

SPRINGER BRIEFS IN STATISTICS

Stefan Bedbur

Udo Kamps

# Multivariate Exponential Families: A Concise Guide to Statistical Inference



Springer

# **SpringerBriefs in Statistics**

SpringerBriefs present concise summaries of cutting-edge research and practical applications across a wide spectrum of fields. Featuring compact volumes of 50 to 125 pages, the series covers a range of content from professional to academic. Typical topics might include:

- A timely report of state-of-the art analytical techniques
- A bridge between new research results, as published in journal articles, and a contextual literature review
- A snapshot of a hot or emerging topic
- An in-depth case study or clinical example
- A presentation of core concepts that students must understand in order to make independent contributions

SpringerBriefs in Statistics showcase emerging theory, empirical research, and practical application in Statistics from a global author community.

SpringerBriefs are characterized by fast, global electronic dissemination, standard publishing contracts, standardized manuscript preparation and formatting guidelines, and expedited production schedules.

More information about this series at <http://www.springer.com/series/8921>

Stefan Bedbur • Udo Kamps

# Multivariate Exponential Families: A Concise Guide to Statistical Inference

 Springer

Stefan Bedbur  
Institute of Statistics  
RWTH Aachen University  
Aachen, Germany

Udo Kamps  
Institute of Statistics  
RWTH Aachen University  
Aachen, Germany

ISSN 2191-544X

SpringerBriefs in Statistics

ISBN 978-3-030-81899-9

ISSN 2191-5458 (electronic)

ISBN 978-3-030-81900-2 (eBook)

<https://doi.org/10.1007/978-3-030-81900-2>

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2021

This work is subject to copyright. All rights are solely and exclusively licensed 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.

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.

The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG.  
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

# Preface

Parametric families of probability distributions and their properties are extensively studied in the literature on statistical modeling and inference. Exponential families of distributions comprise density functions of a particular form, which enables general assertions and leads to nice features. A literature review of books in the field of Mathematical Statistics discussing exponential families and their properties is given in the first chapter. With a focus on parameter estimation and hypotheses testing, we introduce the reader to properties of multivariate and multiparameter exponential families and show a variety of detailed examples. The material is widely self-contained and written in a mathematical setting. It may serve as a concise course on exponential families in a systematic structure. Supplements and comments can be found on the website [https://www.isw.rwth-aachen.de/EF\\_SpringerBriefs.php](https://www.isw.rwth-aachen.de/EF_SpringerBriefs.php).

Dear reader, we would very much appreciate receiving your remarks, criticism, and hints! We hope you will benefit from our introduction to exponential families.

Aachen, Germany  
May 2021

Stefan Bedbur  
Udo Kamps

# Contents

<b>1</b>	<b>Introduction: Aims and Outline</b>	1
<b>2</b>	<b>Parametrizations and Basic Properties</b>	5
2.1	Definition and Examples	5
2.2	Order and Minimal Representation	17
2.3	Structural Properties and Natural Parametrization	23
2.4	Analytical Properties and Mean Value Parametrization	32
<b>3</b>	<b>Distributional and Statistical Properties</b>	43
3.1	Generating Functions	43
3.2	Marginal and Conditional Distributions	47
3.3	Product Measures	49
3.4	Sufficiency and Completeness	54
3.5	Score Statistic and Information Matrix	57
3.6	Divergence and Distance Measures	60
<b>4</b>	<b>Parameter Estimation</b>	65
4.1	Maximum Likelihood Estimation	66
4.2	Constrained Maximum Likelihood Estimation	71
4.3	Efficient Estimation	78
4.4	Asymptotic Properties	84
<b>5</b>	<b>Hypotheses Testing</b>	93
5.1	One-Sided Test Problems	93
5.2	Two-Sided Test Problems	101
5.3	Testing for a Selected Parameter	105
5.4	Testing for Several Parameters	109
<b>6</b>	<b>Exemplary Multivariate Applications</b>	117
6.1	Negative Multinomial Distribution	117
6.2	Dirichlet Distribution	125
6.3	Generalized Order Statistics	131

