.r.t α :

$$\alpha^2 + 2\alpha + \frac{M_1}{M_1 - M_2} = 0. \quad (2.20)$$

If $\frac{M_1}{M_1 - M_2} < 0$ (or equivalently $M_1 < M_2$) then (2.20) has one positive solution

$$\hat{\alpha} = -1 + \sqrt{1 - \frac{M_1}{M_1 - M_2}}.$$

The first equation of system (2.19) gives

$$\hat{\beta} = \frac{\hat{\alpha} + 1}{\hat{\alpha}} M_1.$$

So, the estimate by the method of moments is

$$(\hat{\alpha}, \hat{\beta}) = \left(-1 + \sqrt{1 - \frac{M_1}{M_1 - M_2}}, \quad \frac{\sqrt{1 - \frac{M_1}{M_1 - M_2}}}{-1 + \sqrt{1 - \frac{M_1}{M_1 - M_2}}} M_1 \right),$$

where M_1 and M_2 are given by (2.18).

Exercise 2.18. Let $\{X_i\}_{i=1}^n$ be an i.i.d. sample from a distribution with the Lebesgue density that depends on the parameter $\theta \in \mathbb{R}$ (σ is a fixed positive number):

$$p(x, \theta) = (2\sigma)^{-1} e^{-|x-\theta|/\sigma}.$$

Compute the maximum likelihood estimate for the parameter θ .

This model is known as a shift of a Laplace law.

The maximum likelihood approach leads to maximizing the sum

$$L(\theta) = -n \log(2\sigma) - \sum_{i=1}^n |X_i - \theta|/\sigma,$$

or equivalently to minimizing the sum $\sum_{i=1}^n |X_i - \theta|$:

$$\tilde{\theta} = \underset{\theta}{\operatorname{argmin}} \sum_{i=1}^n |X_i - \theta|.$$

Order the observations $X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}$ and consider two cases.

1. Suppose that n is even. Denote $k = n/2 \in \mathbb{N}$. It is worth mentioning that

$$|X_{(1)} - \theta| + |X_{(n)} - \theta| \geq |X_{(n)} - X_{(1)}|, \quad (2.21)$$

where equality takes place if and only if $\theta \in [X_{(1)}, X_{(n)}]$ Analogously,

$$|X_{(2)} - \theta| + |X_{(n-1)} - \theta| \geq |X_{(n-1)} - X_{(2)}| \quad (2.22)$$

...

$$|X_{(k)} - \theta| + |X_{(k+1)} - \theta| \geq |X_{(k+1)} - X_{(k)}| \quad (2.23)$$

This yields that

$$\begin{aligned} \sum_{i=1}^n |X_i - \theta| &= \sum_{i=1}^n |X_{(i)} - \theta| = \sum_{j=1}^k (|X_{(j)} - \theta| + |X_{(n-j+1)} - \theta|) \\ &\geq \sum_{j=1}^k |X_{(n-j+1)} - X_{(j)}|. \end{aligned} \quad (2.24)$$

Equality in (2.24) takes place if and only if all the inequalities (2.21)–(2.23) are in fact equalities. This means that $\operatorname{argmin} \sum |X_i - \theta|$ is minimized by any $\theta \in [X_{(k)}, X_{(k+1)}]$, in particular by

$$\tilde{\theta} = \operatorname{med} X_i = \frac{X_{(k)} + X_{(k+1)}}{2}.$$

2. Suppose that n is odd. Denote $k = (n - 1)/2 \in \mathbb{N}$. Equalities (2.21)–(2.23) are still true. This yields the analogue for (2.24):

$$\begin{aligned} \sum_{i=1}^n |X_i - \theta| &= \sum_{i=1}^n |X_{(i)} - \theta| = \underbrace{|X_{(k+1)} - \theta|}_{\geq 0} + \sum_{j=1}^k \underbrace{|X_{(j)} - \theta| + |X_{(n-j+1)} - \theta|}_{\geq |X_{(n-j+1)} - X_{(j)}|} \\ &\geq \sum_{j=1}^k |X_{(n-j+1)} - X_{(j)}|. \end{aligned} \quad (2.25)$$

Note that the following two equalities take place only in the case of $\tilde{\theta} = \operatorname{med} X_i = X_{(k+1)}$:

$$\begin{aligned} |X_{(k+1)} - \theta| &= 0 \\ \sum_{j=1}^k [|X_{(j)} - \theta| + |X_{(n-j+1)} - \theta|] &= \sum_{j=1}^k |X_{(n-j+1)} - X_{(j)}| \end{aligned}$$

This completes the proof.

Exercise 2.19. Consider the volatility model with parameter θ :

$$Y = \xi^2, \quad \xi \sim \mathcal{N}(0, \theta).$$

1. Prove that θ is a natural parameter.
2. Find a canonical parameter for this model.
3. Compute the Fisher information for this model with canonical parameter.

1. The proof is straightforward:

$$\mathbb{E}Y = \mathbb{E}\xi^2 = \underbrace{\text{Var}\xi}_{=\theta} + \underbrace{(\mathbb{E}\xi)^2}_{=0} = \theta.$$

2. Denote by $p_\xi(x)$ the pdf of ξ :

$$p_\xi(x) = \frac{1}{\sqrt{2\pi\theta}} \exp\left(-\frac{x^2}{2\theta}\right).$$

The density function of Y can be derived from $p_\xi(x)$:

$$\begin{aligned} p_Y(y, \theta) &= \frac{1}{2\sqrt{y}} p_\xi(\sqrt{y}) = \frac{1}{2\sqrt{2\pi\theta y}} \exp\left(-\frac{y}{2\theta}\right) \\ &= \frac{1}{2\sqrt{2\pi y}} \exp\left(-\frac{y}{2\theta} - \frac{1}{2} \log \theta\right). \end{aligned} \quad (2.26)$$

This density representation means that $C(\theta) = -(2\theta)^{-1}$. The canonical parameter is determined by the equality $v \stackrel{\text{def}}{=} C(\theta)$, i.e. $v = -(2\theta)^{-1}$. This yields

$$p_Y(y, v) = \frac{1}{2\sqrt{2\pi y}} \exp\{yv - d(v)\},$$

where $d(v) = 1/2 \log\{-1/(2v)\}$.

3. According to the general theory,

$$I(v) = d''(v) = \frac{1}{2v^2}.$$

Exercise 2.20. Let (P_v) be a Gaussian shift experiment, that is $P_v = \mathcal{N}(v, 1)$, $v \in \mathbb{R}$. Let $\{X_i\}_{i=1}^n$ be an i.i.d. sample from a distribution P_{v^*} .

1. Is the parameter v a natural parameter? Is it a canonical parameter?
2. Check that

$$\mathcal{K}(v_1, v_2) = (v_1 - v_2)^2 / 2.$$

3. Check that for any v_0 and any $C > 0$, the equation

$$\mathcal{K}(v_0 + u, v_0) = C \quad (2.27)$$

has only one positive (u^+) and only one negative (u^-) solution.

4. Compute the maximum likelihood estimator $\tilde{v}$ and check that

$$L(\tilde{v}, v) = (\tilde{v} - v)^2 n / 2.$$

5. Fix some $\zeta > 0$. Consider the equation (2.27) with $v_0 = v^*$ and $C = \zeta/n$. According to item 3, this equation has two solutions: denote the positive solution by u^+ , and the negative solution by u^- . Denote also $v^+ = v^* + u^+$, and $v^- = v^* + u^-$.

- (a) Compute the sets $\{L(\tilde{v}, v^*) \geq \zeta\}$, $\{L(v^+, v^*) \geq \zeta\}$, $\{L(v^-, v^*) \geq \zeta\}$.
 (b) Check that

$$\{L(\tilde{v}, v^*) \geq \zeta\} \subseteq \{L(v^+, v^*) \geq \zeta\} \cup \{L(v^-, v^*) \geq \zeta\}.$$

Note that the last item is fulfilled for any v^* (not necessary the true value).

1. Parameter v is a natural parameter, because the expected value of a r.v. with distribution $\mathcal{N}(v, 1)$ is equal to v . The parameter v is also a canonical parameter, because the density function can be represented in the following way

$$p(x, v) = \frac{1}{\sqrt{2\pi}} \exp \left\{ -\frac{(x - v)^2}{2} \right\} = p(x) \exp \{xv - d(v)\},$$

where

$$p(x) = \varphi(x), \quad d(v) = \frac{v^2}{2}.$$

2. According to the formula for the canonical parametrization,

$$\mathcal{K}(v_1, v_2) = d'(v_1)(v_1 - v_2) - \{d(v_1) - d(v_2)\}. \quad (2.28)$$

In the case of a Gaussian shift, (2.28) yields

$$\mathcal{K}(v_1, v_2) = v_1(v_1 - v_2) - \frac{v_1^2 - v_2^2}{2} = (v_1 - v_2) \left(v_1 - \frac{v_1 + v_2}{2} \right) = \frac{(v_1 - v_2)^2}{2}.$$

3. The statement is a straightforward corollary from the previous item:

$$\mathcal{K}(u, v_0) = \frac{(v_0 + u - v_0)^2}{2} = \frac{u^2}{2} = C.$$

This equation has two solutions: one positive $u^+ = \sqrt{2C}$ and one negative $u^- = -\sqrt{2C}$.

4. The maximum likelihood approach leads to maximizing the sum

$$L(v) = n \log \frac{1}{2\pi} - \sum_{i=1}^n \frac{(X_i - v)^2}{2}.$$

Then the maximum likelihood estimator is equal to $\tilde{v} = \sum_i X_i / n$. Consider the difference between $L(\tilde{v})$ and $L(v)$:

$$\begin{aligned} L(\tilde{v}, v) &= L(\tilde{v}) - L(v) = - \sum_i \frac{(X_i - \tilde{v})^2}{2} + \sum_i \frac{(X_i - v)^2}{2} \\ &= \frac{1}{2} \sum_i \left\{ (X_i - v)^2 - (X_i - \tilde{v})^2 \right\} = \frac{1}{2} \sum_i (2X_i - \tilde{v} - v) (\tilde{v} - v) \\ &= \frac{1}{2} \left(\underbrace{2 \sum_i X_i}_{2n\tilde{v}} - n\tilde{v} - nv \right) (\tilde{v} - v) = \frac{(\tilde{v} - v)^2 n}{2}. \quad (2.29) \end{aligned}$$

5. (a) Formula (2.29) yields

$$\begin{aligned} \{L(\tilde{v}, v^*) \geq \xi\} &= \left\{ \frac{(\tilde{v} - v^*)^2 n}{2} \geq \xi \right\} \\ &= \left\{ \tilde{v} \geq \sqrt{\frac{2\xi}{n}} + v^* \right\} \cup \left\{ \tilde{v} \leq -\sqrt{\frac{2\xi}{n}} + v^* \right\} \end{aligned}$$

From (2.27) (in item 3) we know that $v^+ = v^* + \sqrt{2\xi/n}$ and $v^- = v^* - \sqrt{2\xi/n}$. Then

$$\begin{aligned} \{L(v^+, v^*) \geq \xi\} &= \left\{ \frac{n}{2} (2\tilde{v} - v^+ - v^*) (\tilde{v} - v^*) \geq \xi \right\} \\ &= \left\{ \frac{n}{2} (2\tilde{v} - 2v^* - \sqrt{\frac{2\xi}{n}}) \sqrt{\frac{2\xi}{n}} \geq \xi \right\} \\ &= \left\{ \tilde{v} \geq v^* + \sqrt{\frac{2\xi}{n}} \right\}. \end{aligned}$$

Analogously,

$$\{L(v^-, v^*) \geq \xi\} \supset \left\{ \tilde{v} \leq v^* - \sqrt{\frac{2\xi}{n}} \right\}.$$

(b) The required embedding is trivial.

Natural parametrization has some “nice” properties:

1.

$$\tilde{\theta} = \frac{1}{n} \sum_{i=1}^n Y_i.$$

2.

$$L(\tilde{\theta}, \theta) = n\mathcal{K}(\mathbb{P}_{\tilde{\theta}}, \mathbb{P}_{\theta}).$$

The following exercise shows, that the choice of parametrization is crucial for the first property, but the second one is fulfilled for any parametrization.

Exercise 2.21. *Let (P_{θ}) be an exponential family (θ – **any** parameter). Let $\{X_i\}_{i=1}^n$ be an i.i.d. sample from distribution that belongs to (P_{θ}) , and X be a random variable with the same distribution.*

Show that the maximum likelihood estimator $\tilde{\theta}$ has the following properties:

1.

$$\mathbb{E}_{\tilde{\theta}} X = \frac{1}{n} \sum_{i=1}^n X_i.$$

2.

$$L(\tilde{\theta}, \theta) = n\mathcal{K}(\mathbb{P}_{\tilde{\theta}}, \mathbb{P}_{\theta}).$$

1. $\tilde{\theta}$ is a point of maximum of the function

$$L(\theta) = \sum_{i=1}^n \log p(X_i, \theta) = C(\theta) \sum_{i=1}^n X_i - nB(\theta).$$

Differentiating w.r.t θ yields the equation for $\tilde{\theta}$:

$$C'(\tilde{\theta}) \sum_{i=1}^n X_i - nB'(\tilde{\theta}) = 0. \quad (2.30)$$

On the other hand, differentiating both sides of the equality

$$\int p(x, \theta) dx = 1$$

w.r.t. θ yields

$$\begin{aligned}
 0 &= \int \frac{\partial}{\partial \theta} \{p(x, \theta)\} dx = \int \frac{\partial}{\partial \theta} \{\log p(x, \theta)\} p(x, \theta) dx \\
 &= \int \{xC'(\theta) - B'(\theta)\} p(x, \theta) dx \\
 &= C'(\theta) \underbrace{\int xp(x, \theta) dx}_{=\mathbb{E}_\theta X} - B'(\theta) \underbrace{\int p(x, \theta) dx}_{=1}.
 \end{aligned}$$

This means that the equality

$$C'(\theta)\mathbb{E}_\theta X - B'(\theta) = 0$$

holds for any parameter θ , in particular for $\theta = \tilde{\theta}$:

$$C'(\tilde{\theta})\mathbb{E}_{\tilde{\theta}} X - B'(\tilde{\theta}) = 0. \quad (2.31)$$

Comparison of the equations (2.30) and (2.31) (using positivity of the first derivative of function $C(\theta)$) completes the proof.

2. Transformation of the left-hand side yields:

$$\begin{aligned}
 L(\tilde{\theta}, \theta) &= \sum_{i=1}^n \left\{ \log p(X_i, \tilde{\theta}) - \log p(X_i, \theta) \right\} \\
 &= \{C(\tilde{\theta}) - C(\theta)\} \sum_{i=1}^n X_i - n\{B(\tilde{\theta}) - B(\theta)\}. \quad (2.32)
 \end{aligned}$$

The Kullback-Leibler divergence in the right-hand side can be transformed in the following way:

$$\begin{aligned}
 \mathcal{K}(\mathbb{P}_{\tilde{\theta}}, \mathbb{P}_\theta) &= \int \log \left\{ \frac{p(x, \tilde{\theta})}{p(x, \theta)} \right\} P_{\tilde{\theta}}(dx) \\
 &= \{C(\tilde{\theta}) - C(\theta)\} \int x P_{\tilde{\theta}}(dx) - \{B(\tilde{\theta}) - B(\theta)\} \\
 &= \{C(\tilde{\theta}) - C(\theta)\} \mathbb{E}_{\tilde{\theta}} X - \{B(\tilde{\theta}) - B(\theta)\}. \quad (2.33)
 \end{aligned}$$

Comparison of the equalities (2.32) and (2.33) using the first item completes the proof.

Exercise 2.22 (Suhov and Kelbert 2005). *There is widespread agreement amongst the managers of the Reliable Motor Company that the number x of faulty cars produced in a month has a binomial distribution*

$$\mathbb{P}(x = s) = \binom{n}{s} p^s (1 - p)^{n-s}, s = 0, 1, \dots, n; 0 \leq p \leq 1.$$

There is, however, some dispute about the parameter p . The general manager has a prior distribution for p which is uniform (i.e. with the pdf $f_p(x) = \mathbf{1}(0 \leq x \leq 1)$), while the more pessimistic production manager has a prior distribution with density $f_p(x) = 2x\mathbf{1}(0 \leq x \leq 1)$. Both pdfs are concentrated on $(0, 1)$.

- (i) In a particular month, s faulty cars are produced. Show that if the general manager's loss function is $(\hat{p} - p)^2$, where $\hat{p}$ is her estimate and p is the true value, then her best estimate of p is

$$\hat{p} = \frac{s + 1}{n + 2}$$

- (ii) The production manager has responsibilities different from those of the general manager, and a different loss function given by $(1 - p)(\hat{p} - p)^2$. Find his best estimator of p and show that it is greater than that of the general manager unless $s \geq n/2$.

You may assume that, for non-negative integers α, β ,

$$\int_0^1 p^\alpha (1 - p)^\beta dp \approx \frac{\alpha! \beta!}{(\alpha + \beta + 1)!}$$

As $\mathbb{P}_p(X = s) = \alpha p^s (1 - p)^{n-s}$, $s = 0, 1, \dots, n$, the posterior for the general manager (GM) is

$$\pi^{GM}(p | s) = \alpha p^s (1 - p)^{n-s} \mathbf{1}(0 < p < 1),$$

and for the production manager (PM)

$$\pi^{PM}(p | s) = \alpha p p^s (1 - p)^{n-s} \mathbf{1}(0 < p < 1).$$

Then the expected loss for the GM is minimized at the posterior mean:

$$\begin{aligned} \hat{p}^{GM} &= \frac{\int_0^1 p p^s (1 - p)^{n-s} dp}{\int_0^1 p^s (1 - p)^{n-s} dp} \\ &= \frac{(s + 1)!(n - s)!}{(n - s + s + 2)!} \frac{(n - s + s + 1)!}{s!(n - s)!} = \frac{s + 1}{n + 2}. \end{aligned}$$

For the PM, the expected loss

$$\int_0^1 (1 - p) (p - a)^2 \pi^{PM}(p | s) dp$$

is minimized at

$$a = \frac{\int_0^1 p(1-p)\pi^{PM}(p|s) dp}{\int_0^1 (1-p)\pi^{PM}(p|s) dp},$$

which yields

$$\begin{aligned} \hat{p}^{PM} &= \frac{\int_0^1 p(1-p)pp^s(1-p)^{n-s} dp}{\int_0^1 p(1-p)pp^s(1-p)^{n-s} dp} \\ &= \frac{(s+2)!(n-s+1)!}{(n-s+s+4)!} \frac{(n-s+s+3)!}{(s+1)!(n-s+1)!} = \frac{s+2}{n+4}. \end{aligned}$$

We see that $(s+2)/(n+4) > (s+1)/(n+2)$, i.e., $s < n/2$.

Exercise 2.23. Denote the number of incoming telecom signals between $[0, t]$ as $C(0, t)$. Assume that $C(0, t)$ satisfies

- (a) The number of arrivals in disjoint time intervals are independent;
- (b) The distribution of $C(s, t)$ depends on $t - s$;
- (c) For $h > 0$ small, $P\{C(0, h) = 1\} = \lambda h + o(h)$, where $\lambda > 0$ is a constant;
- (d) $P\{C(0, h) \geq 2\} = o(h)$.

Please answer the following questions:

1. Prove that $C(0, t)$ follows a Poisson distribution with mean λt .
2. Find the function $p(y)$, $C(\theta)$ and $B(\theta)$ of the natural parametrization

$$p(y, \theta) \stackrel{\text{def}}{=} p(y)e^{yC(\theta)-B(\theta)}$$

and function $d(\theta)$ of the canonical parametrization

$$p(y, \theta) \stackrel{\text{def}}{=} e^{y\theta-d(\theta)}$$

for this Poisson distribution with mean λt .

3. Find an estimator for constant λ .

1. Let $X_m^n = C\{(m-1)t/n, mt/n\}$, $1 \leq m \leq n$, X_m^n are i.i.d. by assumption (a). Define Y_m^n be i.i.d. Bernoulli random variable such that $Y_m^n = 1$ with probability $/n$, $1 \leq m \leq n$. Define

$$S_n = X_1^n + \dots + X_n^n$$

and

$$T_n = Y_1^n + \dots + Y_n^n.$$

Suppose $P\{C(0, h) = 1\} = \lambda h + g_1(h)$ and $P\{C(0, h) \geq 2\} = g_2(h)$ where $g_1(h)$ and $g_2(h)$ are of order $\mathcal{O}(h)$. We claim the following lemma:

Lemma 2.1. *Let $a_1, \dots, a_n$ and $b_1, \dots, b_n$ be complex numbers with modulus $\leq c$, then*

$$\left| \prod_{m=1}^n a_m - \prod_{m=1}^n b_m \right| \leq c^{n-1} \sum_{m=1}^n |a_m - b_m|.$$

The proof of this simple lemma is left to the reader (hint: use induction). The modulus of $\varphi_Y(\xi) = \exp(\mathbf{i}Y_m^n \xi)$ and $\varphi_{X_m^n}(\xi) = \exp(\mathbf{i}X_m^n \xi)$ are less than 1, $|\varphi_{X_m^n}(\xi) - \varphi_Y(\xi)| \leq 2g_1(t/n) + 2g_2(t/n)$ (verify!). By the lemma,

$$\begin{aligned} & |\mathbb{E} \exp(\mathbf{i}T_n \xi) - \mathbb{E} \exp(\mathbf{i}S_n \xi)| \\ &= \left| \prod_{m=1}^n \varphi_{X_m^n}(\xi) - \prod_{m=1}^n \varphi_{Y_m}(\xi) \right| \\ &\leq \sum_{m=1}^n |\varphi_{X_m^n}(\xi) - \varphi_Y(\xi)| \\ &\leq \sum_{m=1}^n 2 \left\{ \left| g_1\left(\frac{t}{n}\right) \right| + \left| g_2\left(\frac{t}{n}\right) \right| \right\} \\ &\rightarrow 0, \quad \text{as } n \rightarrow \infty. \end{aligned}$$

Now we show that $\mathbb{E} \exp(\mathbf{i}T_n \xi) \rightarrow \exp\{\lambda t(\exp(\mathbf{i}\xi) - 1)\}$, the characteristic function of the Poisson distribution with mean λt and finish the proof. Observe that $|\mathbb{E} \exp(\mathbf{i}Y_m^n \xi)| = (1 - \lambda t/n) + (\lambda t/n) \exp(\mathbf{i}\xi) = 1 + (\lambda t/n)\{\exp(\mathbf{i}\xi) - 1\}$ and $|\exp(\mathbf{i}\xi) - 1| \leq 2$. When n large, $\lambda t/n \leq 1/2$. Using the lemma again,

$$\begin{aligned} & \left| \exp(\lambda t\{\exp(\mathbf{i}\xi) - 1\}) - \prod_{m=1}^n [1 + (\lambda t/n)\{\exp(\mathbf{i}\xi) - 1\}] \right| \\ &\leq \sum_{m=1}^n \left| \exp\left[\frac{\lambda t}{n}\{\exp(\mathbf{i}\xi) - 1\}\right] - \left[1 + \frac{\lambda t}{n}\{\exp(\mathbf{i}\xi) - 1\}\right] \right| \\ &\leq \sum_{m=1}^n \left(\frac{\lambda t}{n}\right)^2 |\exp(\mathbf{i}\xi) - 1|^2 \\ &\leq 4 \left(\frac{\lambda t}{n}\right) \lambda t \\ &\rightarrow 0, \end{aligned}$$

as $n \rightarrow \infty$. This finishes the proof.

2. The Poisson density with mean λt is

$$p(y, \lambda t) = \exp(-\lambda t)(\lambda t)^y / y!.$$

The $p(y)$, $C(\lambda)$, $B(\lambda)$ of the natural parametrization is

$$\begin{aligned} p(y) &= \frac{1}{y!}; \\ C(\lambda t) &= \log(\lambda t); \\ B(\lambda t) &= \lambda t. \end{aligned}$$

The $d(\lambda t)$ for the canonical parametrization is

$$d(\lambda t) = -\lambda t(y + 1) + y \log(\lambda t) - \log y!.$$

3. Suppose we have an observation of the number of signal y between time 0 and t .

The maximizer for the log natural parametrization is $\hat{\lambda} = y/t$.

Exercise 2.24. Let Y be an i.i.d. sample from $P_{\theta^*} \in (P_\theta)$, where (P_θ) is a regular parametric family. The fundamental exponential bound for the maximum likelihood is given by the fact that for any $0 < \varrho < 1$, $0 < s < 1$, $\mu > 0$, the log-likelihood process $L(\theta, \theta^*)$ fulfills for a fixed constant $\varOmega(\varrho, s)$

$$\mathbb{E} \exp \left[\varrho \sup_{\theta \in \Theta} \{ \mu L(\theta, \theta^*) + s \mathcal{M}(\mu, \theta, \theta^*) \} \right] \leq \varOmega(\varrho, s), \quad (2.34)$$

see Spokoiny and Dickhaus (2014). Denote the set $\mathcal{A}(\mathfrak{z}, \theta^*) = \{ \theta : \mathcal{M}(\mu, \theta, \theta^*) \leq \mathfrak{z} \}$, where $\mathfrak{z}$ is positive, and $\mathcal{M}(\mu, \theta, \theta^*)$ is the rate function defined for $\mu > 0$ by

$$\mathcal{M}(\mu, \theta, \theta^*) \stackrel{\text{def}}{=} -\log \mathbb{E}_{\theta^*} \exp \{ \mu L(\theta, \theta^*) \}.$$

Using (2.34), prove that for any $\varrho' < \varrho$,

1.

$$\mathbb{E} \left[\exp \left\{ \varrho' s \mathcal{M}(\mu, \tilde{\theta}, \theta^*) \right\} \mathbf{1} \left\{ \tilde{\theta} \notin \mathcal{A}(\mathfrak{z}, \theta^*) \right\} \right] \leq \varOmega(\varrho, s) \exp \{ -(\varrho - \varrho') s \mathfrak{z} \};$$

in particular,

$$\mathbb{P} \{ \tilde{\theta} \notin \mathcal{A}(\mathfrak{z}, \theta^*) \} \leq \varOmega(\varrho, s) \exp(-\varrho s \mathfrak{z}).$$

2.

$$\mathbb{E} \left[\mathcal{M}(\mu, \tilde{\theta}, \theta^*) \mathbf{1}(\tilde{\theta} \notin \mathcal{A}(\mathfrak{z}, \theta^*)) \right] \leq \frac{1}{\varrho' s} \varOmega(\varrho, s) \exp \{ -(\varrho - \varrho') s \mathfrak{z} \}.$$

1. The inequalities $L(\tilde{\theta}, \theta^*) \geq 0$ and $\mathcal{M}(\mu, \tilde{\theta}, \theta^*) > \mathfrak{z}$ for $\tilde{\theta} \notin \mathcal{A}(\mathfrak{z}, \theta^*)$ imply

$$\begin{aligned}
 & \mathbb{E} \left[\exp\{(\varrho - \varrho')s\mathfrak{z}\} \exp\{\varrho's \mathcal{M}(\mu, \tilde{\theta}, \theta^*)\} \mathbf{1}\{\tilde{\theta} \notin \mathcal{A}(\mathfrak{z}, \theta^*)\} \right] \\
 & \leq \mathbb{E} \left[\exp\{\varrho s \mathcal{M}(\mu, \tilde{\theta}, \theta^*)\} \mathbf{1}\{\tilde{\theta} \notin \mathcal{A}(\mathfrak{z}, \theta^*)\} \right] \\
 & \leq \mathbb{E} \left[\exp\{\varrho s \mathcal{M}(\mu, \tilde{\theta}, \theta^*)\} \right] \\
 & \leq \mathbb{E} \left[\exp\{\varrho \mu L(\tilde{\theta}, \theta^*) + \varrho s \mathcal{M}(\mu, \tilde{\theta}, \theta^*)\} \right] \\
 & \leq \mathfrak{Q}(\varrho, s),
 \end{aligned}$$

and the assertion follows.

2. The second item directly follows from the first one, because $x < e^x$ for any positive x .

Exercise 2.25. Consider a multivariate normal rv $\mathbf{Y} \sim \mathcal{N}(\theta^*, \Sigma)$, where $\Sigma = (nD^2)^{-1}$ for some matrix D . In other words, $\mathbf{Y} = \theta^* + \xi$ with $\xi \sim \mathcal{N}\{0, (nD^2)^{-1}\}$.

1. Check that the log-likelihood ratio computed on one observation of $\mathbf{Y}$ is equal to

$$L(\theta, \theta^*) = n(\theta - \theta^*)^\top D^2 \xi - n\|D(\theta - \theta^*)\|^2/2. \quad (2.35)$$

2. Prove that the r.v. ξ is equal to

$$\xi = (nD^2)^{-1} \nabla L(\theta^*).$$

1. The log-likelihood is equal to

$$\begin{aligned}
 & L(\theta, \theta^*) \\
 & = L(\theta) - L(\theta^*) \\
 & = -\frac{1}{2}(\mathbf{Y} - \theta)^\top \Sigma^{-1}(\mathbf{Y} - \theta) + \frac{1}{2}(\mathbf{Y} - \theta^*)^\top \Sigma^{-1}(\mathbf{Y} - \theta^*) \\
 & = -\frac{1}{2}(\mathbf{Y} - \theta^* + \theta^* - \theta)^\top \Sigma^{-1}(\mathbf{Y} - \theta^* + \theta^* - \theta) \\
 & \quad + \frac{1}{2}(\mathbf{Y} - \theta^*)^\top \Sigma^{-1}(\mathbf{Y} - \theta^*) \\
 & = -(\theta^* - \theta)^\top \Sigma^{-1}(\mathbf{Y} - \theta^*) - \frac{1}{2}(\theta^* - \theta)^\top \Sigma^{-1}(\theta^* - \theta).
 \end{aligned}$$

To conclude the proof, it is sufficient to note that

$$\Sigma^{-1}(\mathbf{Y} - \theta^*) = nD^2 \xi,$$

and

$$\begin{aligned} \frac{1}{2}(\boldsymbol{\theta}^* - \boldsymbol{\theta})^\top \Sigma^{-1}(\boldsymbol{\theta}^* - \boldsymbol{\theta}) &= \frac{1}{2}(\boldsymbol{\theta}^* - \boldsymbol{\theta})^\top nD^2(\boldsymbol{\theta}^* - \boldsymbol{\theta}) \\ &= n\|D(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|^2/2. \end{aligned}$$

2. The proof is straightforward:

$$\begin{aligned} \nabla L(\boldsymbol{\theta}^*) &= \nabla \left\{ -\frac{n}{2} \log |2\pi \Sigma| - \frac{1}{2}(\mathbf{Y} - \boldsymbol{\theta}^*)^\top \Sigma^{-1}(\mathbf{Y} - \boldsymbol{\theta}^*) \right\} \\ &= \Sigma^{-1}(\mathbf{Y} - \boldsymbol{\theta}^*) = nD^2 \boldsymbol{\xi}. \end{aligned}$$

Exercise 2.26. Consider the model from the previous exercise, $\mathbf{Y} \sim \mathcal{N}(\boldsymbol{\theta}^*, \Sigma)$ with $\Sigma = (nD^2)^{-1}$ for some matrix D .

Using the formula (2.35), simulate the log-likelihood ratio for $D^2 = \begin{pmatrix} 3 & 1 \\ 1 & 2 \end{pmatrix}$,

$$\boldsymbol{\theta}^* = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \in \mathbb{R}^2 \text{ and } \boldsymbol{\theta} = \boldsymbol{\theta}_1 \stackrel{\text{def}}{=} \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \boldsymbol{\theta} = \boldsymbol{\theta}_2 \stackrel{\text{def}}{=} \begin{pmatrix} 1.2 \\ 1 \end{pmatrix}.$$

Draw a plot for $L(\boldsymbol{\theta}, \boldsymbol{\theta}^*)$ as a function of $\boldsymbol{\xi}$ and a plot for an estimator of the density function of $L(\boldsymbol{\theta}, \boldsymbol{\theta}^*)$.

Define $\mu \stackrel{\text{def}}{=} -n\|D(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|^2/2$, and note that $nD^2 \boldsymbol{\xi} \sim \mathcal{N}(0, nD^2)$. Therefore, by formula (2.35), the rv $L(\boldsymbol{\theta}, \boldsymbol{\theta}^*)$ has the distribution

$$L(\boldsymbol{\theta}, \boldsymbol{\theta}^*) \sim \mathcal{N}\{\mu, (\boldsymbol{\theta} - \boldsymbol{\theta}^*)^\top (nD^2)(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\}.$$

The square root of D^2 can be found via the Jordan decomposition $D^2 = \Gamma \Lambda \Gamma^\top$, where Γ is the eigenvector matrix and Λ is the diagonal matrix of eigenvalues of D^2 . In our case, the diagonal entries of the matrix Λ are $\lambda_1 = (5 + \sqrt{5})/2$ and $\lambda_2 = (5 - \sqrt{5})/2$.

Figure 2.5 describes the simulation of the r.v. $\boldsymbol{\xi}$ for $n = 1,000$, $\boldsymbol{\theta} = \boldsymbol{\theta}_1$ and $\boldsymbol{\theta} = \boldsymbol{\theta}_2$.

Exercise 2.27 (Shao 2005). Let $(X_1, \dots, X_n)$ be a random sample from a distribution on $\mathbb{R}$ with the Lebesgue density $\frac{1}{\sigma} f(\frac{x-\mu}{\sigma})$, where $f(x) > 0$ is a known Lebesgue density and $f'(x)$ exists for all $x \in \mathbb{R}$, $\mu \in \mathbb{R}$, and $\sigma > 0$. Let $\boldsymbol{\theta} = (\mu, \sigma)$. Show that the Fisher information about $\boldsymbol{\theta}$ contained in $X_1, \dots, X_n$ is

$$I(\boldsymbol{\theta}) = \frac{n}{\sigma^2} \begin{pmatrix} \int \frac{\{f'(x)\}^2}{f(x)} dx & \int \frac{f'(x)\{xf'(x)+f(x)\}}{f(x)} dx \\ \int \frac{f'(x)\{xf'(x)+f(x)\}}{f(x)} dx & \int \frac{\{xf'(x)+f(x)\}^2}{f(x)} dx \end{pmatrix},$$

assuming that all integrals are finite.

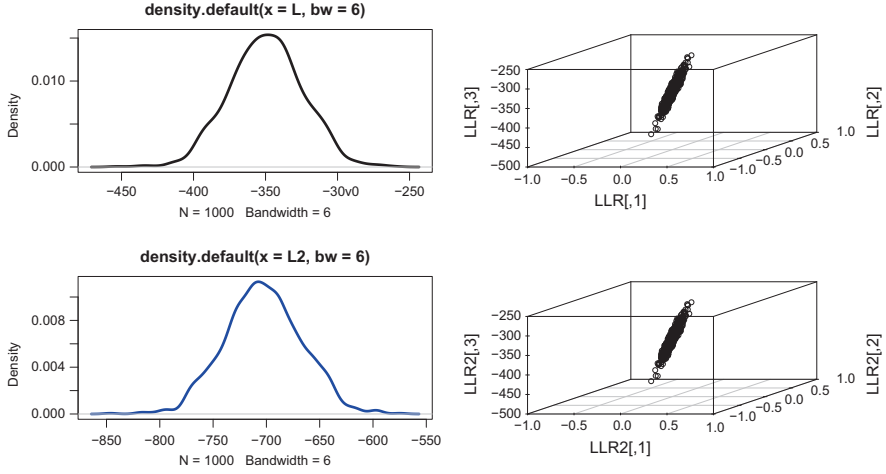


Fig. 2.5 Plots of density estimator and log-likelihood ratio function.  MSEloglikelihood

Denote $g(\mu, \sigma, x) \stackrel{\text{def}}{=} \log \frac{1}{\sigma} f\left(\frac{x-\mu}{\sigma}\right)$. Then

$$\frac{\partial}{\partial \mu} g(\mu, \sigma, x) = -\frac{f'\left(\frac{x-\mu}{\sigma}\right)}{\sigma f\left(\frac{x-\mu}{\sigma}\right)}$$

$$\frac{\partial}{\partial \sigma} g(\mu, \sigma, x) = -\frac{(x-\mu) f'\left(\frac{x-\mu}{\sigma}\right)}{\sigma f\left(\frac{x-\mu}{\sigma}\right)} - \frac{1}{\sigma}.$$

By the direct computation,

$$\begin{aligned} \mathbb{E} \left\{ \frac{\partial}{\partial \mu} g(\mu, \sigma, X_1) \right\}^2 &= \frac{1}{\sigma^2} \int \left\{ \frac{f'\left(\frac{x-\mu}{\sigma}\right)}{f\left(\frac{x-\mu}{\sigma}\right)} \right\}^2 \frac{1}{\sigma} f\left(\frac{x-\mu}{\sigma}\right) dx \\ &= \frac{1}{\sigma^2} \int \frac{\{f'\left(\frac{x-\mu}{\sigma}\right)\}^2}{f\left(\frac{x-\mu}{\sigma}\right)} d\left(\frac{x-\mu}{\sigma}\right) \\ &= \frac{1}{\sigma^2} \int \frac{\{f'(x)\}^2}{f(x)} dx, \end{aligned}$$

$$\mathbb{E} \left\{ \frac{\partial}{\partial \sigma} g(\mu, \sigma, X_1) \right\}^2 = \frac{1}{\sigma^2} \int \left\{ \frac{x-\mu}{\sigma} \frac{f'\left(\frac{x-\mu}{\sigma}\right)}{f\left(\frac{x-\mu}{\sigma}\right)} + 1 \right\}^2 \frac{1}{\sigma} f\left(\frac{x-\mu}{\sigma}\right) dx$$

$$\begin{aligned}
&= \frac{1}{\sigma^2} \int \left\{ x \frac{f'(x)}{f(x)} + 1 \right\}^2 f(x) dx \\
&= \frac{1}{\sigma^2} \int \frac{\{x f'(x) + f(x)\}^2}{f(x)} dx,
\end{aligned}$$

and

$$\begin{aligned}
&\mathbb{E} \left\{ \frac{\partial}{\partial \mu} g(\mu, \sigma, x) \frac{\partial}{\partial \sigma} g(\mu, \sigma, x) \right\} \\
&= \frac{1}{\sigma^2} \int \frac{f' \left(\frac{x-\mu}{\sigma} \right)}{f \left(\frac{x-\mu}{\sigma} \right)} \left\{ \frac{x-\mu}{\sigma} \frac{f' \left(\frac{x-\mu}{\sigma} \right)}{f \left(\frac{x-\mu}{\sigma} \right)} + 1 \right\} \frac{1}{\sigma} f \left(\frac{x-\mu}{\sigma} \right) dx \\
&= \frac{1}{\sigma^2} \int \frac{f'(x) \{x f'(x) + f(x)\}}{f(x)} dx.
\end{aligned}$$

The result follows since

$$I(\theta) = n \mathbb{E} \left\{ \frac{\partial}{\partial \theta} \log \frac{1}{\sigma} f \left(\frac{X_1 - \mu}{\sigma} \right) \right\} \left\{ \frac{\partial}{\partial \theta} \log \frac{1}{\sigma} f \left(\frac{X_1 - \mu}{\sigma} \right) \right\}^\top.$$

Exercise 2.28 (Shao 2005). Let X be a random variable having a cumulative distribution function F . Show that if $\mathbb{E}X$ exists, then

$$\mathbb{E}X = \int_0^\infty \{1 - F(x)\} dx - \int_{-\infty}^0 F(x) dx.$$

By Fubini's theorem,

$$\begin{aligned}
\int_0^\infty \{1 - F(x)\} dx &= \int_0^\infty \int_{(x, \infty)} dF(y) dx \\
&= \int_0^\infty \int_{(0, y)} dx dF(y) \\
&= \int_0^\infty y dF(y).
\end{aligned}$$

Similarly,

$$\int_{-\infty}^0 F(x) dx = \int_{-\infty}^0 \int_{(-\infty, x]} dF(y) dx = - \int_{-\infty}^0 y dF(y).$$

If $\mathbb{E}X$ exists, then at least one of $\int_0^\infty y dF(y)$ and $\int_{-\infty}^0 y dF(y)$ is finite and

$$\mathbb{E}X = \int_{-\infty}^\infty yF(y) = \int_0^\infty \{1 - F(x)\} dx - \int_{-\infty}^0 F(x) dx.$$

Exercise 2.29 (Shao 2005). Let $(X_1, \dots, X_n)$ be a random sample from the exponential distribution on (a, ∞) with scale parameter 1, where $a \in \mathbb{R}$ is unknown.

1. Construct $(1 - \alpha)$ – confidence interval for a using the cumulative distribution function of the smallest order statistic $X_{(1)}$.
 2. Show that the confidence interval in (i) can also be obtained using a pivotal quantity.
1. The cumulative distribution function of $X_{(1)}$ is

$$F_a(t) = \begin{cases} 0 & t \leq a \\ 1 - \exp^{-n(t-a)} & t > a, \end{cases}$$

which is decreasing in a for fixed $t > a$. A $(1 - \alpha)$ – confidence interval for a has upper limit equal to the unique solution of $F_a(T) = \alpha_1$ and lower limit equal to the unique solution of $F_a(T) = 1 - \alpha_2$, where $\alpha_1 + \alpha_2 = \alpha$. Then, $[T + n^{-1} \log(\alpha_2), T + n^{-1} \log(1 - \alpha_1)]$ is the resulting confidence interval.

2. Note that $W(a) = n(X_{(1)} - a)$ has the exponential distribution on $(0, \infty)$ with scale parameter 1. Therefore the distribution of $W(a)$ doesn't depend on the parameter and, hence, $W(a)$ is a pivotal quantity. The $1 - \alpha$ confidence interval for a constructed this random variable is the same as that derived in item (i).

Exercise 2.30 (Shao 2005). Let F_n be the edf based on a random sample of size n from cdf F on $\mathbb{R}$ having Lebesgue density f . Let $\varphi_n(t)$ be the Lebesgue density of the p th sample quantile $F_n^{-1}(p)$.

Denote by m_p the integer part of np . Introduce also the quantity ℓ_p , which is equal to m_p if np is an integer and is equal to $m_p + 1$ if np is not an integer. Prove that

$$\varphi_n(t) = n \binom{n-1}{\ell_p-1} \{F(t)\}^{\ell_p-1} \{1 - F(t)\}^{n-\ell_p} f(t),$$

1. Using the fact that $nF_n(t)$ has a binomial distribution;
2. Using the Lebesgue density of the j -th order statistic.

1. Since $nF_n(t)$ has the binomial distribution with size n and probability $F(t)$, for any $t \in \mathbb{R}$,

$$\begin{aligned} \mathbb{P}\{F_n^{-1}(p) \leq t\} &= \mathbb{P}\{F_n(t) \geq p\} \\ &= \sum_{i=\ell_p}^n \binom{n}{i} \{F(t)\}^i \{1 - F(t)\}^{n-i}. \end{aligned}$$

Differentiating term by term leads to

$$\begin{aligned}
 \varphi_n(t) &= \sum_{i=l_p}^n \binom{n}{i} i \{F(t)\}^{i-1} \{1-F(t)\}^{n-i} f(t) \\
 &\quad - \sum_{i=l_p}^n \binom{n}{i} (n-i) \{F(t)\}^i \{1-F(t)\}^{n-i-1} f(t) \\
 &= \binom{n}{l_p} l_p \{F(t)\}^{l_p-1} \{1-F(t)\}^{n-l_p} f(t) \\
 &\quad + n \sum_{i=l_p+1}^n \binom{n-1}{i-1} \{F(t)\}^{i-1} \{1-F(t)\}^{n-i} f(t) \\
 &\quad - n \sum_{i=l_p}^{n-1} \binom{n-1}{i} \{F(t)\}^i \{1-F(t)\}^{n-i-1} f(t) \\
 &= n \binom{n-1}{l_p-1} \{F(t)\}^{l_p-1} \{1-F(t)\}^{n-l_p} f(t).
 \end{aligned}$$

2. The Lebesgue density of the j -th order statistic is

$$n \binom{n-1}{j-1} \{F(t)\}^{j-1} \{1-F(t)\}^{n-j} f(t).$$

Then, the result follows from the fact that

$$F_n^{-1}(p) = \begin{cases} X_{(m_p)} & \text{if } np \text{ is an integer,} \\ X_{(m_p+1)} & \text{if } np \text{ is not an integer.} \end{cases}$$

Exercise 2.31. Consider samples $\{Y_i\}_{i=1}^n$, where Y_i are i.i.d. with distribution function $F_Y(y)$. We want to estimate the τ th quantile of the distribution function $F_Y^{-1}(\tau)$:

$$F_Y^{-1}(\tau) \stackrel{\text{def}}{=} \inf \{y \in \mathbb{R} : \tau \leq F_Y(y)\}.$$

This problem can be seen as in a location model:

$$Y_i = \theta^* + \varepsilon_i, \quad \varepsilon_i \sim \text{ALD}(\tau),$$

where $F_{\varepsilon}^{-1}(\tau) = 0$ and ε_i 's are i.i.d. The QMLE estimation follows the framework with ALD likelihood, where ALD stands for "Asymmetric Laplace Distribution", and has probability density function

$$f(u|\tau) = \tau(1 - \tau)\exp\{-\rho_{\tau}(u)\},$$

with $\rho_{\tau}(u) = u\{\tau\mathbf{1}(u \geq 0) - (1 - \tau)\mathbf{1}(u < 0)\}$.

1. Prove that

$$\operatorname{argmin}_{\theta} \mathbb{E}\rho_{\tau}(Y_i - \theta) = F_Y^{-1}(\tau) = \theta^*. \quad (2.37)$$

2. Please write the empirical loss function for the estimation of $F_Y^{-1}(\tau)$.

1. To prove (2.37),

$$\begin{aligned} & \frac{\partial \mathbb{E}\rho_{\tau}(Y_i - \theta)}{\partial \theta} \\ &= \frac{\partial \int \{\tau(Y_i - \theta)\mathbf{1}(Y_i - \theta > 0) - (1 - \tau)\mathbf{1}(Y_i - \theta \leq 0)\} dF_Y(u)}{\partial \theta} \\ &= -\tau \theta f_Y(\theta) - \tau \{1 - F_Y(\theta)\} + \tau \theta f(\theta) - (1 - \tau) f_Y(\theta) + (1 - \tau)(F_Y(\theta) + \theta f_Y(\theta)) \\ &= (1 - \tau)F_Y(\theta) - \tau \{1 - F_Y(\theta)\} \\ &= F_Y(\theta) - \tau \end{aligned}$$

Solve

$$\frac{\partial \mathbb{E}\rho_{\tau}(Y_i - \theta)}{\partial \theta} = 0,$$

we get

$$F(\theta^*) = \tau.$$

Thus, $\theta^* = F_Y^{-1}(\tau)$.

2. An estimator of θ^* would be

$$\operatorname{argmin}_{\theta} \sum_{i=1}^n \{\tau \mathbf{1}(Y_i > \theta) - (1 - \tau) \mathbf{1}(Y_i < \theta)\}.$$

Exercise 2.32. Consider samples $\{(X_i, Y_i)\}_{i=1}^n$ i.i.d., in a regression framework, we now want to estimate the conditional τ th quantile of the conditional distribution function $F_{Y|X}^{-1}(\tau)$. If we believe in the following linear model:

$$Y_i = X_i^\top \theta^* + \epsilon_i, \quad \epsilon_i \sim \text{ALD}(\tau),$$

where $F_{\epsilon|X}^{-1}(\tau) = 0$ and ϵ_i s are i.i.d. Similarly we take a QMLE in an ALD likelihood.

1. Prove that

$$\theta^* = \underset{\theta}{\operatorname{argmin}} \mathbb{E}_{Y|X} \rho_\tau(Y_i - X_i^\top \theta) \quad (2.38)$$

$$F_{Y|X_i}^{-1}(\tau) = X_i^\top \theta^* \quad (2.40)$$

2. Suppose now $\{(X_i, Y_i)\}_{i=1}^n$ is a bivariate i.i.d. sequence from a joint normal distribution $N(\mu, \Sigma)$, where

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix}.$$

Please write down the theoretical form of $F_{Y|X}^{-1}(\tau)$. (Hint: Observe that the conditional distribution is again normally distributed, with $\mu_{Y|X=x} = \mu_1 + \sigma_{12}\sigma_{22}^{-1}(x - \mu_2)$ and $\sigma_{Y|X} = \sigma_{11} - \sigma_{12}^2/\sigma_{22}$.)