**References** ..... 137

**Index** ..... 141



# Symbols

$\mathbb{N}$	$\{1, 2, 3, \dots\}$ set of natural numbers
$\mathbb{N}_0$	$\{0, 1, 2, \dots\}$ set of natural numbers including zero
$\mathbb{Z}$	$\{\dots, -2, -1, 0, 1, 2, \dots\}$ set of integers
$\mathbb{R}$	$(-\infty, \infty)$ set of real numbers
$\mathbb{C}$	$\{a + ib : a, b \in \mathbb{R}\}$ set of complex numbers
$A^n$	$\times_{i=1}^n A$ $n$ -fold Cartesian product of set $A$
$\mathbb{1}_A$	indicator function of set $A$ $\mathbb{1}_A(x) = 1$ if $x \in A$ and zero otherwise
$\text{int}(A)$	interior of set $A$
$\mathbf{a}^t, \mathbf{A}^t$	transpose of vector $\mathbf{a}$ and matrix $\mathbf{A}$
$\mathbf{A} > 0$ ( $\mathbf{A} \geq 0$ )	matrix $\mathbf{A}$ is positive (semi)definite
$\mathbf{A} \geq \mathbf{B}$	matrix $\mathbf{A} - \mathbf{B}$ is positive semidefinite
$\nabla g$	$\left(\frac{\partial}{\partial x_1} g, \dots, \frac{\partial}{\partial x_k} g\right)^t$ vector of partial derivatives (gradient) of $g : A \subset \mathbb{R}^k \rightarrow \mathbb{R}$
$\mathbf{H}_g$	$\left(\frac{\partial^2}{\partial x_i \partial x_j} g\right)_{1 \leq i, j \leq k}$ Hessian matrix of $g : A \subset \mathbb{R}^k \rightarrow \mathbb{R}$
$\mathbf{D}_g$	$(\nabla g_1, \dots, \nabla g_n)^t$ Jacobian matrix of $g = (g_1, \dots, g_n) : A \subset \mathbb{R}^k \rightarrow \mathbb{R}^n$
$(\mathcal{X}, \mathcal{B})$	measurable space of observations sample space $\mathcal{X}$ endowed with $\sigma$ -algebra $\mathcal{B}$
$T : (\mathcal{X}, \mathcal{B}) \rightarrow (\mathcal{Y}, \mathcal{C})$	mapping $T : \mathcal{X} \rightarrow \mathcal{Y}$ is $(\mathcal{B} - \mathcal{C})$ -measurable
$\mathbb{B}^k$	Borel sets of $\mathbb{R}^k$
$\mu_{\mathcal{X}}$	counting measure on the power set of $\mathcal{X}$
$\lambda_{\mathcal{X}}$	Lebesgue measure on the Borel sets of $\mathcal{X}$

$\text{poi}(\bullet)$	$\{\text{poi}(a) : a \in (0, \infty)\}$ family of Poisson distributions
$\text{b}(n_0, \bullet)$	$\{\text{b}(n_0, p) : p \in (0, 1)\}$ family of binomial distributions with $n_0 \in \mathbb{N}$ fixed
$\text{nb}(r_0, \bullet)$	$\{\text{nb}(r_0, p) : p \in (0, 1)\}$ family of negative binomial distributions with $r_0 \in \mathbb{N}$ fixed
$\text{geo}(\bullet)$	$\{\text{nb}(1, p) : p \in (0, 1)\}$ family of geometric distributions
$\text{kn}(\bullet, \bullet)$	$\{\text{kn}(\alpha, p) : \alpha \in (0, \infty), p \in (0, 1)\}$ family of Kemp discrete normal distributions
$\text{g}(\bullet, \bullet)$	$\{\text{g}(p, \tau) : p \in (0, 1), \tau \in \mathbb{R}\}$ family of Good distributions
$\text{m}(n_0, \bullet)$	$\{\text{m}(n_0, \mathbf{p}) : \mathbf{p} = (p_1, \dots, p_k) \in (0, 1)^k, \sum_{i=1}^k p_i = 1\}$ family of multinomial distributions with $n_0 \in \mathbb{N}$ fixed
$\text{nm}(r_0, \bullet)$	$\{\text{nm}(r_0, \mathbf{p}) : \mathbf{p} = (p_1, \dots, p_k) \in (0, 1)^k, \sum_{i=1}^k p_i < 1\}$ family of negative multinomial distributions with $r_0 \in \mathbb{N}$ fixed
$\text{mps}(h, \bullet)$	$\{\text{mps}(h, \boldsymbol{\vartheta}) : \boldsymbol{\vartheta} \in \Theta\}$ family of multivariate power series distributions with $\Theta \subset (0, \infty)^k$ and function $h$ fixed
$N(\bullet, \bullet)$	$\{N(\mu, \sigma^2) : \mu \in \mathbb{R}, \sigma^2 \in (0, \infty)\}$ family of univariate normal distributions
$N(\bullet, \sigma_0^2)$	$\{N(\mu, \sigma_0^2) : \mu \in \mathbb{R}\}$ family of univariate normal distributions with $\sigma_0^2 \in (0, \infty)$ fixed
$N(\mu_0, \bullet)$	$\{N(\mu_0, \sigma^2) : \sigma^2 \in (0, \infty)\}$ family of univariate normal distributions with $\mu_0 \in \mathbb{R}$ fixed
$N_k(\bullet, \boldsymbol{\Sigma}_0)$	$\{N_k(\boldsymbol{\mu}, \boldsymbol{\Sigma}_0) : \boldsymbol{\mu} \in \mathbb{R}^k\}$ family of multivariate normal distributions with $\boldsymbol{\Sigma}_0 \in \mathbb{R}^{k \times k}$ fixed
$IN(\bullet, \bullet)$	$\{IN(\mu, \eta) : \mu, \eta \in (0, \infty)\}$ family of inverse normal distributions
$\mathcal{G}(\bullet, \bullet)$	$\{\mathcal{G}(\alpha, \beta) : \alpha, \beta \in (0, \infty)\}$ family of gamma distributions
$\mathcal{E}(\bullet)$	$\{\mathcal{G}(\alpha, 1) : \alpha \in (0, \infty)\}$ family of exponential distributions
$\mathcal{D}(\bullet)$	$\{\mathcal{D}(\boldsymbol{\beta}) : \boldsymbol{\beta} \in (0, \infty)^{k+1}\}$ family of Dirichlet distributions
$\mathcal{B}(\bullet, \bullet)$	$\{\mathcal{D}(\beta_1, \beta_2) : \beta_1, \beta_2 \in (0, \infty)\}$ family of beta distributions
$\mathcal{L}(h, \bullet)$	$\{\mathcal{L}(h, \mathbf{a}) : \mathbf{a} \in (0, \infty)^k\}$ family of Liouville distributions with function $h$ fixed

# Chapter 1

## Introduction: Aims and Outline



The study of parametric families of probability distributions is a main foundation of statistical modeling and inference. Although some probability distributions may be quite different at first sight with respect to their density functions, they can be observed to share a similar structure and common nice properties. Such findings always motivate to look for explanations and to examine the reasons behind the similarities.

Let us take a look at three distributions to illustrate a starting point for the topic of exponential families.

A basic discrete, one-parameter distribution is the binomial distribution with counting density function

$$f_p(x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x \in \{0, 1, \dots, n\}, \quad (1.1)$$

for some fixed natural number  $n$  and parameter  $p \in (0, 1)$ .

An important continuous two-parameter distribution is the gamma distribution with Lebesgue density function

$$f_{\alpha, \beta}(x) = \frac{1}{\Gamma(\beta) \alpha^\beta} x^{\beta-1} \exp\left\{-\frac{x}{\alpha}\right\}, \quad x \in (0, \infty), \quad (1.2)$$

and parameters  $\alpha \in (0, \infty)$  and  $\beta \in (0, \infty)$ .

As a third distribution, we consider a multivariate normal distribution with Lebesgue density function

$$f_{\boldsymbol{\mu}}(\mathbf{x}) = \frac{1}{(2\pi)^{k/2}} \exp\left\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^t(\mathbf{x} - \boldsymbol{\mu})\right\}, \quad \mathbf{x} \in \mathbb{R}^k, \quad (1.3)$$

for  $k \in \mathbb{N}$ , its covariance matrix given by the  $k$ -dimensional unity matrix, and parameter vector  $\boldsymbol{\mu} = (\mu_1, \dots, \mu_k)^t \in \mathbb{R}^k$ .

These distributions are different with respect to their support, dimension, and number of parameters.

On the other hand, by rewriting the density functions in formulas (1.1)–(1.3) according to

$$f_p(x) = (1-p)^n \exp \left\{ \ln \left( \frac{p}{1-p} \right) x \right\} \binom{n}{x}, \quad (1.4)$$

$$f_{\alpha,\beta}(x) = \frac{1}{\Gamma(\beta) \alpha^\beta} \exp \left\{ -\frac{x}{\alpha} + \beta \ln(x) \right\} \frac{1}{x}, \quad (1.5)$$

and

$$f_{\boldsymbol{\mu}}(\mathbf{x}) = \exp \left\{ -\frac{\boldsymbol{\mu}^t \mathbf{x}}{2} \right\} \exp \left\{ \sum_{j=1}^k \mu_j x_j \right\} \frac{\exp\{-\mathbf{x}^t \mathbf{x}/2\}}{(2\pi)^{k/2}}, \quad (1.6)$$

all of the density functions exhibit a common multiplicative structure

$$f_{\vartheta}(x) = C(\vartheta) \exp \left\{ \sum_{j=1}^k Z_j(\vartheta) T_j(x) \right\} h(x), \quad x \in \mathcal{X}, \quad (1.7)$$

where  $\vartheta$  denotes the respective parameter (vector) and  $\mathcal{X}$  is the support of the distribution. Here, the first factor  $C(\vartheta)$  is a normalizing constant depending only on  $\vartheta$ . In the exponent of the second factor,  $Z_1(\vartheta), \dots, Z_k(\vartheta)$  are constants depending only on  $\vartheta$ , and  $T_1, \dots, T_k$  are mappings on  $\mathcal{X}$ , which are free of  $\vartheta$  (with  $k = 1$  in formula (1.4) and  $k = 2$  in formula (1.5)). The third factor  $h$  is a function on  $\mathcal{X}$  not depending on  $\vartheta$ .

Distribution families with density functions of the form (1.7) are called *exponential families*. Having seen that the aforementioned distributions constitute exponential families (see formulas (1.4)–(1.6)), several questions arise at this point, such as the following:

- Are there other distributions (uni- and multivariate, one- and multiparameter distributions) possessing exponential family structure?
- What are special properties of an exponential family, and are there advantages when having statistical inference in mind?
- What is the use of a unified approach and a general theory of exponential families?

It will turn out that the exponential family structure comprises many well known distributions and that there are major benefits when considering exponential families of distributions in statistical inference. Various properties are found to hold true

under fairly general assumptions on the underlying exponential family, which gives rise for a unified approach.