1. To prove (2.40),

$$\begin{aligned} & \frac{\partial \mathbb{E} \rho_\tau(Y_i - X_i^\top \theta)}{\partial \theta_j} \\ &= \frac{\partial \int \{\tau(Y_i - X_i^\top \theta) \mathbf{1}(Y_i - X_i^\top \theta > 0)\} dF_{Y|X}(u)}{\partial \theta_j} \\ & \quad - \frac{(1 - \tau) \int \{(Y_i - X_i^\top \theta) \mathbf{1}(Y_i - X_i^\top \theta \leq 0)\} dF_{Y|X}(u)}{\partial \theta_j} \\ &= -\tau X_{ij} X_i^\top \theta f_Y(X_i^\top \theta) - X_{ij} \tau \{1 - F_Y(X_i^\top \theta)\} + X_{ij} \tau X_i^\top \theta f_Y(X_i^\top \theta) \\ & \quad - (1 - \tau) X_{ij} f_Y(X_i^\top \theta) + (1 - \tau) X_{ij} (F_Y(X_i^\top \theta) + X_i^\top \theta f_Y(X_i^\top \theta)) \\ &= (1 - \tau) X_{ij} F_{Y|X}(X_i^\top \theta) - \tau X_{ij} \{1 - F_{Y|X}(X_i^\top \theta)\} \\ &= X_{ij} F_{Y|X}(X_i^\top \theta) - \tau X_{ij} \end{aligned}$$

Solve

$$\frac{\partial \mathbb{E}_{Y|X} \rho_\tau(Y_i - X_i^\top \theta)}{\partial \theta_j} = 0, \forall j \in 1, \dots, d$$

we get

$$F_{Y|X}(X_i^\top \theta^*) = \tau, \forall i, 1, \dots, n$$

Thus, $F_{Y|X_i}^{-1}(\tau) = X_i^\top \theta^*$.

2. Use the hint, we have the normal conditional distribution. Given $X = x$, $(Y_i - u_{Y|X=x})/\sigma_{Y|X} \sim \mathbf{N}(0, 1)$. Denote $\Phi^{-1}(\tau)$ as the τ th quantile of a standard normal distribution. Then we have,

$$F_{Y|X=x}^{-1}(\tau) = \sigma_{Y|X} \Phi^{-1}(\tau) + u_{Y|X=x}$$

References

- Shao, J. (2005). *Mathematical statistics: Exercises and solutions*. New York: Springer
- Spokoiny, V., & Dickhaus, T. (2014). *Basics of modern parametric statistics*. Berlin: Springer.
- Suhov, Y., & Kelbert, M. (2005). *Probability and statistics by example, 1 basic probability and statistics*. New York: Cambridge University Press.

Chapter 3

Parameter Estimation for a Regression Model

Bir regresyon modeli için parametre tahmini

Ayağını yorganına göre uzat.

Stretch your legs according to the length of the quilt.

Exercise 3.1. Let a regression function $f(\cdot)$ be represented by a linear combination of basis functions $\Psi_1(\cdot), \dots, \Psi_p(\cdot)$.

Suppose that for $x \in \mathbb{R}^d$ the regression function $f(\cdot)$ is quadratic in x . Describe the basis and the corresponding vector of coefficients in these cases.

The function $f(\cdot)$ being quadratic in x , means when $d = 1$,

$f(x) = \theta_1 + \theta_2 x + \theta_3 x^2$, which obviously leads to

$$\Psi_1(x) = 1, \quad \Psi_2(x) = x, \quad \Psi_3(x) = x^2$$

When $d > 1$, $f(x) = \theta_1 + A^\top x + x^\top Bx$, where $A \in \mathbb{R}^d$, B is a $d \times d$ matrix. Then we can write

$$f(x) = \theta_1 + \sum_{j=1}^d A_j x_j + \sum_{j=1}^d \sum_{k=1}^d B_{jk} x_j x_k \quad (3.1)$$

where A_j is the j th element in A , and B_{jk} is the element in j th row and k th column of B .

Define $e_j = \underbrace{(0, \dots, 1, \dots, 0)}_j^\top_{d \times 1}$, the j th unit vector. The second term in (3.1) can

be rewritten as: $\sum_{j=1}^d A_j e_j^\top x$, similarly the quadratic term in (3.1) can be rewritten as: $\sum_{j=1}^d \sum_{k=1}^d e_j^\top x x^\top e_k B_{jk}$.

Defining now: $\Psi_1(x) = 1$, $\Psi_2(x) = e_1^\top x$, $\Psi_3(x) = e_2^\top x, \dots$

$$\begin{aligned}\Psi_{d+2}(x) &= e_1^\top x x^\top e_1 = x_1^2, \\ \Psi_{d+3}(x) &= e_1^\top x x^\top e_2 = x_1 x_2, \\ &\vdots \\ \Psi_{2d+2}(x) &= e_2^\top x x^\top e_2 = x_2^2, \\ \Psi_{2d+3}(x) &= e_2^\top x x^\top e_3 = x_2 x_3, \\ &\vdots \\ \Psi_{\frac{d^2}{2} + \frac{3d}{2} + 2}(x) &= e_d^\top x x^\top e_d = x_d^2\end{aligned}$$

We see that (3.1) can be written as a linear combination of Ψ_j , $j = 1, \dots, J$, $J = \frac{d^2}{2} + \frac{3d}{2} + 2$.

Exercise 3.2. Let X be a continuous rv with cdf $F(x)$. The median $\text{med}(x)$ is defined as $P\{X \geq \text{med}(x)\} = \frac{1}{2} = P\{X \leq \text{med}(x)\}$.

Suppose that $\text{med}(x) = 0$, show that

$$\forall z \in \mathbb{R} \quad \mathbb{E}|X - z| \geq \mathbb{E}|X| \quad (3.2)$$

Interpret (3.2) in terms of a loss function framework.

$$\begin{aligned}\mathbb{E}|X - z| &= \int_{-\infty}^z (X - z) dF(x) + \int_z^{+\infty} (z - X) dF(x) \\ &= \int_{-\infty}^z F(x) dx + \int_z^{+\infty} \{1 - F(x)\} dx \\ &= \begin{cases} \int_{-\infty}^0 F(x) dx + \int_0^z F(x) dx + \int_0^{+\infty} \{1 - F(x)\} dx \\ \quad - \int_0^z \{1 - F(x)\} dx & (z \geq 0) \\ \int_{-\infty}^0 F(x) dx - \int_z^0 F(x) dx + \int_0^{+\infty} \{1 - F(x)\} dx \\ \quad + \int_z^0 \{1 - F(x)\} dx & (z < 0) \end{cases}\end{aligned}$$

$$\begin{aligned}
&= \begin{cases} \mathbb{E}|X| + 2 \int_0^z \{F(x) - 0.5\} dx & (z \geq 0) \\ \mathbb{E}|X| - 2 \int_z^0 \{F(x) - 0.5\} dx & (z < 0) \end{cases} \\
&\geq \mathbb{E}|X|
\end{aligned}$$

since the second terms are in both cases positive.

Define the loss function $\rho(u) = |u|$. The above inequality (3.2) can be rewritten as:

$$\mathbb{E}\rho(X - z) \geq \mathbb{E}\rho(X) \quad \forall z \in \mathbb{R},$$

meaning that $\text{med}(x) = 0$ is the minimum loss (contrast) parameter w.r.t $\rho(u) = |u|$.

Exercise 3.3. *Specify the estimating equation for the generalized EFn (exponential family) and find the solution for the case of the constant regression function $f(X_i, \theta) \equiv \theta$.*

Recall we say that $\mathcal{P}$ is an EF if all measures $P_\theta \in \mathcal{P}$ are dominated by a σ -finite measure μ_0 on $\mathcal{Y}$ and the density functions $p(y, \theta) = dP_\theta/d\mu_0(y)$ are of the form

$$p(y, \theta) \stackrel{\text{def}}{=} \frac{dP_\theta}{d\mu_0}(y) = p(y)e^{yC(\theta) - B(\theta)}.$$

where $C(\theta)$ and $B(\theta)$ are some given nondecreasing functions on θ and $p(y)$ is a negative function on $\mathcal{Y}$.

Also we know $B'(\theta) = \theta = C'(\theta)$

Consider $Y_i = f(X_i, \theta) + \varepsilon_i$, ε_i s are i.i.d, $\theta \in \mathbb{R}^p$ the parameter θ can be estimated via maximum likelihood with

$$\begin{aligned}
L(\theta) &\stackrel{\text{def}}{=} \sum_i \ell\{Y_i, f(X_i, \theta)\} \\
&= \sum_i \{\log p(Y_i)\} + Y_i C\{f(X_i, \theta)\} - B\{f(X_i, \theta)\}
\end{aligned}$$

The corresponding MLE $\tilde{\theta}$ maximizes $L(\theta)$:

$$\tilde{\theta} = \underset{\theta}{\operatorname{argmax}} \sum_i \ell\{Y_i, f(X_i, \theta)\}.$$

The estimating equation $\nabla L(\theta) = 0$ reads as

$$\begin{aligned}
\sum_i \ell' \{Y_i, f(X_i, \theta)\} \nabla f(X_i, \theta) &= \sum_i [Y_i C' \{f(X_i, \theta)\} - B' \{f(X_i, \theta)\}] \nabla f(X_i, \theta) \\
&= \sum_i \left[Y_i C' \{f(X_i, \theta)\} - f(X_i, \theta) C' \{f(X_i, \theta)\} \right] \nabla f(X_i, \theta)
\end{aligned}$$

$$\begin{aligned}
&= \sum_i \{Y_i - f(X_i, \theta)\} [C' \{f(X_i, \theta)\} \nabla f(X_i, \theta)] \\
&= 0
\end{aligned}$$

When $f(X_i, \theta) = \theta$,

$$\begin{aligned}
\sum_i \ell' \{Y_i, f(X_i, \theta)\} \bar{v} f(X_i, \theta) &= \sum_i \ell' \{Y_i, \theta\} \\
&= \sum_i (Y_i - \theta) C'(\theta) \\
&= 0
\end{aligned}$$

then we have

$$\tilde{\theta} = \sum_i Y_i / n$$

Exercise 3.4. *Specify the estimating equation for generalized EFC regression and find the solution for the case of constant regression with $f(X_i, v) \equiv v$. Relate the natural and the canonical representation.*

Recall from Exercise 3.3, the natural parametrization of an EF distribution has the likelihood

$$\ell(y, v) = C(v)y - B(v) + \log P(y),$$

while the canonical parametrization:

$$\ell(y, v) = yv - d(v)$$

therefore

$$\begin{aligned}
\ell'(y, v) &= C'(v)y - B'(v) = (y - v)C'(v) \\
&= y - d'(v),
\end{aligned}$$

so

$$d(v) = y - (y - v)C'(v).$$

Thus the estimating equation

$$\sum_i [Y_i - d' \{f(X_i, \theta)\}] \nabla f(X_i, \theta) = 0$$

can be written as

$$\sum_i (Y_i - [Y_i - \{Y_i - f(X_i, \theta)\} C' \{f(X_i, \theta)\}]) \nabla f(X_i, \theta) = 0, \quad (3.3)$$

when $f(X_i, \theta) \equiv \theta$, (3.3) is

$$\sum_i \{Y_i - \theta\} C'(\theta) = 0.$$

and the solution is $\tilde{\theta} = \sum_i Y_i / n$

Exercise 3.5. *Specify the estimating equation for the case of logit regression.*

The log likelihood with canonical parametrization equals

$$\ell(y, v) = yv - \log(1 + e^v).$$

Therefore $\tilde{\theta} = \operatorname{argmax}_{\theta} L(\theta) = \operatorname{argmax}_{\theta} \sum_i \{Y_i \psi_i^{\top} \theta - \log(1 + e^{\psi_i^{\top} \theta})\}$.
Differentiating w.r.t. θ yields:

$$\frac{\partial}{\partial \theta} L(\theta) = \sum_i \left(Y_i \psi_i - \frac{1}{1 + e^{\psi_i^{\top} \theta}} e^{\psi_i^{\top} \theta} \psi_i \right).$$

Therefore, the estimation equation is:

$$\sum_i \left(Y_i - \frac{e^{\psi_i^{\top} \theta}}{1 + e^{\psi_i^{\top} \theta}} \right) \psi_i = 0.$$

Exercise 3.6.

Credit scoring is a method used to evaluate the credit risk of loan applications. In this example, demographic and credit history variables are used in a logit regression to isolate the effects of various applicant characteristics on credit defaults.

The data is obtained from Fahrmeir and Tutz (1994). A total of $n = 1,000$ observations is used, in which 700 of the individuals have no problem with paying the credit. The response variable $Y \in \{0, 1\}$ is binary, where $Y = 0$ and $Y = 1$ represent “no default” and “default”, respectively. Explanatory variables are the age of the applicant, amount of loan, and some dummy variables are used indicating that:

- Previous loans were okay
- Savings of the applicant is more than 1,000 EUR
- Loan is for a car
- The applicant is a house owner

Table 3.1 GLM results and overall model fit.  MSEglmmest

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	−0.7238	0.3395	−2.1320	0.0330
Age	−0.0178	0.0072	−2.4760	0.0133
Amount	−0.0001	0.0000	2.0040	0.0450
Previous loan	−0.8220	0.1617	−5.0840	0.0000
d9.12	0.4904	0.2821	1.7380	0.0822
d12.18	0.7513	0.2848	2.6380	0.0083
d18.24	0.7522	0.2803	2.6830	0.0073
d24	1.2080	0.3006	4.0190	0.0001
Savings	−1.0040	0.2209	−4.5450	0.0000
Purpose (car)	−0.4389	0.1684	−2.6060	0.0092
House	0.6852	0.2049	3.3450	0.0008
Overall model fit				
Null model −2 log likelihood	1,221.7			
Full model −2 log likelihood	1,104.5			
Chi-square	117.2			
Degrees of freedom	10			

Table 3.2 The goodness of the model.  MSEperformance

	Bankrupt (estimated)	Non-bankrupt (estimated)	Total
Bankrupt (data)	658	42	700
Non-bankrupt (data)	236	64	300
Total	894	106	1,000

- The durations of the desired loans are; 9–12 months, 12–18 months, 18–24 months and more than 24 months

In the first step, scores denoted by “s” are calculated by $\beta_0 + \beta^\top x$ and then the probability of default of each individual credit applicant is found by $G(\beta_0 + \beta^\top x)$, where $G : \mathbb{R} \rightarrow [0, 1]$ is a known function that only takes on a value between 0 and 1. In this example, G is a logistic function ψ :

$$G(t) = \psi(t) = \{1 + \exp(-t)\}^{-1}$$

The results of the model can be summarised as (Table 3.1):
To test the goodness of fit we check the difference in deviance residuals for the model used above versus the null model. The large value of the chi-square test statistics of 117.2 indicates that the model as a whole fits significantly better than an empty model. Moreover, the comparison of real data set and logit estimation is given in the Table 3.2.
Additionally, so as to visualize the model we plot the scores with respect to probability of default and response variable as Fig. 3.1. The goodness of the model can be checked with Lorenz curve (Fig. 3.2), the plot of $P(S < s)$ against $P(S < s \mid Y = 1)$.

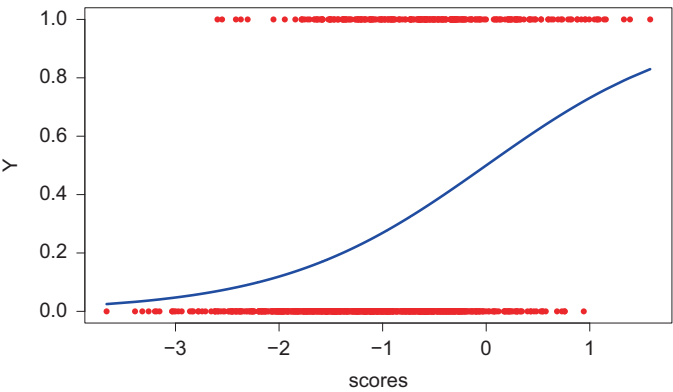


Fig. 3.1 The plot of scores with respect to response variable.  MSElogit

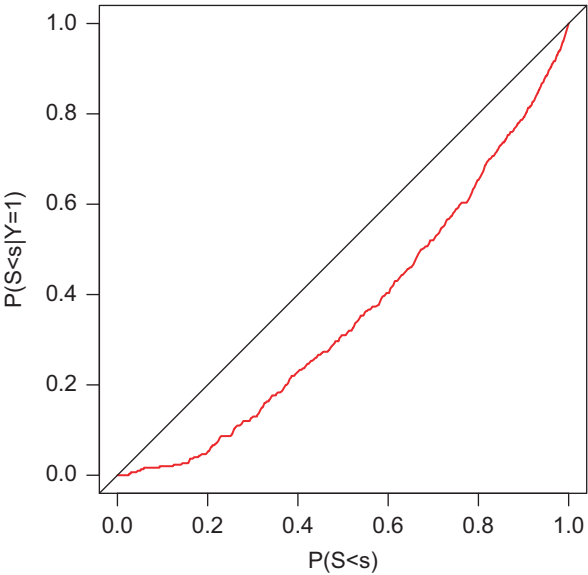


Fig. 3.2 Lorenz curve.  MSElorenz

Given the results, the older the applicant and the shorter the duration of the desired loan, the less the probability of default is. Creditworthiness is higher for the applicants having savings more than 1,000 EUR, less problems about paying back the previous loans and demanding a loan for a car, whereas the applicants owning a house have higher probability of default. On the other hand, the model yields a positive relation between the amount of the desired loan and the ability of applicants to pay the loan back which can be explained by the fact that high levels of credits are given to reliable applicants.

Exercise 3.7. Simulate an i.i.d. random sample $\{X_i\}_{i=1}^n$ which follows a specific distribution ($n = 300$).

1. Assume the true distribution of X_i 's is $t(3)$. Estimate the kernel density function $\hat{f}_h(x)$ using a Gaussian kernel and plot the kernel density curve.
2. Let f_0 be the true density of the sample. Since $\hat{f}_h(x)$ is biased in a finite sample, we can not compare it with f_0 directly. We rather compare it with $\mathbb{E}_{f_0}[\hat{f}_h(x)]$ which is the expectation of $\hat{f}_h(x)$ under f_0 , where

$$\mathbb{E}_{f_0}[\hat{f}_h(x)] = g(x) = \frac{1}{h} \int K\left(\frac{x-u}{h}\right) f_0(u) du. \quad (3.4)$$

Let Z_j 's be the random variables which come from a specific distribution, assume that $H_0: Z_j \sim f_0 = t(3)$. Then we can approximate $g(x)$ via

$$\hat{g}(x) = \frac{1}{Nh} \sum_{j=1}^N K\left(\frac{x-Z_j}{h}\right), \quad (3.5)$$

where $N = 10^6$. Plot $\hat{g}(x)$ and compare it with the kernel density curve.

3. Assume that $H_0: Z_j \sim f_0 = \mathcal{N}(\hat{\mu}, \hat{\sigma}^2)$, where $\hat{\mu}$ and $\hat{\sigma}$ are the mean and standard deviation estimated from the sample respectively. Approximate $g(x)$ as in question 2. Plot the curve of it and compare it with the kernel density estimate.
 4. Assume now the true distribution of X_i 's is $\mathcal{N}(0, 1)$, perform the same procedure as in question 2 and 3. Compare the resulting curves.
1. The kernel density estimator is as follows:

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x-X_i}{h}\right),$$

where we use the Gaussian kernel function:

$$K(u) = \varphi(u).$$

In Fig. 3.3, the solid line denotes the kernel density estimator $\hat{f}_h(x)$.

2. In Fig. 3.3, the dashed line denotes $\hat{g}(x)$, where $f_0 = t(3)$. We find that $\hat{f}_h(x)$ is very close to $\hat{g}(x)$. If we had compared $\hat{f}_h(x)$ with the true density f_0 , we would not see such degree of closeness due to the finite sample bias.
3. The mean and standard deviation can be estimated as follows:

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n X_i,$$

$$\hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \hat{\mu})^2.$$

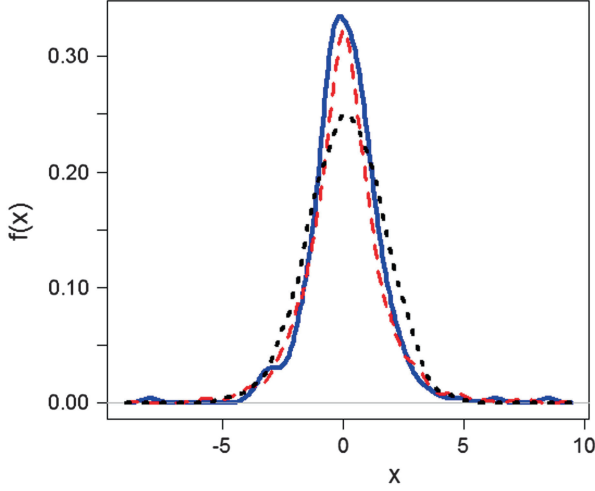


Fig. 3.3 The kernel density estimator $\hat{f}_h(x)$ (solid line), $\hat{g}(x)$ with $f_0 = t(3)$ (dashed line), and $\hat{g}(x)$ with $f_0 = \mathcal{N}(\hat{\mu}, \hat{\sigma}^2)$ (dotted line), for $n = 300$.  MSEnonpara1

In Fig. 3.3, the dotted line denotes $\hat{g}(x)$, where $f_0 = \mathcal{N}(\hat{\mu}, \hat{\sigma}^2)$. From Fig. 3.3 we find that $\hat{f}_h(x)$ is closer to $\hat{g}(x)$ with $f_0 = t(3)$, it provides the evidence for $f_0 = t(3)$.

4. In Fig. 3.4, the solid line denotes the kernel density estimator $\hat{f}_h(x)$, the dashed line denotes $\hat{g}(x)$ with $f_0 = t(3)$, and the dotted line denotes $\hat{g}(x)$ with $f_0 = \mathcal{N}(\hat{\mu}, \hat{\sigma}^2)$. From Fig. 3.4 we find that $\hat{f}_h(x)$ is closer to $\hat{g}(x)$ with $f_0 = \mathcal{N}(\hat{\mu}, \hat{\sigma}^2)$, it provides the evidence for $f_0 = \mathcal{N}(\hat{\mu}, \hat{\sigma}^2)$.

Exercise 3.8. Consider the error in design model

$$Y_i = \Psi_i^\top \theta^* + \varepsilon_i, \quad Z_i = \Psi_i + U_i, \quad i = 1, \dots, n,$$

where both Ψ_i and θ^* are $p \times 1$ vectors. Assume that $p = 1$, Ψ_i 's are unobservable. Instead, (Y_i, Z_i) 's are observable. $\text{Cov}(\varepsilon_i, \Psi_i) = \text{Cov}(\varepsilon_i, U_i) = \text{Cov}(\Psi_i, U_i) = 0$, $\text{Var}(\Psi_i) \stackrel{\text{def}}{=} \sigma_\psi^2$, $\text{Var}(U_i) \stackrel{\text{def}}{=} \sigma_u^2$, $\text{Var}(\varepsilon_i) \stackrel{\text{def}}{=} \sigma_\varepsilon^2$.

1. Let b be the regression coefficient by regressing Y_i on Z_i , show that $b \leq \theta^*$.
 2. Let $\theta = 2$, $\Psi_i \sim U(0, 1)$, $\varepsilon_i \sim \mathcal{N}(0, 0.01)$, $U_i \sim \mathcal{N}(0, 0.09)$. Estimate the coefficient b and verify the result in question 1.
 3. Let $n = 300$, where the value of θ and the distributions of the variables are the same as in question 2. Plot the regression line of Y_i on Z_i , then plot the linear regression line of Y_i on Ψ_i on the same graph, interpret the result.
1. Assume that the regression equation of regressing Y_i on Z_i is $Y_i = bZ_i + v_i$, then $\min_b \mathbb{E}(Y_i - bZ_i)^2$ has solution (3.6):

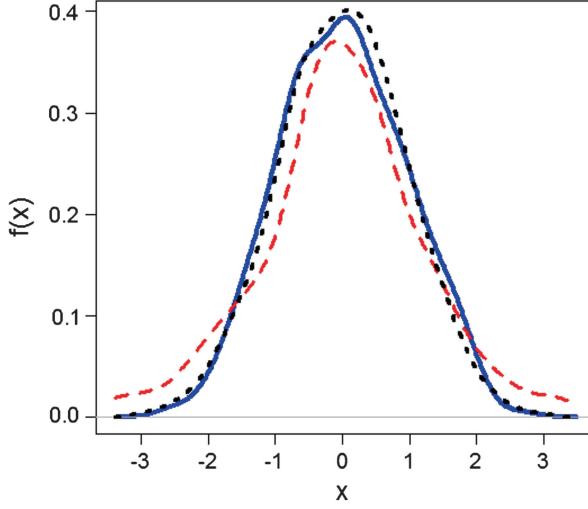


Fig. 3.4 The kernel density estimator $\hat{f}_h(x)$ (solid line), $\hat{g}(x)$ with $f_0 = t(3)$ (dashed line), and $\hat{g}(x)$ with $f_0 = \mathcal{N}(\hat{\mu}, \hat{\sigma}^2)$ (dotted line), for $n = 300$.  MSEnonpara2

$$b = \frac{\text{Cov}(Y_i, Z_i)}{\text{Var}(Z_i)}, \quad (3.6)$$

where by assumption:

$$\begin{aligned} \text{Var}(Z_i) &= \text{Var}(\Psi_i + U_i) \\ &= \text{Var}(\Psi_i) + \text{Var}(U_i) + 2 \underbrace{\text{Cov}(\Psi_i, U_i)}_{=0} \\ &= \sigma_\psi^2 + \sigma_u^2, \end{aligned}$$

and

$$\begin{aligned} \text{Cov}(Y_i, Z_i) &= \text{Cov}(\Psi_i \theta^* + \varepsilon_i, Z_i) \\ &= \text{Cov}(\Psi_i \theta^* + \varepsilon_i, \Psi_i + U_i) \\ &= \text{Cov}(\Psi_i \theta^*, \Psi_i) + \text{Cov}(\Psi_i \theta^*, U_i) + \underbrace{\text{Cov}(\varepsilon_i, \Psi_i)}_{=0} + \underbrace{\text{Cov}(\varepsilon_i, U_i)}_{=0} \\ &= \theta^* \text{Var}(\Psi_i) + \theta^* \underbrace{\text{Cov}(\Psi_i, U_i)}_{=0} \\ &= \theta^* \sigma_\psi^2. \end{aligned}$$

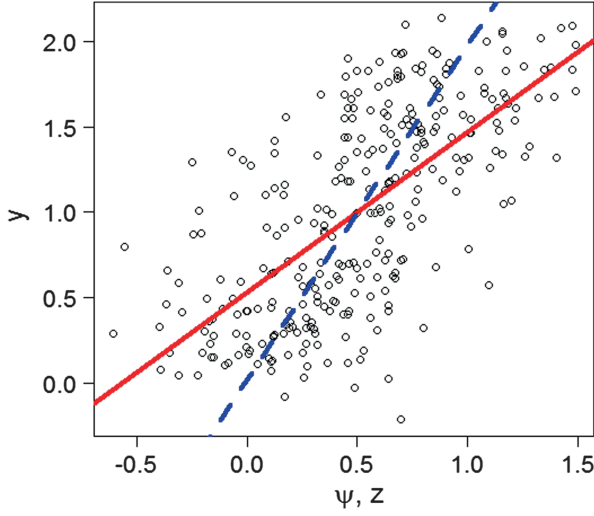


Fig. 3.5 The linear regression line of Y_i on Z_i (solid line) and the linear regression line of Y_i on Ψ_i (dashed line), for $n = 300$.  MSE regression

Therefore

$$b = \frac{\text{Cov}(Y_i, Z_i)}{\text{Var}(Z_i)} = \frac{\theta^* \sigma_\psi^2}{\text{Var}(Z_i)} = \theta^* \frac{\sigma_\psi^2}{\sigma_\psi^2 + \sigma_u^2} = \theta^* \frac{1}{1 + \sigma_u^2 / \sigma_\psi^2} \leq \theta^*.$$

2. Since $\Psi_i \sim U(0, 1)$, $\varepsilon_i \sim \mathcal{N}(0, 0.01)$, $U_i \sim \mathcal{N}(0, 0.09)$ and $\theta = 2$, then $\sigma_\psi^2 = \frac{1}{12}(1 - 0)^2 = \frac{1}{12}$, $\sigma_\varepsilon^2 = 0.01$, $\sigma_u^2 = 0.09$,

$$\hat{b} = \theta \frac{1}{1 + \sigma_u^2 / \sigma_\psi^2} = 2 \times \frac{1}{1 + 0.09 \times 12} = 2 \times \frac{1}{2.08} \approx 0.9615 < 2.$$

3. We generate $n = 300$ samples for Ψ_i , ε_i , and U_i , then perform the linear regression of Y_i on Z_i and the linear regression of Y_i on Ψ_i . In Fig. 3.5, the solid line denotes the linear regression line of Y_i on Z_i , where $\hat{b}$ is the slope of the solid line. The dashed line denotes the linear regression line of Y_i on Ψ_i , where θ is the slope of the dashed line. In Fig. 3.5, we can see that the solid line is gentler than the dashed line, it can be concluded that the value of $\hat{b}$ is smaller than the value of θ . Furthermore, we simulate the sampling 400 times and estimate the coefficient $\tilde{b}$ as follows:

$$\tilde{b} = \frac{1}{400} \sum_{j=1}^{400} \hat{b}_j = 0.9614.$$

Exercise 3.9. Consider the regression model

$$Y = \Psi^\top \theta^* + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \Sigma),$$

where Ψ is an $p \times n$ matrix of stochastic regressors s.t.

$$\frac{\Psi \varepsilon}{n} \xrightarrow{\mathbb{P}} C, \quad C \neq 0.$$

Assume that W is an $n \times l$ matrix of l instruments ($l \geq p$) with:

$$\frac{W^\top \varepsilon}{n} \xrightarrow{\mathbb{P}} 0. \quad (3.7)$$

$$\frac{W^\top \Psi^\top}{n} \xrightarrow{\mathbb{P}} \Omega_{W\Psi} < \infty. \quad (3.8)$$

$$\frac{W^\top W}{n} \xrightarrow{\mathbb{P}} \Omega_{WW} \quad (\text{positive definite}), \quad (3.9)$$

where $\text{rank}(\Omega_{W\Psi}) = \text{rank}(W^\top \Psi^\top) = p$.

1. Motivate assumptions (3.7) and (3.8).
 2. Propose an instrumental variable estimation (IVE) for θ^* and show that it is consistent.
 3. Derive the simple IVE when the number of instruments equals the number of regressors, i.e. $l = p$.
1. According to the weak law of large numbers we know that $\frac{W^\top \varepsilon}{n}$ is the sample analogue of $\mathbb{E}(W_i^\top \varepsilon_i)$, so $\frac{W^\top \varepsilon}{n} \xrightarrow{\mathbb{P}} 0$ implies $\mathbb{E}(W_i^\top \varepsilon_i) = 0$, and $\frac{W^\top \Psi^\top}{n} \xrightarrow{\mathbb{P}} \Omega_{W\Psi}$ implies $\mathbb{E}(W_i^\top \Psi_i^\top) = \Omega_{W\Psi}$.
 2. Let $P_W \stackrel{\text{def}}{=} W(W^\top W)^{-1}W^\top$, and $P_W^\top = P_W$, $P_W^\top P_W = P_W$, let the new covariates be $P_W \Psi^\top$, then

$$\begin{aligned} \hat{\theta}_{IV} &= (\Psi P_W^\top P_W \Psi^\top)^{-1} \Psi P_W^\top Y \\ &= (\Psi P_W \Psi^\top)^{-1} \Psi P_W Y \\ &= (\Psi P_W \Psi^\top)^{-1} \Psi P_W (\Psi^\top \theta^* + \varepsilon) \\ &= \theta^* + (\Psi P_W \Psi^\top)^{-1} \Psi P_W \varepsilon, \end{aligned}$$

since

$$\frac{W^\top \varepsilon}{n} \xrightarrow{\mathbb{P}} 0,$$

so that

$$\begin{aligned}
\Psi P_W \varepsilon &= \Psi W (W^\top W)^{-1} W^\top \varepsilon \\
&= \Psi W \underbrace{\left(\frac{W^\top W}{n} \right)^{-1}}_{\rightarrow \Omega_{WW}} \underbrace{\frac{W^\top \varepsilon}{n}}_{\rightarrow 0},
\end{aligned}$$

then

$$\Psi P_W \varepsilon \xrightarrow{\mathbb{P}} 0_p,$$

therefore

$$\hat{\theta}_{IV} \xrightarrow{\mathbb{P}} \theta^*.$$

3. When $l = p$,

$$\begin{aligned}
\hat{\theta}_{IV} &= (\Psi P_W \Psi^\top)^{-1} \Psi P_W Y \\
&= (\Psi W (W^\top W)^{-1} W^\top \Psi^\top)^{-1} \Psi W (W^\top W)^{-1} W^\top Y \\
&= (W^\top \Psi^\top)^{-1} (W^\top W) (\Psi W)^{-1} \Psi W (W^\top W)^{-1} W^\top Y \\
&= (W^\top \Psi^\top)^{-1} W^\top Y.
\end{aligned}$$

Exercise 3.10. We know that the income of people is affected by many factors, for example education level and ability. Suppose we omit the variable which measures ability. But we know that education level is correlated with ability, which means that if we omit it, then there would be an endogeneity problem in the regression function (i.e. the regressor “education” is correlated with the error term). To solve this problem we need to find an instrumental variable which is correlated with education level but uncorrelated with ability. Consider the following model

$$Y = \theta_0 + \theta_1 X + \varepsilon,$$

where Y is the log-transformation of income, X is the highest year of school completed, ε is the error term and contains the ability. Then choose W (the number of brothers and sisters) as instrumental variable which means that $\text{Cov}(X, W) \neq 0$, and $\text{Cov}(\varepsilon, W) = 0$, then consider the following model

$$X = m_0 + m_1 W + v.$$

1. Use 2010 GSS data which comes from the website of The General Social Survey: <http://www3.norc.org/GSS+Website/>. For convenience, the missing values of X and Y have been deleted from the data. Perform the linear regression of Y on X , estimate the coefficients, write down the equation and interpret the result.

2. Perform the linear regression of X on W , estimate the coefficients, use t test to test $H_0 : m_1 = 0$, write down the equation and interpret the result.
3. Take W as instrumental variable, estimate θ_{IV} , write down the equation. Compare the results before and after using the instrumental variable.

1.

$$\hat{Y} = 3.3042 + 0.0626X.$$

It means that 1 year more education increases the income by 6.26 %.  MSEivgss

2.

$$\hat{X} = 15.0783 - 0.3169W.$$

From result of t test for m_1 , we can see that p-value is less than $2e^{-16}$, which is statistically significant. Then the $H_0 : m_1 = 0$ is rejected. We can conclude that there is a significant negative correlation between X and W .

3.

$$\hat{Y} = 3.5443 + 0.0525X.$$

Compared with OLS estimator $\hat{\theta}_1$, the IV estimator $\hat{\theta}_{IV}$ is a little lower, i.e. 1 year more education increases the income by 5.25 %. From Exercise 3.9 we know that if our assumptions $Cov(X, W) \neq 0$, and $Cov(\varepsilon, W) = 0$ are true, then the IV estimator is consistent.

Exercise 3.11. Consider an infinite dimensional model of continuously stratified random sampling in which one has i.i.d. observations $W_i = (X_i, R_i, Z_i)$ with $X_i \in [0, 1]^d$, $Z_i = R_i Y_i$, and $R_i, Y_i \in [0, 1]$ and are conditionally independent given X_i , with $g(X) = \mathbb{E}(R|X)$ known and $h(X) = \mathbb{E}(Y|X)$ unknown. The parameter of interest is $\theta = \mathbb{E}(Y)$.

Prove that the Horvitz-Thompson estimator

$$\hat{\theta} = \frac{1}{n} \sum_{i=1}^n \frac{Z_i}{g(X_i)}$$

is a consistent estimator for θ .

Because R_i, Y_i are conditionally independent,

$$\begin{aligned} \mathbb{E}\left\{\frac{RY}{g(X)}\right\} &= \mathbb{E}\left\{\frac{\mathbb{E}(RY|X)}{g(X)}\right\} \quad (\text{by the law of iterated expectation}) \\ &= \mathbb{E}\left\{\frac{\mathbb{E}(R|X)\mathbb{E}(Y|X)}{g(X)}\right\} \end{aligned}$$

$$\begin{aligned}
&= \mathbb{E}\left\{\frac{g(X)h(X)}{g(X)}\right\} \\
&= \mathbb{E}\{h(X)\} \\
&= \mathbb{E}\{\mathbb{E}(Y|X)\} = \mathbb{E}(Y) = \theta,
\end{aligned}$$

therefore

$$\mathbb{E}\left\{\frac{RY}{g(X)}\right\} = \theta. \quad (3.10)$$

Since W_i are i.i.d., according to the weak law of large numbers, the empirical counterpart of the left hand side of (3.10) is:

$$\frac{1}{n} \sum_{i=1}^n \frac{Z_i}{g(X_i)} \xrightarrow{\mathbb{P}} \mathbb{E}\left\{\frac{RY}{g(X)}\right\} = \theta.$$

Exercise 3.12. Let a sequence of i.i.d. random variables $\{X_i\}_{i=1}^n \sim \mathcal{N}(0, \sigma^2)$, and $\sum_{i=1}^n a_i^2 = n$. Prove the Chernoff bound

$$\mathbb{P}\left(\left|\sum_{i=1}^n a_i X_i\right| > t\right) \leq 2 \exp\left(-\frac{t^2}{2n\sigma^2}\right).$$

Since

$$\mathbb{P}\left(\left|\sum_{i=1}^n a_i X_i\right| > t\right) = \mathbb{P}\left(\sum_{i=1}^n a_i X_i > t\right) + \mathbb{P}\left(\sum_{i=1}^n a_i X_i < -t\right),$$

without loss of generality, let us derive

$$\mathbb{P}\left(\sum_{i=1}^n a_i X_i > t\right) \leq \exp\left(-\frac{t^2}{2n\sigma^2}\right),$$

the argument is symmetric for $\mathbb{P}(\sum_{i=1}^n a_i X_i < -t)$. Then for any $s > 0$:

$$\begin{aligned}
\mathbb{P}\left(\sum_{i=1}^n a_i X_i > t\right) &= \mathbb{P}\left(s \sum_{i=1}^n a_i X_i > st\right) \\
&= \mathbb{P}\left\{\exp\left(s \sum_{i=1}^n a_i X_i\right) > \exp(st)\right\} \\
&\leq \frac{\mathbb{E}[\exp(s \sum_{i=1}^n a_i X_i)]}{\exp(st)} \quad (\text{by Markov's inequality})
\end{aligned}$$

$$\begin{aligned}
&= \frac{\prod_{i=1}^n \mathbb{E}[\exp(s a_i X_i)]}{\exp(st)} \quad (\text{by independence of } X_i \text{'s}) \\
&= \exp(-st) \prod_{i=1}^n \mathbb{E}[\exp(s a_i X_i)].
\end{aligned}$$

From the moment-generating function of a normal distribution we know that if $X \sim \mathcal{N}(\mu, \sigma^2)$, then $\mathbb{E}[\exp(tX)] = \exp(t\mu + \frac{1}{2}\sigma^2 t^2)$. In our case, since X_i 's are i.i.d. random variables and $X_i \sim \mathcal{N}(0, \sigma^2)$, then

$$\mathbb{E}[\exp(s a_i X_i)] = \exp\left(\frac{\sigma^2 s^2 a_i^2}{2}\right),$$

thus

$$\mathbb{P}\left(\sum_{i=1}^n a_i X_i > t\right) \leq \exp(-st) \prod_{i=1}^n \mathbb{E}[\exp(s a_i X_i)] \quad (3.11)$$

$$= \exp\left(-st + \frac{\sigma^2 s^2 \sum_{i=1}^n a_i^2}{2}\right) \quad (3.12)$$

$$= \exp\left(-st + \frac{\sigma^2 s^2 n}{2}\right), \quad (3.13)$$

minimizing $(-st + \sigma^2 s^2 n/2)$ for $s > 0$, we get

$$-t + \sigma^2 s n = 0,$$

then

$$s = \frac{t}{\sigma^2 n},$$

we insert s into (3.13)

$$\exp\left(-st + \frac{\sigma^2 s^2 n}{2}\right) = \exp\left(-\frac{t^2}{n\sigma^2} + \frac{t^2}{2n\sigma^2}\right) = \exp\left(-\frac{t^2}{2n\sigma^2}\right),$$

thus

$$\mathbb{P}\left(\sum_{i=1}^n a_i X_i > t\right) \leq \exp\left(-\frac{t^2}{2n\sigma^2}\right),$$

therefore

$$\mathbb{P}\left(\left|\sum_{i=1}^n a_i X_i\right| > t\right) \leq 2 \exp\left(-\frac{t^2}{2n\sigma^2}\right).$$

Exercise 3.13. From Exercise 3.12 we have the Chernoff bound for i.i.d. normal variables:

$$\mathbb{P}\left(\left|\sum_{i=1}^n a_i X_i\right| > t\right) \leq 2 \exp\left(-\frac{t^2}{2n\sigma^2}\right).$$

Assume Ψ is a $p \times n$ dimensional design matrix in a regression problem, and ε is the $n \times 1$ dimension i.i.d. normal noise, where $i = 1, \dots, n$, $j = 1, \dots, p$. Ψ_j is j th row of Ψ . Assume that Ψ_j 's are normalized and orthogonal.

1. Prove that

$$\max_{1 \leq j \leq p} |\Psi_j \varepsilon| = \mathcal{O}_p\{\sigma \sqrt{2n \log(2p)}\}.$$

2. Prove that

$$\mathbb{E}\left(\max_{1 \leq j \leq p} |\Psi_j \varepsilon|\right) \leq \sigma \sqrt{2n \log(2p)}.$$

1.

$$\begin{aligned} \mathbb{P}\left(\max_{1 \leq j \leq p} |\Psi_j \varepsilon| > \lambda\right) &\leq \sum_{j=1}^p \mathbb{P}(|\Psi_j \varepsilon| > \lambda) \\ &= \sum_{j=1}^p \mathbb{P}\left(\left|\sum_{i=1}^n \Psi_{ji} \varepsilon_i\right| > \lambda\right) \\ &\leq \sum_{j=1}^p 2 \exp\left(-\frac{\lambda^2}{2n\sigma^2}\right) \quad (\text{since } \Psi_j \text{'s are normalized}), \end{aligned}$$

take $\lambda = \sqrt{2n \log(2p/\delta)\sigma^2}$, where δ is a constant, then we get

$$\begin{aligned} \sum_{j=1}^p 2 \exp\left\{-\frac{2n \log(2p/\delta)\sigma^2}{2n\sigma^2}\right\} &= \sum_{j=1}^p 2 \exp\{-\log(2p/\delta)\} \\ &= 2p \cdot \frac{1}{\exp\{\log(2p/\delta)\}} \\ &= 2p \cdot \frac{1}{2p/\delta} \\ &= \delta, \end{aligned}$$

i.e.

$$\mathbb{P}\left(\max_{1 \leq j \leq p} |\Psi_j \varepsilon| > \lambda\right) = \mathbb{P}\left\{\max_{1 \leq j \leq p} |\Psi_j \varepsilon| > \sigma \sqrt{2n \log(2p/\delta)}\right\} \leq \delta,$$

therefore

$$\max_{1 \leq j \leq p} |\Psi_j \varepsilon| = \mathcal{O}_p\{\sigma \sqrt{2n \log(2p)}\}.$$

2. From Jensen's inequality we know that if $g(X)$ is convex, then

$$g[\mathbb{E}(X)] \leq \mathbb{E}[g(X)].$$

In our case, since $\exp\left(\max_{1 \leq j \leq p} |\Psi_j \varepsilon|\right)$ is a convex function, then for any $s > 0$, we get

$$\begin{aligned} \exp\left\{s \mathbb{E}\left(\max_{1 \leq j \leq p} |\Psi_j \varepsilon|\right)\right\} &\leq \mathbb{E}\left\{\exp\left(s \cdot \max_{1 \leq j \leq p} |\Psi_j \varepsilon|\right)\right\} \\ &\leq \sum_{j=1}^p \mathbb{E}\left\{\exp\left(s \cdot |\Psi_j \varepsilon|\right)\right\} \\ &\leq 2p \exp\left(\frac{n\sigma^2 s^2}{2}\right), \end{aligned}$$

we take log for both sides of the inequality, then

$$\mathbb{E}\left(\max_{1 \leq j \leq p} |\Psi_j \varepsilon|\right) \leq \frac{\log(2p)}{s} + \frac{n\sigma^2 s}{2}$$

we minimize $\log(2p)/s + n\sigma^2 s/2$, then

$$(-1)(s^{-2}) \log(2p) + \frac{n\sigma^2}{2} = 0,$$

thus

$$s = \frac{\sqrt{2 \log(2p)}}{\sigma \sqrt{n}},$$

therefore

$$\begin{aligned} \mathbb{E}\left(\max_{1 \leq j \leq p} |\Psi_j \varepsilon|\right) &\leq \frac{\log(2p)}{s} + \frac{n\sigma^2 s}{2} \\ &= \log(2p) \cdot \frac{\sigma \sqrt{n}}{\sqrt{2 \log(2p)}} + \frac{n\sigma^2}{2} \cdot \frac{\sqrt{2 \log(2p)}}{\sigma \sqrt{n}} \end{aligned}$$

$$\begin{aligned}
&= 2 \cdot \frac{\sigma \sqrt{n \log(2p)}}{\sqrt{2}} \\
&= \sigma \sqrt{2n \log(2p)}.
\end{aligned}$$

Exercise 3.14. Consider the error in design model

$$Y_i = m(\Psi_i) + \varepsilon_i, \quad Z_i = \Psi_i + U_i, \quad i = 1, \dots, n,$$

where ε_i 's are the regression errors, U_i 's are the measurement errors, Ψ_i 's are $p \times 1$ dimensional vectors. Assume that $p = 1$, $\mathbb{E}(\varepsilon_i | \Psi_i) = 0$, the true function is $m(\psi) = 5\psi^2$, $\Psi_i \sim \mathcal{N}(3, 4)$, $\varepsilon_i \sim \mathcal{N}(0, 0.01)$, $U_i \sim \mathcal{N}(0, 0.81)$.

1. Write down the kernel regression estimator (Nadaraya-Watson estimator) and the deconvoluted kernel regression estimator of $m(\psi)$.
2. Generate a random sample with $n = 3,000$, download and use the R package "decon", then plot the deconvoluted kernel regression curve, the kernel regression curve from the sample without measurement errors (i.e. kernel regression based on ψ) and the kernel regression curve from the sample with measurement errors (i.e. kernel regression based on z). Determine which estimator is better.

1. The Nadaraya-Watson estimator is as follows:

$$\hat{m}_{NW}(\psi) = \frac{\sum_{i=1}^n Y_i K\left(\frac{\psi - \Psi_i}{h}\right)}{\sum_{i=1}^n K\left(\frac{\psi - \Psi_i}{h}\right)},$$

we use a Gaussian kernel as the kernel function:

$$K(u) = \varphi(u).$$

The deconvoluted kernel regression estimator is as follows:

$$\hat{m}_D(\psi) = \frac{\sum_{i=1}^n Y_i L\left(\frac{\psi - Z_i}{h}\right)}{\sum_{i=1}^n L\left(\frac{\psi - Z_i}{h}\right)},$$

where

$$L(s) = \frac{1}{2\pi} \int e^{-its} \frac{\varphi_K(t)}{\varphi_U(t/h)} dt.$$

2. From Fig. 3.6 we can conclude that the deconvoluted kernel regression estimator is closer to the kernel regression estimator from the sample without measurement errors. The deconvoluted kernel regression estimator performs better than the kernel regression estimator from the sample with measurement errors.

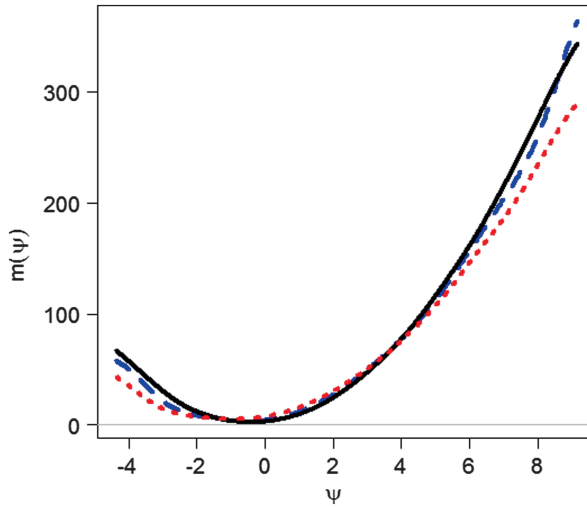


Fig. 3.6 The kernel regression curve from the sample without measurement errors (*solid line*), the deconvoluted kernel regression curve (*dashed line*), and the kernel regression curve from the sample with measurement errors (*dotted line*), for $n = 3,000$.  MSEdecon

Reference

Fahrmeir, L., & Tutz, G. (1994). *Multivariate statistical modelling based on generalized linear models*. Heidelberg: Springer.