Exponential families are discussed in textbooks and monographs on Mathematical Statistics, usually. There are respective examples, recurring examples or chapters, such as in Bickel and Doksum [16], Ferguson [24], Keener [35], Kotz et al. [37, Chapter 54 by M. Casalis], Lehmann and Casella [39], Lehmann and Romano [40], Liese and Miescke [42], Lindsey [44], Pfanzagl [50], Shao [55] and Witting [63] or books dedicated to exponential families, such as the fundamental accounts of Brown [20], Jørgensen [29], Letac [41], Barndorff-Nielsen [4], and Sundberg [56]. In particular, the books of Witting [63] and Brown [20] influenced the present text on exponential families.

In the sequel, we focus on properties of multivariate and multiparameter exponential families along with a variety of detailed examples as well as on the use of exponential families in parameter estimation and hypotheses testing, mainly based on the maximum-likelihood principle. Nevertheless, the present introductory material also serves as a solid basis for other methods of statistical inference within exponential families.

Chapters 2–5 offer a widely self-contained, mathematically rigorous and concise course on exponential families in a systematic structure for Master- and Ph.D.-students. The text can be considered an introduction to Mathematical Statistics restricted to using exponential families. As prerequisites, the reader should have a solid knowledge of calculus, linear algebra, and of an introduction to probability and measure theory. For the chapter on hypotheses testing, a former introductory course on statistics may be helpful.

In Chaps. 2 and 3 on parametrizations, distributional and statistical properties of exponential families, the proofs of assertions are not detailed; however, ideas and references are given, throughout. Thereafter, in Chaps. 4 and 5 on parameter estimation and hypotheses testing within exponential families, proofs and derivations are worked out. In Chap. 6, three multiparameter and multivariate families of distributions are exemplarily examined in order to utilize and illustrate the results on exponential families, and, by this, the benefit of a general approach is demonstrated.

In the present concise introduction to multiparameter exponential families, several topics have been left out, such as the field of variance functions, cuts, advanced structural results, confidence regions, others than maximum-likelihood approaches, and Bayesian inference. For further reading on the theory and applications of exponential families, we refer to, for example, Brown [20], Lehmann and Casella [39], and Liese and Miescke [42] regarding Bayesian methods, Brown [20], Letac [41], Barndorff-Nielsen [4], Witting [63], Kotz et al. [37, Chapter 54] and Liese and Mieschke [42] regarding structural considerations and statistical inference, as well as to the recent monograph Sundberg [56] highlighting different fields of applications in statistics.

# Chapter 2

## Parametrizations and Basic Properties



### 2.1 Definition and Examples

Motivated by the examples in the introduction, we start with the formal definition of an exponential family by specifying the form of a probability density function.

**Definition 2.1** Let  $(\mathcal{X}, \mathcal{B})$  be a measurable space and  $\mathcal{P} = \{P_{\vartheta} : \vartheta \in \Theta\}$  be a parametric family of probability distributions on  $\mathcal{B}$ , where  $\Theta \neq \emptyset$  denotes a set of real parameters or parameter vectors. If there exist

- a  $\sigma$ -finite measure  $\mu$  on  $\mathcal{B}$  dominating  $\mathcal{P}$ ,
- mappings  $C, Z_1, \dots, Z_k : \Theta \rightarrow \mathbb{R}$  and
- mappings  $h, T_1, \dots, T_k : (\mathcal{X}, \mathcal{B}) \rightarrow (\mathbb{R}^1, \mathbb{B}^1)$  with  $h \geq 0$

for some integer  $k \in \mathbb{N}$  in such a way that

$$f_{\vartheta}(x) = \frac{dP_{\vartheta}}{d\mu}(x) = C(\vartheta) \exp \left\{ \sum_{j=1}^k Z_j(\vartheta) T_j(x) \right\} h(x), \quad x \in \mathcal{X}, \quad (2.1)$$

is a  $\mu$ -density of  $P_{\vartheta}$  for every  $\vartheta \in \Theta$ , then  $\mathcal{P}$  is said to form an *exponential family* (EF).

By the definition of the mappings involved in formula (2.1), the functions  $C, Z_1, \dots, Z_k$  do not depend on  $x \in \mathcal{X}$ , and the functions  $h, T_1, \dots, T_k$  do not depend on  $\vartheta \in \Theta$ .

It is clear from the definition that an EF may have various different representations. The mappings  $C, Z_1, \dots, Z_k$  and  $h, T_1, \dots, T_k$  as well as the number  $k$  and the dominating measure  $\mu$  in formula (2.1) are not uniquely determined. For instance, for some  $j \in \{1, \dots, k\}$ , we may replace  $Z_j(\vartheta)$  by  $Z_j(\vartheta) - a$  and  $h$  by  $h \exp\{aT_j\}$  for some constant  $a \in \mathbb{R}$  to obtain another representation of the densities. Likewise, a shift of  $T_j$  and a respective modification of  $C$  gives a new representation. Moreover, although being artificial, setting  $Z_{k+1} = 0$  or  $T_{k+1} = 0$  yields another representation with  $k + 1$  summands in the exponent.

Since  $\int f_\vartheta d\mu = 1$  for all  $\vartheta \in \Theta$ ,  $C(\vartheta)$  is just a normalizing constant.

**Lemma 2.2** *In the situation of Definition 2.1, we have for every  $\vartheta \in \Theta$*

$$0 < C(\vartheta) = \left[ \int \exp \left\{ \sum_{j=1}^k Z_j(\vartheta) T_j(x) \right\} h(x) d\mu(x) \right]^{-1} < \infty.$$

Many well known univariate and multivariate distributions form EFs when varying one or more of their parameters. Here, the dominating measure  $\mu$  is usually the counting measure (in the discrete case) or the Lebesgue measure (in the continuous case). We show some prominent examples; in particular, the distributions in the introduction turn out to form EFs when varying the parameters. Throughout the text, superscript  $t$  at a vector or matrix denotes transposition.

**Examples:**  $\text{poi}(\bullet)$ ,  $\text{b}(n_0, \bullet)$ ,  $\text{nb}(r_0, \bullet)$ ,  $\text{geo}(\bullet)$ ,  $\text{fn}(\bullet, \bullet)$ ,  $\text{g}(\bullet, \bullet)$ ,  $\text{m}(n_0, \bullet)$

In the following examples and throughout, let  $\mu_X$  denote the counting measure on the power set of  $X$ .

*Example 2.3 (Poisson Distribution)* The counting density of the Poisson distribution  $\text{poi}(a)$ ,  $a \in (0, \infty)$ , is given by

$$f_a(x) = e^{-a} \frac{a^x}{x!} = e^{-a} e^{\ln(a)x} \frac{1}{x!}, \quad x \in \mathbb{N}_0.$$

Hence,  $\text{poi}(\bullet) = \{\text{poi}(a) : a \in (0, \infty)\}$  forms an EF with representation (2.1) given by  $\mu = \mu_{\mathbb{N}_0}$ ,  $k = 1$ ,

$$C(a) = e^{-a}, \quad Z_1(a) = \ln(a), \quad a \in (0, \infty),$$

$$\text{and} \quad T_1(x) = x, \quad h(x) = \frac{1}{x!}, \quad x \in \mathbb{N}_0.$$

*Example 2.4 (Binomial Distribution)* The counting density of the binomial distribution  $\mathfrak{b}(n, p)$ ,  $n \in \mathbb{N}$ ,  $p \in (0, 1)$ , is given by

$$\begin{aligned} f_{n,p}(x) &= \binom{n}{x} p^x (1-p)^{n-x} \\ &= (1-p)^n \exp \left\{ \ln \left( \frac{p}{1-p} \right) x \right\} \binom{n}{x}, \quad x \in \{0, 1, \dots, n\} = \mathcal{X}. \end{aligned}$$

Hence, for fixed  $n = n_0 \in \mathbb{N}$ ,  $\mathfrak{b}(n_0, \bullet) = \{\mathfrak{b}(n_0, p) : p \in (0, 1)\}$  forms an EF with representation (2.1) given by  $\mu = \mu_{\mathcal{X}}$ ,  $k = 1$ ,

$$C(p) = (1-p)^{n_0}, \quad Z_1(p) = \ln \left( \frac{p}{1-p} \right), \quad p \in (0, 1),$$

$$\text{and} \quad T_1(x) = x, \quad h(x) = \binom{n_0}{x}, \quad x \in \mathcal{X}.$$

*Example 2.5 (Negative Binomial Distribution)* The counting density of the negative binomial distribution  $\mathfrak{nb}(r, p)$ ,  $r \in \mathbb{N}$ ,  $p \in (0, 1)$ , is given by

$$\begin{aligned} f_{r,p}(x) &= \binom{x+r-1}{x} p^r (1-p)^x \\ &= p^r e^{\ln(1-p)x} \binom{x+r-1}{x}, \quad x \in \mathbb{N}_0. \end{aligned}$$

Hence, for fixed  $r = r_0 \in \mathbb{N}$ ,  $\mathfrak{nb}(r_0, \bullet) = \{\mathfrak{nb}(r_0, p) : p \in (0, 1)\}$  forms an EF with representation (2.1) given by  $\mu = \mu_{\mathbb{N}_0}$ ,  $k = 1$ ,

$$C(p) = p^{r_0}, \quad Z_1(p) = \ln(1-p), \quad p \in (0, 1),$$

$$\text{and} \quad T_1(x) = x, \quad h(x) = \binom{x+r_0-1}{x}, \quad x \in \mathbb{N}_0.$$

In particular, for  $r_0 = 1$ , the EF  $\mathfrak{geo}(\bullet) = \mathfrak{nb}(1, \bullet)$  of *geometric distributions* results.