Chapter 4

Estimation in Linear Models

Estimarea modelelor liniare

Stai strâmb, vorbește drept.

Walk wryly, speak straight.

Exercise 4.1. *A company decides to compare the effect of three marketing strategies*

- 1. Advertisement in local newspaper,*
- 2. Presence of sales assistant,*
- 3. Special presentation in shop windows,*

on the sales of their portfolio in 30 shops. The 30 shops were divided into 3 groups of 10 shops. The sales using the strategies 1, 2, and 3 were $y_1 = (9, 11, 10, 12, 7, 11, 12, 10, 11, 13)^\top$, $y_2 = (10, 15, 11, 15, 15, 13, 7, 15, 13, 10)^\top$, and $y_3 = (18, 14, 17, 9, 14, 17, 16, 14, 17, 15)^\top$, respectively. Define x_i as the index of the shop, i.e., $x_i = i, i = 1, 2, \dots, 30$. Using this notation, the null hypothesis corresponds to a constant regression line, $\mathbb{E}Y = \mu$. What does the alternative hypothesis involving a regression curve look like?

There are $p = 3$ factors and $n = 30$ observations in the data set. The company wants to know whether all three marketing strategies have the same effect or whether there is a difference. The null hypothesis is $H_0 : \mu_1 = \mu_2 = \mu_3$ and the alternative hypothesis is $H_1 : \mu_l \neq \mu_{l'}$ for some l and l' . The standard approach to this problem is the analysis of variance (ANOVA) technique which leads to an F -test.

In this exercise, we use an alternative and in fact equivalent approach based on the regression model. The null hypothesis can be tested in a regression model that has explanatory variables defined as $z_{2i} = (x_i \in (11, 20))$ and $z_{3i} = (x_i \in (21, 30))$.

These two variables now allow to describe the difference in sales due to the marketing strategies.

The regression model can be written as

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 1_{10} & 0_{10} & 0_{10} \\ 1_{10} & 1_{10} & 0_{10} \\ 1_{10} & 0_{10} & 1_{10} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} + \boldsymbol{\varepsilon}.$$

Here, the regression curve corresponding to the alternative hypothesis in the ANOVA model looks like three horizontal lines, each of them corresponding to one marketing strategy.

The F -test for testing the null hypothesis $H_0 : \beta_2 = \beta_3 = 0$ corresponds to the test of the null hypothesis that the effect of the three marketing strategies is the same.

A N O V A	SS	df	MSS	F-test	P-value
Regression	102.600	2	51.300	8.783	0.0012
Residuals	157.700	27	5.841		
Total Variation	260.300	29	8.976		

Multiple R = 0.62782
 R^2 = 0.39416
Adjusted R^2 = 0.34928
Standard Error = 2.41676

PARAMETERS	Beta	SE	StandB	t-test	P-value
b[0,]=	10.6000	0.7642	0.0000	13.870	0.0000
b[1,]=	1.8000	1.0808	0.2881	1.665	0.1074
b[2,]=	4.5000	1.0808	0.7202	4.164	0.0003

 MSEanovapull

The above computer output shows that the value of the F -statistic for our null hypothesis is 8.783, the corresponding p-value is smaller than 0.05. Thus, on the usual confidence level 95 %, the null hypothesis is rejected.

The computer output also contains the mean sales of all three marketing strategies. The mean sales for the first marketing strategy were 10.6, for the second strategy $10.6 + 1.8 = 12.4$, and for the third strategy $10.6 + 4.5 = 15.1$.

Exercise 4.2. Consider the linear model $\mathbf{Y} = \Psi^\top \boldsymbol{\theta}^* + \boldsymbol{\varepsilon}$ where $\hat{\boldsymbol{\theta}} = \arg \min_{\boldsymbol{\theta}^*} \boldsymbol{\varepsilon}^\top \boldsymbol{\varepsilon}$ is subject to the linear constraints $A\hat{\boldsymbol{\theta}} = \mathbf{a}$ where $A(\mathbf{q} \times p)$, $(\mathbf{q} \leq p)$ is of rank $\mathbf{q}$ and $\mathbf{a}$ is of dimension $(\mathbf{q} \times 1)$.

Show that

$$\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}_{OLS} - (\Psi\Psi^\top)^{-1}A^\top \{A(\Psi\Psi^\top)^{-1}A^\top\}^{-1} (A\hat{\boldsymbol{\theta}}_{OLS} - \mathbf{a}),$$

where $\hat{\boldsymbol{\theta}}_{OLS} = (\Psi\Psi^\top)^{-1}\Psi\mathbf{Y}$ is the unconstrained (ordinary) least squares estimator.

We define

$$f(\boldsymbol{\theta}^*, \lambda) = (\mathbf{Y} - \Psi^\top \boldsymbol{\theta}^*)^\top (\mathbf{Y} - \Psi^\top \boldsymbol{\theta}^*) - \lambda^\top (A\boldsymbol{\theta}^* - \mathbf{a}),$$

where $\lambda \in \mathbb{R}^q$ and solve the system of equations:

$$\frac{\partial f(\boldsymbol{\theta}^*, \lambda)}{\partial \boldsymbol{\theta}^*} = 0,$$

$$\frac{\partial f(\boldsymbol{\theta}^*, \lambda)}{\partial \lambda} = 0.$$

Evaluating the derivatives, we obtain the system of equations:

$$\frac{\partial f(\boldsymbol{\theta}^*, \lambda)}{\partial \boldsymbol{\theta}^*} = -2\Psi\mathbf{Y} + 2\Psi\Psi^\top \hat{\boldsymbol{\theta}} - A^\top \hat{\lambda} = 0, \quad (4.1)$$

$$\frac{\partial f(\boldsymbol{\theta}^*, \lambda)}{\partial \lambda} = -(A\hat{\boldsymbol{\theta}} - \mathbf{a})^\top = 0, \quad (4.2)$$

rearranging (4.1) with respect to $\hat{\boldsymbol{\theta}}$ leads to

$$\hat{\boldsymbol{\theta}} = (\Psi\Psi^\top)^{-1}\Psi\mathbf{Y} + \frac{1}{2}(\Psi\Psi^\top)^{-1}A^\top \hat{\lambda}, \quad (4.3)$$

$$A\hat{\boldsymbol{\theta}} = A\hat{\boldsymbol{\theta}}_{OLS} + \frac{1}{2}A(\Psi\Psi^\top)^{-1}A^\top \hat{\lambda}. \quad (4.4)$$

Next, rearranging (4.4) with respect to $\hat{\lambda}$ implies that

$$\hat{\lambda} = 2\{A(\Psi\Psi^\top)^{-1}A^\top\}^{-1}(\mathbf{a} - A\hat{\boldsymbol{\theta}}_{OLS}). \quad (4.5)$$

Set (6.28) in (4.3)

$$\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}_{OLS} - (\Psi\Psi^\top)^{-1}A^\top\{A(\Psi\Psi^\top)^{-1}A^\top\}^{-1}(A\hat{\boldsymbol{\theta}}_{OLS} - \mathbf{a}).$$

Exercise 4.3. Denote by $\mathbf{Y} = (Y_1, \dots, Y_n)^\top$ (resp. $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_n)^\top$) the vector of observations (resp. of errors) and by Ψ the $p \times n$ design matrix. Consider the linear Gaussian model under the homogeneous noise assumption

$$\mathbf{Y} = \Psi^\top \boldsymbol{\theta}^* + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 I_n),$$

where $\boldsymbol{\theta}^* = (\theta_1^*, \dots, \theta_p^*)^\top \in \mathbb{R}^p$ is an unknown parameter vector.

Prove that the maximum likelihood estimator for the parameter θ^* is equal to

$$\tilde{\theta} = (\Psi\Psi^\top)^{-1}\Psi Y. \quad (4.6)$$

The log-likelihood function equals

$$L(\theta) = -\frac{n}{2} \log(2\pi) - \frac{\log\{\det(\Sigma)\}}{2} - \frac{1}{2}(Y - \Psi^\top\theta)^\top \Sigma^{-1}(Y - \Psi^\top\theta).$$

In the case of homogenous noise, the last formula boils down to:

$$L(\theta) = -\frac{n}{2} \log(2\pi) - \frac{2n \log \sigma}{2} - \frac{1}{2}\sigma^{-2}(Y - \Psi^\top\theta)^\top (Y - \Psi^\top\theta).$$

The maximum likelihood estimator can be found as a solution of the equation

$$\frac{\partial}{\partial \theta} L(\theta) = -\frac{1}{2}\sigma^{-2}(-2\Psi Y + 2\Psi\Psi^\top\theta) = 0,$$

and (4.6) follows.

Exercise 4.4. Consider the model from the previous exercise

$$Y = \Psi^\top\theta^* + \varepsilon,$$

but with colored noise, i.e. $\varepsilon \sim \mathcal{N}(0, \Sigma)$. Prove that for any θ

1. $\text{Var}\{\nabla L(\theta)\} = \Psi\Sigma^{-1}\Psi^\top$,
2. $\nabla^2 L(\theta) = -\Psi\Sigma^{-1}\Psi^\top$.

(So $\text{Var}\{\nabla L(\theta)\}$ and $\nabla^2 L(\theta)$ don't depend on θ).

1. The log-likelihood for this model is equal to

$$L(\theta) = -\frac{1}{2}(Y - \Psi^\top\theta)^\top \Sigma^{-1}(Y - \Psi^\top\theta) - \frac{1}{2} \log \{(2\pi)^n \det \Sigma\}.$$

This yields for its gradient $\nabla L(\theta)$:

$$\nabla L(\theta) = \Psi\Sigma^{-1}(Y - \Psi^\top\theta), \quad (4.7)$$

and in view of $\text{Var}(Y) = \Sigma$, it holds

$$\text{Var}\{\nabla L(\theta)\} = \text{Var}(\Psi\Sigma^{-1}Y) = \Psi\Sigma^{-1} \underbrace{\text{Var} Y}_{\Sigma} \Sigma^{-1}\Psi^\top = \Psi\Sigma^{-1}\Psi^\top.$$

as required.

2. The required formula directly follows from (4.7).

Exercise 4.5. Consider univariate polynomial regression of degree $p - 1$

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

where X_i are fixed points, errors ε_i are assumed to be i.i.d. normal, and the function f can be represented as

$$f(x) = \theta_1^* + \theta_2^*x + \dots + \theta_p^*x^{p-1}.$$

At the same time, for any fixed point x_0 , this function can also be written as

$$f(x) = u_1^* + u_2^*(x - x_0) + \dots + u_p^*(x - x_0)^{p-1}.$$

1. Write the matrices Ψ and $\check{\Psi}$ such that for any given design points X_i , $i = 1, \dots, n$,

$$\mathbf{f} = \Psi^\top \boldsymbol{\theta}^* = \check{\Psi}^\top \mathbf{u}^*, \quad (4.8)$$

where

$$\begin{aligned} \mathbf{f} &= (f(X_1), f(X_2), \dots, f(X_n))^\top, \\ \boldsymbol{\theta}^* &= (\theta_1^*, \theta_2^*, \dots, \theta_n^*)^\top, \\ \mathbf{u}^* &= (u_1^*, u_2^*, \dots, u_n^*)^\top. \end{aligned}$$

Compute also the matrices $\Psi\Psi^\top$ and $\check{\Psi}\check{\Psi}^\top$.

2. Describe an orthogonal transformation A such that

$$\check{\Psi} = A\Psi$$

- For $p = 1$,
- For $p > 1$ (with assumption that $n \geq p$).

1.

$$\Psi = \begin{pmatrix} 1 & 1 & \dots & 1 \\ X_1 & X_2 & \dots & X_n \\ \vdots & \vdots & \ddots & \vdots \\ X_1^{p-1} & X_2^{p-1} & \dots & X_n^{p-1} \end{pmatrix}.$$

Denote $B \stackrel{\text{def}}{=} \Psi\Psi^\top$. Denote also elements of B by b_{ij} , $i, j = 1 \dots p$, and elements of Ψ by ψ_{ij} , $i = 1 \dots p$, $j = 1 \dots n$. Note that $\psi_{ij} = X_j^{i-1}$. By the definition of the product of matrices,

$$b_{ij} = \sum_{s=1}^n \psi_{is} \psi_{js} = \sum_{s=1}^n X_s^{i+j-2}, \quad i, j = 1, \dots, p.$$

So,

$$\Psi \Psi^\top = \begin{pmatrix} n & \sum X_s & \dots & \sum X_s^{p-1} \\ \sum X_s & \sum X_s^2 & \dots & \sum X_s^p \\ \vdots & \vdots & \ddots & \vdots \\ \sum X_s^{p-1} & \sum X_s^p & \dots & \sum X_s^{2p-2} \end{pmatrix}.$$

Analogously,

$$\check{\Psi} = \begin{pmatrix} 1 & 1 & \dots & 1 \\ X_1 - x_0 & X_2 - x_0 & \dots & X_n - x_0 \\ \vdots & \vdots & \ddots & \vdots \\ (X_1 - x_0)^{p-1} & (X_2 - x_0)^{p-1} & \dots & (X_n - x_0)^{p-1} \end{pmatrix}.$$

and

$$\check{\Psi} \check{\Psi}^\top = \begin{pmatrix} n & \sum (X_s - x_0) & \dots & \sum (X_s - x_0)^{p-1} \\ \sum (X_s - x_0) & \sum (X_s - x_0)^2 & \dots & \sum (X_s - x_0)^p \\ \vdots & \vdots & \ddots & \vdots \\ \sum (X_s - x_0)^{p-1} & \sum (X_s - x_0)^p & \dots & \sum (X_s - x_0)^{2p-2} \end{pmatrix}.$$

- 2.(a) In the case of $p = 1$, Ψ and $\check{\Psi}$ are real numbers equal to 1. Then A is an identical transformation.
 (b) Let now $p > 1$. First we prove two lemmas.

Lemma 4.1.

$$\mathbf{u}^* = A^{-1} \boldsymbol{\theta}^*.$$

Proof.

$$\mathbf{f} = \Psi^\top \boldsymbol{\theta}^* = \Psi^\top (A A^{-1}) \boldsymbol{\theta}^* = \check{\Psi}^\top (A^{-1} \boldsymbol{\theta}^*).$$

On the other hand, (4.8) yields

$$\mathbf{f} = \check{\Psi}^\top \mathbf{u}^*.$$

Hence,

$$\check{\Psi}^\top (A^{-1} \boldsymbol{\theta}^* - \mathbf{u}^*) = 0. \quad (4.9)$$

Note that the matrix $\check{\Psi}^T$ is a $n \times p$ matrix with $n > p$. This matrix has a rank p , because first p columns form a Vandermonde matrix with determinant

$$\det(\check{\Psi}) = \prod_{1 \leq i \leq j \leq p} (X_j - X_i) \neq 0.$$

Hence, equality (4.9) yields $A^{-1}\theta^* - u^* = 0$ as required.

Lemma 4.2.

$$u_m^* = \frac{1}{(m-1)!} f^{(m-1)}(x_0), \quad m = 1, \dots, p. \quad (4.10)$$

Proof. Recall that

$$f(x) = u_1^* + u_2^*(x - x_0) + \dots + u_p^*(x - x_0)^{p-1}$$

Then for $m = 1, \dots, p$

$$f^{(m-1)}(x) = 1 \dots (m-1) u_m^* + 2 \dots m u_{m+1}^*(x - x_0) + \dots$$

Substitution $x = x_0$ gives

$$f^{(m-1)}(x_0) = (m-1)! u_m^*$$

and the statement of the lemma follows.

Now substitute the expression

$$f(x) = \theta_1^* + \theta_2^* x + \dots + \theta_p^* x^{p-1} = \sum_{k=1}^p \theta_k^* x^{k-1}.$$

into (4.10):

$$u_m^* = \frac{1}{(m-1)!} f^{(m-1)}(x_0) = \frac{1}{(m-1)!} \sum_{k=m}^p \theta_k^* x_0^{k-m}, \quad m = 1, \dots, p$$

According to Lemma 4.1,

$$A^{-1} = \begin{pmatrix} 1/0! & x_0/0! & x_0^2/0! & \dots & x_0^{p-1}/0! \\ 0 & 1/1! & x_0/1! & \dots & x_0^{p-2}/1! \\ 0 & 0 & 1/2! & \dots & x_0^{p-3}/2! \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 1/(p-1)! \end{pmatrix}.$$

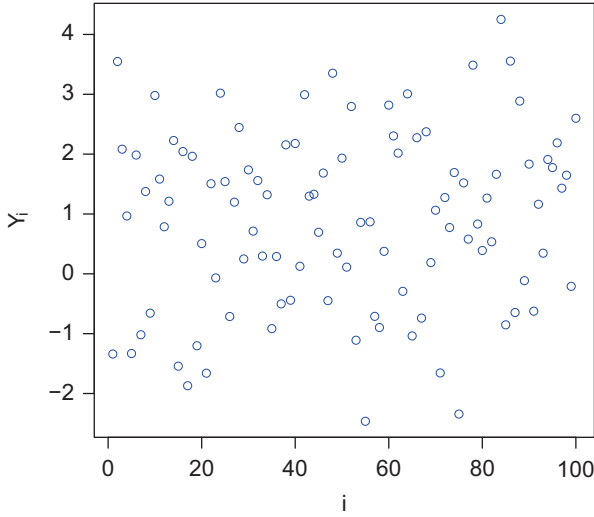


Fig. 4.1 Consider the model on a sample (i, Y_i) with $\theta^* = (1, 1)^\top$ and $\sigma = 1$. $\tilde{\theta} = (1.012115, 1.099624)^\top$. [MSEExercise471](#)

Exercise 4.6. Consider the model

$$Y_i = \cos(2X_i) \theta_1^* + \sin(X_i/2) \theta_2^* + \varepsilon_i, \quad i = 1, \dots, n,$$

where

- $\theta^* = (\theta_1^*, \theta_2^*)^\top$ is an unknown parameter vector;
- $\{\varepsilon_i\}_{i=1}^n$ are i.i.d. $\mathcal{N}(0, \sigma^2)$, $X_i = (-1)^i \pi$, $i = 1, \dots, n$;
- n is even.

1. Rewrite this model as the linear Gaussian model, and show that the design is orthogonal.
2. Compute the maximum likelihood estimator for the parameter θ^* (Figs. 4.1 and 4.2).

1. This model can be rewritten as

$$Y = \Psi^\top \theta^* + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \Sigma),$$

where $Y = (Y_1, \dots, Y_n)^\top$, $\theta^* = (\theta_1^*, \theta_2^*)^\top$, $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)^\top \sim \mathcal{N}(0, \sigma^2 I_n)$, and

$$\Psi^\top = \begin{pmatrix} \cos(2X_1) & \sin(X_1/2) \\ \cos(2X_2) & \sin(X_2/2) \\ \vdots & \vdots \\ \cos(2X_n) & \sin(X_n/2) \end{pmatrix}.$$

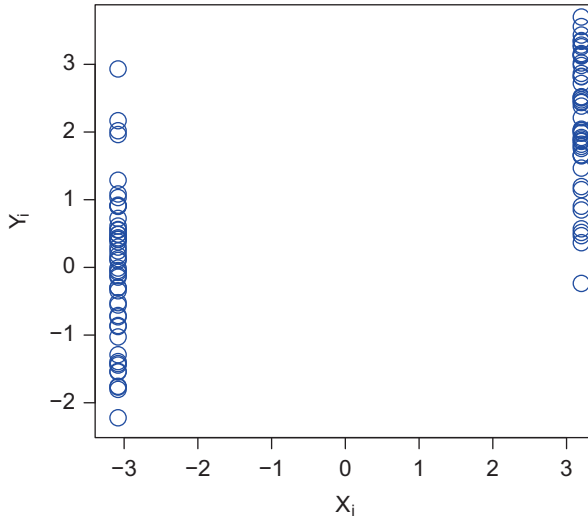


Fig. 4.2 Consider the model on a sample (X_i, Y_i) with $\theta^* = (1, 1)^\top$ and $\sigma = 1$. $\tilde{\theta} = (1.012115, 1.099624)^\top$. [MSEExercise472](#)

Substitution $X_i = (-1)^i \pi, i = 1, \dots, n$ yields

$$\Psi^\top = \begin{pmatrix} \cos(-2\pi) & \sin(-\pi/2) \\ \cos(2\pi) & \sin(\pi/2) \\ \vdots & \vdots \\ \cos\{(-1)^{n-1}2\pi\} & \sin\{(-1)^{n-1}\pi/2\} \\ \cos\{(-1)^n 2\pi\} & \sin\{(-1)^n \pi/2\} \end{pmatrix} = \begin{pmatrix} 1 & -1 \\ 1 & 1 \\ \vdots & \vdots \\ 1 & -1 \\ 1 & 1 \end{pmatrix}.$$

If n is even then the columns of the matrix $\Psi^\top$ are orthogonal:

$$\Psi \Psi^\top = \begin{pmatrix} n & 0 \\ 0 & n \end{pmatrix} = n I_2.$$

2. The general formula for estimation under the homogeneous noise assumption simplifies drastically for this design:

$$\tilde{\theta} = (\Psi \Psi^\top)^{-1} \Psi Y = n^{-1} \Psi Y.$$

Thus,

$$\begin{aligned} \tilde{\theta}_1 &= \frac{1}{n} (Y_1 + Y_2 + \dots + Y_n), \\ \tilde{\theta}_2 &= \frac{1}{n} (-Y_1 + Y_2 - \dots + Y_n). \end{aligned}$$

Exercise 4.7. *Let*

$$\Psi_1 = \begin{pmatrix} -1/2 \\ 1/\sqrt{2} \\ 1/2 \end{pmatrix}, \quad \Psi_2 = \begin{pmatrix} 1/2 \\ 1/\sqrt{2} \\ -1/2 \end{pmatrix} \quad \text{and} \quad Y = \begin{pmatrix} 1 \\ \sqrt{2} \\ -1 \end{pmatrix}.$$

Consider the model

$$Y = f + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 I_n).$$

1. *Let $\Psi^\top$ be a 3×2 matrix with columns Ψ_1 and Ψ_2 , and let f be:*

$$f = \Psi^\top \theta^*,$$

for some $\theta^ \in \mathbb{R}^2$. Compute the MLE estimator of θ^* .*

2. *Let*

$$\max_{\theta} \|f - \Psi^\top \theta\| > 0.$$

Find the explicit formula for $\theta^\dagger$ as linear transformation of the vector f .

1. First note that the design matrix Ψ is orthonormal.

$$\begin{pmatrix} -1/2 & 1/\sqrt{2} & 1/2 \\ 1/2 & 1/\sqrt{2} & -1/2 \end{pmatrix} \begin{pmatrix} -1/2 & 1/2 \\ 1/\sqrt{2} & 1/\sqrt{2} \\ 1/2 & -1/2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

This fact simplifies the computation:

$$\tilde{\theta} = \Psi Y = \frac{1}{2} \begin{pmatrix} -1 & \sqrt{2} & 1 \\ 1 & \sqrt{2} & -1 \end{pmatrix} \begin{pmatrix} 1 \\ \sqrt{2} \\ -1 \end{pmatrix} = \begin{pmatrix} 0 \\ 2 \end{pmatrix}.$$

2. For the orthonormal design, the computation also simplifies drastically:

$$\theta^\dagger = \Psi f = \frac{1}{2} \begin{pmatrix} -1 & \sqrt{2} & 1 \\ 1 & \sqrt{2} & -1 \end{pmatrix} f.$$

Exercise 4.8. 1. *Consider the model*

$$Y = f + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \Sigma)$$

with $f = \Psi_1 \theta_1$, $\Psi_1 \in \mathbb{R}^n$, $\theta_1 \in \mathbb{R}$.

Find the formula for $\theta_1^\dagger$ as linear transformation of the vector f .

2. Assume now that the true stochastic is

$$Y = \Psi_1 \theta_1^* + \Psi_2 \theta_2^* + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \Sigma),$$

for some design Ψ_1, Ψ_2 , find the conditions when $\theta_1^\dagger = \theta_1^*$.

1. In this case matrix Ψ is a matrix with one string Ψ_1 . According to the general formula (see the proof of Theorem 4.4.3 of Spokoiny and Dickhaus 2014),

$$\theta_1^\dagger = \left(\Psi_1^\top \Sigma^{-1} \Psi_1 \right)^{-1} \Psi_1^\top \Sigma^{-1} f = \frac{\Psi_1^\top \Sigma^{-1} f}{\Psi_1^\top \Sigma^{-1} \Psi_1}.$$

2. Now we should put $f = \Psi_1 \theta_1^* + \Psi_2 \theta_2^*$ in the formula for $\theta_1^\dagger$:

$$\theta_1^\dagger = \frac{\Psi_1^\top \Sigma^{-1} f}{\Psi_1^\top \Sigma^{-1} \Psi_1} = \frac{\Psi_1^\top \Sigma^{-1} (\Psi_1 \theta_1^* + \Psi_2 \theta_2^*)}{\Psi_1^\top \Sigma^{-1} \Psi_1} = \theta_1^* + \theta_2^* \frac{\Psi_1^\top \Sigma^{-1} \Psi_2}{\Psi_1^\top \Sigma^{-1} \Psi_1}.$$

This means, that $\theta_1^\dagger = \theta_1^*$ if and only if

$$\theta_2^* \Psi_1^\top \Sigma^{-1} \Psi_2 = 0,$$

or, equivalently, if and only if (a) $\theta_2^* = 0$ or (b) $\Psi_1^\top \Sigma^{-1} \Psi_2 = 0$. Condition (a) means that the model considered in the first item is true (note that $\theta_1^\dagger = \theta_1^*$ is obviously fulfilled in this case). Condition (b) is a condition on the design.

Exercise 4.9. Consider the model

$$Y = \Psi^\top \theta + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 I), \quad (4.11)$$

and let the true stochastic be

$$Y = \Psi^\top \theta^* + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \Sigma_0). \quad (4.12)$$

with a fixed covariance matrix Σ_0 . Prove that in this case

$$\theta^\dagger = \theta^*.$$

In the model (4.11), $f = \Psi^\top \theta$ and $\Sigma = \sigma^2 I$. Substituting these values in the general formula for solving MLE estimator gives

$$\theta^\dagger = (\Psi \Psi^\top)^{-1} \Psi f.$$

According to the true model (4.12), $f = \Psi^\top \theta^*$. Hence,

$$\theta^\dagger = (\Psi \Psi^\top)^{-1} \Psi \Psi^\top \theta^* = \theta^*.$$

Exercise 4.10. Let ξ be the stochastic component of $\tilde{\theta}$ built for the misspecified linear model $Y = \Psi^\top \theta^* + \varepsilon$ with $\text{Var}(\varepsilon) = \Sigma$. Let the true noise variance be Σ_0 .

1. Prove that the variance of $\tilde{\theta}$ is equal to

$$\text{Var}_{\Sigma_0}(\tilde{\theta}) = (\Psi \Sigma^{-1} \Psi^\top)^{-1} \Psi \Sigma^{-1} \Sigma_0 \Sigma^{-1} \Psi^\top (\Psi \Sigma^{-1} \Psi^\top)^{-1}. \quad (4.13)$$

2. Check that the matrix in the right hand side of (4.13) is of dimension $p \times p$.

1. Note that

$$\tilde{\theta} = \Xi Y = \Xi f + \Xi \varepsilon,$$

where $\Xi = (\Psi \Sigma^{-1} \Psi^\top)^{-1} \Psi \Sigma^{-1}$. Then

$$\text{Var}_{\Sigma_0}(\tilde{\theta}) = \text{Var}_{\Sigma_0}(\Xi \varepsilon) = \Xi \mathbb{E}_{\Sigma_0} [\varepsilon \varepsilon^\top] \Xi^\top = \Xi \Sigma_0 \Xi^\top,$$

and (4.13) follows.

2. Recall that Ψ is a $p \times n$ matrix, Σ and Σ_0 $n \times n$ matrices. Then $\Psi \Sigma^{-1} \Psi^\top$ is a $p \times p$ matrix, and the required fact follows:

$$\underbrace{(\Psi \Sigma^{-1} \Psi^\top)^{-1}}_{p \times p} \underbrace{\Psi}_{p \times n} \underbrace{\Sigma^{-1}}_{n \times n} \underbrace{\Sigma_0}_{n \times n} \underbrace{\Sigma^{-1}}_{n \times n} \underbrace{\Psi^\top}_{n \times p} \underbrace{(\Psi \Sigma^{-1} \Psi^\top)^{-1}}_{p \times p}.$$

Exercise 4.11. Assume $Y = \Psi^\top \theta^* + \varepsilon$ with $\varepsilon \sim \mathcal{N}(0, \Sigma)$. Then for any $\mu < 1$

$$\mathbb{E}_{\theta^*} \exp\{\mu L(\tilde{\theta}, \theta^*)\} = (1 - \mu)^{-p/2},$$

where p is the dimension of the vector θ^* .

The distribution of $2L(\tilde{\theta}, \theta^*)$ is chi-squared with p degrees of freedom. This means that there exist p independent standard normal distributed variables ξ_i , $i = 1 \dots p$ such that

$$2L(\tilde{\theta}, \theta^*) = \sum_{i=1}^p \xi_i^2.$$

Then (under $\mathbb{P}_{\theta^*}$)

$$\mathbb{E} \exp\{\mu L(\tilde{\theta}, \theta^*)\} = \mathbb{E} \exp\left(\frac{1}{2} \mu \sum_{i=1}^p \xi_i^2\right) = \prod_{i=1}^p \mathbb{E} \exp\left(\frac{1}{2} \mu \xi_i^2\right)$$

So one has to compute $\mathbb{E} \exp(\frac{1}{2}\mu\xi^2)$ for a standard normal ξ :

$$\mathbb{E} \exp(\frac{1}{2}\mu\xi^2) = \int e^{\mu x^2/2} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx = \frac{1}{\sqrt{2\pi}} \int \exp(-\frac{1-\mu}{2}x^2) dx.$$

The change $t = \sqrt{1-\mu} x$ yields

$$\mathbb{E} \exp(\frac{1}{2}\mu\xi^2) = (1-\mu)^{-1/2} \frac{1}{\sqrt{2\pi}} \int \exp(-\frac{t^2}{2}) dt = (1-\mu)^{-1/2}.$$

and completes the proof.

Exercise 4.12. Consider the model

$$\mathbf{Y} = \Psi^\top \boldsymbol{\theta} + \boldsymbol{\varepsilon}. \quad (4.14)$$

with homogeneous errors $\boldsymbol{\varepsilon}$: $\mathbb{E}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top = \sigma^2 I_n$.

1. Prove that there exists an orthogonal transformation $U : \mathbb{R}^p \rightarrow \mathbb{R}^p$ leading to the spectral representation

$$\mathbf{Z} = \Lambda \mathbf{u} + \boldsymbol{\xi},$$

where $\mathbf{Z} = U\Psi\mathbf{Y} \in \mathbb{R}^p$, Λ is a diagonal $p \times p$ matrix, $\mathbf{u} = U\boldsymbol{\theta} \in \mathbb{R}^p$, and errors $\boldsymbol{\xi} = U\Psi\boldsymbol{\varepsilon} \in \mathbb{R}^p$ are uncorrelated: $\mathbb{E}\boldsymbol{\xi}\boldsymbol{\xi}^\top = \sigma^2 \Lambda$.

2. Prove that if $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 I_n)$, then the vector $\boldsymbol{\xi}$ is also normal, i.e.: $\boldsymbol{\xi} \sim \mathcal{N}(0, \sigma^2 \Lambda)$.
1. The operator $\Psi\Psi^\top$ is self – adjoint, therefore there exists an orthogonal transformation U such that

$$U\Psi\Psi^\top U^\top = \text{diag}(\lambda_1, \dots, \lambda_p) = \Lambda,$$

where λ_i , $i = 1, \dots, p$ are the eigenvalues of the operator $\Psi\Psi^\top$. Applying the transformation $U\Psi$ to both sides of (4.14), we arrive at

$$\underbrace{U\Psi\mathbf{Y}}_{\mathbf{Z}} = \underbrace{U\Psi\Psi^\top U^\top}_{\Lambda} \underbrace{U\boldsymbol{\theta}}_{\mathbf{u}} + \underbrace{U\Psi\boldsymbol{\varepsilon}}_{\boldsymbol{\xi}}.$$

The errors $\boldsymbol{\xi}$ are uncorrelated, because

$$\mathbb{E}\boldsymbol{\xi}\boldsymbol{\xi}^\top = U\Psi \underbrace{\mathbb{E}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top}_{\sigma^2 I_n} \Psi^\top U^\top = \sigma^2 U\Psi\Psi^\top U^\top = \sigma^2 \Lambda.$$

2. If $\boldsymbol{\varepsilon}$ is normal, then $\boldsymbol{\xi}$ is also normal as a result of the linear transformation of $\boldsymbol{\varepsilon}$:

$$\boldsymbol{\xi} = U\Psi\boldsymbol{\varepsilon}.$$

This completes the proof (Härdle and Simar 2011).

Exercise 4.13. 1. Find the matrix Λ for the designs from Exercises 4.4 and 4.5.
2. Let the matrix Λ be equal to the identity matrix. What can you say about the design matrix?

1. The design Ψ in both cases is orthonormal. Thus, we conclude that

$$\Lambda = U \underbrace{\Psi\Psi^\top}_{I_p} U^\top = I_p.$$

2. Assume now that

$$U\Psi\Psi^\top U^\top = I_p. \quad (4.15)$$

Multiplying (4.15) by $U^\top$ from the left side and by U to the right side, we arrive at

$$\Psi\Psi^\top = U^\top U = I_p,$$

because the matrix U is orthogonal. So, the design is orthonormal.

Exercise 4.14. Consider the model

$$\mathbf{Y} = \Psi^\top \boldsymbol{\theta} + \boldsymbol{\varepsilon} \quad \text{with} \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(0, \Sigma).$$

Check that the linear transformation $\check{\mathbf{Y}} = \Sigma^{-1/2} \mathbf{Y}$ of the data does not change the value of the log-likelihood ratio $L(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ for any $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2$.

Recall that

$$L(\boldsymbol{\theta}) = -\frac{1}{2}(\mathbf{Y} - \Psi^\top \boldsymbol{\theta})^\top \Sigma^{-1}(\mathbf{Y} - \Psi^\top \boldsymbol{\theta}) + R,$$

where $R = -n \log(2\pi)/2 - \log(\det \Sigma)/2$ does not depend on $\mathbf{Y}$ and $\boldsymbol{\theta}$. Then

$$\begin{aligned} L(\boldsymbol{\theta}) &= -\frac{1}{2}(\Sigma^{-1/2} \mathbf{Y} - \Sigma^{-1/2} \Psi^\top \boldsymbol{\theta})^\top (\Sigma^{-1/2} \mathbf{Y} - \Sigma^{-1/2} \Psi^\top \boldsymbol{\theta}) + R \\ &= -\frac{1}{2}(\check{\mathbf{Y}} - \check{\Psi}^\top \boldsymbol{\theta})^\top (\check{\mathbf{Y}} - \check{\Psi}^\top \boldsymbol{\theta}) + R, \end{aligned}$$

where $\check{\Psi} = \Psi \Sigma^{-1/2}$.

The transformed data $\check{Y}$ follows:

$$\check{Y} = \check{\Psi}^\top \theta^* + \xi$$

with $\xi = \Sigma^{-1/2} \epsilon \sim \mathcal{N}(0, I_p)$ yielding the log-likelihood

$$\begin{aligned} \check{L}(\theta) &= -\frac{1}{2}(\check{Y} - \check{\Psi}^\top \theta)^\top (\check{Y} - \check{\Psi}^\top \theta) + \check{R} \\ &= L(\theta) + \check{R} - R \end{aligned}$$

where $\check{R} = -n \log(2\pi)/2$. Thus, we conclude that for any θ_1 and θ_2 ,

$$\begin{aligned} L(\theta_1, \theta_2) &= L(\theta_1) - L(\theta_2) \\ &= \check{L}(\theta_1) - \check{L}(\theta_2). \end{aligned}$$

Exercise 4.15. Consider the model from Exercise 4.5:

$$Y = \Psi^\top \theta^* + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2 I_n),$$

where $\Psi^\top = (\Psi_1 \ \Psi_2)^\top$ is a 3×2 -matrix with rows:

$$\Psi_1 = \begin{pmatrix} -1/2 \\ 1/\sqrt{2} \\ 1/2 \end{pmatrix}, \quad \Psi_2 = \begin{pmatrix} 1/2 \\ 1/\sqrt{2} \\ -1/2 \end{pmatrix}.$$

1. Explain the Wilks' phenomenon in this case.
2. Compute the likelihood-based confidence ellipsoids for the parameter θ^* , if $Y = (1, \sqrt{2}, -1)^\top$ and $\sigma = 1$.

1. From Theorem 4.5.1 of Spokoiny and Dickhaus (2014), we know that

$$L(\tilde{\theta}, \theta^*) = \frac{1}{2}(\tilde{\theta} - \theta^*)^\top \Psi \Sigma^{-1} \Psi^\top (\tilde{\theta} - \theta^*).$$

In our case, $\Psi \Sigma^{-1} \Psi^\top = \sigma^{-2} I_2$, and therefore:

$$L(\tilde{\theta}, \theta^*) = \frac{1}{2\sigma^2} \|\tilde{\theta} - \theta^*\|^2$$

The Wilks' phenomenon tells us that the distribution of $L(\tilde{\theta}, \theta^*)$ is χ_2^2 for any σ and any θ^* .

2. An α – confidence set for the parameter θ^* may be constructed as follows:

$$\mathcal{E}(\mathfrak{z}) = \{\theta : L(\tilde{\theta}, \theta) \leq \mathfrak{z}_\alpha\},$$

where $\mathfrak{z}_\alpha$ is defined by $P\{U > 2\mathfrak{z}_\alpha\} = \alpha$, $U \sim \chi_2^2$.

As it was already shown in Exercise 4.5, $\tilde{\theta} = (0, 2)^\top$. Therefore:

$$\mathcal{E}(\mathfrak{z}) = \{\theta : \theta_1^2 + (\theta_2 - 2)^2 \leq 2\mathfrak{z}_\alpha\}.$$

is an α – confidence set for the parameter θ^* .

Exercise 4.16. Consider the estimate obtained by the method of Tikhonov regularization

$$\tilde{\theta}_\alpha = (\Psi\Psi^\top + \alpha I_p)^{-1}\Psi Y.$$

1. Prove that the bias of this estimate

$$B(\alpha) = \|\mathbb{E}\tilde{\theta}_\alpha - \theta^*\|$$

grows with the regularization parameter α .

2. Prove that the trace of the variance matrix of $\tilde{\theta}_\alpha$,

$$V(\alpha) = \text{tr } \mathbb{E}\left\{\left(\tilde{\theta}_\alpha - \mathbb{E}\tilde{\theta}_\alpha\right)\left(\tilde{\theta}_\alpha - \mathbb{E}\tilde{\theta}_\alpha\right)^\top\right\},$$

decreases in α .

1. Note that

$$\mathbb{E}\tilde{\theta}_\alpha = (\Psi\Psi^\top + \alpha I_p)^{-1}\Psi\Psi^\top\theta^*,$$

resulting in the bias:

$$B(\alpha) = \left\| \left\{ (\Psi\Psi^\top + \alpha I_p)^{-1}\Psi\Psi^\top - I_p \right\} \theta^* \right\|.$$

The matrix $\Psi\Psi^\top$ is positive definite. The Jordan decomposition yields an orthogonal matrix U and positive numbers $\lambda_1, \dots, \lambda_p$:

$$\Psi\Psi^\top = U \text{diag}(\lambda_1, \dots, \lambda_p) U^\top.$$

Then

$$\begin{aligned} (\Psi\Psi^\top + \alpha I_p)^{-1}\Psi\Psi^\top - I_p &= U \text{diag}\left(\frac{\lambda_1}{\lambda_1 + \alpha} - 1, \dots, \frac{\lambda_p}{\lambda_p + \alpha} - 1\right) U^\top \\ &= U \text{diag}\left(\frac{-1}{1 + \lambda_1/\alpha}, \dots, \frac{-1}{1 + \lambda_p/\alpha}\right) U^\top. \end{aligned}$$

This yields:

$$\begin{aligned} B^2(\alpha) &= \boldsymbol{\theta}^{*\top} U \operatorname{diag} \left\{ \frac{1}{(1 + \lambda_1/\alpha)^2}, \dots, \frac{1}{(1 + \lambda_p/\alpha)^2} \right\} U^\top \boldsymbol{\theta}^* \\ &\stackrel{\text{def}}{=} \boldsymbol{\theta}^{*\top} U D(\alpha) U^\top \boldsymbol{\theta}^*, \end{aligned}$$

where the matrix $D(\alpha)$ is diagonal. Let now α_1 and α_2 be two positive numbers such that $\alpha_1 > \alpha_2$. From

$$\frac{1}{(1 + \lambda_i/\alpha_1)^2} \geq \frac{1}{(1 + \lambda_i/\alpha_2)^2}, \quad i = 1, \dots, p$$

we conclude that $D(\alpha_1) \geq D(\alpha_2)$. Then

$$v^\top D(\alpha_1) v \geq v^\top D(\alpha_2) v$$

for any $v \in \mathbb{R}^p$, in particular for $v = U^\top \boldsymbol{\theta}^*$. This observation completes the proof.

2. Note that

$$\begin{aligned} V(\alpha) &= \operatorname{tr} \mathbb{E} \left\{ \left(\tilde{\boldsymbol{\theta}}_\alpha - \mathbb{E} \tilde{\boldsymbol{\theta}}_\alpha \right) \left(\tilde{\boldsymbol{\theta}}_\alpha - \mathbb{E} \tilde{\boldsymbol{\theta}}_\alpha \right)^\top \right\} \\ &= \operatorname{tr} \left\{ (\Psi \Psi^\top + \alpha I_p)^{-1} \Psi^\top \underbrace{\mathbb{E} (Y - \mathbb{E} Y) (Y - \mathbb{E} Y)^\top}_{=\sigma^2 I_p} \Psi (\Psi \Psi^\top + \alpha I_p)^{-1} \right\} \\ &= \sigma^2 \operatorname{tr} \left\{ (\Psi \Psi^\top + \alpha I_p)^{-1} \Psi^\top \Psi (\Psi \Psi^\top + \alpha I_p)^{-1} \right\}. \end{aligned}$$

Computation in the basis of the eigenvectors of $\Psi \Psi^\top$ yields:

$$(\Psi \Psi^\top + \alpha I_p)^{-1} \Psi^\top \Psi (\Psi \Psi^\top + \alpha I_p)^{-1} = U \operatorname{diag} \left\{ \frac{\lambda_1}{(\alpha + \lambda_1)^2}, \dots, \frac{\lambda_p}{(\alpha + \lambda_p)^2} \right\} U^\top,$$

and we arrive at

$$V(\alpha) = \sigma^2 \sum_{k=1}^p \frac{\lambda_p}{(\alpha + \lambda_p)^2}. \quad (4.16)$$

From (4.16), it directly follows that $V(\alpha)$ is a decreasing function of $\alpha > 0$.

Exercise 4.17. Let $L_G(\boldsymbol{\theta})$ be the penalized log-likelihood

$$L_G(\boldsymbol{\theta}) \stackrel{\text{def}}{=} L(\boldsymbol{\theta}) - \frac{1}{2} \|G\boldsymbol{\theta}\|^2,$$

where G is a symmetric $p \times p$ -matrix. Denote $\tilde{\boldsymbol{\theta}}_G = \operatorname{argmax}_{\boldsymbol{\theta}} L_G(\boldsymbol{\theta})$.

1. Prove that for any parameter θ

$$2L_G(\tilde{\theta}_G, \theta) = (\tilde{\theta}_G - \theta)^\top (\sigma^{-2}\Psi\Psi^\top + G^2)(\tilde{\theta}_G - \theta).$$

2. Denote also $\theta_G = \operatorname{argmax}_\theta \mathbb{E}L_G(\theta)$. Prove that

$$2L_G(\tilde{\theta}_G, \theta_G) = \sigma^{-2}\varepsilon^\top \Pi_G \varepsilon \quad (4.17)$$

with $\Pi_G = \Psi^\top (\Psi\Psi^\top + \sigma^2 G^2)^{-1} \Psi$.

1. Recall that the penalized log-likelihood is equal to

$$\begin{aligned} L_G(\theta) &= L(\theta) - \frac{1}{2} \|G\theta\|^2 \\ &= -\frac{1}{2\sigma^2} \|Y - \Psi^\top \theta\|^2 - \frac{1}{2} \|G\theta\|^2 - \frac{n}{2} \log(2\pi\sigma^2). \end{aligned}$$

Consider $L_G(\tilde{\theta}_G, \theta) = L_G(\tilde{\theta}_G) - L_G(\theta)$ as a function of the second argument θ . This is a quadratic function satisfying $L_G(\tilde{\theta}_G, \tilde{\theta}_G) = 0$. Next, by definition of $\tilde{\theta}_G$, this function attains its minimum exactly at the point $\theta = \tilde{\theta}_G$ implying

$$dL_G(\tilde{\theta}_G, \theta)/d\theta|_{\theta=\tilde{\theta}_G} = 0.$$

Moreover, simple algebra yields

$$d^2 L_G(\tilde{\theta}_G, \theta)/d\theta^2 = \sigma^{-2}\Psi\Psi^\top + G^2$$

for any θ . The Taylor expansion at $\theta = \tilde{\theta}_G$ gives:

$$L_G(\tilde{\theta}_G, \theta) = L_G(\tilde{\theta}_G, \tilde{\theta}_G + \theta - \tilde{\theta}_G) = \frac{1}{2}(\tilde{\theta}_G - \theta)^\top \frac{d^2 L_G(\tilde{\theta}_G, \theta)}{d\theta^2}(\tilde{\theta}_G - \theta)$$

and the required formula for the likelihood ratio follows.

2. A straightforward calculus leads to the following expression for $\tilde{\theta}_G$:

$$\tilde{\theta}_G \stackrel{\text{def}}{=} (\Psi\Psi^\top + \sigma^2 G^2)^{-1} \Psi Y.$$

This gives that

$$\tilde{\theta}_G - \theta_G = (\Psi\Psi^\top + \sigma^2 G^2)^{-1} \Psi \underbrace{(Y - \mathbb{E}Y)}_{\varepsilon},$$

and (4.17) is proven.

Exercise 4.18. Consider the model

$$\mathbf{Y} = \Psi^\top \boldsymbol{\theta}^* + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 I) \quad (4.18)$$

with a two-dimensional parameter $\boldsymbol{\theta}^*$ and orthonormal design, i.e. $\Psi\Psi^\top = I_2$. Consider the penalized log-likelihood

$$L_G(\boldsymbol{\theta}) \stackrel{\text{def}}{=} L(\boldsymbol{\theta}) - \frac{1}{2} \|G\boldsymbol{\theta}\|^2$$

with diagonal matrix G , $G = \text{diag}(\alpha, \beta)$.

- (i) Find the bias of the penalized MLE as a function of α , β , σ , and $\boldsymbol{\theta}^*$.
- (ii) Find the trace of the variance matrix of the penalized MLE as a function of α , β and σ .
- (iii) Show that the bias monotonously increases while the trace of the variance matrix monotonously decreases in each parameter α , β , σ .

1. The penalized MLE is equal $\tilde{\boldsymbol{\theta}}_G = \Xi_G \mathbf{Y}$ with $\Xi_G = (\Psi\Psi^\top + \sigma^2 G^2)^{-1} \Psi$ (see Spokoiny and Dickhaus 2014, Sect. 4.6.2). In our case,

$$\Xi_G = \text{diag}\left(\frac{1}{1 + \alpha^2 \sigma^2}, \frac{1}{1 + \beta^2 \sigma^2}\right) \Psi.$$

The calculation of the bias of this estimate is straightforward:

$$\begin{aligned} B(\alpha, \beta, \sigma, \boldsymbol{\theta}^*) &= \|\mathbb{E} \tilde{\boldsymbol{\theta}}_G - \boldsymbol{\theta}^*\| = \|\Xi_G \Psi \Psi^\top \boldsymbol{\theta}^* - \boldsymbol{\theta}^*\| = \|(\Xi_G - I_2) \boldsymbol{\theta}^*\| \\ &= \left\| \text{diag}\left(-\frac{\alpha^2 \sigma^2}{1 + \alpha^2 \sigma^2}, -\frac{\beta^2 \sigma^2}{1 + \beta^2 \sigma^2}\right) \boldsymbol{\theta}^* \right\| \\ &= \sqrt{\left(\frac{\alpha^2 \sigma^2}{1 + \alpha^2 \sigma^2} \theta_1^*\right)^2 + \left(\frac{\beta^2 \sigma^2}{1 + \beta^2 \sigma^2} \theta_2^*\right)^2}, \end{aligned} \quad (4.19)$$

where θ_1^* , θ_2^* are the components of the vector $\boldsymbol{\theta}^*$.