*Example 2.6 (Kemp Discrete Normal Distribution)* The counting density of the Kemp discrete normal distribution  $\mathfrak{k}n(\alpha, p)$ ,  $\alpha \in (0, \infty)$ ,  $p \in (0, 1)$ , is given by

$$\begin{aligned} f_{\alpha, p}(x) &= \left( \sum_{j=-\infty}^{\infty} \alpha^j p^{j(j-1)/2} \right)^{-1} \alpha^x p^{x(x-1)/2} \\ &= \left( \sum_{j=-\infty}^{\infty} \alpha^j p^{j(j-1)/2} \right)^{-1} \exp \left\{ \ln(\alpha)x + \ln(p) \frac{x(x-1)}{2} \right\}, \quad x \in \mathbb{Z}. \end{aligned}$$

Hence,  $\mathfrak{k}n(\bullet, \bullet) = \{\mathfrak{k}n(\alpha, p) : \alpha \in (0, \infty), p \in (0, 1)\}$  forms an EF with representation (2.1) given by  $\mu = \mu_{\mathbb{Z}}$ ,  $k = 2$ ,

$$C(\alpha, p) = \left( \sum_{j=-\infty}^{\infty} \alpha^j p^{j(j-1)/2} \right)^{-1}$$

$$Z_1(\alpha, p) = \ln(\alpha), \quad Z_2(\alpha, p) = \ln(p), \quad \alpha \in (0, \infty), \quad p \in (0, 1),$$

$$\text{and} \quad T_1(x) = x, \quad T_2(x) = \frac{x(x-1)}{2}, \quad h(x) = 1, \quad x \in \mathbb{Z}.$$

*Example 2.7 (Good Distribution)* The counting density of the Good distribution  $\mathfrak{g}(p, \tau)$ ,  $p \in (0, 1)$ ,  $\tau \in \mathbb{R}$ , is given by

$$\begin{aligned} f_{p, \tau}(x) &= \left( \sum_{j=1}^{\infty} p^j j^{\tau} \right)^{-1} p^x x^{\tau} \\ &= \left( \sum_{j=1}^{\infty} p^j j^{\tau} \right)^{-1} \exp \{ \ln(p)x + \tau \ln(x) \}, \quad x \in \mathbb{N}. \end{aligned}$$

Hence,  $\mathfrak{g}(\bullet, \bullet) = \{\mathfrak{g}(p, \tau) : p \in (0, 1), \tau \in \mathbb{R}\}$  forms an EF with representation (2.1) given by  $\mu = \mu_{\mathbb{N}}$ ,  $k = 2$ ,

$$C(p, \tau) = \left( \sum_{j=1}^{\infty} p^j j^{\tau} \right)^{-1}$$

$$Z_1(p, \tau) = \ln(p), \quad Z_2(p, \tau) = \tau, \quad p \in (0, 1), \quad \tau \in \mathbb{R},$$

$$\text{and} \quad T_1(x) = x, \quad T_2(x) = \ln(x), \quad h(x) = 1, \quad x \in \mathbb{N}.$$

*Example 2.8 (Multinomial Distribution)* The counting density of the multinomial distribution  $\mathfrak{m}(n, \mathbf{p})$ ,  $n \in \mathbb{N}$ ,  $\mathbf{p} = (p_1, \dots, p_k)^t \in \Theta = \{(q_1, \dots, q_k)^t \in (0, 1)^k : \sum_{i=1}^k q_i = 1\}$ , where  $k \geq 2$ , is given by

$$\begin{aligned} f_{n,\mathbf{p}}(\mathbf{x}) &= \frac{n!}{\prod_{i=1}^k x_i!} \left( \prod_{j=1}^k p_j^{x_j} \right) \\ &= \exp \left\{ \sum_{j=1}^k \ln(p_j) x_j \right\} \frac{n!}{\prod_{i=1}^k x_i!} \end{aligned}$$

for  $\mathbf{x} = (x_1, \dots, x_k)^t \in \mathcal{X} = \{(y_1, \dots, y_k)^t \in \mathbb{N}_0^k : \sum_{i=1}^k y_i = n\}$ . Hence, for fixed  $n = n_0 \in \mathbb{N}$ ,  $\mathfrak{m}(n_0, \bullet) = \{\mathfrak{m}(n_0, \mathbf{p}) : \mathbf{p} \in \Theta\}$  forms an EF with representation (2.1) given by  $\mu = \mu_{\mathcal{X}}$ ,

$$C(\mathbf{p}) = 1, \quad Z_j(\mathbf{p}) = \ln(p_j), \quad \mathbf{p} \in \Theta,$$

$$\text{and} \quad T_j(\mathbf{x}) = x_j, \quad h(\mathbf{x}) = \frac{n_0!}{\prod_{i=1}^k x_i!}, \quad \mathbf{x} \in \mathcal{X}, \quad 1 \leq j \leq k.$$

More examples of discrete multivariate EFs can be found in [28, pp. 12–16] including, among others, the family of multivariate power series distributions. The negative multinomial distribution, which forms a particular multivariate power series distribution, is studied in Sect. 6.1.