2. The trace of the variance matrix is equal to $\sigma^2 \text{tr}(\Xi_G \Xi_G^\top)$ in this case (See Spokoiny and Dickhaus 2014, Theorem 4.6.2); therefore

$$\begin{aligned} V(\alpha, \beta, \sigma) &= \sigma^2 \text{tr} \left\{ \text{diag}\left(\frac{1}{1 + \alpha^2 \sigma^2}, \frac{1}{1 + \beta^2 \sigma^2}\right) \Psi \Psi^\top \text{diag}\left(\frac{1}{1 + \alpha^2 \sigma^2}, \frac{1}{1 + \beta^2 \sigma^2}\right) \right\} \\ &= \sigma^2 \left\{ \left(\frac{1}{1 + \alpha^2 \sigma^2}\right)^2 + \left(\frac{1}{1 + \beta^2 \sigma^2}\right)^2 \right\}. \end{aligned} \quad (4.20)$$

3. From (4.19) we conclude the monotonicity in α . Indeed by dividing the first term in (4.19) by α^2 one obtains an increasing function. Similarly the monotonicity

with respect to β , σ can be seen. The same mechanism applies to (4.20), completes the proof.

Exercise 4.19. Consider the model

$$Y = \Psi^\top \theta^* + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 I_p)$$

with an orthonormal design, i.e. $\Psi\Psi^\top = I_p$. Define the shrinkage estimate of $\theta^* = (\theta_1, \dots, \theta_p)^\top$:

$$\tilde{\theta}_{\alpha,j} = \alpha_j \psi_j^\top Y$$

where $\alpha_j \in (0, 1)$, $\alpha = (\alpha_1, \dots, \alpha_p)^\top$, and $\psi_j^\top$ is the j -th column of the matrix $\Psi^\top$, $j = 1, \dots, p$. Denote the estimate of $f = \Psi^\top \theta^*$ by $\tilde{f}_\alpha = \Psi^\top \tilde{\theta}_\alpha$.

1. Prove that the risk $R(\tilde{f}_\alpha) \stackrel{\text{def}}{=} \mathbb{E} \|\tilde{f}_\alpha - f\|^2$ of this estimate fulfills

$$R(\tilde{f}_\alpha) = \sum_{j=1}^p f_j^2 (1 - \alpha_j)^2 + \sigma^2 \sum_{j=1}^p \alpha_j^2,$$

where $f_j = \psi_j^\top f$.

2. Specify the risk for the case of projection estimate, i.e. $\alpha_j = 1$ ($j \leq m$) with fixed m .

1. The estimate $\tilde{f}_\alpha$ allows the following representation:

$$\tilde{f}_\alpha = \Psi^\top \tilde{\theta}_\alpha = \sum_{j=1}^p \tilde{\theta}_{\alpha,j} \psi_j = \sum_{j=1}^p \alpha_j \psi_j \psi_j^\top Y.$$

The bias – variance decomposition gives

$$R(\tilde{f}_\alpha) = B^2(\tilde{f}_\alpha) + V(\tilde{f}_\alpha), \quad (4.21)$$

where

$$\begin{aligned} B^2(\tilde{f}_\alpha) &= \|\mathbb{E} \tilde{f}_\alpha - f\|^2 = \left\| \sum_{j=1}^p \alpha_j \psi_j \psi_j^\top f - f \right\|^2 \\ &= \left\| \sum_{j=1}^p (\alpha_j - 1) \psi_j \psi_j^\top f \right\|^2 \\ &= \sum_{j=1}^p (\alpha_j - 1)^2 (\psi_j^\top f)^2, \end{aligned}$$

and

$$\begin{aligned}
 V(\tilde{\mathbf{f}}_\alpha) &= \mathbb{E} \|\tilde{\mathbf{f}}_\alpha - \mathbb{E} \tilde{\mathbf{f}}_\alpha\|^2 = \mathbb{E} \|\tilde{\boldsymbol{\theta}}_\alpha - \mathbb{E} \tilde{\boldsymbol{\theta}}_\alpha\|^2 \\
 &= \sum_{j=1}^p \text{Var } \theta_{\alpha,j}^2 = \sum_{j=1}^p \alpha_j^2 \psi_j^\top \text{Var } Y \psi_j \\
 &= \sigma^2 \sum_{j=1}^p \alpha_j^2.
 \end{aligned}$$

2. For the case of projection estimation, the formula for the risk boils down to

$$R(\tilde{\mathbf{f}}_\alpha) = \sum_{j=1}^m f_j^2 + m\sigma^2.$$

Exercise 4.20. Consider the model

$$\mathbf{Y} = \Psi^\top \boldsymbol{\theta}^* + \Phi^\top \boldsymbol{\eta}^* + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_n), \quad (4.22)$$

where $\boldsymbol{\theta}^* \in \mathbb{R}^p$ is the target parameter, $\boldsymbol{\eta}^* \in \mathbb{R}^k$ is the nuisance parameter, Ψ is the $p \times n$ matrix, while Φ is the $k \times n$ matrix.

Let some value $\boldsymbol{\eta}^\circ$ of the nuisance parameter be fixed. Define the estimate $\tilde{\boldsymbol{\theta}}(\boldsymbol{\eta}^\circ)$ by partial optimization of the joint log-likelihood $L(\boldsymbol{\theta}, \boldsymbol{\eta}^\circ)$ w.r.t. the first parameter $\boldsymbol{\theta}$:

$$\tilde{\boldsymbol{\theta}}(\boldsymbol{\eta}^\circ) = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} L(\boldsymbol{\theta}, \boldsymbol{\eta}^\circ).$$

1. Prove that if the adaptivity condition is fulfilled

$$\Psi \Phi^\top = \mathbf{0}, \quad (4.23)$$

then the partial estimate $\tilde{\boldsymbol{\theta}}(\boldsymbol{\eta}^\circ)$ does not depend on $\boldsymbol{\eta}^\circ$:

$$\tilde{\boldsymbol{\theta}}(\boldsymbol{\eta}^\circ) = (\Psi \Psi^\top)^{-1} \Psi \mathbf{Y}. \quad (4.24)$$

2. Prove that the likelihood ratio is equal to

$$L(\tilde{\boldsymbol{\theta}}(\boldsymbol{\eta}^\circ), \boldsymbol{\eta}^\circ) - L(\boldsymbol{\theta}, \boldsymbol{\eta}^\circ) = \frac{1}{2\sigma^2} \left\| \Psi^\top (\tilde{\boldsymbol{\theta}}(\boldsymbol{\eta}^\circ) - \boldsymbol{\theta}) \right\|^2 \quad (4.25)$$

for any value of the parameter $\boldsymbol{\theta}$.

Remark 4.1. It immediately follows from (4.24) and (4.25) that if (4.23) is fulfilled then the likelihood ratio is independent of $\boldsymbol{\eta}^\circ$.

1. Note that $\tilde{\theta}(\eta^\circ)$ is the MLE of the residual model $Y - \Phi^\top \eta^\circ = \Psi^\top \theta^* + \varepsilon$:

$$\tilde{\theta}(\eta^\circ) = (\Psi \Psi^\top)^{-1} \Psi (Y - \Phi^\top \eta^\circ).$$

Taking into account the adaptivity condition (4.23), we conclude that the partial estimate $\tilde{\theta}(\eta^\circ)$ is equal to

$$\tilde{\theta}(\eta^\circ) = (\Psi \Psi^\top)^{-1} \Psi Y,$$

and therefore does not depend on the nuisance parameter η°

2. The proof follows the same lines as the proof of Theorem 4.5.1 from Spokoiny and Dickhaus (2014). Consider $L(\tilde{\theta}(\eta^\circ), \eta^\circ) - L(\theta, \eta^\circ)$ as a function of θ ; denote this function by $f(\theta)$. This is a quadratic function satisfying $f(\tilde{\theta}(\eta^\circ)) = 0$. Next, by definition of the MLE, this function attains its minimum exactly at the point $\theta = \tilde{\theta}(\eta^\circ)$ implying $df(\theta)/d\theta|_{\theta=\tilde{\theta}(\eta^\circ)} = 0$. Since

$$L(\theta, \eta^\circ) = -\frac{1}{2\sigma^2} (Y - \Psi^\top \theta - \Phi^\top \eta^\circ)^\top (Y - \Psi^\top \theta - \Phi^\top \eta^\circ) + R,$$

where R does not depend on θ , we conclude that

$$d^2 f(\tilde{\theta}(\eta^\circ), \theta)/d\theta^2 = \sigma^{-2} \Psi \Psi^\top.$$

The Taylor expansion at the point $\theta = \tilde{\theta}(\eta^\circ)$ yields

$$\begin{aligned} f(\theta) &= \frac{1}{2} \{\tilde{\theta}(\eta^\circ) - \theta\}^\top \frac{d^2 f(\tilde{\theta}(\eta^\circ))}{d\theta^2} \{\tilde{\theta}(\eta^\circ) - \theta\} \\ &= \frac{1}{2\sigma^2} (\tilde{\theta}(\eta^\circ) - \theta)^\top \Psi \Psi^\top (\tilde{\theta}(\eta^\circ) - \theta). \end{aligned}$$

This completes the proof.

Exercise 4.21. 1. Let L be a likelihood of the linear model that depends on a parameter $\mathbf{v} \in \mathbb{R}^p$. Let P be a linear operator, $P : \mathbb{R}^p \rightarrow \mathbb{R}^k$. Prove that

$$P \operatorname{argmax}_{\mathbf{v}} L(\mathbf{v}) = \operatorname{argmax}_{\theta \in \mathbb{R}^k} \sup_{\mathbf{v}: P\mathbf{v}=\theta} L(\mathbf{v}).$$

2. Let L be a likelihood of the model that depends on two parameters θ and η . Denote $(\tilde{\theta}, \tilde{\eta}) \stackrel{\text{def}}{=} \operatorname{argmax} L(\theta, \eta)$. Prove that

$$\operatorname{argmax}_{\theta} L(\theta, \tilde{\eta}) = \operatorname{argmax}_{\theta} \sup_{\eta} L(\theta, \eta). \quad (4.26)$$

Remark 4.2. This exercise yields the equivalence of different definitions of the profile estimation, see Sect. 4.8.3 of Spokoiny and Dickhaus (2014).

1. Denote by $\tilde{\mathbf{v}}$ the MLE of the parameter $\mathbf{v}$. For any fixed $\boldsymbol{\theta} \in \mathbb{R}^k$,

$$L(\tilde{\mathbf{v}}) \geq \sup_{\mathbf{v}: P\mathbf{v}=\boldsymbol{\theta}} L(\mathbf{v}),$$

where the equality holds iff the set $\{\mathbf{v} : P\mathbf{v} = \boldsymbol{\theta}\}$ includes $\tilde{\mathbf{v}}$. This is fulfilled only in the case $\boldsymbol{\theta} = P\tilde{\mathbf{v}}$. In other words, the maximum value of $\sup_{\{\mathbf{v}: P\mathbf{v}=\boldsymbol{\theta}\}} L(\mathbf{v})$ is attained at the point $\boldsymbol{\theta} = P\tilde{\mathbf{v}}$.

2. Obviously,

$$L(\tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{\eta}}) \geq \sup_{\boldsymbol{\eta}} L(\boldsymbol{\theta}, \boldsymbol{\eta}),$$

where the equality is possible if $\boldsymbol{\theta} = \tilde{\boldsymbol{\theta}}$. The observation that the expression in the left side of (4.26) equals $\tilde{\boldsymbol{\theta}}$ concludes the proof.

Exercise 4.22. Consider the model (4.22) with $p = k = 2$, even n , and

$$\Psi^{\top} = \begin{pmatrix} 1 & 1 \\ 1 & -1 \\ \vdots & \vdots \\ 1 & 1 \\ 1 & -1 \end{pmatrix}, \quad \Phi^{\top} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ \vdots & \vdots \\ 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

1. Show that the adaptivity condition

$$\Psi\Phi^{\top} = \mathbf{0} \tag{4.27}$$

is not fulfilled in this case.

2. Find the $p \times k$ matrix C such that the linear transformation

$$\boldsymbol{\eta}' = \boldsymbol{\eta} + C^{\top}\boldsymbol{\theta} \tag{4.28}$$

leads to the model

$$\mathbf{Y} = \check{\Psi}^{\top}\boldsymbol{\theta} + \Phi^{\top}\boldsymbol{\eta}' + \boldsymbol{\varepsilon}. \tag{4.29}$$

that satisfies the adaptivity condition.

1. By direct calculation,

$$\Psi\Phi^{\top} = \frac{n}{2} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}. \tag{4.30}$$

Therefore the condition (4.27) is violated.

2. Substituting (4.28) into the original model (4.22), we arrive at

$$\mathbf{Y} = \Psi^\top \boldsymbol{\theta} + \Phi^\top (\boldsymbol{\eta}' - \mathbf{C}^\top \boldsymbol{\theta}) + \boldsymbol{\varepsilon} = (\Psi - \mathbf{C}\Phi)^\top \boldsymbol{\theta} + \Phi^\top \boldsymbol{\eta}' + \boldsymbol{\varepsilon}.$$

Selecting $\mathbf{C}$ to ensure the adaptivity leads to the equation

$$(\Psi - \mathbf{C}\Phi)\Phi^\top = 0$$

or $\mathbf{C} = \Psi\Phi^\top(\Phi\Phi^\top)^{-1}$. In our case,

$$\Phi\Phi^\top = \frac{n}{2}I_2.$$

Together with (4.30), this yields

$$\mathbf{C} = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}.$$

Exercise 4.23. Consider the model (4.22) with $p = 2, k = 1, n = 4$, and

$$\Psi^\top = \begin{pmatrix} 1 & 1 \\ 1 & -1 \\ 1 & 1 \\ 1 & -1 \end{pmatrix}, \quad \Phi^\top = \begin{pmatrix} 1 \\ 0 \\ -1 \\ 0 \end{pmatrix}.$$

The sample $\mathbf{Y} = (Y_1, Y_2, Y_3, Y_4)^\top$ is given. Compute the partial estimates for the parameters $\boldsymbol{\theta}^*$ and $\boldsymbol{\eta}^*$.

Notice that Ψ and Φ satisfy the adaptivity condition (4.27)

$$\Psi\Phi^\top = (0, 0)^\top$$

is fulfilled in this case. This means that the partial estimate $\tilde{\boldsymbol{\theta}}(\boldsymbol{\eta}^\circ)$ doesn't depend on $\boldsymbol{\eta}^\circ$ and is equal to

$$\begin{aligned} \tilde{\boldsymbol{\theta}}(\boldsymbol{\eta}^\circ) &= (\Psi\Psi^\top)^{-1} \Psi\mathbf{Y} = \frac{1}{4} \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \end{pmatrix} \mathbf{Y} \\ &= \frac{1}{4} \begin{pmatrix} Y_1 + Y_2 + Y_3 + Y_4 \\ Y_1 - Y_2 + Y_3 - Y_4 \end{pmatrix}. \end{aligned}$$

Similarly we can invert the role of $\boldsymbol{\theta}^*$ and $\boldsymbol{\eta}^*$. Since the adaptivity condition holds, the partial estimate $\tilde{\boldsymbol{\eta}}(\boldsymbol{\theta}^\circ)$ does not depend on $\boldsymbol{\theta}^\circ$ and is the least square estimator

$$\tilde{\boldsymbol{\eta}}(\boldsymbol{\theta}^\circ) = (\Phi\Phi^\top)^{-1} \Phi\mathbf{Y} = \frac{1}{2} (Y_2 - Y_4).$$

Exercise 4.24. Consider the model $\mathbf{Y} = \mathbf{v}^* + \boldsymbol{\varepsilon}^*$ in $\mathbb{R}^p$, $\mathbf{v}^* = (v_1^*, \dots, v_p^*)^\top$ and let the target estimation be the sum of coefficients $\theta^* = v_1^* + \dots + v_p^*$.

1. Find matrices Υ and P such that the model can be viewed as

$$\mathbf{Y} = \Upsilon^\top \mathbf{v}^* + \boldsymbol{\varepsilon}^*, \quad (4.31)$$

and the problem is to estimate

$$\theta^* = P \mathbf{v}^*. \quad (4.32)$$

2. Reduce to $(\boldsymbol{\theta}, \boldsymbol{\eta})$ – setup (see (4.22)) by an orthogonal change of the basis.

1. The model can be considered in the form (4.31) with the identity $p \times p$ – matrix Υ . The target of estimation can be represented the form (4.32) with a linear operator P from $\mathbb{R}^p$ to $\mathbb{R}$ given by $P \mathbf{v}^* \stackrel{\text{def}}{=} (1 \ 1 \ \dots \ 1) \mathbf{v}^*$.
2. Consider the orthogonal matrix

$$U = \begin{pmatrix} 1/\sqrt{p} & 1/\sqrt{p} & 1/\sqrt{p} & 1/\sqrt{p} & \dots & 1/\sqrt{p} \\ 1/\sqrt{2} * 1 & -1/\sqrt{2} * 1 & 0 & 0 & \dots & 0 \\ 1/\sqrt{3} * 2 & 1/\sqrt{3} * 2 & -2/\sqrt{3} * 2 & 0 & \dots & 0 \\ 1/\sqrt{4} * 3 & 1/\sqrt{4} * 3 & 1/\sqrt{4} * 3 & -3/\sqrt{4} * 3 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{1}{\sqrt{p(p-1)}} & \frac{1}{\sqrt{p(p-1)}} & \frac{1}{\sqrt{p(p-1)}} & \frac{1}{\sqrt{p(p-1)}} & \dots & \frac{-(p-1)}{\sqrt{p(p-1)}} \end{pmatrix}, \quad (4.33)$$

i.e., the first row of the matrix $U = (u_{ij})_{i,j=1}^p$ is equal to $1/\sqrt{p}$, and for $i \leq 2$

$$u_{ij} = \begin{cases} 1/\sqrt{i(i-1)}, & j < i, \\ -(i-1)/\sqrt{i(i-1)}, & j = i, \\ 0, & j > i. \end{cases}$$

Note that U from (4.32) can be decomposed into

$$U = U_1 + U_2,$$

where

$$U_1 = \frac{1}{\sqrt{p}} \begin{pmatrix} 1 & 1 & 1 & \dots & 1 \\ 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 0 & 0 & 0 & \dots & 0 \end{pmatrix},$$

$$U_2 = \begin{pmatrix} 0 & 0 & 0 & \dots & 0 \\ 1/\sqrt{2*1} & -1/\sqrt{2*1} & 0 & \dots & 0 \\ 1/\sqrt{3*2} & 1/\sqrt{3*2} & -2/\sqrt{3*2} & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{1}{\sqrt{p(p-1)}} & \frac{1}{\sqrt{p(p-1)}} & \frac{1}{\sqrt{p(p-1)}} & \dots & \frac{-(p-1)}{\sqrt{p(p-1)}} \end{pmatrix}.$$

So, the vector $\mathbf{Y}$ is transformed to

$$\begin{aligned} U\mathbf{Y} &= U\mathbf{v}^* + U\boldsymbol{\varepsilon}^* \\ &= U_1\mathbf{v}^* + U_2\mathbf{v}^* + U\boldsymbol{\varepsilon}^*. \end{aligned}$$

Note that

$$U_1\mathbf{v}^* = \Psi^\top \theta^*,$$

where $\Psi = (1 \ 0 \ \dots \ 0)/\sqrt{p}$. Since $\varepsilon_1^* \stackrel{\text{def}}{=} U\varepsilon^* \sim \mathcal{N}(\mathbf{0}, \sigma^2 I_p)$, we conclude that the model can be reduced to the $(\boldsymbol{\theta}, \boldsymbol{\eta})$ – setup in the following form:

$$U\mathbf{Y} = \Psi^\top \theta^* + U_2\mathbf{v}^* + \boldsymbol{\varepsilon}_1^*.$$

Exercise 4.25. Consider the general linear model

$$\mathbf{Y} = \Upsilon^\top \mathbf{v}^* + \boldsymbol{\varepsilon}$$

with

$$\Upsilon^\top = \begin{pmatrix} -1/2 & 1/\sqrt{2} & 1/2 \\ 1/2 & 1/\sqrt{2} & -1/2 \end{pmatrix}.$$

The target of estimation is the sum of the components of the vector $\mathbf{v}^*$, i.e., $\theta^* = v_1^* + \dots + v_p^*$.

1. Find the estimate for the parameter θ^* as a profile estimate.
2. Compute $\mathbb{E}\|\tilde{\theta} - \theta^*\|^2$ by the Gauss-Markov theorem.

1. Exercise 4.22 yields that this problem can be considered as the profile estimation problem with $P \stackrel{\text{def}}{=} (1 \ 1 \ 1)$. Taking into account that $\Upsilon\Upsilon^\top = I_2$, we conclude that

$$\begin{aligned} \tilde{\theta} &= P(\Upsilon\Upsilon^\top)^{-1}\Upsilon\mathbf{Y} \\ &= (1 \ 1 \ 1) \begin{pmatrix} -1/2 & 1/2 \\ 1/\sqrt{2} & 1/\sqrt{2} \\ 1/2 & -1/2 \end{pmatrix} \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \frac{1}{\sqrt{2}} (Y_1 + Y_2). \end{aligned}$$

2. According to the Gauss-Markov theorem,

$$\mathbb{E}\|\tilde{\theta} - \theta^*\|^2 = \sigma^2 \operatorname{tr} \left\{ P (\Upsilon \Upsilon^\top)^{-1} P^\top \right\}.$$

So $\mathbb{E}\|\tilde{\theta} - \theta^*\|^2 = 3\sigma^2$.

If we are looking at

$$\theta^* = \sum_{j=1}^p \frac{1}{j} v_j^*,$$

can one also calculate $\mathbb{E}\|\tilde{\theta} - \theta^*\|^2$? We leave this as a training exercise.

Exercise 4.26. 1. Consider the model (4.22) with the matrices Ψ and Φ from the Exercise 4.23. Construct the profile estimate of θ^* .

2. Consider the model (4.22) with matrixes Ψ and Φ from the Exercise 4.22 and $n = 4$. Is it possible to construct the estimate of θ^* as the profile MLE by the linear transformation (4.28) such that the model (4.29) satisfies the adaptivity condition (4.27)?

1. The adaptivity condition holds for this model (see Exercise 4.23); therefore, the estimate of θ coincides with the MLE estimate for the model

$$Y = \Upsilon^\top v^* + \varepsilon^*,$$

where $\Upsilon^\top \stackrel{\text{def}}{=} (\Psi^\top, \Phi^\top)$ is a 4×3 -matrix, and the object of estimation is the vector obtained by the projection P to the first two coordinates. This gives the profile MLE

$$\tilde{\theta} = P (\Upsilon \Upsilon^\top)^{-1} \Upsilon Y.$$

2. Exercise 4.22 shows that the linear transformation (4.28) with the matrix

$$C = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$$

leads to the model (4.29) that satisfies the adaptivity condition. But the matrix $\tilde{\Psi} \stackrel{\text{def}}{=} \Psi - C\Phi$ is a zero matrix in this case. Therefore, the model (4.29) doesn't include θ^* , and therefore, this parameter cannot be estimated using such approach.

Exercise 4.27. Consider the model (4.22). Fix the nuisance parameter as η° , denote the estimate $\tilde{\theta}(\eta^\circ)$ obtained by partial optimization of the joint log-likelihood $L(\theta, \eta^\circ)$ w.r.t. the first parameter θ :

$$\tilde{\theta}(\eta^\circ) = \operatorname{argmax}_{\theta} L(\theta, \eta^\circ).$$

Let $\hat{\boldsymbol{\eta}}$ be a pilot estimate of the parameter $\boldsymbol{\eta}$. Denote the estimate obtained by the plug-in method by $\hat{\boldsymbol{\theta}}$,

$$\hat{\boldsymbol{\theta}} \stackrel{\text{def}}{=} \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \|\hat{\mathbf{Y}} - \Psi^\top \boldsymbol{\theta}\|^2 = (\Psi \Psi^\top)^{-1} \Psi \hat{\mathbf{Y}},$$

where $\hat{\mathbf{Y}} \stackrel{\text{def}}{=} \mathbf{Y} - \Phi^\top \hat{\boldsymbol{\eta}}$.

1. Prove that $\hat{\boldsymbol{\theta}} = \tilde{\boldsymbol{\theta}}(\hat{\boldsymbol{\eta}})$.
2. Write down the formulae for the pilot estimates for $\hat{\boldsymbol{\theta}}$

(a) $\hat{\boldsymbol{\eta}}_1 = \mathbf{0}$

(b) $\hat{\boldsymbol{\eta}}_2 = \mathbf{Y}_1 (k = 1)$

(c) General linear estimate $\hat{\boldsymbol{\eta}}_3 \stackrel{\text{def}}{=} A\mathbf{Y}$, where A is a $k \times n$ matrix.

1. Note that the estimate $\hat{\boldsymbol{\theta}}$ is the MLE in the model

$$\hat{\mathbf{Y}} \stackrel{\text{def}}{=} \Psi^\top \boldsymbol{\theta}^* + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 I_n).$$

In other words,

$$\hat{\boldsymbol{\theta}} \stackrel{\text{def}}{=} \left\{ \underset{\boldsymbol{\theta}}{\operatorname{argmax}} L(\boldsymbol{\theta}, \boldsymbol{\eta}) \right\} \Big|_{\boldsymbol{\eta}=\hat{\boldsymbol{\eta}}},$$

so (i) follows.

2. The definition of the estimate $\hat{\boldsymbol{\theta}}$ can be rewritten as:

$$\hat{\boldsymbol{\theta}} \stackrel{\text{def}}{=} (\Psi \Psi^\top)^{-1} \Psi (\mathbf{Y} - \Phi^\top \hat{\boldsymbol{\eta}}).$$

Plug in $\hat{\boldsymbol{\eta}} = \hat{\boldsymbol{\eta}}_3 = A\mathbf{Y}$, we have

$$\hat{\boldsymbol{\theta}} \stackrel{\text{def}}{=} (\Psi \Psi^\top)^{-1} \Psi (I_n - \Phi^\top A) \mathbf{Y}.$$

The other two cases $\hat{\boldsymbol{\eta}} = \hat{\boldsymbol{\eta}}_1$ and $\hat{\boldsymbol{\eta}} = \hat{\boldsymbol{\eta}}_2$ are the special cases with $A = \mathbf{0}$ and $A = (1, 0, \dots, 0)$ respectively.

Exercise 4.28. Consider the model (4.22). With the initial guess $\boldsymbol{\theta}^\circ$ for the target $\boldsymbol{\theta}^*$, consider the following two-step procedure:

- (i) Compute the partial MLE for the model

$$\mathbf{Y}(\boldsymbol{\theta}^\circ) = \Phi^\top \boldsymbol{\eta}^* + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 I_n) \tag{4.34}$$

with $\mathbf{Y}(\boldsymbol{\theta}^\circ) = \mathbf{Y} - \Psi^\top \boldsymbol{\theta}^\circ$. This leads to the estimate

$$\tilde{\boldsymbol{\eta}}(\boldsymbol{\theta}^\circ) = (\Phi \Phi^\top)^{-1} \Phi \mathbf{Y}(\boldsymbol{\theta}^\circ). \tag{4.35}$$

(ii) Estimate the target parameter θ^* by fitting $\Psi^\top \theta^*$ to the residuals

$$\hat{Y}(\theta^\circ) = Y - \Phi^\top \tilde{\eta}(\theta^\circ). \quad (4.36)$$

This method results in the estimate

$$\hat{\theta}(\theta^\circ) = (\Psi \Psi^\top)^{-1} \Psi \hat{Y}(\theta^\circ). \quad (4.37)$$

1. Compute the mean and the variance of $\hat{\theta}(\theta^\circ)$.
2. Consider the adaptive case with $\Psi \Phi^\top = 0$. Show that the two step estimate $\hat{\theta}(\theta^\circ)$ coincides with the partial MLE $\tilde{\theta} = (\Psi \Psi^\top)^{-1} \Psi Y$.
3. Let Ψ be orthogonal, i.e. $\Psi \Psi^\top = I_p$. Show that

$$\text{Var}\{\hat{\theta}(\theta^\circ)\} = \sigma^2(I_p - \Psi \Pi_\eta \Psi^\top),$$

where $\Pi_\eta = \Phi^\top (\Phi \Phi^\top)^{-1} \Phi$.

1. Combining (4.34)–(4.37) we have

$$\begin{aligned} \hat{\theta}(\theta^\circ) &= (\Psi \Psi^\top)^{-1} \Psi \hat{Y}(\theta^\circ) \\ &= (\Psi \Psi^\top)^{-1} \Psi \{Y - \Phi^\top \tilde{\eta}(\theta^\circ)\} \\ &= (\Psi \Psi^\top)^{-1} \Psi \{Y - \Pi_\eta Y(\theta^\circ)\} \\ &= (\Psi \Psi^\top)^{-1} \Psi \{Y - \Pi_\eta (Y - \Psi^\top \theta^\circ)\} \\ &= (\Psi \Psi^\top)^{-1} \Psi (I_n - \Pi_\eta) Y + (\Psi \Psi^\top)^{-1} \Psi \Pi_\eta \Psi^\top \theta^\circ. \end{aligned}$$

It follows that

$$\mathbb{E}\{\hat{\theta}(\theta^\circ)\} = (\Psi \Psi^\top)^{-1} \Psi (I_n - \Pi_\eta) \mathbb{E}Y.$$

Taking into account that $\text{Var} Y = \sigma^2 I_n$ and that Π_η is a projector, we conclude also that

$$\text{Var}\{\hat{\theta}(\theta^\circ)\} = \sigma^2 (\Psi \Psi^\top)^{-1} \Psi (I_n - \Pi_\eta) \Psi^\top (\Psi \Psi^\top)^{-1}.$$

2. Substituting (4.36) to (4.37) gives

$$\begin{aligned} \hat{\theta}(\theta^\circ) &= (\Psi \Psi^\top)^{-1} \Psi \{Y - \Phi^\top \tilde{\eta}(\theta^\circ)\} \\ &= (\Psi \Psi^\top)^{-1} \Psi Y. \end{aligned}$$

3. From (4.38) and $\Psi\Psi^T = I_p$,

$$\begin{aligned} & \sigma^2(\Psi\Psi^T)^{-1}\Psi(I_n - \Pi_\eta)\Psi^T(\Psi\Psi^T)^{-1} \\ &= \sigma^2(\Psi\Psi^T)^{-1} - \sigma^2(\Psi\Psi^T)^{-1}\Psi\Pi_\eta\Psi^T(\Psi\Psi^T)^{-1} \\ &= \sigma^s - \sigma^2\Psi\Pi_\eta\Psi^T. \end{aligned}$$

Therefore, (4.38) is right.

Exercise 4.29. Consider the iterative procedure based on the two-step procedure from the Exercise 4.27. One starts with the initial guess θ° for the target θ^* . Set

$$\hat{\theta}_1 \equiv \theta^\circ, \quad \hat{\eta}_1 \stackrel{\text{def}}{=} \tilde{\eta}(\theta^\circ) = \left\{ \operatorname{argmax}_\eta L(\theta, \eta) \right\} \Big|_{\theta=\theta^\circ}.$$

Then recompute the estimates in the iterative way ($k = 1, 2, \dots$):

$$\begin{aligned} \hat{\theta}_{k+1} &\stackrel{\text{def}}{=} \tilde{\theta}(\hat{\eta}_k) = \left\{ \operatorname{argmax}_\theta L(\theta, \eta) \right\} \Big|_{\eta=\hat{\eta}_k}, \\ \hat{\eta}_{k+1} &\stackrel{\text{def}}{=} \tilde{\eta}(\hat{\theta}_{k+1}) = \left\{ \operatorname{argmax}_\eta L(\theta, \eta) \right\} \Big|_{\theta=\hat{\theta}_{k+1}}. \end{aligned}$$

1. Consider the adaptive situation with $\Psi^\top\Phi = 0$. Prove that the above procedure stabilizes in one step.
2. Denote the operators $\Pi_\theta \stackrel{\text{def}}{=} \Psi^\top(\Psi\Psi^\top)^{-1}\Psi$ and $\Pi_\eta \stackrel{\text{def}}{=} \Phi^\top(\Phi\Phi^\top)^{-1}\Phi$. Prove the following recurrent formula for $\Psi^\top\hat{\theta}_k$ and $\Phi^\top\hat{\eta}_k$ ($k \geq 1$):

$$\Psi^\top\hat{\theta}_{k+1} = (\Pi_\theta - \Pi_\theta\Pi_\eta)Y + \Pi_\theta\Pi_\eta\Psi^\top\hat{\theta}_k, \quad (4.38)$$

$$\Phi^\top\hat{\eta}_{k+1} = (\Pi_\eta - \Pi_\eta\Pi_\theta)Y + \Pi_\eta\Pi_\theta\Phi^\top\hat{\eta}_k. \quad (4.39)$$

1. Note that the estimates $\hat{\theta}_k, \hat{\eta}_k$ ($k = 1, 2, \dots$) are equal to

$$\hat{\eta}_1 = (\Phi\Phi^\top)^{-1}\Phi(Y - \Psi^\top\theta^\circ), \quad (4.40)$$

$$\hat{\theta}_{k+1} = (\Psi\Psi^\top)^{-1}\Psi(Y - \Phi^\top\hat{\eta}_k), \quad k = 1, 2, \dots \quad (4.41)$$

$$\hat{\eta}_{k+1} = (\Phi\Phi^\top)^{-1}\Phi(Y - \Psi^\top\hat{\theta}_{k+1}), \quad k = 1, 2, \dots \quad (4.42)$$

This yields that in the adaptive situation with $\Psi^\top\Phi = 0$, we have

$$\hat{\eta}_1 = (\Phi\Phi^\top)^{-1}\Phi Y, \quad \hat{\theta}_2 = (\Psi\Psi^\top)^{-1}\Psi Y,$$

and further iterations don't change the values of the estimates.

2. Let us prove the formulae (4.38) and (4.39) by induction on k . From (4.40), it follows that

$$\begin{aligned}\Phi^\top \hat{\eta}_1 &= \Phi^\top (\Phi \Phi^\top)^{-1} \Phi (Y - \Psi^\top \theta^\circ) \\ &= \Pi_\eta (Y - \Psi^\top \hat{\theta}_1).\end{aligned}$$

From (4.41) we conclude that

$$\begin{aligned}\Psi^\top \hat{\theta}_2 &= \Psi^\top (\Psi \Psi^\top)^{-1} \Psi (Y - \Phi^\top \hat{\eta}_1) \\ &= \Pi_\theta (Y - \Phi^\top \hat{\eta}_1) \\ &= \Pi_\theta \{Y - \Pi_\eta (Y - \Psi^\top \hat{\theta}_1)\} \\ &= (\Pi_\theta - \Pi_\theta \Pi_\eta) Y + \Pi_\theta \Pi_\eta \Psi^\top \hat{\theta}_1.\end{aligned}$$

Analogously, formula (4.42) yields

$$\begin{aligned}\Phi^\top \hat{\eta}_2 &= \Phi^\top (\Phi \Phi^\top)^{-1} \Phi (Y - \Psi^\top \hat{\theta}_2) \\ &= \Pi_\eta (Y - \Psi^\top \hat{\theta}_2) \\ &= \Pi_\eta \{Y - \Pi_\theta (Y - \Phi^\top \hat{\eta}_1)\} \\ &= (\Pi_\eta - \Pi_\eta \Pi_\theta) Y + \Pi_\eta \Pi_\theta \Phi^\top \hat{\eta}_1.\end{aligned}$$

Therefore, the formulas (4.38) and (4.39) are proven for $k = 1$. To make the induction step, we note that (4.41) and (4.42) imply that for any $k > 1$

$$\begin{aligned}\Psi^\top \hat{\theta}_{k+1} &= \Pi_\theta (Y - \Phi^\top \hat{\eta}_k), \\ \Phi^\top \hat{\eta}_{k+1} &= \Pi_\eta (Y - \Psi^\top \hat{\theta}_{k+1}).\end{aligned}$$

The further proof follow the same lines.

Exercise 4.30. Show that for any self-adjoint matrices A and B ,

$$\|AB\|_\infty = \|BA\|_\infty.$$

Recall that by $\|A\|_\infty$ we denote the spectral norm of the matrix A , i.e., the largest singular value of the matrix A . Denote by λ any singular value of the matrix AB . By the definition of the singular value, there exist unit length vectors $\mathbf{e}_1$ and $\mathbf{e}_2$ such that

$$AB\mathbf{e}_1 = \lambda\mathbf{e}_2, \quad (AB)^* \mathbf{e}_2 = \lambda\mathbf{e}_1. \quad (4.43)$$

It is worth mentioning that $(AB)^* = B^*A^* = BA$ where A^* is the adjoint of A . Therefore, we conclude that λ is also a singular value of the matrix BA . So, the matrices AB and BA have the same spectral values and hence the spectral norms of these matrices coincide.

Exercise 4.31. Consider the set-up from the Exercise 4.22. Suppose that

$$\lambda \stackrel{\text{def}}{=} \|\Pi_\eta \Pi_\theta\|_\infty < 1.$$

1. Show by induction arguments that for $k \geq 1$

$$\Phi^\top \hat{\eta}_{k+1} = A_{k+1}Y + (\Pi_\eta \Pi_\theta)^k \Phi^\top \hat{\eta}_1, \quad (4.44)$$

where the linear operator A_k fulfills $A_1 = 0$ and

$$A_{k+1} = \Pi_\eta - \Pi_\eta \Pi_\theta + \Pi_\eta \Pi_\theta A_k = \sum_{i=0}^{k-1} (\Pi_\eta \Pi_\theta)^i (\Pi_\eta - \Pi_\eta \Pi_\theta).$$

2. Show that A_k converges to $A \stackrel{\text{def}}{=} (I_n - \Pi_\eta \Pi_\theta)^{-1}(\Pi_\eta - \Pi_\eta \Pi_\theta)$.

3. Prove that

$$\Phi^\top \hat{\eta} = (I_n - \Pi_\eta \Pi_\theta)^{-1}(\Pi_\eta - \Pi_\eta \Pi_\theta)Y,$$

where the value $\hat{\eta}$ is the limiting value for the sequence $\hat{\eta}_k$.

Remark 4.3. Analogously, one can prove the same formulas for the estimate $\hat{\theta}_k$ and for the limiting value $\hat{\theta}$ by changing the role of θ and η .

1. The first item trivially follows from (4.39). In fact, for $k = 1$ the formula (4.44) coincides with (4.39):

$$\begin{aligned} \Phi^\top \hat{\eta}_2 &= (\Pi_\eta - \Pi_\eta \Pi_\theta)Y + \Pi_\eta \Pi_\theta \Phi^\top \hat{\eta}_1 \\ &= A_2Y + \Pi_\eta \Pi_\theta \Phi^\top \hat{\eta}_1. \end{aligned}$$

The induction step is also straightforward:

$$\begin{aligned} \Phi^\top \hat{\eta}_{k+1} &= (\Pi_\eta - \Pi_\eta \Pi_\theta)Y + \Pi_\eta \Pi_\theta \Phi^\top \hat{\eta}_k \\ &= (\Pi_\eta - \Pi_\eta \Pi_\theta)Y + \Pi_\eta \Pi_\theta \left\{ A_kY + (\Pi_\eta \Pi_\theta)^{k-1} \Phi^\top \hat{\eta}_1 \right\} \\ &= \underbrace{(\Pi_\eta - \Pi_\eta \Pi_\theta + \Pi_\eta \Pi_\theta A_k)}_{=A_{k+1}}Y + (\Pi_\eta \Pi_\theta)^k \Phi^\top \hat{\eta}_1. \end{aligned}$$

2. The aim is to show that

$$\begin{aligned} & \|A_{k+1} - A\|_\infty \\ &= \left\| \left\{ \sum_{i=0}^{k-1} (\Pi_\eta \Pi_\theta)^i - (I_n - \Pi_\eta \Pi_\theta)^{-1} \right\} (\Pi_\eta - \Pi_\eta \Pi_\theta) \right\|_\infty \longrightarrow 0, \end{aligned}$$

as $k \rightarrow \infty$. This fact follows from the observations that $\|\Pi_\eta - \Pi_\eta \Pi_\theta\|_\infty \leq 1$ and $\|(\Pi_\eta \Pi_\theta)^i\|_\infty \leq \|\Pi_\eta \Pi_\theta\|_\infty^i \leq \lambda^i$. (to be continued)

3. Since $\|A_k - A\|_\infty \rightarrow 0$, the sequence $\Phi^\top \hat{\eta}_k$ converge to $\Phi^\top \hat{\eta} \stackrel{\text{def}}{=} AY$, because

$$\begin{aligned} \|\Phi^\top \hat{\eta}_k - AY\|_\infty &= \|(A_k - A)Y + (\Pi_\eta \Pi_\theta)^{k-1} \Phi^\top \hat{\eta}_1\|_\infty \\ &\leq \|A_k - A\|_\infty \|Y\|_\infty + \lambda^{k-1} \|\Phi^\top \hat{\eta}_1\|_\infty \longrightarrow 0, \quad k \rightarrow \infty. \end{aligned}$$

Inserting $\Phi^\top \hat{\eta}$ in place of $\Phi^\top \hat{\eta}_k$ and $\Phi^\top \hat{\eta}_{k+1}$ in (4.39) completes the proof.

Exercise 4.32. (This exercise is based on the ideas from Csiszár and Tusnády 1984) Let $\mathcal{P}$ and $\mathcal{Q}$ be two arbitrary sets and let D be a function depending on two parameters, $P \in \mathcal{P}$ and $Q \in \mathcal{Q}$. Denote by (P^*, Q^*) the point of global maximum, i.e.,

$$\max_{P, Q} D(P, Q) = D(P^*, Q^*).$$

Consider the following procedure for estimating the pair (P^*, Q^*) : starting with an initial value $P^{(0)}$, one iteratively computes the estimates $(k = 0, 1, \dots)$

$$\begin{aligned} Q^{(k+1)} &= \operatorname{argmax}_{Q \in \mathcal{Q}} D(P^{(k)}, Q), \\ P^{(k+1)} &= \operatorname{argmax}_{P \in \mathcal{P}} D(P, Q^{(k+1)}). \end{aligned}$$

Let the following inequality (so-called 5-point property) be fulfilled for any $k \geq 0$:

$$D(P^*, Q^*) - D(P^{(k+1)}, Q^{(k+1)}) \leq D(P^*, Q^{(k+1)}) - D(P^*, Q^{(k)}). \quad (4.45)$$

Prove that

$$\lim_{k \rightarrow +\infty} D(P^{(k)}, Q^{(k)}) = D(P^*, Q^*).$$

Hint. Prove the following fact:

Let two upper bounded real sequences $\{a_k\}_{k=1}^\infty, \{b_k\}_{k=1}^\infty$ satisfy the inequality

$$a_{k+1} + (b_{k+1} - b_k) \geq c \geq a_k \quad (4.46)$$

for some $c \in \mathbb{R}$ and any $k \in \mathbb{N}$. Then a_k converges to c as $k \rightarrow \infty$.

First note that the statement of the exercise follows from (4.46). In fact, set

$$\begin{aligned} a_k &= D(P^{(k)}, Q^{(k)}), \\ b_k &= D(P^*, Q^{(k)}), \\ c &= D(P^*, Q^*). \end{aligned}$$

Both sequences $\{a_k\}, \{b_k\}$ are bounded by c , and moreover the 5-point property (4.45) yields

$$a_{k+1} + (b_{k+1} - b_k) \geq c.$$

So, our aim is to prove (4.46). For any natural N ,

$$0 \leq \sum_{k=1}^N (c - a_{k+1}) \leq \sum_{k=1}^N (b_{k+1} - b_k) = b_{N+1} - b_1.$$

This means that the series $\sum_{k=1}^N (c - a_{k+1})$ converges and therefore $a_k \rightarrow c$ as $k \rightarrow +\infty$.

References

- Csiszár, I., & Tusnády, G. (1984). Information geometry and alternating minimization procedures. *Statistics & Decisions, Supplement Issue, 1*, 205–237.
- Härdle, W., & Simar, L. (2011). *Applied multivariate statistical analysis* (3rd ed.). Berlin: Springer.
- Spokoiny, V., & Dickhaus, T. (2014). *Basics of modern parametric statistics*. Berlin: Springer.

Chapter 5

Bayes Estimation

Estimation par la méthode de Bayes

Qui ne risque rien n'a rien.

Nothing ventured, nothing gained.

Exercise 5.1. Consider the Bernoulli experiment $\mathbf{Y} = (Y_1, \dots, Y_n)^\top$ with $n = 10$ and let

$$\pi(0.5) = \pi(0.9) = 1/2.$$

1. Compute the posterior distribution of θ if

(a) We observe $\mathbf{y} = (1, \dots, 1)^\top$. Which value of θ has the highest probability?

(b) We observe a sample $\mathbf{y} = (y_1, \dots, y_n)^\top$ with the number of successes $y_1 + \dots + y_n = 5$. Which value of θ has the highest probability?

2. Show that the posterior density $p(\theta|\mathbf{y})$ depends only on the number of successes S .

1. (a) Denote the probability of observing $\mathbf{y}$ by $p(\mathbf{y})$. Then

$$\begin{aligned} p(\mathbf{y}) &= \pi(0.5) p(\mathbf{y}|\theta = 0.5) + \pi(0.9) p(\mathbf{y}|\theta = 0.9) \\ &= \frac{1}{2} \left\{ (0.5)^{10} + (0.9)^{10} \right\}. \end{aligned}$$

By the Bayes formula,

$$\begin{aligned}
 p(\theta = 0.5|\mathbf{y}) &= \frac{p(\mathbf{y}|\theta = 0.5) p(\theta = 0.5)}{p(\mathbf{y})} \\
 &= \frac{\frac{1}{2} (0.5)^{10}}{\frac{1}{2} \{(0.5)^{10} + (0.9)^{10}\}} = \frac{1}{1 + (1.8)^{10}}, \\
 p(\theta = 0.9|\mathbf{y}) &= \frac{p(\mathbf{y}|\theta = 0.9) p(\theta = 0.9)}{p(\mathbf{y})} \\
 &= \frac{\frac{1}{2} (0.9)^{10}}{\frac{1}{2} \{(0.9)^{10} + (0.5)^{10}\}} = \frac{1}{1 + (\frac{5}{9})^{10}},
 \end{aligned}$$

and we conclude that $p(\theta = 0.9|\mathbf{y})$ is larger than $p(\theta = 0.5|\mathbf{y})$.

(b) Let now the number of successes $y_1 + \dots + y_n$ be equal to 5. In this case,

$$\begin{aligned}
 p(\mathbf{y}|\theta = 0.5) &= \binom{10}{5} (0.5)^5 (0.5)^{10-5}, \\
 p(\mathbf{y}|\theta = 0.9) &= \binom{10}{5} (0.9)^5 (0.1)^{10-5}.
 \end{aligned}$$

The posterior probabilities can be computed by Bayes formula:

$$\begin{aligned}
 p(\theta = 0.5|\mathbf{y}) &= \frac{p(\mathbf{y}|\theta = 0.5) p(\theta = 0.5)}{p(\mathbf{y})} \\
 &= \frac{\frac{1}{2} p(\mathbf{y}|\theta = 0.5)}{\frac{1}{2} \{p(\mathbf{y}|\theta = 0.5) + p(\mathbf{y}|\theta = 0.9)\}} \\
 &= \frac{(0.5)^{10}}{(0.5)^{10} + (0.9)^5 (0.1)^{10-5}} \\
 &= \frac{1}{1 + (1.8)^5 (0.2)^{10-5}},
 \end{aligned}$$

and

$$\begin{aligned}
 p(\theta = 0.9|\mathbf{y}) &= \frac{p(\mathbf{y}|\theta = 0.9) p(\theta = 0.9)}{p(\mathbf{y})} \\
 &= \frac{(0.9)^5 (0.1)^{10-5}}{(0.5)^{10} + (0.9)^5 (0.1)^{10-5}} \\
 &= \frac{1}{1 + (\frac{5}{9})^5 (5)^{10-5}}.
 \end{aligned}$$

Comparing $p(\theta = 0.5|\mathbf{y})$ with $p(\theta = 0.9|\mathbf{y})$ leads to

$$p(\theta = 0.5|\mathbf{y})^{-1} < p(\theta = 0.9|\mathbf{y})^{-1}$$

$$(1.8)^5 (0.2)^5 < \left(\frac{5}{9}\right)^5 (5)^5$$

$$\frac{18 \times 2}{100} < \frac{25}{9}$$

and the clear conclusion that for $\theta = 0.5$ the posterior density is maximized.

2. Let the number of successes be equal to S . Then

$$p(\mathbf{y}|\theta = 0.5) = \binom{n}{S} (0.5)^S (0.5)^{n-S} = \binom{n}{S} (0.5)^n, \quad (5.1)$$

$$p(\mathbf{y}|\theta = 0.9) = \binom{n}{S} (0.9)^S (0.1)^{n-S}. \quad (5.2)$$

The Bayes formula yields

$$\begin{aligned} p(\theta = 0.5|\mathbf{y}) &= \frac{\frac{1}{2} p(\mathbf{y}|\theta = 0.5)}{\frac{1}{2} p(\mathbf{y}|\theta = 0.5) + \frac{1}{2} p(\mathbf{y}|\theta = 0.9)} \\ &= \frac{p(\mathbf{y}|\theta = 0.5)}{p(\mathbf{y}|\theta = 0.5) + p(\mathbf{y}|\theta = 0.9)}. \end{aligned}$$

Thus $p(\theta = 0.5|\mathbf{y})$ depends on $p(\mathbf{y}|\theta = 0.5)$ and $p(\mathbf{y}|\theta = 0.9)$, both of which depend only on the numbers of successes S , and don't depend on the exact realisations $y_1, \dots, y_n$, see (5.1)–(5.2).

Exercise 5.2. Let the conditional distribution of Y given θ be $\mathcal{N}(\theta, \sigma^2)$, and the prior distribution of the parameter θ be $\mathcal{N}(\nu, \eta^2)$. Using the Bayes formula, prove that

$$\theta | Y \sim \mathcal{N}\left(\frac{\nu\sigma^2 + Y\eta^2}{\sigma^2 + \eta^2}, \frac{\sigma^2\eta^2}{\sigma^2 + \eta^2}\right).$$

Denote the marginal distribution of Y by $p(Y)$, the prior density of θ by $\pi(\theta)$, and the density of the conditional distribution of Y given θ by $p(Y|\theta)$. We know that

$$\pi(\theta) = \eta^{-1} \varphi\{(\theta - \nu)/\eta\}, \quad (5.3)$$

$$p(Y|\theta) = \sigma^{-1} \varphi\{(Y - \theta)/\sigma\}. \quad (5.4)$$

Note that $Y = \theta + \varepsilon$, where $\theta \sim \mathcal{N}(\nu, \eta^2)$ and $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ independent of θ . Therefore Y is normal with mean ν and variance $\eta^2 + \sigma^2$, i.e.

$$p(Y) = (\eta^2 + \sigma^2)^{-1/2} \varphi\left\{(Y - \nu)/(\eta^2 + \sigma^2)^{1/2}\right\}. \quad (5.5)$$

In our notation, the Bayes formula is

$$\theta | Y \sim p(\theta|Y) = \frac{p(Y, \theta)}{p(Y)} = \frac{p(Y|\theta)\pi(\theta)}{p(Y)}. \quad (5.6)$$

Substituting (5.3), (5.4), and (5.5) into (5.6), we arrive at

$$\begin{aligned} p(\theta|Y) &= \frac{p(Y|\theta)\pi(\theta)}{p(Y)} \\ &= 2\pi\left(\frac{\sigma^2\eta^2}{\sigma^2 + \eta^2}\right)^{-1/2} \exp\left[-\frac{1}{2}\left\{\frac{(Y - \theta)^2}{\sigma^2} + \frac{(\theta - \nu)^2}{\eta^2} - \frac{(Y - \nu)^2}{\sigma^2 + \eta^2}\right\}\right]. \end{aligned}$$

For completing the proof, it is sufficient to note that

$$\frac{(Y - \theta)^2}{\sigma^2} + \frac{(\theta - \nu)^2}{\eta^2} - \frac{(Y - \nu)^2}{\sigma^2 + \eta^2} = A\theta^2 - 2B\theta + C,$$

where the values A , B and C are equal to

$$\begin{aligned} A &= \frac{1}{\sigma^2} + \frac{1}{\eta^2} = \frac{\sigma^2 + \eta^2}{\eta^2\sigma^2}, \\ B &= \frac{Y}{\sigma^2} + \frac{\nu}{\eta^2} = \frac{\sigma^2\nu + \eta^2Y}{\eta^2\sigma^2}, \\ C &= \frac{Y^2}{\sigma^2} + \frac{\nu^2}{\eta^2} - \frac{Y^2 - 2Y\nu + \nu^2}{\sigma^2 + \eta^2}, \end{aligned}$$

and hence $p(\theta|Y)$ is a density of the normal distribution with mean $(\nu\sigma^2 + Y\eta^2)/(\sigma^2 + \eta^2)$ and variance $\sigma^2\eta^2/(\sigma^2 + \eta^2)$.

Exercise 5.3. Let $\mathbf{Y} = (Y_1, \dots, Y_n)^\top$ be i.i.d. and for each Y_i

$$\begin{aligned} Y_i | \theta &\sim \mathcal{N}(\theta, \sigma^2), \\ \theta &\sim \mathcal{N}(\nu, \eta^2). \end{aligned}$$

Prove that for $S = Y_1 + \dots + Y_n$

$$\theta | \mathbf{Y} \sim \mathcal{N}\left(\frac{\nu\sigma^2 + S\eta^2}{\sigma^2 + n\eta^2}, \frac{\sigma^2\eta^2}{\sigma^2 + n\eta^2}\right).$$

Hint: Consider $Y_i = \theta + \varepsilon_i$, $\bar{Y} = S/n$, and define $\zeta = \theta - v - (1 - \rho)(\bar{Y} - v)$, where $\rho = \sigma^2/(n\eta^2 + \sigma^2)$. Check that ζ and each Y_i are uncorrelated and hence independent.

Note that $\bar{Y} = \theta + \bar{\varepsilon}$, with $\theta \sim \mathcal{N}(v, \eta^2)$ and $\bar{\varepsilon} = n^{-1} \sum_{i=1}^n \varepsilon_i \sim \mathcal{N}(0, \sigma^2/n)$ independent of θ . Therefore $\bar{Y}$ is normal with mean $\mathbb{E}\bar{Y} = \mathbb{E}\theta + \mathbb{E}\bar{\varepsilon} = v$ and the variance

$$\text{Var}(\bar{Y}) = \text{Var} \theta + \text{Var} \bar{\varepsilon} = \eta^2 + \sigma^2/n.$$

Next observe that

$$\begin{aligned} \mathbb{E}\{(\theta - v)(\bar{Y} - v)\} &= \mathbb{E}\{(\theta - v)(\theta + \bar{\varepsilon} - v)\} = \mathbb{E}(\theta - v)^2 = \eta^2 \\ &= (1 - \rho) \text{Var}(\bar{Y}) \end{aligned}$$

with $\rho = \sigma^2/(n\eta^2 + \sigma^2)$. Thus the rv's $\bar{Y} - v$ and

$$\begin{aligned} \zeta &= \theta - v - (1 - \rho)(\bar{Y} - v) \\ &= \rho(\theta - v) - (1 - \rho)\bar{\varepsilon} \end{aligned}$$

are Gaussian and uncorrelated and therefore independent. The conditional distribution $\mathcal{L}(\zeta|\bar{Y})$ of ζ given $\bar{Y}$ (or $S = \bar{Y}n$) coincides with the unconditional distribution and hence, it is normal with mean zero. The variance of ζ is equal to

$$\begin{aligned} \text{Var}(\zeta) &= \rho^2 \text{Var}(\theta) + (1 - \rho)^2 \text{Var}(\bar{\varepsilon}) \\ &= \rho^2 \eta^2 + (1 - \rho)^2 \sigma^2/n \\ &= \frac{\sigma^4}{(\sigma^2 + n\eta^2)^2} \eta^2 + (\sigma^2/n) - 2 \left(\frac{\sigma^2}{\sigma^2 + n\eta^2} \right) (\sigma^2/n) \\ &\quad + \frac{\sigma^4}{(\sigma^2 + n\eta^2)^2} (\sigma^2/n) \\ &= \frac{(\sigma^4/n) - 2(\sigma^4/n) + (\sigma^2/n)(\sigma^2 + n\eta^2)}{\sigma^2 + n\eta^2} \\ &= \frac{\sigma^2 \eta^2}{\sigma^2 + n\eta^2}. \end{aligned}$$

This yields the result because with (5.7)

$$\theta = \zeta + \rho v + (1 - \rho)\bar{Y}.$$

Exercise 5.4. *Non-informative priors give equal probability weight to all possible parameter values. For $\Theta = \{\theta_1, \dots, \theta_M\}$, the non-informative prior is $\pi(\theta_j) = 1/M$, $j = 1, \dots, M$. Check that the posterior measure:*

$$p(\theta_k | y) = \frac{p(y | \theta_k) \pi(\theta_k)}{p(y)}$$

is non informative if and only if all the measures $\mathbb{P}_{\theta_m}$ coincide.

1. Prove if all the measures $\mathbb{P}_{\theta_m}$ coincide, $p(\theta_k | y)$ is non informative. The marginal density of y is:

$$p(y) = M^{-1} \sum_{m=1}^M p(y | \theta_m)$$

The posterior measure is therefore

$$\begin{aligned} p(\theta_k | y) &= \frac{p(y | \theta_k) \pi(\theta_k)}{p(y)} = \frac{p(y | \theta_k)}{\sum_{m=1}^M p(y | \theta_m)} = \frac{p(y | \theta_k)}{M p(y | \theta_k)} \\ &= M^{-1} \end{aligned}$$

Thus $p(\theta_k | y)$ is a non-informative measure.

2. If $p(\theta_k | y)$ are the same for any k , $p(y | \theta_k) = \frac{p(\theta_k | y) p(y)}{\pi(\theta_k)}$ also coincide for any k . Therefore, all the measures $\mathbb{P}_{\theta_m}$ coincide.

Exercise 5.5. *A classical example for Bayes risk is testing for a disease D or the presence of certain genetic markers on DNA sequences. Every test T has a certain false alarm rate:*

$$\rho_{1,-1} = \mathbb{P}(T = 1 | D = -1)$$

and a false negative rate:

$$\rho_{-1,1} = \mathbb{P}(T = -1 | D = 1)$$

Suppose that for the test under consideration $\rho_{1,-1} = 0.05$, $\rho_{-1,1} = 0.05$. From the population screening we know $\mathbb{P}(D = 1) = 0.01$.

Calculate the $\mathbb{P}(T = 1 | D = 1)$ and calculate the probability of having a disease given that the test is positive. Also calculate $\mathbb{P}(D = 1 | T = -1)$.

The probability $\rho_{1,1} = \mathbb{P}(T = 1 | D = 1) = 1 - \rho_{-1,1} = 0.95$. Bayes' formula yields

$$\mathbb{P}(D = 1 | T = 1) = \mathbb{P}(T = 1 | D = 1) \mathbb{P}(D = 1) / \mathbb{P}(T = 1)$$

$$\begin{aligned}
&= \rho_{1,1} \mathbb{P}(D = 1) / \{ \rho_{1,1} \mathbb{P}(D = 1) + \rho_{1,-1} \mathbb{P}(D = -1) \} \\
&= (0.95 * 0.01) / (0.95 * 0.01 + 0.05 * 0.99) \\
&= 0.0095 / 0.06 = 0.161
\end{aligned}$$

Hence the probability of having actually the disease is just about 16 %! How can this be such a low number? This can be elucidated by noting that with the marginal distribution of D one person out of 100 has actually this disease. Given the value of $\rho_{1,-1}$ -the false alarm rate- one expects another 5 people. In total we have 6 people testing positive but only 1 to have the disease. This ratio $1/6$ is roughly 16 % as calculated above.

It is also interesting to investigate the chance of actually having the disease given that the test is negative. This is calculated as:

$$\begin{aligned}
\mathbb{P}(D = 1 | T = -1) &= \mathbb{P}(T = -1 | D = 1) \mathbb{P}(D = 1) / \mathbb{P}(T = -1) \\
&= (0.05 * 0.01) / (0.05 * 0.01 + 0.95 * 0.99) \\
&= 0.0005 / 0.94 = 0.00053 = 0.053
\end{aligned}$$

In terms of this chance pattern we may conclude that this test is acceptable.

Exercise 5.6. *The daily business of an investment bank is to decide upon credit worthiness based on rating techniques. Two types of customers (firms) demande credit: good ones and bad ones. Denote similar to Example 5.5 the probability of successful credit repayment as $\rho_1 = \mathbb{P}(T = 1)$ and $D = 1 / -1$ a good/bad customer. Suppose that $\rho_{1,1} = \mathbb{P}(T = 1 | D = 1) = 80\%$ and that $\rho_{1,-1} = \mathbb{P}(T = 1 | D = -1) = 10\%$. From macroeconomic news and rating companies we observe $\rho_1 = 70\%$. Show that the success probability is 94.9 %.*

For a change of argument we give a finite population version of the proof. Suppose there are 10^6 credit applicants. Given $\rho_1 = 0.7$ there are 700,000 good clients and 300,000 bad clients. Of these $560,000 = 0.8 * 700,000$ respectively $30,000 = 0.1 * 300,000$ are successfully repaying their credit. So in total there are 590,000 successful clients giving the success probability of 94.9 %.

From the investment bank point of view credits are issued to bad clients in 5.1 % of the cases.

Exercise 5.7. *Consider the univariate Gaussian shift model $Y_i | \theta \sim \mathcal{N}(\theta, \sigma^2)$ and $\vartheta \sim \mathcal{N}(\nu, \eta^2)$.*

1. Check that for the situation with only one observation ($n = 1$) the value $\int_{-\infty}^{\infty} p(y|\theta) d\theta$ is finite for every y and the posterior distribution of θ coincides with the distribution of Y .
2. Compute the posterior for $n > 1$.

1. Recall that

$$p(y|\theta) = \sigma^{-1} \varphi\{(y - \theta)/\sigma\}$$

It is now easy to see that

$$\int_{-\infty}^{\infty} p(y|\theta) d\theta = \int_{-\infty}^{\infty} \sigma^{-1} \varphi\{(y - \theta)/\sigma\} d\theta = 1 \quad (5.7)$$

since we may interpret the integrand for all y as the pdf of $\theta \sim \mathcal{N}(y, \sigma^2)$. Suppose now that $\theta \in \Theta$ a compact subset of $\mathbb{R}$. Define $\pi(\Theta) = (\int_{\Theta} d\theta)^{-1}$ then

$$\begin{aligned} p(y) &= \pi(\Theta)^{-1} \int_{\Theta} p(y|\theta) d\theta \\ p(\theta|y) &= \frac{p(y|\theta)\pi(\theta)}{p(y)} = \frac{p(y|\theta)}{\int_{\Theta} p(y|\theta) d\theta} \end{aligned} \quad (5.8)$$

Using (5.7) one sees that (5.8) yields the identity of the posterior with the pdf of $(Y|\theta)$.

2. As alternative of proving (5.8) is to recall that in the situation that $\vartheta \sim \mathcal{N}(\nu, \eta^2)$ the posterior is:

$$\theta | Y \sim \mathcal{N}\left(\frac{\nu\sigma^2 + Y\eta^2}{\sigma^2 + \eta^2}, \frac{\sigma^2\eta^2}{\sigma^2 + \eta^2}\right). \quad (5.9)$$

see Exercise 5.2. Let now the prior $\mathcal{N}(\nu, \eta^2)$ become informative in the sense that $\eta^2 \rightarrow \infty$. Then (5.9) will behave asymptotically as $\mathcal{N}(Y, \sigma^2)$ with pdf $\sigma^{-1}\varphi\{(\theta - y)/\sigma\}$.

Applying symmetry of the normal pdf one sees again (5.8). Using this same argument in the situation of Exercise 5.3 when we calculated the posterior for $n > 1$ leads us to:

$$(\theta | Y) \sim \mathcal{N}(S/n, \sigma^2/n)$$

where $S = Y_1 + \dots + Y_n$.

Exercise 5.8. Let $\{X_i\}_{i=1}^n$ be an i.i.d. sample from the normal distribution $\mathcal{N}(\theta, \sigma^2)$, where θ and σ^2 are unknown. Let $\eta = \frac{1}{2\sigma^2}$ and the prior $p(\eta)$ be the Gamma distribution: $p(\eta) \sim \Gamma(\alpha, \lambda)$. The prior distribution $p(\theta|\sigma^2)$ of θ given σ^2 is also the normal distribution:

$$p(\theta|\sigma^2) \sim \mathcal{N}(\theta_0, \frac{\sigma_0^2}{\eta})$$

1. Compute the joint posterior distribution of (θ, η) .
2. Compute the marginal posterior distribution of η .
3. Compute the marginal posterior distribution of θ given η .
4. Compute Bayes estimates of σ^2 and θ .

1. The joint density of (θ, η) is

$$\begin{aligned} p(\theta, \eta) &= p(\theta|\sigma^2)p(\eta) \\ &= \frac{\sqrt{\eta}}{\sqrt{2\pi\sigma_0^2}} \exp\left\{-\eta(\theta - \theta_0)^2/(2\sigma_0^2)\right\} \frac{\alpha^\lambda}{\Gamma(\lambda)} e^{-\alpha\eta} \eta^{\lambda-1} \end{aligned}$$

Therefore, the joint posterior pdf given $x = \{X_i\}_{i=1}^n$ is

$$\begin{aligned} p(\theta, \eta|x) &\propto (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left\{-\eta \sum_{i=1}^n (x_i - \theta)^2\right\} p(\theta, \eta) \\ &\propto (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left\{-\eta \sum_{i=1}^n (x_i - \theta)^2\right\} \frac{\sqrt{\eta}}{\sqrt{2\pi\sigma_0^2}} \\ &\quad \exp\left\{\frac{-\eta(\theta - \theta_0)^2}{(2\sigma_0^2)}\right\} \frac{\alpha^\lambda}{\Gamma(\lambda)} e^{-\alpha\eta} \eta^{\lambda-1} \\ &\propto \frac{\sqrt{\eta}}{\sqrt{2\pi\sigma_0^2}} (2\pi\sigma^2)^{-\frac{n}{2}} \\ &\quad \exp\left\{-\eta \sum_{i=1}^n (x_i - \theta)^2 - \alpha\eta - \frac{\eta(\theta - \theta_0)^2}{(2\sigma_0^2)}\right\} \frac{\alpha^\lambda}{\Gamma(\lambda)} \eta^{\lambda-1} \end{aligned}$$

Plug in $\eta = \frac{1}{2\sigma^2}$

$$p(\theta, \eta|x) \propto \eta^{\frac{n+1}{2}+\lambda-1} \exp\left[-\eta\left\{\alpha + S^2 + n(\theta - \bar{x})^2 + \frac{(\theta - \theta_0)^2}{2\sigma_0^2}\right\}\right]$$

where,

$$\sum_{i=1}^n (x_i - \theta)^2 = \sum_{i=1}^n (x_i - \bar{x})^2 + n(\theta - \bar{x})^2$$

In the right hand side expression, α and $S^2 = \sum_{i=1}^n (x_i - \bar{x})^2$ does not depend on parameters θ and y , and it is easy to find:

$$\begin{aligned} n(\theta - \bar{x})^2 + \frac{(\theta - \theta_0)^2}{2\sigma_0^2} &= \left(n + \frac{1}{2\sigma_0^2}\right)\theta^2 - 2\left(n\bar{x} + \frac{\theta_0}{2\sigma_0^2}\right)\theta + n\bar{x}^2 + \frac{\theta_0^2}{2\sigma_0^2} \\ &= \left(n + \frac{1}{2\sigma_0^2}\right)(\theta - \xi)^2 + m(x), \end{aligned}$$

where,

$$\xi = (n + \frac{1}{2\sigma_0^2})^{-1}(n\bar{x} + \frac{\theta_0}{2\sigma_0^2})$$

$$m(x) = \frac{(\theta_0 - \bar{x})^2}{2\sigma_0^2 + \frac{1}{n}}$$

The joint posterior density function of (θ, η) can therefore be written as:

$$p(\theta, \eta|x) \propto \eta^{\frac{n+1}{2}+\lambda-1} \exp \left[-\eta \left\{ \alpha + S^2 + m(x) + (n + \frac{1}{2\sigma_0^2})(\theta - \xi)^2 \right\} \right]$$

2. Rewrite the expression of the joint posterior density function as follows:

$$\begin{aligned} p(\theta, \eta|x) &\propto \eta^{\frac{n}{2}+\lambda-1} \exp \left[-\eta \left\{ \alpha + S^2 + m(x) \right\} \right] \frac{1}{\sigma} \\ &\quad \exp \left\{ -\eta \left(n + \frac{1}{2\sigma_0^2} \right) (\theta - \xi)^2 \right\} \end{aligned}$$

The posterior marginal distribution is

$$\begin{aligned} p(\eta|x) &= \int_{-\infty}^{\infty} p(\theta, \sigma^2|x) d\theta \\ &\propto \eta^{\frac{n}{2}+\lambda-1} \exp \left[-\eta \left\{ \alpha + S^2 + m(x) \right\} \right] \end{aligned}$$

This is the pdf of a $\Gamma(\alpha + S^2 + m(x), \frac{n}{2} + \lambda)$ distribution.

3. Note that

$$p(\theta|\eta, x) = \frac{p(\theta, \eta|x)}{p(\eta|x)}$$

The posterior marginal distribution of θ is

$$p(\theta|\eta, x) \propto \frac{1}{\sigma} \exp \left\{ -\eta \left(n + \frac{1}{2\sigma_0^2} \right) (\theta - \xi)^2 \right\}$$

which is proportional to the pdf of a Normal distribution

$$\mathcal{N}(\xi, \{(2n + \sigma_0^{-2})\eta\}^{-1}).$$

4. (a) The joint posterior density function is

$$p(\theta, \eta|x) \propto \eta^{\frac{n+1}{2}+\lambda-1} \exp[-\eta \{\alpha + S^2 + m(x)\}] \\ \exp \left\{ -\eta \left(n + \frac{1}{2\sigma_0^2} \right) (\theta - \xi)^2 \right\}$$

The posterior marginal distribution $p(\eta|x)$ is proportional to the pdf of a $\Gamma \left\{ \alpha + S^2 + m(x), \frac{n}{2} + \lambda \right\}$ distributed random variable, where

$$\xi = \left(n + \frac{1}{2\sigma_0^2} \right)^{-1} \left(n\bar{x} + \frac{\theta_0}{2\sigma_0^2} \right),$$

$$\eta = \frac{1}{2\sigma^2},$$

$$m(x) = \frac{(\theta_0 - \bar{x})^2}{2\sigma_0^2 + \frac{1}{n}},$$

$$S^2 = \sum_{i=1}^n (x_i - \bar{x})^2.$$

The Bayes estimate of σ^2 is

$$\begin{aligned} \tilde{\sigma}^2 &= \frac{1}{2} \int_0^\infty \int_{-\infty}^\infty \eta^{-1} p(\theta, \eta|x) d\eta d\theta \\ &= \frac{1}{2} \int_0^\infty \eta^{-1} p(\eta|x) d\eta \\ &= \frac{\alpha + S^2 + m(x)}{n + 2\lambda - 2} \end{aligned}$$

(b) In order to compute the posterior marginal density of θ , it is necessary to calculate $p(\theta|x) = \int_0^\infty p(\theta, \eta|x) d\eta$. This integration could be expressed as follows:

$$p(\theta|x) = \int_0^\infty \eta^{\frac{n+1}{2}+\lambda-1} \\ \exp \left[-\eta \left\{ \alpha + S^2 + m(x) + \left(n + \frac{1}{2\sigma_0^2} \right) (\theta - \xi)^2 \right\} \right] d\eta$$

Remark 5.1.

$$\eta^{\frac{n+1}{2}+\lambda-1} \exp \left[-\eta \left\{ \alpha + S^2 + m(x) + \left(n + \frac{1}{2\sigma^2} \right) (\theta - \xi)^2 \right\} \right]$$

is the pdf of a

$$\Gamma \left(s, \frac{n+1}{2} + p \right)$$

distributed random variable, here, $s = \alpha + S^2 + m(x) + \left(n + \frac{1}{2\sigma^2} \right) (\theta - \xi)^2$. Therefore, the integration of this density function is 1.

It is now easy to get:

$$\begin{aligned} \int_0^\infty \eta^{\frac{n+1}{2}+\lambda-1} \exp \left\{ -\eta \left[\alpha + S^2 + m(x) + \left(n + \frac{1}{2\sigma^2} \right) (\theta - \xi)^2 \right] \right\} d\eta \\ = \Gamma \left(\frac{n+1}{2} + p \right) / s^{\frac{n+1}{2}+\lambda} \end{aligned}$$

So the posterior marginal density of θ could be expressed as

$$\begin{aligned} p(\theta|x) &= \int_0^\infty p(\theta, \eta|x) d\eta \\ &\propto \left\{ \alpha + S^2 + m(x) + \left(n + \frac{1}{2\sigma^2} \right) (\theta - \xi)^2 \right\}^{-\frac{n+2p+1}{2}} \\ &\propto \left\{ 1 + \frac{\left(n + \frac{1}{2\sigma_0^2} \right) (\theta - \xi)^2}{\alpha + S^2 + m(x)} \right\}^{-\frac{(n+2p)+1}{2}} \\ &\propto \left(1 + \frac{1}{n+2\lambda} t^2 \right)^{-\frac{(n+2p)+1}{2}} \end{aligned}$$

which is proportional to the pdf of a $t(n+2\lambda)$ distributed random variable.

Here, we suppose

$$\begin{aligned} u^2 &= (\theta - \xi)^2 \left\{ \left(n + 2\lambda \right) \left(n + \frac{1}{2\sigma_0^2} \right) / \alpha + S^2 + m(x) \right\} \\ &= K^2 (\theta - \xi)^2, u = K(\theta - \xi) \text{ is a linear transformation of } \theta. \text{ Denote } p(\theta|x) \propto f(u). \end{aligned}$$

As $p(\theta|x)$ is symmetric about $u = 0$, the Bayes estimate of θ could be written as

$$\begin{aligned}
\tilde{\theta} &= \int_{-\infty}^{\infty} \int_0^{\infty} \theta p(\theta, \eta|x) d\eta d\theta \\
&= \int_{-\infty}^{\infty} \theta p(\theta|x) d\theta = \int_{-\infty}^{\infty} (\theta - \xi) + \xi p(\theta|x) d\theta \\
&\propto \int_{-\infty}^{\infty} \frac{1}{K} u f(u) du + \xi \int_{-\infty}^{\infty} p(\theta|x) d\theta
\end{aligned}$$

we have $u = K(\theta - x_i)$, $\theta - x_i = \frac{u}{K}$, $\theta = \frac{u}{K} + x_i$, the first integration is symmetric about $t = 0$, so the result of $p(\theta|x)$ should be zero. The second integration is the pdf of Gamma distribution, the result is 1. Therefore, we have

$$\tilde{\theta} = 0 + \xi = \xi$$

Exercise 5.9. Let $X \sim f(x, \theta)$, $\theta = (\theta_1, \theta_2)$, $\theta_i \in \Theta_i$, $i = 1, 2$; $\theta \sim p(\theta) = p(\theta_1|\theta_2)p(\theta_2)$, $p(\theta_2)$ is density function on Θ_2 . Given any θ_2 , $p(\theta_1|\theta_2)$ is the probability density function of θ_1 on Θ_1 . If θ_2 is given, and the Bayes estimate under quadratic loss of $h(\theta_1) = g(\theta_1, \theta_2)$ is $\eta(X, \theta_2)$, then the Bayes estimate under quadratic loss of $g(\theta_1, \theta_2)$ is $\eta(X)$, which satisfies the following relationship: $\eta(X) = \int_{\Theta_2} \eta(X, \theta_2) p(\theta_2|X) d\theta_2$. $p(\theta_2|X)$ is the posterior density function of θ_2

Prove this result and apply it to 5.7 to find Bayes estimate of $\mu(\theta$ in 5.7), σ^2 , $g(\mu, \delta) = \mu\sigma^2$ with quadratic loss.

From the conditional distribution function formula we have

$$p(\theta_1, \theta_2|x) = p(\theta_1|x, \theta_2)p(\theta_2|x)$$

$p(\theta_2|x)$ is the posterior density function of θ_2 , $p(\theta_1|x, \theta_2)$ is the posterior density function of θ_1 given θ_2 . As the Bayes estimate of $h(\theta_1)$ with quadratic loss is $\eta(X, \theta_2)$, $\eta(X, \theta_2) = \int_{\Theta_1} h(\theta_1) p(\theta_1|X, \theta_2) d\theta_1$. The Bayes estimate of $g(\theta_1, \theta_2)$ is

$$\begin{aligned}
\delta(X) &= \int_{\Theta_2} \int_{\Theta_1} g(\theta_1, \theta_2) p(\theta_1, \theta_2|X) d\theta_1 d\theta_2 \\
&= \int_{\Theta_2} \left\{ \int_{\Theta_1} g(\theta_1, \theta_2) p(\theta_1|X, \theta_2) d\theta_1 \right\} p(\theta_2|X) d\theta_2 \\
&= \int_{\Theta_2} \eta(X, \theta_2) p(\theta_2|X) d\theta_2.
\end{aligned}$$

In 5.7, $\{X_i\}_{i=1}^n$ is an i.i.d. sample from the normal distribution $N(\mu, \sigma^2)$, $\theta_1 = \mu$, $\theta_2 = \eta = \frac{1}{2\sigma^2}$. $\theta_2 = \eta \sim p(\theta_2) \sim \Gamma(a, \lambda)$; $p(\theta_1|\theta_2) = p(\mu|\eta)$ is the pdf of a $N(\mu_0, \frac{\sigma^2}{\eta})$ distributed random variables, $\eta = \frac{1}{2\sigma^2}$, and the posterior distribution of $\theta_2 = \delta$ is $p(\theta_2|X)$ which is the pdf of a $\Gamma(\alpha + S^2 + m(x), \frac{n}{2} + \lambda)$ distributed

random variables, $\eta = \frac{1}{2\sigma^2}$, $\xi = (n + \frac{1}{2\sigma^2})^{-1}(n\bar{x} + \frac{\mu_0}{2\sigma_0^2})$, $m(x) = \frac{(\theta_0 - \bar{x})^2}{2\sigma_0^2 + \frac{1}{n}}$, $S^2 = \sum_{i=1}^n (x_i - \bar{x})^2$. Now we shall compute the Bayes estimate of $g(\theta_1, \theta_2) = \mu, \sigma^2$ and $\mu\sigma^2$ with quadratic loss:

1. $g(\theta_1, \theta_2) = \mu = \theta_1$. $\{X_i\}_{i=1}^n$ is an i.i.d. sample from the normal distribution $\mathcal{N}(\mu, \sigma^2)$, the prior π is also the normal distribution $\mathcal{N}(\mu_0, \frac{\sigma_0^2}{\eta})(\theta_2 = \eta = \frac{1}{2\sigma^2})$, the posterior distribution is normal as well. So the Bayes estimate of $\mu = \theta_1$ given θ_2 is posterior mean.

$$\begin{aligned}\eta(X, \theta_2) &= \left(\frac{n}{\sigma^2} + \frac{\eta}{\sigma_0^2}\right)^{-1} \left(\frac{n}{\sigma^2} \bar{x} + \frac{\eta}{\sigma_0^2} \mu_0\right) \\ &= \left(n + \frac{1}{2\sigma_0^2}\right)^{-1} \left(n\bar{x} + \frac{1}{2\sigma_0^2} \mu_0\right) \\ &= \xi\end{aligned}$$

This expression does not depend on $\theta_2 = \eta$. Thus, the Bayes estimation of $\mu = \theta_1$ is $\tilde{\mu} = \eta(X) = \int_{\Theta_2} \xi p(\theta_2|X) d\theta_2 = \xi$, which is consistent with the result in 5.7.

2. $g(\theta_1, \theta_2) = \sigma^2 = (2\theta_2)^{-1}$, $g(\theta_1, \theta_2) = (2\theta_2)^{-1}$ does not depend on θ_1 . Therefore, the Bayes estimate of $g(\theta_1, \theta_2)$ is

$$\eta(X, \theta_2) = \int_{\Theta_1} (2\theta_2)^{-1} p(\theta_1|X, \theta_2) d\theta_1 = (2\theta_2)^{-1}$$

As $\theta_2|X$ has a Gamma distribution mentioned above, the Bayes estimate of $\sigma^2 = (2\theta_2)^{-1}$ is

$$\tilde{\sigma}^2 = \frac{1}{2} \int_0^\infty \theta_2^{-1} p(\theta_2|X) d\theta_2 = \frac{\alpha + S^2 + m(X)}{n + 2\lambda - 2} \quad (5.10)$$

3. $g(\theta_1, \theta_2) = \mu\sigma^2 = \theta_1(2\theta_2)^{-1}$. $(2\theta_2)^{-1}$ does not depend on θ_1 . Therefore, the Bayes estimate of $g(\theta_1, \theta_2)$ is $\eta(X, \theta_2) = \xi(2\theta_2)^{-1}$ given θ_2 . As ξ does not depend on θ_2 , from (5.10), the Bayes estimate of $g(\theta_1, \theta_2)$ is

$$\eta(X) = \int_0^\infty \xi(2\theta_2)^{-1} p(\theta_2|X) d\theta_2 = \xi \frac{\alpha + S^2 + m(X)}{n + 2\lambda - 2}$$

Exercise 5.10. Let $\{X_i\}_{i=1}^n$ be an i.i.d. sample, compute the Bayes estimate with quadratic loss and posterior MLE of the corresponding parameters:

1. $X_1 \sim f(x_1, \theta) = 2x_1\theta^{-2}\mathbf{1}\{0 \leq x_1 \leq \theta\}$, θ has a Pareto distribution $PR(\alpha, \mu)$; Hint: Try to prove that the Pareto distribution is the conjugate prior distribution for θ if c is known (c equals 2 here).

2. $X_1 \sim f(x_1, c) = cx_1^{c-1} \mathbf{1}\{0 \leq x_1 \leq 1\} (c > 0)$, c has a Gamma distribution $\Gamma(a, \lambda)$;

Hint: Try to prove that the Gamma distribution is the conjugate prior distribution for c if θ is known (θ equals 1 here).

3. $X_1 \sim f(x_1, b) = b^2 x_1 e^{-bx_1} \mathbf{1}\{x_1 \geq 0\}$, b has a Gamma distribution $\Gamma(a, \lambda)$;
 4. $X_1 \sim \Gamma(\frac{1}{\sigma}, \nu)$, ν is known, σ has a Inverse Gamma distribution $\Gamma^{-1}(a, \lambda)$.

1. (a) θ has a Pareto distribution $PR(\alpha, \mu)$, that is,

$$p(\theta) = \alpha \mu^\alpha \theta^{-(\alpha+1)} \mathbf{1}\{\theta \geq \mu\}$$

we know $\alpha \geq 1$ and $\mu \geq 0$, the posterior distribution density function is

$$\begin{aligned} p(\theta|x) &\propto \theta^{-nc} \mathbf{1}\{\theta \geq x_{(n)}\} \alpha \mu^\alpha \theta^{-(\alpha+1)} \mathbf{1}\{\theta \geq \mu\} \\ &\propto \theta^{-(nc+\alpha+1)} \mathbf{1}\{\theta \geq \theta_0\} \end{aligned}$$

Here, $\theta_0 = \max\{x_{(n)}, \mu\}$, so $p(\theta|x)$ is proportional to the pdf of a $PR(2n + \alpha, \theta_0)$ distributed random variables. Therefore the conjugate prior distribution for θ is Pareto distribution if c is known. Plug in $c = 2$, we easily have

$$p(\theta|x) = (2n + \alpha) \theta_0^{2n+\alpha} \theta^{-(2n+\alpha+1)} \mathbf{1}\{\theta \geq \theta_0\}$$

Which is the pdf of a $PR(2n + \alpha, \theta_0)$ distributed random variables. Thus, the Bayes estimate of θ is

$$\tilde{\theta} = E(\theta|X) = \frac{2n + \alpha}{2n + \alpha - 1} \theta_0$$

- (b) In order to maximize $p(\theta|x)$, it is necessary to minimize θ , in this problem, θ should not be smaller than θ_0 . Therefore, the posterior MLE of θ is $\tilde{\theta} = \theta_0$.
 2. (a) c has a Gamma distribution $\Gamma(a, \lambda)$, that is,

$$p(c) = \frac{a^\lambda}{\Gamma(\lambda)} e^{-ac} c^{\lambda-1} \mathbf{1}\{c \geq 0\}$$

$a > 0, \lambda > 0$, the posterior density function is

$$\begin{aligned} p(c|x) &\propto c^n \left(\prod_{i=1}^n x_i \right)^{c-1} \theta^{-nc} e^{-ac} c^{\lambda-1} \\ &\propto c^{n+\lambda-1} \exp -c[a - \log \sum_{i=1}^n (\log x_i - \log \theta)] \end{aligned}$$

which is proportional to the pdf of a $\Gamma(a - \sum_{i=1}^n (\log x_i - \log \theta), n + \lambda)$ distributed random variable. Thus, the Gamma distribution is the conjugate prior distribution for c if θ is known. Plug in $\theta = 1$, the posterior distribution of c is

$$\begin{aligned} p(c|x) &= c^n \left(\prod_{i=1}^n x_i \right)^{c-1} \frac{a^\lambda}{\Gamma(\lambda)} e^{-ac} c^{\lambda-1} \\ &= c^{n+\lambda-1} \exp -c \left[a - \sum_{i=1}^n (\log x_i - \log \theta) \right] \\ &= c^{n+\lambda-1} \exp -c \left[a - \log \prod_{i=1}^n x_i \right] \end{aligned}$$

which is the pdf of a $\Gamma(a - \sum_{i=1}^n \log x_i, n + \lambda)$ distributed random variable. Thus the Bayes estimate of c is

$$\tilde{c} = \mathbb{E}(c|X) = \frac{n + \lambda}{a - \sum_{i=1}^n \log X_i}$$

- (b) Denote $T = \sum_{i=1}^n \log x_i$, $p(c|x)$ is the pdf of a $\Gamma(a - T, n + \lambda)$ distributed random variable, which could be written as

$$\frac{(a - T)^{n+\lambda}}{\Gamma(n + \lambda)} \exp \{-(a - T)c\} c^{n+\lambda-1}.$$

Now it is easy to write down the posterior MLE of c as follows

$$\tilde{c} = \frac{n + \lambda - 1}{a - T} = \frac{n + \lambda - 1}{a - \sum_{i=1}^n \log X_i}$$

3. (a) $X_1 \sim \Gamma(b, 2)$, the posterior distribution of b is

$$\begin{aligned} p(b|x) &= b^{2n} \prod_{i=1}^n x_i \exp \left\{ -b \sum_{i=1}^n x_i \right\} \frac{a^\lambda}{\Gamma(\lambda)} e^{-ab} b^{\lambda-1} \\ &= \exp \left\{ -b \left(\sum_{i=1}^n x_i + a \right) \right\} b^{2n+\lambda-1} \end{aligned}$$

which is the pdf of a $\Gamma(a + \sum_{i=1}^n x_i, 2n + \lambda)$ distributed random variable.

The Bayes estimate is

$$\begin{aligned}\tilde{b} &= \mathbb{E}(b|X) \\ &= \frac{2n + \lambda}{a + \sum_{i=1}^n \log X_i}\end{aligned}$$

- (b) Denote $T = \sum_{i=1}^n \log x_i$, $p(b|x)$ is proportional to the pdf of a $\Gamma(a + T, 2n + \lambda)$ distributed random variable, which could be written as

$$\frac{(a + T)^{2n+\lambda}}{\Gamma(2n + \lambda)} \exp\{-b(a + T)\} b^{2n+\lambda-1}.$$

The posterior log-likelihood function is

$$L(b|x) = -(a + T)b + (2n + \lambda - 1) \log b + k$$

Here, k is a constant term. From $\frac{\partial L(b|x)}{\partial b} = -(a + T) + (2n + \lambda - 1)/b = 0$, the posterior maximum likelihood estimate of b is

$$\tilde{b} = \frac{2n + \lambda - 1}{a + \sum_{i=1}^n X_i}$$

4. (a) The posterior distribution of σ is

$$\begin{aligned}p(\sigma|x) &= \frac{\sigma^{-nv}}{\Gamma^n(\nu)} \left(\prod_{i=1}^n x_i\right)^{\nu-1} \exp\left\{-\frac{1}{\sigma} \sum_{i=1}^n x_i\right\} \frac{a^\lambda}{\Gamma(\lambda)} \exp\left\{-\frac{a}{\sigma}\right\} \left(\frac{1}{\sigma}\right)^{\lambda+1} \\ &= \exp\left\{-\frac{1}{\sigma} \left(\sum_{i=1}^n x_i + a\right)\right\} \left(\frac{1}{\sigma}\right)^{nv+\lambda+1}\end{aligned}$$

which is the pdf of a $\Gamma^{-1}(\sum_{i=1}^n x_i + a, nv + \lambda)$ distributed random variable.

Therefore, the Bayes estimate of σ is

$$\begin{aligned}\tilde{\sigma} &= \mathbb{E}(\sigma|X) \\ &= \frac{1}{nv + \lambda - 1} \left(a + \sum_{i=1}^n X_i\right)\end{aligned}$$

- (b) Denote $T = \sum_{i=1}^n \log x_i$, $p(\sigma|x)$ is the pdf of the $\Gamma^{-1}(a + T, nv + \lambda)$ distributed random variable, which could be written as

$$\frac{(a + T)^{nv+\lambda}}{\Gamma(nv + \lambda)} \exp\left\{-\frac{1}{\sigma}(a + T)\right\} \left(\frac{1}{\sigma}\right)^{nv+\lambda+1}.$$

Fig. 5.1 A boy is trying to test the Robokeeper which is a machine more reliable than any human goalkeeper



The posterior log-likelihood function is

$$L(\sigma|x) = -(a + T)\frac{1}{\sigma} - (nv + \lambda + 1) \log \sigma + k$$

Here, k is a constant term. $\frac{\partial L(\sigma|x)}{\partial \sigma} = (a + T)/\sigma^2 - (nv + \lambda + 1)/\sigma = 0$, the posterior maximum likelihood estimate of σ is

$$\tilde{\sigma} = \frac{a + \sum_{i=1}^n X_i}{nv + \lambda + 1}$$

Exercise 5.11. *Following a tied soccer game, two teams will have a penalty shoot-out to decide which team shall finally win the tournament. Suppose you are an analyst who is employed by one team and you have the record of the goalkeeper of the other side. Suppose that it is known that in the last two penalty shoot-outs, he has saved the ball 3 times out of $5 + 5 = 10$ shots. Your task is to compute, in the present shoot out, how many times the goalkeeper shall save the ball (Fig. 5.1).*

Hint: Note the record is similar to the Bernoulli experiment $\mathbf{Y} = (Y_1, \dots, Y_n)^\top$ with $n = 10$ in Exercise 5.1.

We denote the event that the goalkeeper saves the ball as A , therefore $p(A) = \theta$. In order to estimate θ , we make n independent observations, among which A occurs x times (Fig. 5.2).

It is necessary to predict the times of success z in the Bernoulli experiment $\mathbf{Y} = (Y_1, \dots, Y_k)^\top$ with $k = 5$. The pdf is

$$p(x|\theta) = \binom{n}{x} \theta^x (1 - \theta)^{n-x}, x = 1, 2, 3 \dots$$

Fig. 5.2 Germany goalkeeper Jens Lehmann's crumpled sheet that helped him save penalties against Argentina in the 2006 World Cup quarter-final shootout raised one million EUR (1.3 million USD) for charity



Assume the prior distribution of θ is Beta distribution $Be(\alpha, \beta)$,

$$p(\theta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1} (1 - \theta)^{\beta-1}, 0 < \theta < 1$$

In order to have the posterior distribution of θ , we should firstly find the joint distribution of x and θ :

$$\begin{aligned} p(x, \theta) &= p(x|\theta)p(\theta) \\ &= \binom{n}{x} \theta^x (1 - \theta)^{n-x} \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1} (1 - \theta)^{\beta-1}, \end{aligned}$$

$$x = 0, 1, \dots, n, 0 < \theta < 1$$

Now determine the marginal distribution $p(x)$

$$p(x) = \int_0^1 p(x, \theta) d\theta \quad (5.11)$$

$$\begin{aligned}
&= \int_0^1 \binom{n}{x} \theta^x (1-\theta)^{n-x} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1} (1-\theta)^{\beta-1} d\theta \\
&= \int_0^1 \binom{n}{x} \theta^{x+\alpha-1} (1-\theta)^{n-x+\beta-1} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} d\theta
\end{aligned}$$

We only need to pay attention to the expression $\theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1}$. We know

$$\int_0^1 \theta^{\alpha'} (1-\theta)^{\beta'} d\theta = \frac{\Gamma(\alpha'+1)\Gamma(\beta'+1)}{\Gamma(\alpha'+\beta'+2)}$$

Applying this to the (5.11), we have the posterior density:

$$\begin{aligned}
p(\theta|x) &= \frac{p(x, \theta)}{p(x)} \\
&= \frac{p(x, \theta)}{\int_0^1 p(x, \theta) d\theta} \\
&= \frac{\Gamma(n+\alpha+\beta)}{\Gamma(x+\alpha)\Gamma(n-x+\beta)} \theta^{x+\alpha-1} (1-\theta)^{n-x+\beta-1}
\end{aligned}$$

which is the pdf of a $Be(\alpha+x, \beta+n-x)$ distribution. The likelihood function of the new sample z is

$$L(z|\theta) = \binom{k}{z} \theta^z (1-\theta)^{k-z}$$

Thus the posterior density of z given x is

$$\begin{aligned}
p(z|x) &= \int_0^1 \binom{k}{z} \theta^z (1-\theta)^{k-z} p(\theta|x) d\theta \\
&= \binom{k}{z} \frac{\Gamma(n+\alpha+\beta)}{\Gamma(x+\alpha)\Gamma(n-x+\beta)} \int_0^1 \theta^{z+x+\alpha-1} (1-\theta)^{k-z+n-x+\beta-1} d\theta \\
&= \binom{k}{z} \frac{\Gamma(n+\alpha+\beta)}{\Gamma(x+\alpha)\Gamma(n-x+\beta)} \frac{\Gamma(z+x+\alpha)\Gamma(k-z+n-x+\beta)}{\Gamma(n+k+\alpha+\beta)}
\end{aligned}$$

Plug in $n = 10$, $x = 3$, $k = 5$. As $p(\theta|x)$ is an expression of α and β , we make $\alpha = \beta = 1$. Take a prior distribution of θ as $Be(1, 1)$, which is also the uniform distribution $U(0, 1)$. The posterior distribution of z is

$$p(z|3) = \binom{5}{z} \frac{\Gamma(12)\Gamma(4+z)\Gamma(13-z)}{\Gamma(17)\Gamma(4)\Gamma(8)}$$

Table 5.1 The posterior probability when $z = 0, 1, 2, 3, 4, 5$

z	0	1	2	3	4	5
$p(z x=3)$	0.1813	0.3022	0.2747	0.1694	0.0641	0.02128

**Fig. 5.3** The Jiao Bei pool

We can choose $z = 0, 1, 2, 3, 4, 5$ in this problem. For example, when $z = 0$, we have

$$p(0|3) = \frac{\Gamma(12)\Gamma(4)\Gamma(13)}{\Gamma(17)\Gamma(4)\Gamma(8)} = \frac{33}{182} = 0.1813$$

When $z = 1$, we have

$$p(1|3) = 5 \frac{\Gamma(12)\Gamma(5)\Gamma(12)}{\Gamma(17)\Gamma(4)\Gamma(8)} = \frac{55}{182} = 0.3022$$

we could calculate all of them as follows (Table 5.1):

From the table we observe that $P(0 \leq z \leq 3) = 0.9231$ and the mode is at 0.3022 when $z = 1$. This says that the goalkeeper has the highest probability to save the ball twice (and higher probability once).

Exercise 5.12. Following the Exercise 1.10, we continue discussing interesting statistical issues of the religious ritual—tossing Jiao Bei. Some temples in Taiwan provide not only a pair of Jiao Bei, but a bowl filled with Jiao Bei, like Fig. 5.3. Worshipers choose one pair from the bowl and perform the ritual introduced in the Exercise 1.10. Worshipers have priors to each Jiao Bei in the pool. This observation inspires this exercise.

Let Y denote the outcome of the Jiao Bei tossing. Y is a Bernoulli random variable with probability p . $Y = 1$ if it is “Sheng-Bei” and $Y = 0$ otherwise.

An experiment is carried out by 20 young statisticians and we obtain 114 observations, with 57 Sheng Bei: Assume that p has a prior distribution $\mathbb{P}(p = 1/2) = 1/3$, $\mathbb{P}(p = 1/3) = 1/3$, $\mathbb{P}(p = 1/4) = 1/3$. What is the posterior density $f(p|\mathbf{y})$?