*Remark 2.9* A general class of discrete multivariate distributions is given by *multivariate power series distributions* with counting densities of the form

$$f_{\boldsymbol{\vartheta}}(\mathbf{x}) = C(\boldsymbol{\vartheta}) h(\mathbf{x}) \prod_{j=1}^k \vartheta_j^{x_j}, \quad \mathbf{x} = (x_1, \dots, x_k)^t \in \mathcal{X} \subset \mathbb{N}_0^k, \quad (2.2)$$

where  $h$  is an appropriate function,  $\boldsymbol{\vartheta} = (\vartheta_1, \dots, \vartheta_k)^t \in \Theta \subset (0, \infty)^k$  is a parameter vector, and  $C(\boldsymbol{\vartheta})$  is a normalizing constant (see [28, Chapter 38]). We denote the multivariate power series distribution with density function (2.2) by  $\text{mps}(h, \boldsymbol{\vartheta})$ . For fixed function  $h$ ,  $\text{mps}(h, \bullet) = \{\text{mps}(h, \boldsymbol{\vartheta}) : \boldsymbol{\vartheta} \in \Theta\}$  forms an EF.

Examples of particular distributions contained in the class of multivariate power series distributions are the multinomial distribution (see Example 2.8), the negative multinomial distribution (see Sect. 6.1), the multivariate logarithmic series distribution ( $C(\vartheta) = -1/\ln(1 - \sum_{j=1}^k \vartheta_j)$ ,  $\vartheta \in (0, 1)^k$  with  $\sum_{j=1}^k \vartheta_j < 1$ , and  $h(\mathbf{x}) = (\sum_{j=1}^k x_j - 1)! / \prod_{j=1}^k x_j!$ ,  $\mathbf{x} \in \mathbb{N}_0^k \setminus \{(0, \dots, 0)^t\}$ ), and the joint distribution of  $k$  independent Poisson-distributed random variables (see Example 2.3).

**Examples:**  $\mathcal{N}(\bullet, \bullet)$ ,  $\mathcal{N}(\mu_0, \bullet)$ ,  $\mathcal{N}(\bullet, \sigma_0^2)$ ,  $\mathcal{IN}(\bullet, \bullet)$ ,  $\mathcal{G}(\bullet, \bullet)$ ,  $\mathcal{E}(\bullet)$ ,  $\mathcal{D}(\bullet)$ ,  $\mathcal{B}(\bullet, \bullet)$

In the following examples and throughout, let  $\lambda_{\mathcal{X}}$  denote the Lebesgue measure on the Borel sets of  $\mathcal{X}$ .

*Example 2.10 (Normal Distribution)* The Lebesgue density of the normal distribution  $\mathcal{N}(\mu, \sigma^2)$  with mean  $\mu \in \mathbb{R}$  and variance  $\sigma^2 \in (0, \infty)$  is given by

$$\begin{aligned} f_{\mu, \sigma^2}(x) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} \\ &= \frac{1}{\sqrt{\sigma^2}} \exp\left\{-\frac{\mu^2}{2\sigma^2}\right\} \exp\left\{\frac{\mu x}{\sigma^2} - \frac{x^2}{2\sigma^2}\right\} \frac{1}{\sqrt{2\pi}}, \quad x \in \mathbb{R}. \end{aligned}$$

Hence,  $\mathcal{N}(\bullet, \bullet) = \{\mathcal{N}(\mu, \sigma^2) : \mu \in \mathbb{R}, \sigma^2 \in (0, \infty)\}$  forms an EF with representation (2.1) given by  $\mu = \lambda_{\mathbb{R}}$ ,  $k = 2$ ,

$$C(\mu, \sigma^2) = \frac{1}{\sqrt{\sigma^2}} \exp\left\{-\frac{\mu^2}{2\sigma^2}\right\},$$

$$Z_1(\mu, \sigma^2) = \frac{\mu}{\sigma^2}, \quad Z_2(\mu, \sigma^2) = -\frac{1}{2\sigma^2}, \quad \mu \in \mathbb{R}, \sigma^2 \in (0, \infty),$$

$$\text{and} \quad T_1(x) = x, \quad T_2(x) = x^2, \quad h(x) = \frac{1}{\sqrt{2\pi}}, \quad x \in \mathbb{R}.$$

Moreover, when fixing either  $\mu = \mu_0$  or  $\sigma^2 = \sigma_0^2$ , the classes  $\mathcal{N}(\mu_0, \bullet) = \{\mathcal{N}(\mu_0, \sigma^2) : \sigma^2 \in (0, \infty)\}$  and  $\mathcal{N}(\bullet, \sigma_0^2) = \{\mathcal{N}(\mu, \sigma_0^2) : \mu \in \mathbb{R}\}$  are easily seen to form one-parameter EFs.

By choosing  $\alpha = \exp\{-(1 - 2\mu)/(2\sigma^2)\}$  and  $p = \exp\{-1/\sigma^2\}$  in Example 2.6 (cf. [26, p. 475]), the Kemp distribution  $\mathfrak{kn}(\alpha, p)$  is seen to be a discrete version of the normal distribution  $\mathcal{N}(\mu, \sigma^2)$ .