0	0	1	1	1	1	0	0	1	1	0	0	1	0	0	1	1	0	1	0
1	0	1	1	0	0	1	1	1	0	1	0	1	0	0	1	1	0	0	1
1	0	1	1	0	1	0	1	1	0	1	0	0	1	1	0	0	0	1	1
0	1	0	0	0	1	0	0	0	1	1	0	0	1	0	1	1	1	1	0
0	0	0	0	1	0	0	1	0	1	0	0	1	0	0	1	0	1	1	1
1	1	0	1	1	1	0	1	0	0	1	0	0	1						

Let $\mathbf{y}$ be the outcome of Jiao Bei tossing experiment. Under the prior distribution,

$$\begin{aligned}\mathbb{P}(\mathbf{y}) &= \frac{1}{3}\mathbb{P}(\mathbf{y} = 57|p = 1/2) + \frac{1}{3}\mathbb{P}(\mathbf{y} = 57|p = 1/3) + \frac{1}{3}\mathbb{P}(\mathbf{y} = 57|p = 1/4) \\ &= \frac{1}{3 * 2^{114}} + \frac{2^{57}}{3^{115}} + \frac{3^{57}}{3 * 4^{114}}.\end{aligned}$$

The posterior probabilities are:

$$\begin{aligned}\mathbb{P}(p = 1/2|\mathbf{y}) &= \frac{\mathbb{P}(\mathbf{y}|p = 1/2)\mathbb{P}(p = 1/2)}{\mathbb{P}(\mathbf{y})} \approx 0.9988; \\ \mathbb{P}(p = 1/3|\mathbf{y}) &= \frac{\mathbb{P}(\mathbf{y}|p = 1/3)\mathbb{P}(p = 1/3)}{\mathbb{P}(\mathbf{y})} \approx 0.0012; \\ \mathbb{P}(p = 1/4|\mathbf{y}) &= \frac{\mathbb{P}(\mathbf{y}|p = 1/4)\mathbb{P}(p = 1/4)}{\mathbb{P}(\mathbf{y})} \approx 0.\end{aligned}$$

Reference

Spokoiny, V., & Dickhaus, T. (2014). *Basics of modern parametric statistics*. Berlin: Springer.

Chapter 6

Testing a Statistical Hypothesis

Provare un'ipotesi

Il segreto del successo è la costanza del proposito.

The secret of success is the perseverance.

Exercise 6.1. Let $\mathbf{X} = \{X_i\}_{i=1}^n$ be an i.i.d. sample from a model of Gaussian shift $\mathcal{N}(\theta, \sigma^2)$ (here σ is a known parameter and θ is a parameter of interest).

(i) Fix some level $\alpha \in (0, 1)$ and find a number $t_\alpha \in \mathbb{R}$ such that the function

$$\phi(\mathbf{X}) \stackrel{\text{def}}{=} \mathbf{1}(\bar{X} \geq t_\alpha)$$

is a test of level α for checking the hypothesis $H_0 : \theta = \theta_0$ against the alternative $H_1 : \theta = \theta_1 < \theta_0$ (θ_0 and θ_1 are two fixed values).

(ii) Find the power function $W(\theta_1)$ for this test.

(iii) Compare α and $W(\theta_1)$. How can you interpret the results of this comparison?

(iv) Why a test in the form

$$\phi(\mathbf{X}) \stackrel{\text{def}}{=} \mathbf{1}(\bar{X} \leq s_\alpha),$$

where $s_\alpha \in \mathbb{R}$ is not appropriate for testing the hypothesis H_0 against the alternative H_1 ?

(i) Observe that $\sqrt{n}(\bar{X} - \theta_0)$ has a standard normal distribution $\mathcal{N}(0, 1)$ under H_0 . Then for any t

$$\mathbb{P}_0(\bar{X} \geq t) = \mathbb{P}_0 \left\{ \frac{\sqrt{n}(\bar{X} - \theta_0)}{\sigma} \geq \frac{\sqrt{n}(t - \theta_0)}{\sigma} \right\} = 1 - \Phi \left\{ \frac{\sqrt{n}(t - \theta_0)}{\sigma} \right\},$$

where $\mathbb{P}_0$ denotes the probability measure of the normal distribution $\mathcal{N}(\theta_0, \sigma^2)$.

Let us fix the parameter t such that $\mathbb{P}_0(\bar{X} \geq t) = \alpha$:

$$\alpha = 1 - \Phi \left\{ \frac{\sqrt{n}(t - \theta_0)}{\sigma} \right\}$$

$$t = t_\alpha = \theta_0 + \sigma z_{1-\alpha} / \sqrt{n},$$

where $z_{1-\alpha}$ is the $(1 - \alpha)$ -quantile of the standard normal distribution.

So, a test of level α is

$$\phi(X) \stackrel{\text{def}}{=} \mathbf{1}(\bar{X} \geq t_\alpha) = \mathbf{1}(\bar{X} \geq \theta_0 + \sigma z_{1-\alpha} / \sqrt{n}).$$

(ii) By the definition of the error of the second kind,

$$\begin{aligned} W(\theta_1) &= 1 - \mathbb{P}_1 \{ \phi(X) = 0 \} = \mathbb{P}_1 \{ \phi(X) = 1 \} \\ &= \mathbb{P}_1 (\bar{X} \geq \theta_0 + \sigma z_{1-\alpha} / \sqrt{n}) \\ &= \mathbb{P}_1 \left\{ \frac{\sqrt{n}(\bar{X} - \theta_1)}{\sigma} \geq z_{1-\alpha} - \sqrt{n}(\theta_0 - \theta_1) \right\} \\ &= 1 - \Phi \left\{ z_{1-\alpha} - \frac{\sqrt{n}(\theta_0 - \theta_1)}{\sigma} \right\} \end{aligned}$$

(iii) One should compare two expressions:

$$\alpha = 1 - \Phi(z_{1-\alpha}) \quad \text{and} \quad W(\theta_1) = 1 - \Phi \left\{ z_{1-\alpha} - \frac{\sqrt{n}(\theta_0 - \theta_1)}{\sigma} \right\}.$$

By assumption, $\theta_0 > \theta_1$. This yields

$$z_{1-\alpha} > z_{1-\alpha} - \frac{\sqrt{n}(\theta_0 - \theta_1)}{\sigma},$$

and therefore $\alpha < W(\theta_1)$ because the function $\Phi(\cdot)$ is monotone increasing. This fact can be interpreted in the following way: the probability of rejecting the hypothesis when it is true is less than the probability of rejecting the hypothesis when it is false. In other words, “true rejection” has larger probability than “false rejection”.

(iv) In the case of the test

$$\phi(\mathbf{X}) \stackrel{\text{def}}{=} \mathbf{1}(\bar{X} \leq s_\alpha),$$

the error of the first level is larger than the power function at any point $\theta_1 < \theta_0$. This means that “false rejection” has larger probability than “true rejection”.

Exercise 6.2. Let a sample $\mathbf{X}$ have only one observation X with density $p(x - \theta)$. Consider the hypothesis $\theta = 0$ against the alternative $\theta = 1$. Describe the critical region of the Neyman-Pearson test for different t_α if p is a density of

- (i) The standard normal distribution $\mathcal{N}(0, 1)$,
 - (ii) The standard Cauchy distribution, i.e. $p(x) = \{\pi(1 + x^2)\}^{-1}$.
- (i) Note that

$$\frac{p(x - 1)}{p(x)} = \frac{\frac{1}{\sqrt{2\pi}} \exp\{-\frac{(x-1)^2}{2}\}}{\frac{1}{\sqrt{2\pi}} \exp(-\frac{x^2}{2})} = \exp\left(x - \frac{1}{2}\right).$$

Thus, the critical region is

$$R_\alpha = \left\{ \frac{p(x - 1)}{p(x)} \geq t_\alpha \right\} = \left\{ \exp\left(x - \frac{1}{2}\right) \geq t_\alpha \right\}.$$

If $t_\alpha \leq 0$ then $R_\alpha = \mathbb{R}$. On the other hand, if t_α is positive then

$$R_\alpha = \{x \geq \log t_\alpha + 1/2\}.$$

(ii) The case of the Cauchy distribution is more complicated.

$$R_\alpha = \left\{ \frac{p(x - 1)}{p(x)} \geq t_\alpha \right\} = \left\{ \frac{1 + x^2}{1 + (x - 1)^2} \geq t_\alpha \right\}.$$

A plot of the function $f(x) = \frac{1+x^2}{1+(x-1)^2}$ is given on the Fig. 6.1.

Note that

- The maximum is attained at the point $x_{\max} = (1 + \sqrt{5})/2$ and is equal to $y_{\max} = (3 + \sqrt{5})/2$
- The minimal value is attained at the point $x_{\min} = (1 - \sqrt{5})/2$ and is equal to $y_{\min} = (3 - \sqrt{5})/2$
- The right and the left “tails” of the function tend to the line $y = 1$.

These three observations yield the following sets R_α :

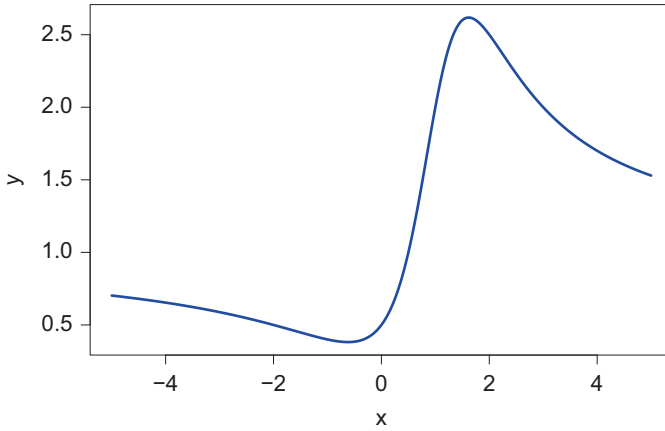


Fig. 6.1 The plot $y = f(x) = (1 + x^2)/(1 + (x - 1)^2)$.  MSEfcauchy

$$R_\alpha = \begin{cases} \emptyset, & \text{if } t_\alpha > y_{\max} \\ x_{\max}, & \text{if } t_\alpha = y_{\max} \\ [x_1, x_2], & \text{if } t_\alpha \in (1, y_{\max}) \\ [x_1, +\infty) & \text{if } t_\alpha = 1 \\ (-\infty, x_1] \cup [x_2, +\infty), & \text{if } t_\alpha \in (y_{\min}, 1) \\ \mathbb{R}, & \text{if } t_\alpha \leq y_{\min} \end{cases}$$

where x_1 and x_2 ($x_1 < x_2$) are two solutions of the quadratic equation

$$\frac{1 + x^2}{1 + (x - 1)^2} = t_\alpha.$$

Exercise 6.3. (Suhov & Kelbert, 2005) Let X_1 be a single observation of a random variable X with the density function $p(x)$.

- (i) Find the form of the most powerful test of fixed size $\alpha = 0.05$ of the hypothesis $H_0 : p(x) = 1/2 \mathbf{1}(x \in [-1, 1])$ against an alternative $H_1 : p(x) = 3/4(1 - x^2)\mathbf{1}(x \in [-1, 1])$.
- (ii) Compute the power of this test.

Let us apply the Neyman-Pearson lemma. The left hand side of the equality

$$\mathbb{P}_0 \{Z(X_1) \geq t_\alpha\} = \alpha = 0.05$$

can be rewritten as

$$\begin{aligned}\mathbb{P}_0\{Z(X_1) \geq t_\alpha\} &= \mathbb{P}_0\{3/2(1 - X_1^2) \geq t_\alpha\} = \mathbb{P}_0\{|X_1| \leq \sqrt{1 - 2/3t_\alpha}\} \\ &= \sqrt{1 - 2/3t_\alpha}.\end{aligned}$$

So, we conclude that the Neyman-Pearson test can be written as

$$\phi(X_1) = \mathbf{1}(|X_1| \leq 0.05).$$

(ii) The calculation of the power is straightforward:

$$\mathbb{P}_1(|X_1| \geq 0.05) = \int_{-0.05}^{0.05} \frac{3}{4}(1 - x^2)dx \approx 0.075.$$

Exercise 6.4. Let $\{X_i\}_{i=1}^n$ be an i.i.d. sample from the exponential distribution

$$p(x, \theta) = \theta^{-1} \exp(-x/\theta), \quad x \geq 0.$$

1. Find the form of the most powerful test of fixed size α of the hypothesis $H_0 : \theta = \theta_0$ against an alternative $H_1 : \theta = \theta_1$, where θ_0 and θ_1 are given (let $\theta_0 < \theta_1$).
2. Compute the power of this test (the cdf of Gamma distribution can be involved).

1. The Neyman-Pearson lemma is applicable in this situation, because the likelihood ratio

$$Z(\mathbf{X}) \stackrel{\text{def}}{=} \frac{\prod_{i=1}^n p(X_i, \theta_1)}{\prod_{i=1}^n p(X_i, \theta_0)} = \left(\frac{\theta_0}{\theta_1}\right)^n \exp\left\{\left(\frac{1}{\theta_0} - \frac{1}{\theta_1}\right) \sum_{i=1}^n X_i\right\}$$

is such that the equation

$$\mathbb{P}_0\{Z(\mathbf{X}) \geq t_\alpha\} = \alpha \tag{6.1}$$

has a solution for any $\alpha > 0$.

In order to find a close form for this solution, note that the random variable $\xi \stackrel{\text{def}}{=} 1/\theta_0 \sum_{i=1}^n X_i$ under hypothesis H_0 has a gamma distribution with parameters n and 1. In fact, X_i is distributed according to the law $\Gamma(1, \theta_0)$; then $\sum_{i=1}^n X_i \sim \Gamma(n, \theta_0)$, and $\xi \sim \Gamma(n, 1)$. Denote the cdf of $\Gamma(n, 1)$ by $G_n(\cdot)$.

(6.1) can be rewritten as

$$\mathbb{P}_0\left\{\left(\frac{\theta_0}{\theta_1}\right)^n \exp\left\{\left(\frac{1}{\theta_0} - \frac{1}{\theta_1}\right) \theta_0 \xi\right\} \geq t_\alpha\right\} = \alpha,$$

or, equivalently (here we use that $\theta_0 < \theta_1$),

$$\mathbb{P}_0\left\{\xi \geq \frac{n \log \frac{\theta_1}{\theta_0} + \log t_\alpha}{1 - \frac{\theta_0}{\theta_1}}\right\} = \alpha.$$

So, t_α should be chosen from the following equation:

$$\frac{n \log \frac{\theta_1}{\theta_0} + \log t_\alpha}{1 - \frac{\theta_0}{\theta_1}} = G_n^{-1}(1 - \alpha).$$

According to the Theorem 8.2.2, we conclude that the Neyman-Pearson test with

$$t_\alpha = \exp \left\{ \left(1 - \frac{\theta_0}{\theta_1} \right) G_n^{-1}(1 - \alpha) + n \log \frac{\theta_0}{\theta_1} \right\}$$

is the most powerful test of size α .

2. For the computation of the power, it is sufficient to mention that under H_1

$$\frac{\theta_0}{\theta_1} \xi \sim \Gamma(n, 1).$$

Hence,

$$\begin{aligned} W &= \mathbb{P}_1 \{Z(X) \geq t_\alpha\} = \mathbb{P}_1 \{\xi \geq G_n^{-1}(1 - \alpha)\} \\ &= \mathbb{P}_1 \left\{ \frac{\theta_0}{\theta_1} \xi \geq \frac{\theta_0}{\theta_1} G_n^{-1}(1 - \alpha) \right\} \\ &= 1 - G_n \left(\frac{\theta_0}{\theta_1} G_n^{-1}(1 - \alpha) \right). \end{aligned}$$

Exercise 6.5 (Pestman & Alberink, 1991). Let $\{X_i\}_{i=1}^n$ be a sample from the distribution with density

$$p(x, \theta) = e^{-x+\theta} I(x > \theta),$$

where $\theta \in \mathbb{R}$. Find a uniformly most powerful (UMP) test with given level α for testing the simple hypothesis $H_0 : \theta = \theta_0$ against the simple alternative $\theta = \theta_1 > \theta_0$.

Firstly, we compute the ratio

$$Z(X) \stackrel{\text{def}}{=} \frac{\prod_{i=1}^n p(X_i, \theta_1)}{\prod_{i=1}^n p(X_i, \theta_0)} = \begin{cases} 0, & \theta_0 < X_{(1)} \leq \theta_1, \\ e^{n(\theta_1 - \theta_0)}, & X_{(1)} > \theta_1. \end{cases}$$

Note that the Neyman-Pearson test is not applicable to this situation, because the equation

$$\mathbb{P}_0 \{Z(X) \geq t_\alpha\} = \alpha$$

does not have solution for $\alpha \neq \mathbb{P}_0(X_{(1)} > \theta_1)$. The aim is to construct a test T such that

$$\mathbb{E}_0 T = \alpha \quad \text{and} \quad \mathbb{E}_1 T = \underset{\phi \in (0,1): \mathbb{E}_0 \phi \leq \alpha}{\operatorname{argmax}} (\mathbb{E}_1 \phi). \quad (6.2)$$

First step. It is worth mentioning that a test in the form

$$T^{(1)} = \begin{cases} \gamma, & \theta_0 < X_{(1)} \leq \theta_1, \\ 1, & X_{(1)} > \theta_1. \end{cases}$$

Any $\gamma \in (0, 1)$ satisfies the second condition from (6.2) because

$$\mathbb{E}_1 T^{(1)} = \gamma \underbrace{\mathbb{P}_1(\theta_0 < X_{(1)} \leq \theta_1)}_{=0} + \underbrace{\mathbb{P}_1(X_{(1)} > \theta_1)}_{=1} = 1,$$

and $\mathbb{E}_1 \phi < 1$ for any test ϕ . Thus, the UMP test can be found by selecting the $\gamma \in (0, 1)$ satisfying $\mathbb{E}_0 T^{(1)} = \alpha$. We have

$$\begin{aligned} \mathbb{E}_0 T^{(1)} &= \gamma \mathbb{P}_0(\theta_0 < X_{(1)} \leq \theta_1) + \mathbb{P}_0(X_{(1)} > \theta_1) \\ &= \gamma \{1 - \mathbb{P}_0(X_{(1)} > \theta_1)\} + \mathbb{P}_0(X_{(1)} > \theta_1) \\ &= \gamma \{1 - e^{n(\theta_0 - \theta_1)}\} + e^{n(\theta_0 - \theta_1)}, \end{aligned}$$

because

$$\mathbb{P}_0(X_{(1)} > \theta_1) = \mathbb{P}_0(X_i > \theta_1, i = 1 \dots, n) = \{\mathbb{P}_0(X_1 > \theta_1)\}^n = e^{n(\theta_0 - \theta_1)}.$$

As the result,

$$\gamma = \frac{\alpha - e^{n(\theta_0 - \theta_1)}}{1 - e^{n(\theta_0 - \theta_1)}}, \quad (6.3)$$

and $\gamma \in (0, 1)$ if and only if

$$e^{n(\theta_0 - \theta_1)} < \alpha. \quad (6.4)$$

So, on the first step we prove that if condition (6.4) is fulfilled than $T^{(1)}$ with γ given by (6.3) is an UMP test. On the second step, we consider a case if the condition (6.4) is not fulfilled.

Second step. For any $\eta \in (0, 1)$, denote

$$T^{(2)} \stackrel{\text{def}}{=} \begin{cases} 0, & \theta_0 < X_{(1)} \leq \theta_1 \\ \eta, & X_{(1)} > \theta_1 \end{cases}$$

Let us find η from the first condition (6.2):

$$\mathbb{P}_0 T^{(2)} = \eta \mathbb{P}_0 (X_{(1)} > \theta_1) = \eta e^{n(\theta_0 - \theta_1)} = \alpha$$

Since condition (6.4) is not fulfilled,

$$\eta = \alpha e^{n(\theta_1 - \theta_0)} \quad (6.5)$$

lies between 0 and 1. The power function for this test is equal to

$$\mathbb{E}_1 T^{(2)} = \eta \mathbb{P}_1 (X_{(1)} > \theta_1) = \eta = \alpha e^{n(\theta_1 - \theta_0)}.$$

It's easy to see that this power function is maximal over all tests with $\mathbb{E}_0 \phi \leq \alpha$. In fact, for any such test ϕ ,

$$\mathbb{E}_1 \phi = \mathbb{E}_0 \left\{ \underbrace{Z(X)}_{\leq e^{n(\theta_1 - \theta_0)}} \phi \right\} \leq \alpha e^{n(\theta_1 - \theta_0)}.$$

So $T^{(2)}$ is an UMP test with η chosen by (6.5) given that the condition (6.4) is not fulfilled.

Exercise 6.6. Consider the model $\{X_i\}_{i=1}^n \sim \mathcal{N}(\theta, \sigma^2)$, where σ is known and θ is the parameter of interest. Define two statistics:

$$T_1 \stackrel{\text{def}}{=} \max_{\theta} L(\theta, \theta_0) \quad \text{and} \quad T_2 \stackrel{\text{def}}{=} \max_{\theta > \theta_0} L(\theta, \theta_0),$$

and two corresponding tests:

$$\phi_1 \stackrel{\text{def}}{=} \mathbf{1}(T_1 > t_{1\alpha}) \quad \text{and} \quad \phi_2 \stackrel{\text{def}}{=} \mathbf{1}(T_2 > t_{2\alpha}),$$

where $t_{i\alpha}$ ($i = 1, 2$) are selected to ensure the level condition

$$\mathbb{E}_0 \phi_i = \mathbb{P}_0 \{T_i > t_{i\alpha}\} = \alpha$$

for a given level α . Both tests, ϕ_1 and ϕ_2 , are used for testing the hypothesis

$$H_0 : \theta = \theta_0.$$

ϕ_1 tests H_0 against the alternative $H_1 : \theta \neq \theta_0$ (two-sided test), and ϕ_2 – against $H_2 : \theta > \theta_0$ (one-sided test).

1. Find the explicit expressions for T_1 and T_2 .
 2. Compute the power functions of the tests ϕ_1 and ϕ_2 .
1. Two methods for finding an explicit form for T_1 are given in the second chapter of Spokoiny and Dickhaus (2014), Theorem 2.9.1. One of the methods is based on deriving the following expression for $L(\theta, \theta_0)$ (here θ, θ_0 are any two points):

$$L(\theta, \theta_0) = \frac{n}{\sigma^2} \left\{ (\tilde{\theta} - \theta_0)(\theta - \theta_0) - \frac{(\theta - \theta_0)^2}{2} \right\}. \quad (6.6)$$

So, $L(\theta, \theta_0)$ is a quadratic polynomial in θ' ; the maximum is attained at the vertex of parabola (at the point $\theta = \tilde{\theta}$):

$$T_1 = \max_{\theta} L(\theta', \theta) = \frac{n}{2\sigma^2} |\tilde{\theta} - \theta_0|^2.$$

For maximizing (6.6) for $\theta > \theta_0$, note that $\theta = \theta_0$ is one of two solutions of the equation $L(\theta, \theta_0) = 0$. Consider two cases:

- (i) If θ_0 is the larger solution. In other words, θ_0 is larger than the x-coordinate of the vertex, i.e. $\theta_0 > \tilde{\theta}$. In this case, $L(\theta, \theta_0) \leq 0$ for any $\theta \geq \theta_0$ with equality iff $\theta = \theta_0$.
- (ii) If θ_0 is the smaller solution. Then the “positive” part of parabola (i.e. $\{\theta : L(\theta, \theta_0) > 0\}$) is in the set $\{\theta : \theta > \theta_0\}$ and maximum is attained at $\theta = \tilde{\theta}$.

To summarize, we conclude that

$$T_2 = \sup_{\theta > \theta_0} L(\theta, \theta_0) = \begin{cases} n|\tilde{\theta} - \theta_0|^2/2\sigma^2 & \text{if } \tilde{\theta} \geq \theta_0, \\ 0 & \text{otherwise.} \end{cases}$$

2. By the definition of the power function,

$$\begin{aligned} \beta_1(\theta) &\stackrel{\text{def}}{=} 1 - \mathbb{E}_{\theta} I\{T_1 > t_{1\alpha}\} = \mathbb{P}_{\theta}(T_1 \leq t_{1\alpha}) \\ &= \mathbb{P}_{\theta}\left(n|\tilde{\theta} - \theta_0|^2/2\sigma^2 \leq t_{1\alpha}\right) \\ &= \mathbb{P}_{\theta}\left(\theta_0 - \sigma\sqrt{\frac{2t_{1\alpha}}{n}} < \tilde{\theta} < \theta_0 + \sigma\sqrt{\frac{2t_{1\alpha}}{n}}\right) \end{aligned}$$

Note that $\xi \stackrel{\text{def}}{=} \sqrt{n}\sigma^{-1}(\tilde{\theta} - \theta)$ is standard normal under $\mathbb{P}_{\theta}$. This yields

$$\begin{aligned}
\beta_1(\theta) &= \mathbb{P}_\theta \left\{ \frac{\sqrt{n}(\theta_0 - \theta)}{\sigma} - \sqrt{2t_{1\alpha}} < \xi < \frac{\sqrt{n}(\theta_0 - \theta)}{\sigma} + \sqrt{2t_{1\alpha}} \right\} \\
&= \Phi \left\{ \frac{\sqrt{n}(\theta_0 - \theta)}{\sigma} + \sqrt{2t_{1\alpha}} \right\} - \Phi \left\{ \frac{\sqrt{n}(\theta_0 - \theta)}{\sigma} - \sqrt{2t_{1\alpha}} \right\}
\end{aligned}$$

Computation of the power function for the test ϕ_2 follows the same lines:

$$\begin{aligned}
\beta_2(\theta) &= \mathbb{P}_\theta \left\{ \frac{n}{\sigma^2} (\tilde{\theta} - \theta_0)^2 / 2 \leq t_{2\alpha} \right\} \\
&= \mathbb{P}_\theta \left(\tilde{\theta} < \theta_0 + \sigma \sqrt{\frac{2t_{1\alpha}}{n}} \right) \\
&= \mathbb{P}_\theta \left\{ \xi < \frac{\sqrt{n}(\theta_0 - \theta)}{\sigma} + \sqrt{2t_{1\alpha}} \right\} \\
&= \Phi \left\{ \frac{\sqrt{n}(\theta_0 - \theta)}{\sigma} + \sqrt{2t_{1\alpha}} \right\}.
\end{aligned}$$

Exercise 6.7. Consider the volatility model with a natural parameter θ :

$$Y = \xi^2, \quad \xi \sim \mathcal{N}(0, \theta)$$

(see Exercise 2.19). Observe an i.i.d. sample $\{Y_i\}_{i=1}^n$ with $n = 10$ and suppose that $\sum_{i=1}^n Y_i = 8$.

- (i) Draw a plot of the function $f(\theta) = \mathcal{K}(\theta, \theta_0)$ for a natural parameter θ and $\theta_0 = 1$. Visually check that for every ζ the set

$$\{\theta : \mathcal{K}(\theta, 1) \leq \zeta\}$$

is a connected subset of $\mathbb{R}$.

- (ii) Change the parameter to the canonical parameter, $v = v(\theta) = -(2\theta)^{-1}$. Draw the similar plot for the canonical parameter v , $g(v) = \mathcal{K}\{v, v(1)\}$ and visually check that the set

$$[\theta : \mathcal{K}(v, v(1)) \leq \zeta]$$

is not a connected subset for some v .

- (i) From Exercise 2.19, we know that

$$p(y, \theta) = \frac{1}{2\sqrt{2\pi y}} \exp \left(-\frac{y}{2\theta} - \frac{1}{2} \log \theta \right).$$

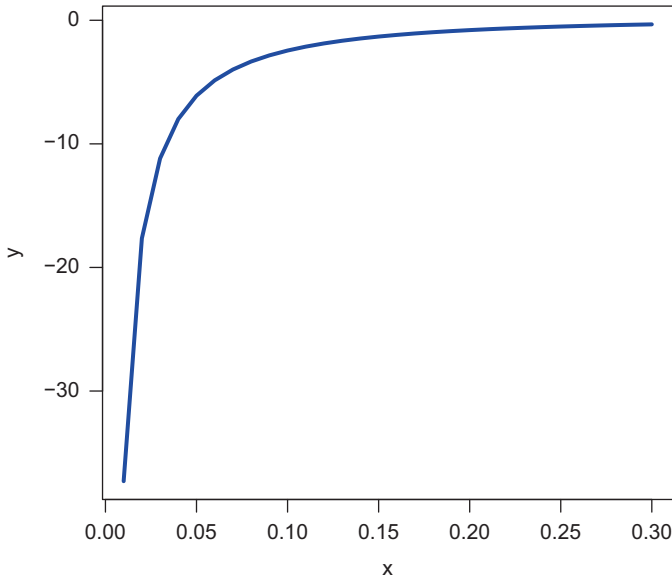


Fig. 6.2 The plot $y = f(\theta)$.  MSEklnatparam

Hence

$$L(\theta) = \sum_{i=1}^n \log p(Y_i, \theta) = -\frac{\sum_{i=1}^n Y_i}{2\theta} - \frac{n}{2} \log \theta - \sum_{i=1}^n \log(2\sqrt{2\pi Y_i})$$

and

$$\mathcal{K}(\theta, 1) = \frac{L(\theta) - L(1)}{n} = -\frac{\sum_{i=1}^n Y_i}{2n} \left(\frac{1}{\theta} - 1 \right) - \frac{1}{2} \log \theta.$$

Substituting the values for n and $\sum Y_i$ yields

$$f(\theta) = -0.4\theta^{-1} - 0.5 \log \theta + 0.4.$$

The graph of the function $f(\theta)$ is given in Fig. 6.2. It is clear that the set

$$\{\theta : f(\theta) \leq \zeta\}$$

is an interval $(0, \theta_\zeta)$ for some $\theta_\zeta > 0$.

(ii) Since $v(\theta) = -(2\theta)^{-1}$; the inverse transform is $\theta(v) = -(2v)^{-1}$.

$$L(v) = v \sum_{i=1}^n Y_i - \frac{n}{2} \log \left(-\frac{1}{2v} \right) - \sum_{i=1}^n \log(2\sqrt{2\pi Y_i})$$

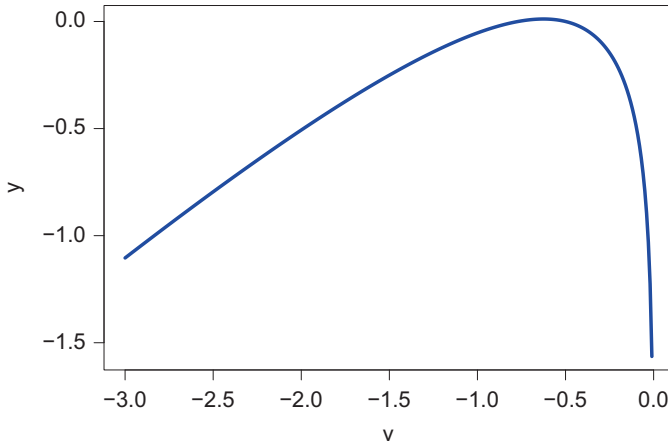


Fig. 6.3 The plot $y = g(v)$.  MSEklcanparam

$$\mathcal{K}\left(v, -\frac{1}{2}\right) = \frac{1}{n} \left(v + \frac{1}{2}\right) \sum_{i=1}^n Y_i - \frac{1}{2} \log\left(-\frac{1}{2v}\right)$$

Note that

$$g(v) = f\{\theta(v)\} = f\left(-\frac{1}{2v}\right) = 0.8v - 0.5 \log\left(-\frac{1}{2v}\right) + 0.4.$$

The graph is given on the Fig. 6.3. It's easy to see that for any $\zeta \in (0, 1)$ the set

$$\{v : g(v) \leq \zeta\}$$

is disconnected.

Exercise 6.8. *This exercise is an illustration of Lemma 8.4.5.*

Consider the model from the previous exercise (a volatility model with a natural parameter θ). Denote by $\tilde{\theta}$ the MLE of θ .

(i) *Prove that for any $\theta_0 \in \mathbb{R}$ and any $t \in \mathbb{R}$, a function*

$$g(\theta) \stackrel{\text{def}}{=} \mathbb{P}_{\theta} \left(\tilde{\theta} > \theta_0 + t \right)$$

is a monotone increasing function on $\mathbb{R}$.

(ii) *Draw a plot of the function $g(\theta)$ for $\theta_0 = 1, t = 0, n = 10$.*

(i) In this exercise, we are faced with an exponential family. Parameter θ is a natural parameter (see Exercise 5.5). According to Theorem 2.11.3, the MLE is equal to

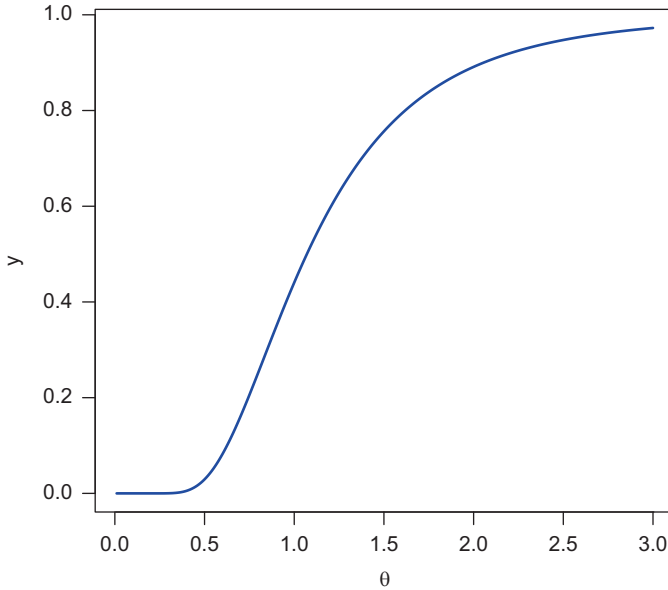


Fig. 6.4 The plot of $g(\theta) = 1 - G_{10}(10/\theta)$.  MSEEX0810

$$\tilde{\theta} = \frac{1}{n} \sum_{i=1}^n Y_i.$$

Note that a random variable $\eta \stackrel{\text{def}}{=} \theta^{-1} \sum_{i=1}^n Y_i$ has a chi-squared distribution with n degrees of freedom; denote the distribution function by G_n . We conclude that

$$g(\theta) = \mathbb{P} \left(\frac{\theta}{n} \eta > \theta_0 + t \right) = 1 - G_n \left\{ \frac{n}{\theta} (\theta_0 + t) \right\},$$

and the monotonicity of the function g is proven.

(ii) In this case, $g(\theta) = 1 - G_{10}(10/\theta)$ (Fig. 6.4).

Exercise 6.9. Let $(\mathbb{P}_\theta)$ be an EFn.

(i) Prove that the α -level LR test for the null $H_0 : \theta \in [\theta_0, \theta_1]$ against the alternative $H_1 : \theta \notin [\theta_0, \theta_1]$ can be written as

$$\phi = \mathbf{1}(\tilde{\theta} < \theta_0 - t_\alpha^-) + \mathbf{1}(\tilde{\theta} > \theta_1 + t_\alpha^+), \quad (6.7)$$

where t_α^+ and t_α^- are selected to ensure

$$\sup_{\theta_0 < \theta < \theta_1} \mathbb{P}_\theta(\phi = 1) \leq \alpha. \quad (6.8)$$

(ii) Let the values t_α^+, t_α^- be selected to ensure

$$\mathbb{P}_{\theta_0}(\tilde{\theta} < \theta_0 - t_\alpha^-) = \alpha/2, \quad \mathbb{P}_{\theta_1}(\tilde{\theta} > \theta_1 + t_\alpha^+) = \alpha/2.$$

Prove that the level condition (6.8) is fulfilled in this case.

(i) The LR test is defined as the test in the form $\mathbf{1}(T > t_\alpha)$, where

$$T \stackrel{\text{def}}{=} \sup_{\theta \notin [\theta_0, \theta_1]} L(\theta) - \sup_{\theta \in [\theta_0, \theta_1]} L(\theta).$$

In order to describe the behavior of the function L , one takes the first derivative of it:

$$L(\theta) = SC(\theta) - nB(\theta) + \sum_i \log p(Y_i),$$

$$\frac{dL}{d\theta} = n(\tilde{\theta} - \theta)C'(\theta),$$

where $S = \sum_i Y_i$ and $\tilde{\theta} = S/n$. This and $C'(\theta) = I(\theta) > 0$ yield that the function $L(\theta)$ monotone increase on the interval $(-\infty, \tilde{\theta}]$ and monotone decrease on $[\tilde{\theta}, +\infty)$. We conclude that

$$T = \begin{cases} L(\tilde{\theta}) - L(\theta_0), & \text{if } \tilde{\theta} \leq \theta_0, \\ \max\{L(\theta_0), L(\theta_1)\} - L(\tilde{\theta}), & \text{if } \theta_0 < \tilde{\theta} < \theta_1, \\ L(\tilde{\theta}) - L(\theta_1), & \text{if } \tilde{\theta} \geq \theta_1. \end{cases}$$

Note that if $\theta_0 < \tilde{\theta} < \theta_1$ then $T < 0$. Thus, a random set $(T > t_\alpha)$ for positive t_α is equal to

$$B_\alpha(\tilde{\theta}) = \left\{ (L(\tilde{\theta}, \theta_0) > t_\alpha) \cap (\tilde{\theta} \leq \theta_0) \right\} \cup \left\{ (L(\tilde{\theta}, \theta_1) > t_\alpha) \cap (\tilde{\theta} \geq \theta_1) \right\}.$$

Below we aim to show that there exist some positive numbers t_α^+ and t_α^- such that

$$\left\{ (L(\tilde{\theta}, \theta_0) > t_\alpha) \cap (\tilde{\theta} \leq \theta_0) \right\} = \left(\tilde{\theta} < \theta_0 - t_\alpha^- \right), \quad (6.9)$$

$$\left\{ (L(\tilde{\theta}, \theta_1) > t_\alpha) \cap (\tilde{\theta} \geq \theta_1) \right\} = \left(\tilde{\theta} > \theta_1 + t_\alpha^+ \right). \quad (6.10)$$

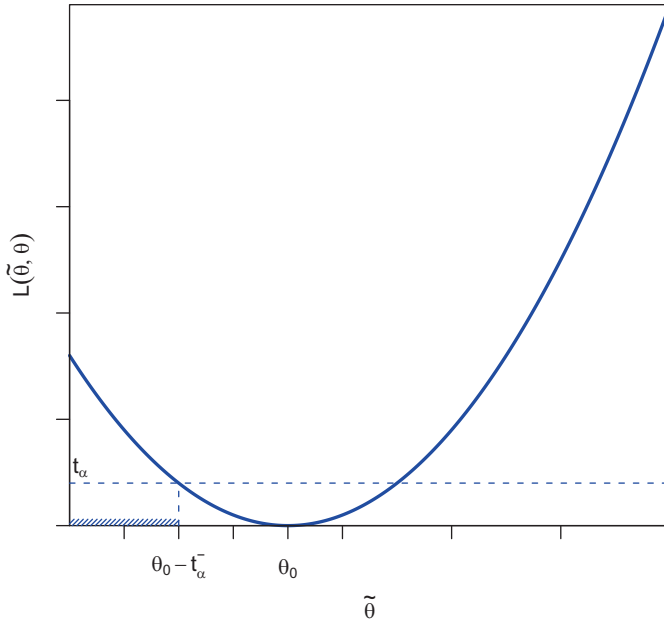


Fig. 6.5 The plot of $\tilde{\theta} < \theta_0 - t_\alpha^-$.  MSEEX0711

Let us concentrate on the first equality. Note that $L(\tilde{\theta}, \theta_0) = n\mathcal{K}(\tilde{\theta}, \theta_0)$. For an exponential family, the Kullback-Leibler divergence $\mathcal{K}(\theta, \theta_0)$ is a monotone increasing continuous function with respect to the first argument:

$$\frac{\partial \mathcal{K}(\theta, \theta_0)}{\partial \theta} = \frac{\partial}{\partial \theta} [\theta \{C(\theta) - C(\theta_0)\} - \{B(\theta) - B(\theta_0)\}] = C(\theta) - C(\theta_0) > 0,$$

because $C'(\theta) = I(\theta) > 0$.

Thus, the equality (6.9) is true, and the form of (6.7) of the LR test is proved. An illustration is given in the Fig. 6.5.

The values t_α^- and t_α^+ should be selected to ensure the condition

$$\sup_{\theta_0 < \theta < \theta_1} \mathbb{E}_\theta \phi = \alpha.$$

The next exercise suggests a way to select these values.

(ii) We aim to show that for any $\theta_0 < \theta < \theta_1$

$$\mathbb{P}_\theta(\tilde{\theta} < \theta_0 - t_\alpha^-) \leq \mathbb{P}_{\theta_0}(\tilde{\theta} < \theta_0 - t_\alpha^-), \quad (6.11)$$

$$\mathbb{P}_\theta(\tilde{\theta} > \theta_1 + t_\alpha^+) \leq \mathbb{P}_{\theta_1}(\tilde{\theta} > \theta_1 + t_\alpha^+). \quad (6.12)$$

The second inequality is proved in Lemma 8.4.5. The first one can be proved analogously. In fact, for any $\mathbb{P}_\theta$ -measurable set A

$$\mathbb{P}_\theta(A) = \int \mathbf{1}(A) d\mathbb{P}_\theta = \int \mathbf{1}(A) \frac{d\mathbb{P}_\theta}{d\mathbb{P}_{\theta_0}} d\mathbb{P}_{\theta_0} = \mathbb{E}_{\theta_0} \exp\{L(\theta, \theta_0)\} \mathbf{1}(A).$$

Function $L(\theta)$ monotone decrease on $\{\theta : \theta > \tilde{\theta}\}$ (it was shown above). Thus, $L(\theta, \theta_0) = L(\theta) - L(\theta_0) < 0$ for any θ, θ_0 such that $\tilde{\theta} < \theta_0 < \theta$, and

$$\mathbb{P}_\theta(\tilde{\theta} < \theta_0 - t_\alpha^-) = \mathbb{E}_{\theta_0} \exp\{L(\theta, \theta_0)\} \mathbf{1}(\tilde{\theta} < \theta_0 - t_\alpha^-) < \mathbb{P}_{\theta_0}(\tilde{\theta} < \theta_0 - t_\alpha^-),$$

and the second inequality (6.11) is proved. The observation

$$[c] \sup_{\theta_0 < \theta < \theta_1} \mathbb{P}_\theta(\phi = 1) \leq \sup_{\theta_0 < \theta < \theta_1} \mathbb{P}_\theta(\tilde{\theta} < \theta_0 - t_\alpha^-) + \sup_{\theta_0 < \theta < \theta_1} \mathbb{P}_\theta(\tilde{\theta} > \theta_1 + t_\alpha^+) \quad (6.13)$$

$$= \mathbb{P}_{\theta_0}(\tilde{\theta} < \theta_0 - t_\alpha^-) + \mathbb{P}_{\theta_1}(\tilde{\theta} > \theta_1 + t_\alpha^+) \quad (6.14)$$

completes the proof.

Exercise 6.10. Consider the time series of weekly DAX returns from Jan 1, 2000 to Dec. 31, 2011. Does the volatility remain constant during this sample period? To test this hypothesis, divide the 12 years data into 3 periods:

- Period 1: Jan. 1, 2000 to Dec. 31, 2003;
- Period 2: Jan. 1, 2004 to Dec. 31, 2007;
- Period 3: Jan. 1, 2008 to Dec. 31, 2011.

Denote the variance of the DAX return in period i as σ_i^2 . Please do the following hypothesis tests:

1. $H_0 : \sigma_1^2 = \sigma_2^2$.
2. $H_0 : \sigma_2^2 = \sigma_3^2$.
3. $H_0 : \sigma_1^2 = \sigma_3^2$.

First we take a look at the data as Fig.6.6. This 11 years DAX return has large volatility clusters between 2000 and 2001, 2003–2004 and 2009–2010. We compute the standard deviations for the three periods: first period 0.0431, second period 0.0213 and third period 0.0401. We expect that the volatility (variance) is nonstationary over time.

We test the three null hypotheses by applying *F-test* on this return time series. For the first hypothesis, the p-value is less than 0.0001, and therefore we reject the first hypothesis. It is similar for the second hypothesis. The third test has p-value 0.2962. This suggests that we cannot reject that the volatility of the period 1 is equal to that of the period 3.

Exercise 6.11. Let $X = \{X_i\}_{i=1}^n$ be i.i.d. $B(\theta)$. Consider the hypothesis $H_0 : \theta_0 = 1/2$ against the alternative $H_1 : \theta_1 = 1/3$. Let for simplicity $n = 2$.

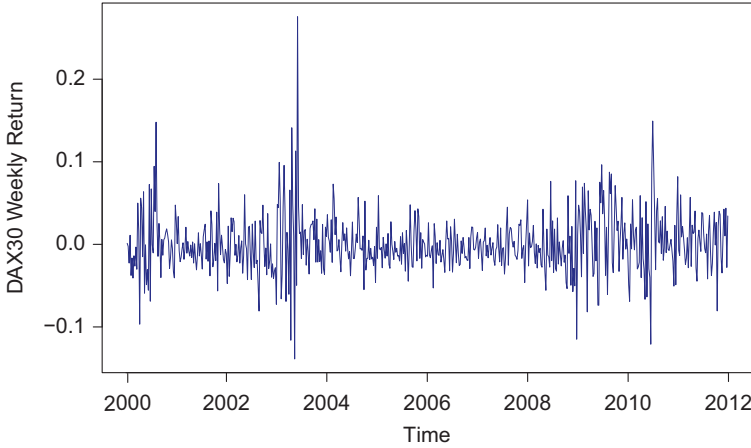


Fig. 6.6 The plot of DAX returns from 20,000,103 to 20,111,227.  MSEDAXre

- (i) Give the explicit formula for the Neyman-Pearson test with the probability of the error of the first kind equal to $\alpha = 0.25$.
- (ii) Provide an example of the weighted sum of the errors of the first and second kind that is minimized in the class of all possible tests by the Neyman-Pearson test.
- (iii) Over which set of tests the Neyman-Pearson test has the largest power?
- (i) For one observation X_i from the Bernoulli sequence,

$$\mathbb{P}_\theta(X_i = x) = (1 - \theta)^{1-x} \theta^x, \quad x \in \{0, 1\}, \quad i = 1, \dots, n.$$

Therefore, with $Z(x)$ being the likelihood ratio:

$$\begin{aligned} \log Z(X) &= \log \frac{\prod_{i=1}^n \mathbb{P}_{\theta_1}(X = x_i)}{\prod_{i=1}^n \mathbb{P}_{\theta_0}(X = x_i)} \\ &= \sum_{i=1}^n \log \left\{ (1 - \theta_1)^{1-x_i} \theta_1^{x_i} \right\} - \sum_{i=1}^n \log \left\{ (1 - \theta_0)^{1-x_i} \theta_0^{x_i} \right\} \\ &= \sum_{i=1}^n x_i \log \frac{\theta_1}{\theta_0} + \sum_{i=1}^n (1 - x_i) \log \frac{1 - \theta_1}{1 - \theta_0} \end{aligned} \quad (6.15)$$

Substituting the values for θ_0 , θ_1 , n , and considering all possible values of X_1 and X_2 , we conclude that $\log Z(X)$ has the following distribution under the hypothesis H_0 (from (6.15))

$$\log Z(X) = \begin{cases} 2 \log(2) - 2 \log(3), & \text{with probability } 1/4 \\ 3 \log(2) - 2 \log(3), & \text{with probability } 1/2 \\ 4 \log(2) - 2 \log(3), & \text{with probability } 1/4. \end{cases}$$

So, the equation

$$\mathbb{P}_{\theta_0} \{ \log Z(X) \geq t_\alpha \} = \alpha$$

has for $\alpha = 1/4$ a solution $t_\alpha = 4 \log(2) - 2 \log(3)$. Since, the Neyman-Pearson test with the error of the first kind equal to $1/4$ has the form $\phi^* \stackrel{\text{def}}{=} \mathbb{I} \{ \log Z(X) \geq 4 \log(2) - 2 \log(3) \}$.

- (ii) By Theorem 6.2.1 (Spokoiny and Dickhaus, 2014), the Neyman-Pearson test minimizes the sum

$$\rho_0 \mathbb{E}_{\theta_0} \phi + \rho_1 \mathbb{E}_{\theta_1} (1 - \phi)$$

over all possible (randomized) tests ϕ , if $t_\alpha = \rho_0 / \rho_1$. Hence, the Neyman-Pearson test ϕ^* minimizes the sum for all pairs of coefficients in the form $(\rho_0, \rho_1) = (a, a / \{4 \log(2) - 2 \log(3)\})$, $a > 0$.

- (iii) The answer for this question follows directly from Theorem 6.2.2 (Spokoiny and Dickhaus, 2014): the Neyman-Pearson test has the largest power over all tests under the level constraint $\mathbb{E}_{\theta_0} \phi \leq 1/4$.

Exercise 6.12. Let $X = \{X_i\}_{i=1}^n$ be a Bernoulli sequence of zeros and ones with a probability of success equal to θ . Consider two functions of the observations X :

$$T^{(1)} \stackrel{\text{def}}{=} \min\{k = 1, \dots, n : X_k = 1\}; \quad T^{(2)} \stackrel{\text{def}}{=} \sum_{i=1}^n X_i.$$

- (i) Consider the simple hypothesis $H_0 : \theta = \theta_0$ against the simple alternative $H_1 : \theta = \theta_1$. Construct the tests (of a fixed level α) in the form $\mathbb{I} \{ T^{(j)} \geq t_\alpha^{(j)} \}$ with some $t_\alpha^{(j)}$ ($j = 1, 2$).
- (ii) Find the power of $T^{(1)}$ and $T^{(2)}$. Check empirically that the power of the Neyman-Pearson test is larger than the power of these tests.
- (i) First note that for any natural m ,

$$\begin{aligned} \mathbb{P}_\theta \{ T^{(1)} \geq m \} &= \mathbb{P}_\theta \{ X_1 = 0, X_2 = 0, \dots, X_{m-1} = 0 \} \\ &= (1 - \theta)^{m-1}. \end{aligned} \tag{6.16}$$

Since the critical value $t_\alpha^{(j)}$ is the solution of the equation

$$\mathbb{P}_{\theta_0} \{ T^{(1)} \geq m \} \leq \alpha,$$

let $\lceil a \rceil$ denote the smallest integer greater than or equal to a , it is equal to

$$t_{\alpha}^{(1)} = \lceil \log_{1-\theta_0}(\alpha) + 1 \rceil.$$

The critical value $t_{\alpha}^{(2)}$ for the test statistic $T^{(2)}$ is equal to the solution of the equation

$$\mathbb{P}_{\theta_0} \{T^{(2)} \geq m\} = \sum_{i=m}^n C_n^m \theta_0^i (1 - \theta_0)^{n-i} = \alpha$$

w.r.t. m .

(ii) According to the formula (6.16), the power for the first test is equal to

$$W^{(1)} = \mathbb{P}_{\theta_1} \{T^{(1)} \geq t_{\alpha}^{(1)}\} = (1 - \theta_1)^{\lceil \log_{1-\theta_0}(\alpha) + 1 \rceil}.$$

The power for the second test can be found from the formula

$$W^{(2)} = \mathbb{P}_{\theta_1} \{T^{(2)} > t_{\alpha}^{(2)}\} = \sum_{i=m}^n C_n^{t_{\alpha}^{(2)}} \theta_1^i (1 - \theta_1)^{n-i} = \alpha.$$

Exercise 6.13. Let $X = \{X_i\}_{i=1}^n$ be an i.i.d. sample from a model of Gaussian shift $\mathcal{N}(\theta, \sigma^2)$. Consider three hypothesis testing problems:

- (i) σ is known; the aim is to test the hypothesis $H_{\theta}^{(0)} : \theta = \theta_0$ against the alternative $H_{\theta}^{(1)} : \theta = \theta_1$, where $\theta_1 \neq \theta_0$;
- (ii) θ is known; the aim is to test the hypothesis $H_{\sigma}^{(0)} : \sigma = \sigma_0$ against the alternative $H_{\sigma}^{(1)} : \sigma = \sigma_1$, where $\sigma_1 \neq \sigma_0$;
- (iii) Both σ and θ are unknown; the aim is to test the hypothesis $H_{\theta, \sigma}^{(0)} : \theta = \theta_0, \sigma = \sigma_0$ against the alternative $H_{\theta, \sigma}^{(1)} : \theta = \theta_1, \sigma = \sigma_1$, where $\theta_1 \neq \theta_0, \sigma_1 \neq \sigma_0$.

Describe the likelihood ratio test for the first and the second situation. Why it is difficult to find the closed form of the likelihood ratio test in the third case?

- (i) The first situation is described in Chap. 6.3.1 from the book by Spokoiny and Dickhaus (2014). The likelihood ratio is equal to

$$\begin{aligned} T^{(1)} &= L(X, \theta_1, \sigma) - L(X, \theta_0, \sigma) \\ &= \sigma^{-2} \{(S - n\theta_0)(\theta_1 - \theta_0) - n(\theta_1 - \theta_0)^2/2\}, \end{aligned} \quad (6.17)$$

where $S = \sum_{i=1}^n X_i$. The likelihood ratio test has the form:

$$\phi^{(1)} = \mathbb{I} \{T^{(1)} \geq \lambda_{\alpha}^{(1)}\},$$

where the critical value $\hat{z}_\alpha^{(1)}$ can be selected from:

$$\mathbb{P}_{\theta_0} \{T^{(1)} \geq \hat{z}_\alpha^{(1)}\} = \alpha. \quad (6.18)$$

Since the sum $S - n\theta_0$ has under H_0 a normal distribution $\mathcal{N}(0, n\sigma^2)$, we conclude that (6.18) can be rewritten as

$$\mathbb{P}_{\theta_0} \left\{ \xi \geq \frac{1}{(\theta_1 - \theta_0) \sigma \sqrt{n}} \{ \sigma^2 \hat{z}_\alpha^{(1)} + n(\theta_1 - \theta_0)^2 / 2 \} \right\} = \alpha,$$

where $\xi = (S - n\theta_0) / \sqrt{n\sigma^2}$ has a standard normal distribution (here we assume for simplicity that $\theta_1 > \theta_0$). Denote an $(1 - \alpha)$ -quantile of the standard normal law by $z_{1-\alpha}$, i.e., $\mathbb{P}(\xi \geq z_{1-\alpha}) = \alpha$. Therefore the critical value can be found from the equation

$$\frac{1}{(\theta_1 - \theta_0) \sigma \sqrt{n}} \{ \sigma^2 \hat{z}_\alpha^{(1)} + n(\theta_1 - \theta_0)^2 / 2 \} = z_{1-\alpha}. \quad (6.19)$$

Finally, we conclude that the likelihood ratio test has the form

$$\phi^{(1)} = \mathbb{I} \left\{ T^{(1)} \geq \sigma^{-2} [z_\alpha \sigma \sqrt{n} (\theta_1 - \theta_0) - n(\theta_1 - \theta_0)^2 / 2] \right\}.$$

(ii) In the second case, we follow the same lines. The likelihood ratio has the form:

$$\begin{aligned} T^{(2)} &= L(X, \theta, \sigma_1) - L(X, \theta, \sigma_0) \\ &= n \log \left(\frac{\sigma_0}{\sigma_1} \right) - \frac{1}{2} \left(\frac{1}{\sigma_1^2} - \frac{1}{\sigma_0^2} \right) \sum_{i=1}^n (X_i - \theta)^2, \end{aligned} \quad (6.20)$$

and the corresponding test can be found as

$$\phi^{(2)} = \mathbb{I} \{ T^{(2)} \geq \hat{z}_\alpha^{(2)} \}.$$

Taking into account that under the hypothesis H_σ^0 , the random variable $\sigma_0^{-2} \sum_{i=1}^n (X_i - \theta)^2$ has a χ -squared distribution with n degrees of freedom, denoting by $w_{1-\alpha}$ the $(1 - \alpha)$ -quantile of this distribution and assuming (for simplicity) that $\sigma_1 > \sigma_2$, we arrive at the following expression for the critical value $\hat{z}_\alpha^{(2)}$:

$$\hat{z}_\alpha^{(2)} = n \log \left(\frac{\sigma_0}{\sigma_1} \right) - \frac{1}{2} \left(\frac{\sigma_0^2}{\sigma_1^2} - 1 \right) w_{1-\alpha}. \quad (6.21)$$

(iii) The likelihood ration in this case is equal to

$$T^{(3)} \stackrel{\text{def}}{=} n \log \left(\frac{\sigma_0}{\sigma_1} \right) - \frac{1}{2\sigma_1^2} \sum_{i=1}^n (X_i - \theta_1)^2 + \frac{1}{2\sigma_0^2} \sum_{i=1}^n (X_i - \theta_0)^2.$$

The likelihood ratio test cannot be written in closed form, because the distribution of $T^{(3)}$ and therefore the quantiles of $T^{(3)}$ are hardly ever known.

Exercise 6.14. Let $X = \{X_i\}_{i=1}^n$ be an i.i.d. sample from a model of Gaussian shift $\mathcal{N}(\theta, \sigma^2)$. Consider the hypothesis $H_0 : \sigma = \sigma_0$ against the alternative $H_1 : \sigma \neq \sigma_0$ if

- (i) θ is known;
- (ii) θ is unknown.

Describe the likelihood-ratio tests in both situations.

- (i) The test statistic is equal to

$$\begin{aligned} T &\stackrel{\text{def}}{=} \sup_{\sigma \neq \sigma_0} L(X, \theta, \sigma) - L(X, \theta, \sigma_0) = L(X, \theta, \tilde{\sigma}) - L(X, \theta, \sigma_0) \\ &= n \log \left(\frac{\sigma_0}{\tilde{\sigma}} \right) - \frac{1}{2} \left(\frac{1}{\tilde{\sigma}^2} - \frac{1}{\sigma_0^2} \right) \sum_{i=1}^n (X_i - \theta)^2, \end{aligned}$$

where $\tilde{\sigma}^2 \stackrel{\text{def}}{=} n^{-1} \sum_{i=1}^n (X_i - \theta)^2$. In order to find the critical value for the test statistic T , note that

- (a) since σ is a natural parameter for this model (see Exercise 2.19),

$$T = n\mathcal{K}(\tilde{\sigma}, \sigma_0);$$

- (b) Lemma 6.4.1 (Spokoiny and Dickhaus, 2014) yields that for every ζ there are two values ζ_α^- and ζ_α^+ such that

$$\{\sigma : \mathcal{K}(\sigma, \sigma_0) < \zeta\} = \{\sigma : \sigma_0 - \zeta_\alpha^- < \sigma < \sigma_0 + \zeta_\alpha^+\}. \quad (6.22)$$

From here it follows that

$$\mathbb{I}(T \geq t_\alpha) = \mathbb{I}(\mathcal{K}(\tilde{\sigma}, \sigma_0) \geq t_\alpha/n) = \mathbb{I}(\tilde{\sigma} \leq \sigma_0 - t_{\alpha,n}^-) + \mathbb{I}(\tilde{\sigma} \geq \sigma_0 + t_{\alpha,n}^+),$$

where the values $t_{\alpha,n}^-$ and $t_{\alpha,n}^+$, which depend on α and n , can be found from the level condition:

$$\mathbb{P}_{\sigma_0}(T \geq t_\alpha) = \mathbb{P}_{\sigma_0}(\tilde{\sigma} \leq \sigma_0 - t_{\alpha,n}^-) + \mathbb{P}_{\sigma_0}(\tilde{\sigma} \geq \sigma_0 + t_{\alpha,n}^+) = \alpha. \quad (6.23)$$

Taking into account that $n\tilde{\sigma}^2/\sigma_0^2$ has a chi-squared distribution with n degrees of freedom (under the hypothesis H_0), we conclude that the values

$$t_{\alpha,n}^- = \sigma_0 \left(1 - \sqrt{\frac{z_{\alpha/2}}{n}} \right), \quad t_{\alpha,n}^+ = \sigma_0 \left(\sqrt{\frac{z_{1-\alpha/2}}{n}} - 1 \right) \quad (6.24)$$

satisfy (6.23), where by z_q we denote the q -quantile of the chi-square distribution with n degrees of freedom.

- (ii) The second case can be viewed as a mirror situation to the paragraph 6.3.3 (Spokoiny and Dickhaus, 2014), where it is discussed the procedure for testing the mean when the variance is unknown. In our case,

$$T^* = \sup_{\theta, \sigma} L(\theta, \sigma) - \sup_{\theta} L(\theta, \sigma_0) = L(\tilde{\theta}, \tilde{\sigma}^*) - L(\tilde{\theta}, \sigma_0),$$

where

$$\tilde{\theta} = n^{-1} \sum_{i=1}^n X_i \quad \text{and} \quad \tilde{\sigma}^* = \sqrt{n^{-1} \sum_{i=1}^n (X_i - \tilde{\theta})^2}.$$

Therefore,

$$T^* = n \log \left(\frac{\sigma_0}{\tilde{\sigma}^*} \right) - \frac{1}{2} \left(\frac{1}{\tilde{\sigma}^{*2}} - \frac{1}{\sigma_0^2} \right) \sum_{i=1}^n (X_i - \tilde{\theta})^2 = n\mathcal{K}(\tilde{\sigma}^*, \sigma_0),$$

and the arguments (a) and (b) from (i) can be applied to this situation also. A unique difference is that the estimate $n\tilde{\sigma}^{*2}/\sigma_0^2$ has a chi-square distribution with $(n-1)$ degrees of freedom. This leads to the critical values as in (6.27), where by z is denoted the quantiles of the chi-square distribution with $n-1$ degrees of freedom.

Exercise 6.15. (This exercise is motivated by Dudewicz and Mishra (1988)) Let $X = \{X_i\}_{i=1}^n$ be an i.i.d. sample from a model of Gaussian shift $\mathcal{N}(\theta, \sigma^2)$, where θ and σ are both unknown; the parameter of interest is θ . Consider the hypothesis $H_0 : \theta = \theta_0$ against the alternative $H_1 : \theta = \theta_1$, where $\theta_1 > \theta_0$. Prove that no test of level α has power larger than α .

First note that the Neyman-Pearson test for a known σ can be written as

$$\phi(X) = \mathbb{I} \{ \bar{X} \geq t_\alpha \} = \mathbb{I} \{ \bar{X} \geq \theta_0 + \sigma z_{1-\alpha} / \sqrt{n} \},$$

see Exercise 6.13 (i). So, this test coincides with the test from Exercise 6.1 (i), and therefore the power of ϕ is equal to

$$W(\theta_1) = 1 - \Phi \left\{ z_{1-\alpha} - \frac{\sqrt{n}(\theta_0 - \theta_1)}{\sigma} \right\},$$

see Exercise 6.1 (ii). This yields that

$$W(\theta_1) = 1 - \Phi \left\{ \Phi^{-1}(1 - \alpha) - \frac{\sqrt{n}(\theta_0 - \theta_1)}{\sigma} \right\} > \alpha.$$

Turning to the case of unknown σ , note that the power of any test of level α cannot exceed the power of the Neymann-Pearson test (see Theorem 6.2.2 from Spokoiny and Dickhaus, 2014). The remark $\inf_{\sigma} W(\theta_1) = \alpha$ completes the proof.

Exercise 6.16 (This exercise is motivated by Dudewicz and Mishra (1988)). In the setup of the previous exercise, assume that the number of observations n can be taken large enough. Consider the following two-stage procedure.

The first step. Fix some $m > 1$ and estimate the mean and the variance of the sample $\{X_i\}_{i=1}^m$ by

$$\bar{X}_m = \frac{1}{m} \sum_{i=1}^m X_i, \quad \hat{\sigma}_m^2 = \frac{1}{m} \sum_{i=1}^m (X_i - \bar{X}_m)^2.$$

The second step. Fix some $\lambda > 1$, calculate

$$n = n(\hat{\sigma}_m) \stackrel{\text{def}}{=} \max \{m + 1, [\lambda \hat{\sigma}_m^2]\} \quad (6.25)$$

and estimate the mean of the subsample $\{X_i\}_{i=m+1}^n$ by

$$\bar{X}_{n-m} = \frac{1}{n-m} \sum_{i=m+1}^n X_i.$$

Next, calculate the weighted sum of the means

$$\tilde{X}_{m,n} \stackrel{\text{def}}{=} w \bar{X}_m + (1 - w) \bar{X}_{n-m},$$

where

$$w = w(\hat{\sigma}_m) \stackrel{\text{def}}{=} \frac{m}{n} \left\{ 1 + \sqrt{1 - \frac{m}{n} \left(1 - \frac{n-m}{\lambda \hat{\sigma}_m^2} \right)} \right\}. \quad (6.26)$$

- (i) Show that under H_0 the random value $\sqrt{\lambda} (\tilde{X}_{m,n} - \theta_0)$ has t -distribution with $m - 1$ degrees of freedom.
- (ii) For any fixed $\alpha, \beta \in (0, 1)$, find the values t_α and λ such that the test $\mathbb{I} \{ \tilde{X}_{m,n} \geq t_\alpha \}$ has level α and power β .
- (i) Consider the random variable

$$U = \frac{\sum_{i=1}^n a_i(\hat{\sigma}_m) X_i - \theta_0}{\hat{\sigma}_m \sqrt{\sum_{i=1}^n \{a_i(\hat{\sigma}_m)\}^2}},$$

where $a_1(\hat{\sigma}_m) = \dots = a_m(\hat{\sigma}_m) = w(\hat{\sigma}_m)$ and $a_{m+1}(\hat{\sigma}_m) = \dots = a_n(\hat{\sigma}_m) = 1 - w(\hat{\sigma}_m)$. By direct calculation, it follows that the choice of w in (6.26) guarantees

$$\hat{\sigma}_m \sqrt{\sum_{i=1}^n (a_i(\hat{\sigma}_m))^2} = \hat{\sigma}_m \sqrt{m \{w(\hat{\sigma}_m)\}^2 + (n - m) \{1 - w(\hat{\sigma}_m)\}^2} = \lambda^{-1}.$$

Recall that the random variable

$$\xi \stackrel{\text{def}}{=} m \frac{\hat{\sigma}_m^2}{\sigma^2}$$

has a chi-squared distribution with $m-1$ degrees of freedom and is independent of $\bar{X}_m$. The first fact yields that the distribution function of U allows the following representation:

$$\mathbb{P}(U \leq u) = \int_0^\infty \mathbb{P} \left\{ \frac{\sum_{i=1}^{n(s)} a_i(s) X_i - \theta_0}{\sigma \sqrt{\sum_{i=1}^{n(s)} (a_i(s))^2}} \leq \frac{\sqrt{v} u}{\sqrt{m}} \mid \xi = v \right\} p_{\chi_{m-1}^2}(v) dv,$$

where by $p_{\chi_{m-1}^2}(v)$ we denote the density function of the chi-squared distribution with $m-1$ degrees of freedom and $s = \hat{\sigma}_m = \sqrt{v}\sigma/\sqrt{m}$. Since $\bar{X}_m$ and ξ are independent, the sum

$$\sum_{i=1}^{n(s)} a_i(s) X_i = w(s)m\bar{X}_m + (1 - w(s)) \sum_{i=m+1}^n X_i$$

is also independent of ξ . This gives

$$\begin{aligned} \mathbb{P}(U \leq u) &= \int_0^\infty \mathbb{P} \left\{ \frac{\sum_{i=1}^{n(s)} a_i(s) X_i - \theta_0}{\sigma \sqrt{\sum_{i=1}^{n(s)} \{a_i(s)\}^2}} \leq \frac{\sqrt{v} u}{\sqrt{m}} \right\} p_{\chi_{m-1}^2}(v) dv \\ &= \int_0^\infty \Phi \left(\frac{\sqrt{v} u}{\sqrt{m}} \right) p_{\chi_{m-1}^2}(v) dv. \end{aligned}$$

The calculation of the last integral is straightforward and remains to the reader.

(ii) The error of the first kind is equal to

$$\alpha = \mathbb{P}_{\theta_0}(\tilde{X}_{m,n} \geq t_\alpha) = \mathbb{P}_{\theta_0} \left\{ \sqrt{\lambda} (\tilde{X}_{m,n} - \theta_0) \geq \sqrt{\lambda} (t_\alpha - \theta_0) \right\}.$$

Therefore,

$$t_\alpha = \frac{z_{1-\alpha}}{\sqrt{\lambda}} + \theta_0, \quad (6.27)$$

where $z_{1-\alpha}$ is a $(1 - \alpha)$ -quantile of the chi-squared distribution with $(n - 1)$ degrees of freedom.

Next, the power of this test is equal to

$$\begin{aligned} \beta &= \mathbb{P}_{\theta_1}(\tilde{X}_{m,n} \geq t_\alpha) = \mathbb{P}_{\theta_1} \left\{ \sqrt{\lambda} (\tilde{X}_{m,n} - \theta_1) \geq \sqrt{\lambda} (t_\alpha - \theta_1) \right\} \\ &= \mathbb{P}_{\theta_1} \left\{ \sqrt{\lambda} (\tilde{X}_{m,n} - \theta_1) \geq \sqrt{\lambda} \left(\frac{z_{1-\alpha}}{\sqrt{\lambda}} + \theta_0 - \theta_1 \right) \right\}. \end{aligned}$$

This yields that

$$\lambda = \left(\frac{z_{1-\beta} - z_{1-\alpha}}{\theta_0 - \theta_1} \right)^2. \quad (6.28)$$

So, the test $\mathbb{I} \{ \tilde{X}_{m,n} \geq t_\alpha \}$, where t_α in the form (6.27) and n is chosen from (6.25) with λ in the form (6.28), has the level α and power β .

Exercise 6.17. (from Pestman and Alberink, 1991) Let $\mathbf{X} = \{X_i\}_{i=1}^m$ and $\mathbf{Y} = \{Y_i\}_{i=1}^n$ be two i.i.d. samples from $\mathcal{N}(\theta_X, \sigma^2)$ and $\mathcal{N}(\theta_Y, \sigma^2)$ respectively, where $\theta_X, \theta_Y, \sigma$ are unknown. Construct the likelihood-ratio test to check the hypothesis $H_0 : \theta_Y - \theta_X = \Delta$ against the alternative $H_1 : \theta_Y - \theta_X \neq \Delta$, where Δ is fixed.

1. By the definition of the Likelihood-ratio test (see Spokoiny and Dickhaus, 2014, Sect. 6.3), the statistic is equal to

$$T \stackrel{\text{def}}{=} I_2 - I_1,$$

where

$$\begin{aligned} I_1 &\stackrel{\text{def}}{=} \sup_{\substack{\theta_X, \theta_Y, \sigma \\ \theta_Y - \theta_X = \Delta}} \{L(\mathbf{X}, \theta_X, \sigma) + L(\mathbf{Y}, \theta_Y, \sigma)\}, \\ I_2 &\stackrel{\text{def}}{=} \sup_{\substack{\theta_X, \theta_Y, \sigma \\ \theta_Y - \theta_X \neq \Delta}} \{L(\mathbf{X}, \theta_X, \sigma) + L(\mathbf{Y}, \theta_Y, \sigma)\}. \end{aligned}$$

It is worth mentioning that one can omit the condition $\theta_Y - \theta_X \neq \Delta$ in the latter supremum.

2. Next, we aim to find the explicit form for the value I_1 . By the Lagrange theorem, the supremum in L_1 can be found by solving the maximization task

$$F(\theta_X, \theta_Y, \sigma, \lambda) \stackrel{\text{def}}{=} L(X, \theta_X, \sigma) + L(Y, \theta_Y, \sigma) + \lambda (\theta_Y - \theta_X - \Delta) \rightarrow \max_{\theta_X, \theta_Y, \sigma, \lambda},$$

where $\lambda \in \mathbb{R}$. Taking into account that

$$F(\theta_X, \theta_Y, \sigma, \lambda) = -\frac{1}{2}(m+n) \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \left\{ \sum_{i=1}^m (X_i - \theta_X)^2 + \sum_{i=1}^n (Y_i - \theta_Y)^2 \right\} + \lambda (\theta_Y - \theta_X - \Delta),$$

we calculate the first derivatives of the function F and consider the equation $\nabla F = \mathbf{0}$:

$$\begin{cases} 0 = \frac{\partial F}{\partial \theta_X} = \frac{1}{\sigma^2} \sum_{i=1}^m (X_i - \theta_X) - \lambda, \\ 0 = \frac{\partial F}{\partial \theta_Y} = \frac{1}{\sigma^2} \sum_{i=1}^n (Y_i - \theta_Y) + \lambda, \\ 0 = \frac{\partial F}{\partial \sigma} = -\frac{m+n}{\sigma} + \frac{1}{\sigma^3} \left\{ \sum_{i=1}^m (X_i - \theta_X)^2 + \sum_{i=1}^n (Y_i - \theta_Y)^2 \right\}, \\ 0 = \frac{\partial F}{\partial \lambda} = \theta_Y - \theta_X - \Delta. \end{cases}$$

From the first two equations, it follows that

$$\sum_{i=1}^m (X_i - \theta_X) + \sum_{i=1}^n (Y_i - \theta_Y) = 0.$$

Together with the fourth equation, this gives

$$\hat{\theta}_X = \frac{m\bar{X} + n\bar{Y} - n\Delta}{m+n}, \quad \hat{\theta}_Y = \frac{m\bar{X} + n\bar{Y} + m\Delta}{m+n}. \quad (6.29)$$

The third equation yields that

$$\hat{\sigma}^2 = \frac{1}{m+n} \left\{ \sum_{i=1}^m (X_i - \hat{\theta}_X)^2 + \sum_{i=1}^n (Y_i - \hat{\theta}_Y)^2 \right\}. \quad (6.30)$$

Substituting (6.29) into the last expression, we get

$$\hat{\sigma}^2 = \frac{1}{m+n} \left\{ \sum_{i=1}^m \left(X_i - \frac{m\bar{X} + n\bar{Y} - n\Delta}{m+n} \right)^2 + \sum_{i=1}^n \left(Y_i - \frac{m\bar{X} + n\bar{Y} + m\Delta}{m+n} \right)^2 \right\}. \quad (6.31)$$

This expression can be simplified. In fact,

$$\begin{aligned} \sum_{i=1}^m \left(X_i - \frac{m\bar{X} + n\bar{Y} - n\Delta}{m+n} \right)^2 &= \sum_{i=1}^m \left\{ (X_i - \bar{X}) + \left(\bar{X} - \frac{m\bar{X} + n\bar{Y} - n\Delta}{m+n} \right) \right\}^2 \\ &= \sum_{i=1}^m (X_i - \bar{X})^2 + mn^2 \left(\frac{\bar{Y} - \bar{X} - \Delta}{m+n} \right)^2, \\ \sum_{i=1}^n \left(Y_i - \frac{m\bar{X} + n\bar{Y} + m\Delta}{m+n} \right)^2 &= \sum_{i=1}^n (Y_i - \bar{Y})^2 + m^2n \left(\frac{\bar{Y} - \bar{X} - \Delta}{m+n} \right)^2. \end{aligned}$$

Substituting the last expression in (6.31), we arrive at

$$\hat{\sigma}^2 = \frac{m\hat{\sigma}_X^2 + n\hat{\sigma}_Y^2}{m+n} + mn \left(\frac{\bar{Y} - \bar{X} - \Delta}{m+n} \right)^2, \quad (6.32)$$

where by $\hat{\sigma}_X^2$ and $\hat{\sigma}_Y^2$ we denote the estimated variances for the first and second samples correspondingly. Finally, using the representations (6.30), we conclude that

$$\begin{aligned} I_1 &= -\frac{1}{2}(m+n) \log(2\pi\hat{\sigma}^2) - \frac{1}{2\hat{\sigma}^2} \left\{ \sum_{i=1}^m (X_i - \hat{\theta}_X)^2 + \sum_{i=1}^n (Y_i - \hat{\theta}_Y)^2 \right\} \\ &= -\frac{1}{2}(m+n) \log(2\pi e\hat{\sigma}^2), \end{aligned}$$

where $\hat{\sigma}^2$ is given by (6.32).

3. The next step is to maximize

$$\begin{aligned} L(X, \theta_X, \sigma) + L(Y, \theta_Y, \sigma) \\ = -\frac{1}{2}(m+n) \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \left\{ \sum_{i=1}^m (X_i - \theta_X)^2 + \sum_{i=1}^n (Y_i - \theta_Y)^2 \right\} \end{aligned}$$

with respect to $(\theta_X, \theta_Y, \sigma) \in \mathbb{R}^2 \times \mathbb{R}_+$. This maximization is straightforward; the maximum is attained at the point

$$(\tilde{\theta}_X, \tilde{\theta}_Y, \tilde{\sigma}^2) \stackrel{\text{def}}{=} (\bar{X}, \bar{Y}, (m\hat{\sigma}_X^2 + n\hat{\sigma}_Y^2)/(m+n))$$

and is equal to

$$I_2 = -\frac{1}{2}(m+n) \log(2\pi e \tilde{\sigma}^2).$$

4. To complete the solution, we note that

$$\begin{aligned} T = I_2 - I_1 &= -\frac{1}{2}(m+n) \log \left(\frac{\tilde{\sigma}^2}{\hat{\sigma}^2} \right) \\ &= \frac{1}{2}(m+n) \log \left(1 + \frac{nm}{n+m} \frac{(\bar{Y} - \bar{X} - \Delta)^2}{m\hat{\sigma}_X^2 + n\hat{\sigma}_Y^2} \right). \end{aligned}$$

Exercise 6.18. Given the S&P 500 index quarterly log returns from Q2 1980 to Q2 2012, which are assumed to be normally distributed with mean θ and standard deviation σ . Consider two hypothesis testing problems:

- (i) $\sigma = 8.03\%$ ($\sigma_{\text{yearly}} = 16\%$) is known; test the null hypothesis $H_\theta^{(0)} : \theta_0 = 1\%$ against the alternative $H_\theta^{(1)} : \theta_1 = 4\%$;
- (ii) $\theta = 1.97\%$ ($\theta_{\text{yearly}} = 8.11\%$) is known; test the null hypothesis $H_\sigma^{(0)} : \sigma_0 = 5\%$ against the alternative $H_\sigma^{(1)} : \sigma_1 = 10\%$;

Perform the likelihood ratio test as given in Exercise 6.13 for the above cases with 5% significance level.

- (i) The time series plot for the S&P 500 index quarterly log returns is shown in Fig. (6.7), where log returns are equal to $\log S_t - \log S_{t-1}$, with S_t is the S&P 500 index at time t . The QQ plot presented in Fig. (6.8) shows that the log-return series is approximately normally distributed.

The likelihood ratio (under normality assumption) is given by Eq. (6.17). Inserting given values $S = \sum_{i=1}^n X_i = 2.52$, $\sigma = 8\%$, $\theta_0 = 1\%$, $\theta_1 = 4\%$, $\alpha = 5\%$ and $n = 128$, we calculate:

$$T^{(1)} = -3.15.$$

The likelihood ratio test for significance level α has the form:

$$\phi^{(1)} = \mathbb{I} \{T^{(1)} \geq \mathfrak{z}_\alpha^{(1)}\},$$

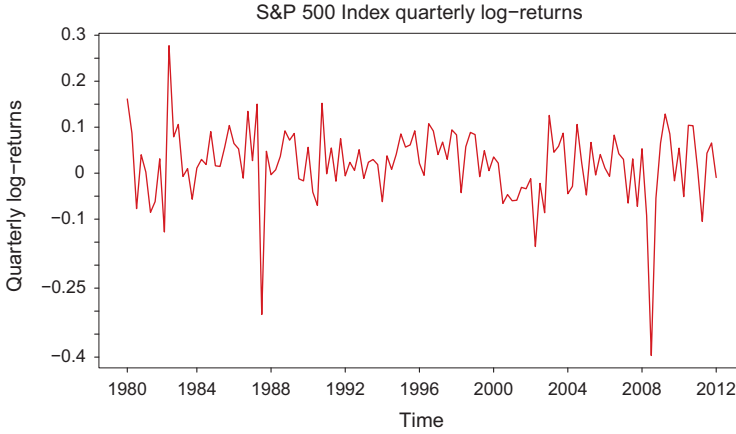


Fig. 6.7 Plot of S&P 500 index quarterly log-returns during the period Q2 1980–Q2 2012.
 MSEspqlogret

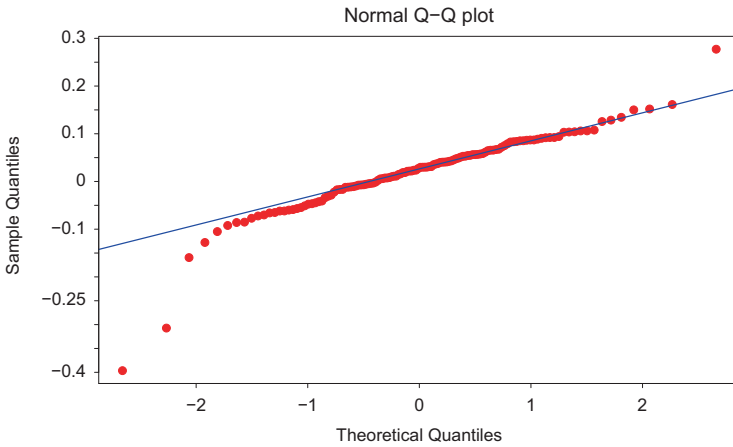


Fig. 6.8 QQ-plot for S&P index quarterly log-returns during the period Q2 1980–Q2 2012.
 MSEqlretqqplot

where the critical value $\hat{z}_{\alpha}^{(1)}$ can be obtained from Eq.(6.19). For given significance level $\alpha = 5\%$, the z-value from the standard normal table is equal to $z_{95\%} = 1.65$. Inserting other inputs, the critical value is calculated equal to $\hat{z}_{5\%}^{(1)} = -1.97$. Since $T^{(1)}$ is not greater than $\hat{z}_{5\%}^{(1)}$, the null hypothesis cannot be not rejected. Thus,

$$\phi^{(1)} = \mathbb{I} \left\{ T^{(1)} \geq \hat{z}_{5\%}^{(1)} \right\} = 0.$$

- (ii) Here we follow the same procedure as in (i), the likelihood ratio is given by Eq. (6.20) is calculated equal to:

$$T^{(2)} = 34.2,$$

where $\theta = 2\%$, $\sigma_0 = 5\%$, $\sigma_1 = 10\%$, $\alpha = 5\%$ and $n = 128$. The corresponding likelihood ratio test is given as:

$$\phi^{(2)} = \mathbb{I} \{ T^{(2)} \geq \mathfrak{z}_{\alpha}^{(2)} \}.$$

The critical value $\mathfrak{z}_{\alpha}^{(1)}$ can be obtained from the Eq. (6.21). Given degrees of freedom $n = 128$, $\alpha = 5\%$ and using χ^2 -squared distribution table $w_{95\%} = 155.4$. We calculate the critical value $\mathfrak{z}_{5\%}^{(2)} = -30.45$. Here $T^{(2)}$ is greater than $\mathfrak{z}_{5\%}^{(2)}$, therefore we reject the null hypothesis,

$$\phi^{(2)} = \mathbb{I} \{ T^{(2)} \geq \mathfrak{z}_{5\%}^{(2)} \} = 1.$$

References

- Dudewicz, E. J., & Mishra, S. N. (1988). *Modern mathematical statistics*. New York: Wiley.
- Pestman, W. R., & Alberink, I. B. (1991). *Mathematical statistics*. Berlin: De Gruyter.
- Spokoiny, V., & Dickhaus, T. (2014). *Basics of modern parametric statistics*. Berlin: Springer.
- Suhov, Y., & Kelbert, M. (2005). *Probability and statistics by example, 1 basic probability and statistics*. New York: Cambridge University Press.

Chapter 7

Testing in Linear Models

מִיִּירָאנִיל מִלְדוּמָב תּוֹרַעֲשָׁה תְּקִידָב

הַיּוֹשֵׁעַ תְּבַצֵּב תְּבַצ

A tong is made from a tong.

Pirkey Avot

Exercise 7.1. Consider the model:

$$Y = f + \varepsilon$$

with the vector of observations Y , response vector f , and vector of mean zero errors $\varepsilon \in \mathbb{R}^n$.

Parametrize the mean of Y as:

$$f = \Psi^\top \theta^*, \quad \theta^* \in \mathbb{R}^p$$

with iid errors $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)^\top$ with covariance matrix I_n . The MLE $\tilde{\theta}$ of θ^* is $\tilde{\theta} = (\Psi\Psi^\top)^{-1}\Psi Y$.

Define the estimated response as:

$$\tilde{f} = \Psi^\top \tilde{\theta}$$

and note that $\tilde{f} = \Pi Y = \Pi(f + \varepsilon)$ where $\Pi = \Psi^\top(\Psi\Psi^\top)^{-1}\Psi$ is a projector into the column space of Ψ . Define $\text{RSS} \stackrel{\text{def}}{=} \|Y - \Psi^\top \tilde{\theta}\|^2$, and note that

$$\begin{aligned} \text{RSS}_0 &\stackrel{\text{def}}{=} \|Y - f_0\|^2 \\ &= \|Y - \Psi^\top \theta^*\|^2 = \text{RSS} + \|\tilde{f} - f_0\|^2 \end{aligned}$$

Estimate σ^2 by:

$$\tilde{\sigma}^2 = \frac{1}{n-p} \text{RSS} = \frac{1}{n-p} \|\mathbf{Y} - \Psi^\top \tilde{\boldsymbol{\theta}}\|^2 \quad (7.1)$$

Show that, $\tilde{\sigma}^2$ is an unbiased, $\sqrt{n}$ consistent estimate of σ^2 :

$$\mathbb{E}\tilde{\sigma}^2 = \sigma^2, \quad \text{Var}\tilde{\sigma}^2 = \frac{2}{n-p}\sigma^4 \quad (7.2)$$

Note that $\sigma^{-2}\|\tilde{\boldsymbol{\epsilon}}\|^2 = \sigma^{-2}\|\mathbf{Y} - \tilde{\mathbf{f}}\|^2 \sim \chi_{n-p}^2$ yielding

$$\begin{aligned} \mathbb{E}\tilde{\sigma}^2 &= (n-p)^{-1} \mathbb{E} \text{RSS} \\ &= (n-p)^{-1} \mathbb{E} \|\mathbf{Y} - \tilde{\mathbf{f}}\|^2 \\ &= (n-p)^{-1} \sigma^2 (n-p) = \sigma^2 \end{aligned}$$

Recall that for $V \sim \chi_d^2$, $\text{Var}(V) = 2d$. Putting $V = \sigma^{-2}\|\mathbf{Y} - \tilde{\mathbf{f}}\|^2$ we see that $\text{Var}(V) = 2(n-p)$ and therefore from (7.1):

$$\text{Var}(\tilde{\sigma}^2) = (n-p)^{-2} \sigma^4 2(n-p) = 2\sigma^4/(n-p).$$

The estimator $\tilde{\sigma}^2$ is $\sqrt{n}$ consistent if $\sqrt{n}(\tilde{\sigma}^2 - \sigma^2) = \mathcal{O}_p(1)$. Using (7.2) one obtains:

$$\mathbb{P}(\sqrt{n}|\tilde{\sigma}^2 - \sigma^2| > \mathfrak{z}) \leq \frac{\text{Var}(\sqrt{n}\tilde{\sigma}^2)}{\mathfrak{z}^2} = \frac{2n\sigma^4}{(n-p)\mathfrak{z}^2}$$

This yields $\sqrt{n}$ consistency by setting $\mathfrak{z} \rightarrow \infty$

Exercise 7.2. Consider the model:

$$\mathbf{Y} = \mathbf{f} + \boldsymbol{\varepsilon}. \quad (7.3)$$

with $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 I_n)$ for an unknown value σ^2 . If $F_{p,n-p}(t_\alpha) = 1 - \alpha$, and $\tilde{\mathfrak{z}}_\alpha = p t_\alpha$, then the test $\tilde{\phi} = \mathbf{1}(\tilde{T} \geq \tilde{\mathfrak{z}}_\alpha)$ where

$$\tilde{T} \stackrel{\text{def}}{=} \frac{1}{2\tilde{\sigma}^2} \|\tilde{\mathbf{f}} - \mathbf{f}_0\|^2 = \frac{(n-p)\|\tilde{\mathbf{f}} - \mathbf{f}_0\|^2}{2\|\mathbf{Y} - \tilde{\mathbf{f}}\|^2} = \frac{\text{RSS}_0 - \text{RSS}}{2\text{RSS}/(n-p)}. \quad (7.4)$$

is an exact level- α test:

$$\mathbb{P}_{\theta_0}(\tilde{\phi} = 1) = \mathbb{P}_{\theta_0}(\tilde{T} \geq \tilde{\mathfrak{z}}_\alpha) = \alpha$$

Observe that the event $\{\tilde{T} \geq \tilde{z}_\alpha\}$ is equivalent to the event $\{p^{-1}\tilde{T} \geq t_\alpha\}$. Note also that $p^{-1}\tilde{T} \sim F_{p, n-p}$ and therefore

$$\mathbb{P}(p^{-1}\tilde{T} \geq t_\alpha) = \alpha.$$

which was to be demonstrated.

Exercise 7.3. Consider the model (7.3) with $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 I_n)$ again with unknown variance σ^2 . Take the critical value $\mathfrak{z}_\alpha$ as $\mathbb{P}(\zeta_p > 2\mathfrak{z}_\alpha) = \alpha$, with $\zeta_p \sim \chi_p^2$ (the known variance case) and define

$$\check{\phi} = \mathbf{1}(\tilde{T} \geq \mathfrak{z}_\alpha).$$

Show that $\check{\phi}$ is an asymptotic level α test:

$$\lim_{n \rightarrow \infty} \mathbb{P}_{\theta_0}(\check{\phi} = 1) = \alpha.$$

We know that from consistency of $\tilde{\sigma}^2 = (n-p)^{-1} \|\mathbf{Y} - \tilde{\mathbf{f}}\|^2$, for all $\delta > 0$

$$\lim_{n \rightarrow \infty} \mathbb{P}\left(\left|\frac{\tilde{\sigma}^2}{\sigma^2} - 1\right| > \varepsilon\right) = 0. \quad (7.5)$$

Define the event:

$$\Omega \stackrel{\text{def}}{=} \left\{ \left| \frac{\tilde{\sigma}^2}{\sigma^2} - 1 \right| < \varepsilon \right\}$$

With the definition of $\tilde{T}$ as in (7.4) we obtain:

$$\begin{aligned} |\mathbb{P}_{\theta_0} \{ \check{\phi} = 1 \} \cap (\Omega \cup \Omega^c) \} - \alpha| &\leq |\mathbb{P}_{\theta_0} \{ \check{\phi} = 1 \} \cap \Omega \} - \alpha| \\ &\quad + \mathbb{P}_{\theta_0} \{ \Omega^c \} \end{aligned} \quad (7.6)$$

$$\mathbb{P}_{\theta_0} \left[\{ \check{\phi} = 1 \} \cap \Omega \right] = \int_{\{ \check{\phi} = 1 \} \cap \Omega} \tilde{T} d\mathbb{P}_{\theta_0} \quad (7.7)$$

Observe that by (7.5) the second term in (7.6) is negligible. We therefore concentrate on the term (7.7).

Note that on Ω , we have with high probability

$$(1 - \varepsilon)\sigma^2 \leq \tilde{\sigma}^2 \leq (1 + \varepsilon)\sigma^2$$

therefore (7.7) can be bounded from above:

$$\int_{\mathfrak{z}\alpha \cap \Omega}^{\infty} \frac{1}{2\mathfrak{z}^2} \|\tilde{\mathbf{f}} - \tilde{\mathbf{f}}_0\|^2 d\mathbb{P}_{\theta_0} \leq \frac{1}{1-\varepsilon} \alpha.$$

and from below by $\frac{1}{1+\varepsilon} \alpha$

Therefore (7.6) lies in the interval

$$\left[\left(1 - \frac{1}{1+\varepsilon} \alpha\right), \left(1 - \frac{1}{1-\varepsilon} \alpha\right) \right]$$

Sending $\varepsilon \rightarrow 0$ we obtain the desired result.

Exercise 7.4. Consider the model (7.3) with $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 I_n)$ with unknown variance σ^2 . Recall the test statistic $\check{\phi}$ from Exercise 7.3. Show that

$$\lim_{n \rightarrow \infty} \sup_f |\mathbb{P}_{\theta_0}(\check{\phi} = 1) - \mathbb{P}_{\theta_0}(\check{\phi} = 1)| = 0. \quad (7.8)$$

Since $\check{\phi}$ is an exact test of level α according to Exercise 7.2, the claim to prove is that

$$\sup_f |\alpha - \mathbb{P}_{\theta_0}(\check{\phi} = 1)| \rightarrow 0.$$

Observe that (7.7) holds independent of f , therefore the claim follows.

Exercise 7.5. Consider the model (7.3) with $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 I_n)$ for an unknown value of σ^2 . Define

$$\tilde{T} = \frac{(n-p) \|\tilde{\mathbf{f}} - \tilde{\mathbf{f}}_0\|^2}{2 \|\mathbf{Y} - \tilde{\mathbf{f}}\|^2} \quad (7.9)$$

as in Chap. 7 of Spokoiny and Dickhaus (2014) where $\tilde{\mathbf{f}} - \tilde{\mathbf{f}}_0 = (\Pi - \Pi_0)\mathbf{Y}$, and Π_0 is the projection on the subspace $\mathcal{L}_0$ spanned by the rows of $\Psi_{\mathbf{y}}$.

It is evident that the numerator of (7.9) equals the rv $\mathfrak{z} = 2(n-p)^{-1} \sigma^2 \zeta_{p-p_0}$, $\zeta_{p-p_0} \sim \chi_{p-p_0}^2$. The denominator is as seen before twice a χ_{n-p}^2 rv.

Adjusting the scaling factor we see that we are actually looking at a ratio of a $\chi_{p-p_0}^2$ and χ_{n-p}^2 rv. This has evidently as $F_{p-p_0, n-p}$ distribution and proves the claim.

Exercise 7.6. Consider a sequence of data generated from

$$f(t) = \theta_1 \cos(\omega_1 t) + \theta_1 \sin(\omega_1 t) + \theta_2 \cos(\omega_2 t) + \theta_2 \sin(\omega_2 t) + \boldsymbol{\varepsilon}_t, \quad (7.10)$$

where $\theta_1, \theta_2, \omega_1$ and ω_2 are constants. $\boldsymbol{\varepsilon}_t \sim \mathcal{N}(0, \sigma^2)$ i.i.d. Suppose we have a data set $\{y_t\}_{t=1}^n$ generated from (7.10). Figure 7.1 illustrates the trajectory of y_t .

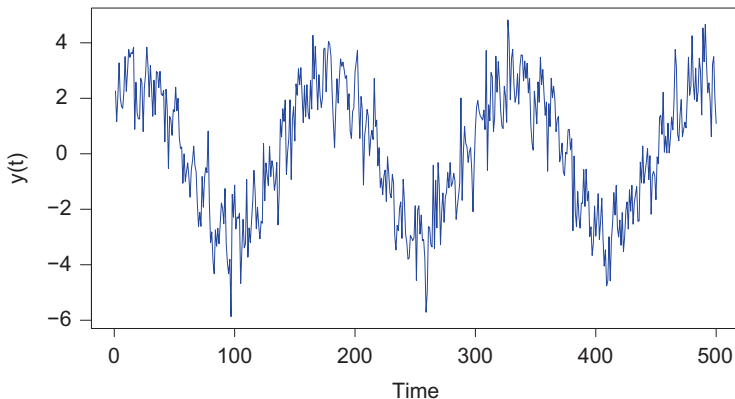


Fig. 7.1 A trajectory of y_t . $\theta_1 = 2$, $\theta_2 = 0.5$, $\omega_1 = 0.04$, $\omega_2 = 0.5$ and $\sigma = 0.8$.  MSE Spectral

1. Taking θ_1, θ_2 as unknown parameters, suggest a linear parametric model for $\{y_t\}_{t=1}^n$ and justify your choice.
2. Suppose ω_1, ω_2 and σ^2 are known, propose a test for the null hypothesis $H_0 : \theta_1 = \theta_2 = 0$.
3. Suppose instead that ω_1, ω_2 are known but σ^2 is unknown, propose a test for the null hypothesis $H_0 : \theta_1 = \theta_2 = 0$.

1. Let $\theta = (\theta_1, \theta_2)^\top$, we suggest the model

$$\begin{aligned} f_\theta(t) &= \theta_1 \cos(\omega_1 t) + \theta_1 \sin(\omega_1 t) + \theta_2 \cos(\omega_2 t) + \theta_2 \sin(\omega_2 t) \\ &= \Psi^\top \theta. \end{aligned}$$

where $\Psi = (\cos(\omega_1 t) + \sin(\omega_1 t), \cos(\omega_2 t) + \sin(\omega_2 t))^\top$. It is clear that $f_\theta(t)$ is in a linear parametric form.

2. Let $\mathbf{Y} = (y_1, y_2, \dots, y_n)^\top$. $H_0 : \theta_1 = \theta_2 = 0$ implies $f_0 = 0$.

$$T = \frac{\|\tilde{f}\|^2}{2\sigma^2} = \frac{\tilde{\theta}^\top \Psi \Psi \tilde{\theta}}{2\sigma^2},$$

where $\tilde{\theta} = (\Psi \Psi^\top)^{-1} \Psi \mathbf{Y}$, because $\boldsymbol{\varepsilon}_t \sim \mathcal{N}(0, \sigma^2)$ i.i.d. The number of parameters $p = 2$. The test is based on the exact distribution

$$2T \sim \chi_2^2.$$

3. Let $\mathbf{Y}$ and $\tilde{f}$ be defined as in the last subexercise.

$$\tilde{\sigma}^2 = \frac{\|\mathbf{Y} - \tilde{f}\|^2}{n - 2}.$$

The test statistics now becomes

$$\tilde{T} = \frac{1}{2\tilde{\sigma}^2} \|\tilde{f}\|^2 = \frac{(n-2)\|\Psi\tilde{\theta}\|^2}{2\|\mathbf{Y} - \Psi\tilde{\theta}\|^2}.$$

For $p^{-1}\tilde{T}$, we have the F-distribution

$$\frac{\tilde{T}}{p} \sim F_{p, n-p}.$$

Exercise 7.7. Consider the panel data model (Badi, 2008):

$$Y_{it} = \alpha + \Psi_{it}\beta + \varepsilon_{it}, \quad i = 1, \dots, N; \quad t = 1, \dots, T.$$

$$\varepsilon_{it} = \mu_i + u_{it},$$

where μ_i stands for the unobservable individual effect, for instance, the ability. And u_{it} is the remaining disturbances with $u_{it} \sim (0, \sigma_u^2)$. Both fixed effects model and random effects model are associated with the assumption of μ_i . For fixed effects model we assume that μ_i is fixed, while for random effects model μ_i is random, i.e. $\mu_i \sim (0, \sigma_\mu^2)$.

1. For fixed effects the matrix representation of the model is:

$$Y = \alpha \mathbf{1}_{NT} + \Psi^\top \beta + \varepsilon, \quad \varepsilon = G\mu + u,$$

where $G = I_N \otimes \mathbf{1}_T$, $\text{rank}(G) = N$. The fixed effects estimator can be denoted by $\hat{\beta}_F$. Let $\hat{\beta}_F = (\Psi Q \Psi^\top)^{-1} \Psi Q Y$, $Q = I_{NT} - P$, and $P = G(G^\top G)^{-1} G^\top$. Assume that $QG = 0$ and $Q\mathbf{1}_{NT} = 0$, show that $\hat{\beta}_F$ is unbiased.

2. The critical assumption of a random effects model is: strict exogeneity of all regressors. The Hausman test helps us to test this assumption, where $H_0 : \mathbb{E}(\varepsilon|\Psi^\top) = 0$, against the alternative $H_1 : \mathbb{E}(\varepsilon|\Psi^\top) \neq 0$. The fixed effects estimator $\hat{\beta}_F$ is consistent under H_0 and H_1 , but not efficient under H_0 . The random effects estimator can be denoted by $\hat{\beta}_R$ which is efficient, consistent and asymptotically efficient under H_0 , but biased and inconsistent under H_1 . The Hausman test statistic is as follows:

$$m = (\hat{\beta}_R - \hat{\beta}_F)^\top \text{Cov}[\hat{\beta}_R - \hat{\beta}_F]^{-1} (\hat{\beta}_R - \hat{\beta}_F).$$

Check the asymptotic behavior of the test statistic under H_0 .

3. Use wages data which come from the website: <http://www.wiley.com>. Regress $\ln \text{wages}$ on all the regressors by using random effects model and fixed effects model. Then perform the Hausman test, interpret the result.

1.

$$\begin{aligned}
\mathbb{E}(\hat{\beta}_F) &= \mathbb{E}[(\Psi Q \Psi^\top)^{-1} \Psi Q Y] \\
&= \mathbb{E}[(\Psi Q \Psi^\top)^{-1} \Psi Q (\alpha 1_{NT} + \Psi^\top \beta + G\mu + u)] \\
&= (\Psi Q \Psi^\top)^{-1} \Psi Q \Psi^\top \beta + (\Psi Q \Psi^\top)^{-1} \Psi Q \mathbb{E}[u] \\
&= \beta
\end{aligned}$$

2.

$$m \xrightarrow{d} \chi_p^2.$$

and p is the number of elements in $\hat{\beta}$.

3. From the result of Hausman test we can see that p -value is less than $2e^{-16}$ which is statistically significant. Therefore H_0 is rejected, i.e. there is a problem of endogeneity. We should apply fixed effects model.  MSEhausman

Exercise 7.8. Consider the model:

$$Y_i = \Psi_i^\top \theta^* + \sigma_i \eta_i, \quad \eta_i \sim \mathcal{N}(0, 1).$$

where $\Psi_i = (\psi_1(X_i), \dots, \psi_p(X_i))$ and $\theta^* = (\theta_1^*, \dots, \theta_p^*)$ are $p \times 1$ vectors, σ_i is a constant parameter, and $\text{Var}(\sigma_i \eta_i) = \sigma_i^2$. To test the heteroscedasticity of the residuals we can apply the White test which is proposed by Halbert White in 1980. The null hypothesis $H_0: \sigma^2 = \sigma_i^2$, against the alternative hypothesis $H_1: \sigma^2 \neq \sigma_i^2$, for $i = 1, \dots, n$. The procedure of the White test can be stated below:

Assume $p = 2$, our model can be written as:

$$Y_i = \theta_1^* \psi_1(X_i) + \theta_2^* \psi_2(X_i) + \sigma_i \eta_i$$

Then perform the ordinary least square regression. The residuals can be obtained by

$$e_i = Y_i - \hat{\theta}_1 \psi_1(X_i) - \hat{\theta}_2 \psi_2(X_i)$$

Then we regress e_i^2 on the regressors which include the original regressors, the cross-products of the regressors and the squared regressors. This auxiliary regression is as follows:

$$e_i^2 = \gamma_1 \psi_1(X_i) + \gamma_2 \psi_1^2(X_i) + \gamma_3 \psi_2(X_i) + \gamma_4 \psi_2^2(X_i) + \gamma_5 \psi_1(X_i) \psi_2(X_i) + u_i$$

Then the White test statistics is as follows:

$$LM = n \times R^2$$

where R^2 comes from the auxiliary regression and is defined as follows:

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

where SSR is the sum of squares of the regression, SSE is the error sum of squared, and SST denotes the total sum of squares.

1. Check the asymptotic behavior of the test statistic under H_0 , and construct the reject region at the critical value $\alpha = 0.05$
2. Use 2010 GSS data which coming from the website of The General Social Survey: <http://www3.norc.ox.ac.uk/GSS+Website/>. Perform the White test and interpret the result.

1.

$$LM \sim \chi_q^2$$

where q denotes the degree of freedom equal to the number of estimated parameters in the auxiliary regression, in our case $q = 5$. If $LM > \chi_q^2$, H_0 is rejected.

2. From the result of White test we can see that p -value is 0.003385 which is statistically significant. Therefore H_0 is rejected, i.e. there is problem of heteroscedasticity.  MSEwhitetest

References

- Badi, H. B. (2008). *Econometric analysis of panel data*. Chichester/Hoboken: Wiley.
- Halbert, W. (1980). A heteroskedasticity-consistent covariance Matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 48(4), 817–838
- Spokoiny, V., & Dickhaus, T. (2014). *Basics of modern parametric statistics*. Berlin: Springer.

Chapter 8

Some Other Testing Methods

Niektoré z ďalších testovacích metód

Nie ten majster, ktorý začne, ale ktorý dokoná.

Not he who begins but he who finishes is the master.

Exercise 8.1. Let $X = (X_1, \dots, X_n)^\top$ be an i.i.d. sample from an unknown distribution P and X be a random variable with this distribution. Let a simple hypothesis H_0 be $P = P_0$ for a given measure P_0 .

Let the observation space (which is a subset of $\mathbb{R}^1$) be split into non-overlapping subsets $A_1, \dots, A_d$. Define for $j = 1, \dots, d$

$$\phi_j(x) = \mathbf{1}(x \in A_j), \quad \psi_j(x) = \frac{1}{\sigma_j} \{\phi_j(x) - p_j\}$$

with

$$p_j = P_0(A_j) = \int_{A_j} P_0(dx) = \mathbb{E}_0 \phi_j(X), \quad \sigma_j^2 = p_j(1 - p_j).$$

- (i) Are these basis functions ψ_i orthonormal under the measure P_0 ?
- (ii) Construct a test statistic $T_{n,d}$ to test H_0 .

(i) Recall that basis functions are orthonormal under the measure P_0 iff

$$\int \psi_j(x) P_0(dx) = 0, \quad \int \psi_j(x) \psi_k(x) P_0(dx) = \delta_{j,k}, \quad \forall j, k,$$

or, equivalently,

$$\mathbb{E}_0 \psi_j(X) = 0, \quad \mathbb{E}_0 \{\psi_j(X) \psi_k(X)\} = \delta_{j,k}, \quad \forall j, k.$$

The first condition is fulfilled

$$\mathbb{E}_0 \psi_j(X) = \mathbb{E}_0 \frac{1}{\sigma_j} \{\phi_j(X) - \mathbb{E}_0 \phi_j(X)\} = 0,$$

but the second one is violated for $j \neq k$:

$$\begin{aligned} \mathbb{E}_0 \{\psi_j(X) \psi_k(X)\} &= \frac{1}{\sigma_j \sigma_k} \left[\underbrace{\mathbb{E}_0 \{\phi_j(X) \phi_k(X)\}}_{=0} - \underbrace{\mathbb{E}_0 \{\phi_j(X)\}}_{=p_j} \underbrace{\mathbb{E}_0 \{\phi_k(X)\}}_{=p_k} \right] \\ &= -\frac{p_j p_k}{\sigma_j \sigma_k} \neq 0. \end{aligned}$$

Hence the functions ψ_j are not orthonormal.

- (ii) The basic idea is to compare observed frequencies $\mathbf{1}(X_i \in A_j)$ with the theoretical ones p_j under H_0 . Direct calculations yield

$$\begin{aligned} T_{n,d} &= n \sum_{j=1}^d M_{j,n}^2 = n \sum_{j=1}^d \left\{ \frac{1}{n} \sum_{i=1}^n \psi_j(X_i) \right\}^2 \\ &= n \sum_{j=1}^d \left[\frac{1}{n} \frac{1}{\sigma_j} \sum_{i=1}^n \{\phi_j(X_i) - p_j\} \right]^2 \\ &= n \sum_{j=1}^d \left[\frac{1}{\sigma_j} \left\{ \frac{1}{n} \sum_{i=1}^n \phi_j(X_i) - p_j \right\} \right]^2 \\ &= \sum_{j=1}^d \frac{n(v_{j,n} - p_j)^2}{\sigma_j^2}, \end{aligned}$$

where

$$v_{j,n} = \frac{1}{n} \sum_{i=1}^n \phi_j(X_i) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}(X_i \in A_j), \quad (8.1)$$

The statistic $T_{n,d}$ results into the test:

$$\phi_d = \mathbf{1}(T_{n,d} > \zeta_\alpha),$$

which is described in Chap. 8.1.1 of Spokoiny and Dickhaus (2014).

Exercise 8.2. Using the CLT, prove that the statistic of the chi-square test converges in law for $d = 2$ to a χ_1^2 rv.

Note that (8.1) for $d = 2$

$$\begin{aligned} v_{2,n} - p_2 &= \frac{1}{n} \sum_{i=1}^n \phi_2(X_i) - p_2 = \frac{1}{n} \sum_{i=1}^n \{1 - \phi_1(X_i)\} - (1 - p_1) \\ &= v_{1,n} - p_1. \end{aligned}$$

The chi-square test (8.1) can now be represented as

$$\begin{aligned} T_\chi &= n \sum_{j=1}^2 \frac{(v_{j,n} - p_j)^2}{p_j} = n (v_{1,n} - p_1)^2 \left(\frac{1}{p_1} + \frac{1}{p_2} \right) \\ &= \left\{ \frac{\sum_{i=1}^n \phi_i(X_j) - np_1}{\sqrt{np_1(1 - p_1)}} \right\}^2, \end{aligned}$$

and the statement of the exercise follows from the CLT. In fact, the rv $\{\sum_{i=1}^n \phi_i(X_j) - np_1\} / \sqrt{np_1(1 - p_1)}$ converges in law to a $\mathcal{N}(0, 1)$ rv and from the continuous mapping theorem we conclude that T_χ converges in law to the squared $\mathcal{N}(0, 1)$ rv.

Exercise 8.3. Let F be the distribution function of a random variable X and let $\{X_i\}_{i=1}^n$ be an i.i.d. sample from F . Denote the empirical cdf as F_n . Show that the distributions of

1. $F(X)$
2. $\sup_x n^{1/2} |F_n(x) - F(x)|$
3. $\int \{F_n(x) - F(x)\}^2 dF(x)$

do not vary with F .

The test statistic 2. is called Kolmogorov-Smirnov and 3. carries the name of Cramer-von Mises.

1.

$$\mathbb{P}\{F(X) \leq x\} = \mathbb{P}\{X \leq F^{-1}(x)\} = F\{F^{-1}(x)\} = x.$$

Thus, the random variable $F(X)$ has a uniform distribution on $[0, 1]$, i.e. is $U(0, 1)$.

2. Denote $F(x) = t$ and rewrite the supremum in the following form:

$$\sup_x n^{1/2} |F_n(x) - F(x)| = \sup_{t \in [0, 1]} |F_n\{F^{-1}(t)\} - t|$$

$$\begin{aligned}
&= \sup_{t \in [0,1]} \left| \frac{1}{n} \sum_{i=1}^n \mathbf{1}\{X_i \leq F^{-1}(t)\} - t \right| \\
&= \sup_{t \in [0,1]} \left| \frac{1}{n} \sum_{i=1}^n \mathbf{1}\{F(X_i) \leq t\} - t \right|.
\end{aligned}$$

As it has been proven in item (i), the random value $F(X_i)$ has a $U(0, 1)$ distribution. This means that the distribution of the random variable $\sup_x n^{1/2}|F_n(x) - F(x)|$ is the same for any F .

3. The proof follows the same lines:

$$\begin{aligned}
\int \{F_n(x) - F(x)\}^2 dF(x) &= \int_0^1 [F_n\{F^{-1}(t)\} - t]^2 dt \\
&= \int_0^1 \left[\frac{1}{n} \sum_{i=1}^n \mathbf{1}\{X_i \leq F^{-1}(t)\} - t \right]^2 dt \\
&= \int_0^1 \left[\frac{1}{n} \sum_{i=1}^n \mathbf{1}\{F(X_i) \leq t\} - t \right]^2 dt,
\end{aligned}$$

and the statement of the exercise is proven.

Exercise 8.4. Let F be the distribution function of a random variable X and let $\{X_i\}_{i=1}^n$ be an i.i.d. sample from F . Denote the edf as F_n . Let H_0 be the hypothesis that the distribution F has the same 4 moments as a $\mathcal{N}(0, 1)$ rv:

$$H_0 : \quad \mathbb{E}X = 0, \mathbb{E}X^2 = 1, \mathbb{E}X^3 = 0, \mathbb{E}X^4 = 3,$$

and the alternative H_1 is that some of these moments differ. Construct the test of method of moments with asymptotic level α .

Hint: use only the first and the second empirical moments.

Consider the function $\mathbf{g} : \mathbb{R} \rightarrow \mathbb{R}^2$, $\mathbf{g}(x) = (x, x^2)^\top$. From the CLT, we know that

$$n^{1/2} V^{-1/2} \left\{ n^{-1} \sum_{i=1}^n \mathbf{g}(X_i) - \mathbb{E}_0 \mathbf{g}(X) \right\} \xrightarrow{\mathcal{L}} \mathbb{N}(0, I_2), \quad (8.2)$$

where $\mathbb{E}_0 \mathbf{g}(X) = (\mathbb{E}_0 X, \mathbb{E}_0 X^2)^\top = (0, 1)^\top$ and

$$V = \mathbb{E}_0 \left[\begin{pmatrix} X - \mathbb{E}_0 X \\ \mathbb{E}_0 X^2 - \mathbb{E}_0 X^2 \end{pmatrix} \begin{pmatrix} X - \mathbb{E}_0 X \\ \mathbb{E}_0 X^2 - \mathbb{E}_0 X^2 \end{pmatrix}^\top \right]$$

$$\begin{aligned}
&= \begin{pmatrix} \mathbb{E}_0 X^2 - (\mathbb{E}_0 X)^2 & \mathbb{E}_0 X^3 - \mathbb{E}_0 X \mathbb{E}_0 X^2 \\ \mathbb{E}_0 X^3 - \mathbb{E}_0 X \mathbb{E}_0 X^2 & \mathbb{E}_0 X^4 - (\mathbb{E}_0 X^2)^2 \end{pmatrix} \\
&= \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}.
\end{aligned}$$

Here the index “0” indicates that we are computing expectations under the null hypothesis H_0 .

Hence the statement (8.2) means that

$$n^{1/2} \left(2^{-1/2} \left(n^{-1} \sum_{i=1}^n X_i^2 - 1 \right) \right) \xrightarrow{\mathcal{L}} \mathbb{N}(0, I_2),$$

and the statistic

$$T_n = n \left\{ \sum_{i=1}^n \left(n^{-1} \sum_{i=1}^n X_i \right)^2 + 2^{-1} \left(n^{-1} \sum_{i=1}^n X_i^2 - 1 \right)^2 \right\}$$

has under H_0 a chi-square distribution with 2 degrees of freedom χ_2^2 . The test

$$\phi = \mathbf{1} \{T_n > \zeta_\alpha\}$$

where ζ_α is a $(1 - \alpha)$ quantile of the χ^2 distribution has the desired asymptotic level α .

The test is also called Jarque Bera Test.

Exercise 8.5. Suppose y_t is the time series of DAX 30, a stock index in Germany. The time series is from December 22, 2009 to December 21, 2011 (as Fig. 8.1).

Define the log return of DAX index:

$$z_t = \log y_t - \log y_{t-1}.$$

Apply Jarque Bera test to z_t .

The test statistics is 99.1888 and the p-value is 2.2×10^{-16} . This suggests that the log returns may not be normally distributed if one takes significant level $\alpha = 0.01$.

Exercise 8.6. Following Exercise 8.5, apply the Kolmogorov-Smirnov test to z_t .

The test statistics is 10.8542 and the p-value is 0.01. This suggests that the log returns may not be normally distributed if one takes the significance level $\alpha = 0.01$.

Exercise 8.7. Following Exercise 8.5, apply the Cramer von Mises test to z_t .

The test statistics is 1.0831 and the p-value is 8.134×10^{-10} . This suggests that the log returns may not be normally distributed if one takes the significance level $\alpha = 0.01$.

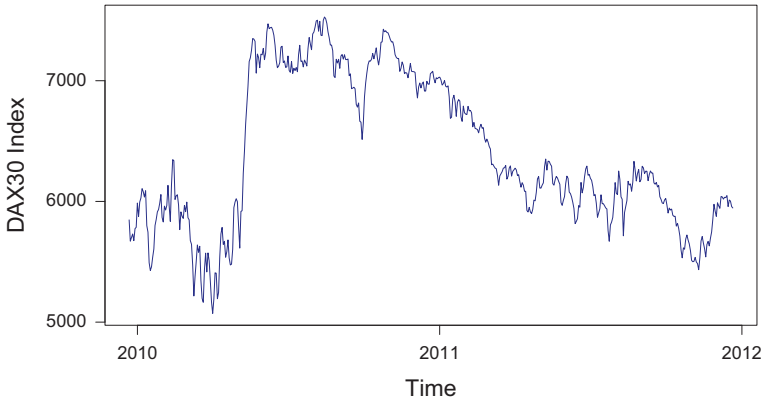


Fig. 8.1 The time series of DAX30.  MSENormalityTests

Exercise 8.8. *Test the hypothesis of the equality of the covariance matrices on two simulated 4-dimensional samples of sizes $n_1 = 30$ and $n_2 = 20$.*

Let $X_{ih} \sim N_p(\mu_h, \Sigma_h)$, $i = 1, \dots, n_h$, $h = 1, 2$, be independent random vectors. The test problem of testing the equality of the covariance matrices can be written as

$$H_0 : \Sigma_1 = \Sigma_2 \text{ versus } H_1 : \text{no constraints.}$$

Both subsamples provide S_h , an estimator of Σ_h , with the Wishart distribution $n_h S_h \sim W_p(\Sigma_h, n_h - 1)$. Under the null hypothesis $H_0 : \Sigma_1 = \Sigma_2$, we have for the common covariance matrix that $\sum_{h=1}^2 n_h S_h \sim W_p(\Sigma, n - 2)$, where $n = \sum_{h=1}^2 n_h$.

Let $S = \frac{n_1 S_1 + n_2 S_2}{n}$ be the weighted average of S_1 and S_2 . The likelihood ratio test leads to the test statistic

$$-2 \log \lambda = n \log |S| - \sum_{h=1}^2 n_h \log |S_h| \quad (8.3)$$

which under H_0 is approximately distributed as a χ_m^2 with $m = \frac{1}{2}(2 - 1)p(p + 1)$ degrees of freedom.

We test the equality of the covariance matrices for the three data sets given in Härdle and Simar (2011) (Example 7.14) who simulated two independent normal distributed samples with $p = 4$ dimensions and the sample sizes of $n_1 = 30$ and $n_2 = 20$ leading to the asymptotic distribution of the test statistics (8.3) with $m = \frac{1}{2}(2 - 1)4(4 + 1) = 10$ degrees of freedom.

- (a) With a common covariance matrix in both populations $\Sigma_1 = \Sigma_2 = I_4$, we obtain the following empirical covariance matrices:

$$S_1 = \begin{pmatrix} 0.812 & -0.229 & -0.034 & 0.073 \\ -0.229 & 1.001 & 0.010 & -0.059 \\ -0.034 & 0.010 & 1.078 & -0.098 \\ 0.073 & -0.059 & -0.098 & 0.823 \end{pmatrix}$$

and

$$S_2 = \begin{pmatrix} 0.559 & -0.057 & -0.271 & 0.306 \\ -0.057 & 1.237 & 0.181 & 0.021 \\ -0.271 & 0.181 & 1.159 & -0.130 \\ 0.306 & 0.021 & -0.130 & 0.683 \end{pmatrix}$$

The determinants are $|S| = 0.590$, $|S_1| = 0.660$ and $|S_2| = 0.356$ leading to the likelihood ratio test statistic:

$$-2 \log \lambda = 50 \log(0.590) - 30 \log(0.660) - 20 \log(0.356) = 6.694$$

The value of the test statistic is smaller than the critical value $\chi^2_{0.95;10} = 18.307$ and, hence, we do not reject the null hypothesis.

- (b) The second simulated samples have covariance matrices $\Sigma_1 = \Sigma_2 = 16I_4$. Now, the standard deviation is 4 times larger than in the previous case. The sample covariance matrices from the second simulation are:

$$S_1 = \begin{pmatrix} 21.907 & 1.415 & -2.050 & 2.379 \\ 1.415 & 11.853 & 2.104 & -1.864 \\ -2.050 & 2.104 & 17.230 & 0.905 \\ 2.379 & -1.864 & 0.905 & 9.037 \end{pmatrix},$$

$$S_2 = \begin{pmatrix} 20.349 & -9.463 & 0.958 & -6.507 \\ -9.463 & 15.502 & -3.383 & -2.551 \\ 0.958 & -3.383 & 14.470 & -0.323 \\ -6.507 & -2.551 & -0.323 & 10.311 \end{pmatrix}$$

and the value of the test statistic is:

$$-2 \log \lambda = 50 \log(40066) - 30 \log(35507) - 20 \log(16233) = 21.693.$$

Since the value of the test statistic is larger than the critical value of the asymptotic distribution, $\chi^2_{0.95;10} = 18.307$, we reject the null hypothesis.

- (c) The covariance matrix in the third case is similar to the second case $\Sigma_1 = \Sigma_2 = 16I_4$ but, additionally, the covariance between the first and the fourth variable is $\sigma_{14} = \sigma_{41} = -3.999$. The corresponding correlation coefficient is $r_{41} = -0.9997$.

The sample covariance matrices from the third simulation are:

$$S_1 = \begin{pmatrix} 14.649 & -0.024 & 1.248 & -3.961 \\ -0.024 & 15.825 & 0.746 & 4.301 \\ 1.248 & 0.746 & 9.446 & 1.241 \\ -3.961 & 4.301 & 1.241 & 20.002 \end{pmatrix}$$

and

$$S_2 = \begin{pmatrix} 14.035 & -2.372 & 5.596 & -1.601 \\ -2.372 & 9.173 & -2.027 & -2.954 \\ 5.596 & -2.027 & 9.021 & -1.301 \\ -1.601 & -2.954 & -1.301 & 9.593 \end{pmatrix}.$$

The value of the test statistic is:

$$-2 \log \lambda = 50 \log(24511) - 30 \log(37880) - 20 \log(6602.3) = 13.175$$

The value of the likelihood ratio test statistic is now smaller than the critical value, $\chi^2_{0.95;10} = 18.307$, and we do not reject the null hypothesis.

Notice that in part (b), we have rejected a valid null hypothesis. One should always keep in mind that a wrong decision of this type (so-called type I error) is possible and it occurs with probability α .  MSEtestcov

Exercise 8.9. Consider two independent iid samples, each of size 10, from two bivariate normal populations. The results are summarized below:

$$\bar{x}_1 = (3, 1)^\top; \bar{x}_2 = (1, 1)^\top$$

$$S_1 = \begin{pmatrix} 4 & -1 \\ -1 & 2 \end{pmatrix}; S_2 = \begin{pmatrix} 2 & -2 \\ -2 & 4 \end{pmatrix}.$$

Provide a solution to the following tests:

- (a) $H_0: \mu_1 = \mu_2 \quad H_1: \mu_1 \neq \mu_2$
- (b) $H_0: \mu_{11} = \mu_{21} \quad H_1: \mu_{11} \neq \mu_{21}$
- (c) $H_0: \mu_{12} = \mu_{22} \quad H_1: \mu_{12} \neq \mu_{22}$

Compare the solutions and comment.

- (a) Let us start by verifying the assumption of equality of the two covariance matrices, i.e., the hypothesis:

$$H_0: \Sigma_1 = \Sigma_2 \quad \text{versus} \quad H_1: \Sigma_1 \neq \Sigma_2.$$

This hypothesis can be tested using the approach described in Exercise 8.8 where we used the test statistic (for $k = 2$ groups):

$$-2 \log \lambda = n \log |S| - \sum_{h=1}^2 n_h \log |S_h|$$

which is under the null hypothesis $H_0 : \Sigma_1 = \Sigma_2$ approximately χ_m^2 distributed, where $m = \frac{1}{2}(k-1)p(p+1) = \frac{1}{2}(2-1)2(2+1) = 3$.

We calculate the average of the observed variance matrices

$$S = \begin{pmatrix} 3 & -1.5 \\ -1.5 & 3 \end{pmatrix}$$

and we get the value of the test statistic

$$-2 \log \lambda = 20 \log |S| - (10 \log |S_1| + 10 \log |S_2|) = 4.8688$$

which is smaller than the critical value $\chi_{0.95;3}^2 = 7.815$. Hence, the value of the test statistic is not significant, we do not reject the null hypothesis, and the assumption of the equality of the variance matrices can be used in testing the equality of the mean vectors.

Now, we can test the equality of the mean vectors:

$$H_0 : \mu_1 = \mu_2 \quad \text{versus} \quad H_1 : \mu_1 \neq \mu_2.$$

The rejection region is given by

$$\frac{n_1 n_2 (n_1 + n_2 - p - 1)}{p(n_1 + n_2)^p} (\bar{x}_1 - \bar{x}_2)^T S^{-1} (\bar{x}_1 - \bar{x}_2) \geq F_{1-\alpha; p, n_1 + n_2 - p - 1}.$$

For $\alpha = 0.05$ we get the test statistic $3.7778 \geq F_{0.95;2,17} = 3.5915$. Hence, the null hypothesis $H_0 : \mu_1 = \mu_2$ is rejected and we can say that the mean vectors of the two populations are significantly different.

- (b) For the comparison of the two mean vectors first components we calculate the 95 % simultaneous confidence interval for the difference. We test the hypothesis

$$H_0 : \mu_{11} = \mu_{21} \quad \text{versus} \quad H_1 : \mu_{11} \neq \mu_{21}.$$

This test problem is only one-dimensional and it can be solved by calculating the common two-sample t -test. The test statistic

$$\frac{\bar{x}_{11} - \bar{x}_{21}}{\sqrt{\frac{4}{n_1} + \frac{2}{n_2}}} = \frac{2}{\sqrt{\frac{6}{10}}} = 2.5820$$

is greater than the corresponding critical value $t_{0.95;18} = 2.1011$ and hence we reject the null hypothesis.

- (c) The comparison of the second component of the mean vectors can be also based on the two-sample t -test. In this case, it is obvious that the value of the test statistic is equal to zero (since $\bar{x}_{12} = \bar{x}_{22} = 1$) and the null hypothesis can not be rejected.

In part (a) we have rejected the null hypothesis that the two mean vectors are equal. From the componentwise test performed in (b) and (c), we observe that the reason for rejecting the equality of the two two-dimensional mean vectors was due mainly to differences in the first component.

Exercise 8.10. *In the vocabulary data set (Bock, 1975) given in the table below, it predicts the vocabulary score of the children in eleventh grade from the results in grades 8–10. Estimate a linear model and test its significance.*

Subjects	Grade 8	Grade 9	Grade 10	Grade 11	Mean
1	1.75	2.60	3.76	3.68	2.95
2	0.90	2.47	2.44	3.43	2.31
3	0.80	0.93	0.40	2.27	1.10
4	2.42	4.15	4.56	4.21	3.83
5	−1.31	−1.31	−0.66	−2.22	−1.38
6	−1.56	1.67	0.18	2.33	0.66
7	1.09	1.50	0.52	2.33	1.36
8	−1.92	1.03	0.50	3.04	0.66
9	−1.61	0.29	0.73	3.24	0.66
10	2.47	3.64	2.87	5.38	3.59
11	−0.95	0.41	0.21	1.82	0.37
12	1.66	2.74	2.40	2.17	2.24
13	2.07	4.92	4.46	4.71	4.04
14	3.30	6.10	7.19	7.46	6.02
15	2.75	2.53	4.28	5.93	3.87
16	2.25	3.38	5.79	4.40	3.96
17	2.08	1.74	4.12	3.62	2.89
18	0.14	0.01	1.48	2.78	1.10
19	0.13	3.19	0.60	3.14	1.77
20	2.19	2.65	3.27	2.73	2.71
21	−0.64	−1.31	−0.37	4.09	0.44
22	2.02	3.45	5.32	6.01	4.20
23	2.05	1.80	3.91	2.49	2.56
24	1.48	0.47	3.63	3.88	2.37
25	1.97	2.54	3.26	5.62	3.35
26	1.35	4.63	3.54	5.24	3.69
27	−0.56	−0.36	1.14	1.34	0.39
28	0.26	0.08	1.17	2.15	0.92
29	1.22	1.41	4.66	2.62	2.47

(continued)

(continued)

Subjects	Grade 8	Grade 9	Grade 10	Grade 11	Mean
30	-1.43	0.80	-0.03	1.04	0.09
31	-1.17	1.66	2.11	1.42	1.00
32	1.68	1.71	4.07	3.30	2.69
33	-0.47	0.93	1.30	0.76	0.63
34	2.18	6.42	4.64	4.82	4.51
35	4.21	7.08	6.00	5.65	5.73
36	8.26	9.55	10.24	10.58	9.66
37	1.24	4.90	2.42	2.54	2.78
38	5.94	6.56	9.36	7.72	7.40
39	0.87	3.36	2.58	1.73	2.14
40	-0.09	2.29	3.08	3.35	2.15
41	3.24	4.78	3.52	4.84	4.10
42	1.03	2.10	3.88	2.81	2.45
43	3.58	4.67	3.83	5.19	4.32
44	1.41	1.75	3.70	3.77	2.66
45	-0.65	-0.11	2.40	3.53	1.29
46	1.52	3.04	2.74	2.63	2.48
47	0.57	2.71	1.90	2.41	1.90
48	2.18	2.96	4.78	3.34	3.32
49	1.10	2.65	1.72	2.96	2.11
50	0.15	2.69	2.69	3.50	2.26
51	-1.27	1.26	0.71	2.68	0.85
52	2.81	5.19	6.33	5.93	5.06
53	2.62	3.54	4.86	5.80	4.21
54	0.11	2.25	1.56	3.92	1.96
55	0.61	1.14	1.35	0.53	0.91
56	-2.19	-0.42	1.54	1.16	0.02
57	1.55	2.42	1.11	2.18	1.82
58	0.04	0.50	2.60	2.61	1.42
59	3.10	2.00	3.92	3.91	3.24
60	-0.29	2.62	1.60	1.86	1.45
61	2.28	3.39	4.91	3.89	3.62
62	2.57	5.78	5.12	4.98	4.61
63	-2.19	0.71	1.56	2.31	0.60
64	-0.04	2.44	1.79	2.64	1.71
Mean	1.14	2.54	2.99	3.47	2.53

Coefficients:

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.4782      0.2999   4.929 6.86e-06 ***
grade8       0.2015      0.1582   1.273  0.2078
grade9       0.2278      0.1152   1.977  0.0526 .
grade10      0.3965      0.1304   3.041  0.0035 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1 1
```

Residual standard error: 1.073 on 60 degrees of freedom
 Multiple R-squared: 0.7042, Adjusted R-squared: 0.6894
 F-statistic: 47.61 on 3 and 60 DF, p-value: 7.144e-16

Regression analysis reveals reasonably high coefficient of determination. Hypothesis of independence (H_0 : all parameters = 0) is rejected on level $\alpha = 0.05$ since the F -statistics is statistically significant (the p -value is smaller than $\alpha = 0.05$).

The vocabulary score from tenth grade (β_3 = grade10) is statistically significant for the forecast of performance in eleventh grade. The other two variables, vocabulary scores from the eighth and ninth grade are not statistically significant at level $\alpha = 0.05$. More formally, the test does not reject the hypothesis that parameters β_2 and β_3 are equal to zero.

One might be tempted to simplify the model by excluding the insignificant variables. Excluding only the score in eighth grade leads to the following result which shows that the variable measuring the vocabulary score in ninth grade has changed its significance.

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.2355      0.2327   5.309 1.63e-06 ***
grade9         0.2893      0.1051   2.752 0.00779 **
grade10        0.5022      0.1011   4.969 5.75e-06 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

Residual standard error: 1.079 on 61 degrees of freedom
Multiple R-squared: 0.6962, Adjusted R-squared: 0.6862
F-statistic: 69.89 on 2 and 61 DF, p-value: < 2.2e-16
```

Hence, the final model explains the vocabulary score in grade eleven using vocabulary scores in the previous two grades.  MSElinregvocab

Exercise 8.11. Assume that we have observations from two p -dimensional normal populations, $x_{i1} \sim N_p(\mu_1, \Sigma)$, $i = 1, \dots, n_1$, and $x_{i2} \sim N_p(\mu_2, \Sigma)$, $i = 1, \dots, n_2$. The mean vectors μ_1 and μ_2 are called profiles. An example of two such 5-dimensional profiles is given in Fig. 8.2. Propose tests of the following hypotheses:

1. Are the profiles parallel?
2. If the profiles are parallel, are they at the same level?
3. If the profiles are parallel, are they also horizontal?

The above questions are easily translated into linear constraints on the means and a test statistic can be obtained accordingly.

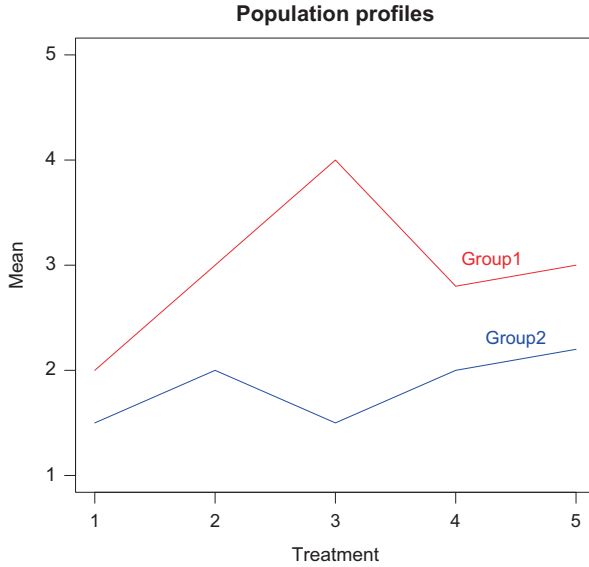


Fig. 8.2 Example of population profiles  MSEprofil

(a) Let C be a $(p-1) \times p$ contrast matrix defined as

$$C = \begin{pmatrix} 1 & -1 & 0 & \cdots & 0 \\ 0 & 1 & -1 & \cdots & 0 \\ 0 & \cdots & 0 & 1 & -1 \end{pmatrix}.$$

The hypothesis of parallel profiles is equivalent to

$$H_0^{(1)} : C\mu_1 - C\mu_2 = C(\mu_1 - \mu_2) = 0_{p-1}.$$

The test of parallel profiles can be based on:

$$C(\bar{x}_1 - \bar{x}_2) \sim N_{p-1} \left(C(\mu_1 - \mu_2), \frac{n_1 + n_2}{n_1 n_2} C \Sigma C^\top \right).$$

Next, for the pooled covariance matrix $S = (n_1 S_1 + n_2 S_2) / (n_1 + n_2)$ we have the Wishart distribution:

$$\begin{aligned} n_1 S_1 + n_2 S_2 &\sim W_p(\Sigma, n_1 + n_2 - 2) \\ C(n_1 S_1 + n_2 S_2)C^\top &\sim W_{p-1}(C \Sigma C^\top, n_1 + n_2 - 2). \end{aligned}$$

Under the null hypothesis, we know that $C(\mu_1 - \mu_2) = 0_{p-1}$ and it follows that the statistic

$$\begin{aligned} & (n_1 + n_2 - 2) \{C(\bar{x}_1 - \bar{x}_2)\}^\top \left\{ \frac{n_1 + n_2}{n_1 n_2} C(n_1 S_1 + n_2 S_2) C^\top \right\}^{-1} C(\bar{x}_1 - \bar{x}_2) \\ &= (n_1 + n_2 - 2) \{C(\bar{x}_1 - \bar{x}_2)\}^\top \left\{ \frac{n_1 + n_2}{n_1 n_2} (n_1 + n_2) C S C^\top \right\}^{-1} C(\bar{x}_1 - \bar{x}_2) \\ &= \frac{(n_1 + n_2 - 2) n_1 n_2}{(n_1 + n_2)^2} \{C(\bar{x}_1 - \bar{x}_2)\}^\top \{C S C\}^{-1} C(\bar{x}_1 - \bar{x}_2) \end{aligned}$$

has the Hotelling T^2 distribution $T^2(p - 1, n_1 + n_2 - 2)$ and the null hypothesis of parallel profiles is rejected if

$$\frac{n_1 n_2 (n_1 + n_2 - p)}{(n_1 + n_2)^2 (p - 1)} \{C(\bar{x}_1 - \bar{x}_2)\}^\top (C S C^\top)^{-1} C(\bar{x}_1 - \bar{x}_2) > F_{1-\alpha; p-1, n_1+n_2-p}. \quad (8.4)$$

- (b) Assuming that the two profiles are parallel, the null hypothesis of the equality of the two levels can be formally written as

$$H_0^{(2)} : 1_p^\top (\mu_1 - \mu_2) = 0.$$

For $1_p^\top (\bar{x}_1 - \bar{x}_2)$, as a linear function of normally distributed random vectors, we have

$$1_p^\top (\bar{x}_1 - \bar{x}_2) \sim N_1 \left(1_p^\top (\mu_1 - \mu_2), \frac{n_1 + n_2}{n_1 n_2} 1_p^\top \Sigma 1_p \right).$$

Since

$$1_p^\top (n_1 S_1 + n_2 S_2) 1_p \sim W_1 \left(1_p^\top \{\Sigma 1_p, n_1 + n_2 - 2\} \right),$$

we have that

$$(n_1 + n_2) 1_p^\top S 1_p \sim W_1(1_p^\top \Sigma 1_p, n_1 + n_2 - 2),$$

where S is the pooled empirical variance matrix. The test of equality can be based on the test statistic:

$$\begin{aligned} & (n_1 + n_2 - 2) \{1_p^\top (\bar{x}_1 - \bar{x}_2)\}^\top \left\{ \frac{n_1 + n_2}{n_1 n_2} C(n_1 S_1 + n_2 S_2) C^\top \right\}^{-1} 1_p^\top (\bar{x}_1 - \bar{x}_2) \\ &= \frac{n_1 n_2 (n_1 + n_2 - 2)}{(n_1 + n_2)^2} \frac{\{1_p^\top (\bar{x}_1 - \bar{x}_2)\}^2}{1_p^\top S 1_p} \sim T^2(1, n_1 + n_2 - 2) \end{aligned}$$

which leads directly the rejection region:

$$\frac{n_1 n_2 (n_1 + n_2 - 2)}{(n_1 + n_2)^2} \frac{\left\{1_p^\top (\bar{x}_1 - \bar{x}_2)\right\}^2}{1_p^\top S 1_p} > F_{1-\alpha; 1, n_1 + n_2 - 2}. \quad (8.5)$$

- (c) If it is accepted that the profiles are parallel, then we can exploit the information contained in both groups to test if the two profiles also have zero slope, i.e., the profiles are horizontal. The null hypothesis may be written as:

$$H_0^{(3)} : C(\mu_1 + \mu_2) = 0.$$

The average profile $\bar{x} = (n_1 \bar{x}_1 + n_2 \bar{x}_2)/(n_1 + n_2)$ has a p -dimensional normal distribution:

$$\bar{x} \sim N_p \left(\frac{n_1 \mu_1 + n_2 \mu_2}{n_1 + n_2}, \frac{1}{n_1 + n_2} \Sigma \right).$$

Now the horizontal, $H_0^{(3)} : C(\mu_1 + \mu_2) = 0_{p-1}$, and parallel, $H_0^{(1)} : C(\mu_1 - \mu_2) = 0_{p-1}$, profiles imply that

$$\begin{aligned} C \left(\frac{n_1 \mu_1 + n_2 \mu_2}{n_1 + n_2} \right) &= \frac{C}{n_1 + n_2} (n_1 \mu_1 + n_2 \mu_2) \\ &= \frac{C}{2(n_1 + n_2)} \{ (n_1 + n_2)(\mu_1 + \mu_2) + (n_1 - n_2)(\mu_1 - \mu_2) \} \\ &= 0_{p-1}. \end{aligned}$$

So, under parallel and horizontal profiles we have

$$C \bar{x} \sim N_{p-1} \left(0_{p-1}, \frac{1}{n_1 + n_2} C \Sigma C^\top \right).$$

and

$$C(n_1 + n_2) S C^\top = C (n_1 S_1 + n_2 S_2) C^\top \sim W_{p-1} (C \Sigma C^\top, n_1 + n_2 - 2).$$

Again, we get under the null hypothesis that

$$(n_1 + n_2 - 2)(C \bar{x})^\top (C \Sigma C^\top)^{-1} C \bar{x} \sim T^2(p - 1, n_1 + n_2 - 2)$$

which leads to the rejection region:

$$\frac{n_1 + n_2 - p}{p - 1} (C\bar{x})^\top (CSC^\top)^{-1} C\bar{x} > F_{1-\alpha; p-1, n_1+n_2-p}. \quad (8.6)$$

References

- Bock, R.D. (1975). Multivariate statistical methods in behavioral research (Vol. 13, p. 623). New York: McGraw-Hill
- Härdle, W., & Simar, L. (2011). *Applied multivariate statistical analysis* (3rd ed.). Berlin: Springer.
- Spokoiny, V., & Dickhaus, T. (2014). *Basics of modern parametric statistics*. Berlin: Springer.

Index

- $\sqrt{n}$ consistent, 160
- 5-point property, 105

- Alternating method, 105

- adaptivity condition, 95
- alternative hypothesis, 129
- asymptotic normality, xvii

- Bayes estimation, 107, 119
- Bayes risk, 112
- Bernoulli, 1
- Bernoulli experiment, 107
- bias, xvii
- Bonferroni rule, 141

- canonical parameter, 33
- Cauchy distribution, 131
- cdf, xii, xviii, xx
 - empirical, xviii
 - joint, xii
 - marginal, xii
- characteristic function, xii
- characteristic polynomial, xviii
- chi-square test, 169
- chi-squared test, orthonormal under the
 - measure basis, test statistic $T_{n,d}$, 167
- χ^2 distribution, xiv
 - quantile, xiv
- CLT, xiv, 169
- conditional distribution, xviii
- conditional expectation, xii
- conditional moments, xviii
- conditional variance, xii
- confidence ellipsoids, 87
- contingency table, xviii
- continuous mapping theorem, 169
- contrast, 26
- contrast matrix, 179
- convergence
 - almost sure, xiv
 - in probability, xiv
- convergence in distribution, xiv
- convergence of the alternating method, spectral
 - norm, 103
- convex hull, xiv
- correlation, xiii
 - empirical, xiii
- correlation matrix
 - empirical, xiii
- covariance, xii
 - empirical, xiii
- covariance matrix, xiii
 - empirical, xiii
- Cramér-Rao inequality, 24
- Cramer-von Mises, 169
- critical value, xviii
- cumulants, xii

- data matrix, xiii
- DAX return, 144
- determinant, xiv
- deviation probabilities for the maximum
 - likelihood, 34
- diagonal, xiii
- distribution, xi
 - χ^2 , xiv

- conditional, xviii
- F -, xiv
- Gaussian, xx
- marginal, xix
- multinormal, xx
- normal, xx
- t -, xiv
- distribution function
 - empirical, xviii
- edf, *see* empirical distribution function
- eigenvalue, xviii
- eigenvector, xviii
- empirical distribution function, xviii, 169
- empirical moments, xix
- error of the first kind, 129
- error of the second kind, 129
- estimate, xix
- estimation under the homogeneous noise
 - assumption, 80
- estimator, xix
- expected value, xix
 - conditional, xii
- Exponential distribution, 133
- exponential family, 33, 37
- F-test, 144
- F -distribution, xiv
 - quantile, xiv
- Fisher information, 23, 24, 33
- Gamma distribution, 120
- Gauss-Markov theorem, 3, 98
- Gaussian distribution, xx
- Gaussian shift, 24, 113
- Glivenko-Cantelli theorem, 9
- Hessian matrix, xix
- horizontal profiles, 178, 181
- indicator, xi
- Jarque Bera Test, 171
- Kolmogorov-Smirnov test, 169
- Kronecker product, xi
- Kullback-Leibler divergence, 23
- likelihood, xix
- likelihood ratio test, 136
- linear constraint, 178
- linear dependence, xix
- linear model, 176
- linear regression, 176
- linear space, xiv
- LLN, xiv
- log-likelihood, xix
- marginal distribution, xix
- marginal moments, xix
- matrix
 - contrast, 179
 - covariance, xiii
 - determinant of, xiv
 - diagonal of, xiii
 - Hessian, xix
 - orthogonal, xx
 - rank of, xiii
 - trace, xiii
- maximum likelihood estimator, 30
- mean, xii, xix
- mean squared error, *see* MSE
- median, xx
- Method of moments, 30
- method of moments for an i.i.d. sample, 170
- ML estimator, 37
- moments, xii, xix
 - empirical, xix
 - marginal, xix
- MSE, xx
- multinormal distribution, xx
- multivariate parameter, 30
- natural parameter, 33
- Neyman-Pearson lemma, 132
- Neyman-Pearson test, 131, 132, 134, 144, 146, 150
- normal distribution, xx
- null hypothesis, 129
- observation, xiii
- One-sided and two-sided tests, 136
- order statistic, xiii
- orthogonal design, 80
- orthogonal matrix, xx
- orthonormal design, 80, 81
- parallel profiles, 178, 179
- Pareto distribution, 120

- pdf, xii
 - conditional, xii
 - joint, xii
 - marginal, xii
- penalized likelihood, bias-variance decomposition, 90
- penalized log-likelihood, ridge regression, 89
- pivotal quantity, xx
- Poisson family, 24
- power function, 129
- profile analysis, 178
- profile estimation, 94
- projection and shrinkage estimates, 92
- p -value, xx

- quantile, xx

- R-efficiency, 24
- random variable, xi, xx
- random vector, xi, xx
- rank, xiii
- Region of rejection (critical region), 131
- regular family, 23

- sample, xiii
- scatterplot, xx
- semi-invariants, xii
- semiparametric estimation, target and nuisance parameters, adaptivity condition, 93

- singular value decomposition, xx
- spectral decomposition, xxi
- spectral representation, 85
- statistical test, 129
- stochastic component, 84
- subspace, xxi
- SVD, *see* singular value decomposition

- Taylor expansion, xxi
- t -distribution, xiv
 - quantile, xiv
- test
 - covariance matrix, 172
 - mean vector, 174
 - two-sample, 175
- test of method of moments, 170
- Tikhonov regularization, 88
- trace, xiii

- uniformly most powerful test, 134

- variance, xiii
 - conditional, xii
 - empirical, xiii
- volatility model, 33

- Wilks phenomenon, 86
- Wishart distribution, 179