*Example 2.11 (Inverse Normal Distribution)* The Lebesgue density of the inverse normal distribution  $\mathcal{IN}(\mu, \eta)$ ,  $\mu, \eta \in (0, \infty)$ , is given by

$$\begin{aligned} f_{\mu, \eta}(x) &= \sqrt{\frac{\eta}{2\pi x^3}} \exp\left\{-\frac{\eta(x - \mu)^2}{2\mu^2 x}\right\} \\ &= \sqrt{\eta} \exp\left\{\frac{\eta}{\mu}\right\} \exp\left\{-\frac{\eta x}{2\mu^2} - \frac{\eta}{2x}\right\} \sqrt{\frac{1}{2\pi x^3}}, \quad x \in (0, \infty). \end{aligned}$$

Hence,  $\mathcal{IN}(\bullet, \bullet) = \{\mathcal{IN}(\mu, \eta) : \mu, \eta \in (0, \infty)\}$  forms an EF with representation (2.1) given by  $\mu = \lambda_{(0, \infty)}$ ,  $k = 2$ ,

$$C(\mu, \eta) = \sqrt{\eta} \exp\left\{\frac{\eta}{\mu}\right\},$$

$$Z_1(\mu, \eta) = -\frac{\eta}{2\mu^2}, \quad Z_2(\mu, \eta) = -\frac{\eta}{2}, \quad \mu, \eta \in (0, \infty),$$

and  $T_1(x) = x, \quad T_2(x) = \frac{1}{x}, \quad h(x) = \sqrt{\frac{1}{2\pi x^3}}, \quad x \in (0, \infty).$

*Example 2.12 (Gamma Distribution)* The Lebesgue density of the gamma distribution  $\mathcal{G}(\alpha, \beta)$ ,  $\alpha, \beta \in (0, \infty)$ , is given by

$$\begin{aligned} f_{\alpha, \beta}(x) &= \frac{1}{\Gamma(\beta) \alpha^\beta} x^{\beta-1} \exp\left\{-\frac{x}{\alpha}\right\} \\ &= \frac{1}{\Gamma(\beta) \alpha^\beta} \exp\left\{-\frac{x}{\alpha} + \beta \ln(x)\right\} \frac{1}{x}, \quad x \in (0, \infty). \end{aligned}$$

Hence,  $\mathcal{G}(\bullet, \bullet) = \{\mathcal{G}(\alpha, \beta) : \alpha, \beta \in (0, \infty)\}$  forms an EF with representation (2.1) given by  $\mu = \lambda_{(0, \infty)}$ ,  $k = 2$ ,

$$C(\alpha, \beta) = \frac{1}{\Gamma(\beta) \alpha^\beta}, \quad Z_1(\alpha, \beta) = -\frac{1}{\alpha}, \quad Z_2(\alpha, \beta) = \beta, \quad \alpha, \beta \in (0, \infty),$$

and  $T_1(x) = x, \quad T_2(x) = \ln(x), \quad h(x) = \frac{1}{x}, \quad x \in (0, \infty).$

Fixing  $\beta = 1$ , the resulting *exponential distributions*  $\mathcal{E}(\alpha)$ ,  $\alpha \in (0, \infty)$ , form the one-parameter EF  $\mathcal{E}(\bullet) = \{\mathcal{G}(\alpha, 1) : \alpha \in (0, \infty)\}$  with representation (2.1) given by  $\mu = \lambda_{(0, \infty)}$ ,  $k = 1$ ,

$$C(\alpha) = \frac{1}{\alpha}, \quad Z_1(\alpha) = -\frac{1}{\alpha}, \quad \alpha \in (0, \infty),$$

$$\text{and} \quad T_1(x) = x, \quad h(x) = 1, \quad x \in (0, \infty).$$

By choosing  $\alpha = 2$  and  $\beta = n/2$ , we arrive at the EF  $\{\chi^2(n) : n \in \mathbb{N}\} = \{\mathcal{G}(2, n/2) : n \in \mathbb{N}\}$  of *chi-square distributions*.

For  $p = \exp\{-1/\alpha\}$  and  $\tau = \beta - 1$ , the Good distribution  $g(p, \tau)$  in Example 2.7 (cf. [26, p. 530]) is seen to be a discrete version of the gamma distribution  $\mathcal{G}(\alpha, \beta)$ .

**Example 2.13 (Dirichlet Distribution)** The Lebesgue density of the Dirichlet distribution  $\mathcal{D}(\boldsymbol{\beta})$ ,  $\boldsymbol{\beta} = (\beta_1, \dots, \beta_{k+1})^t \in (0, \infty)^{k+1}$ , is given by

$$\begin{aligned} f_{\boldsymbol{\beta}}(\mathbf{x}) &= \frac{1}{B(\boldsymbol{\beta})} \left( \prod_{j=1}^k x_j^{\beta_j-1} \right) \left( 1 - \sum_{j=1}^k x_j \right)^{\beta_{k+1}-1} \\ &= \frac{1}{B(\boldsymbol{\beta})} \exp \left\{ \left( \sum_{j=1}^k \beta_j \ln(x_j) \right) + \beta_{k+1} \ln \left( 1 - \sum_{j=1}^k x_j \right) \right\} \left[ \left( 1 - \sum_{j=1}^k x_j \right) \prod_{j=1}^k x_j \right]^{-1} \end{aligned}$$

for  $\mathbf{x} = (x_1, \dots, x_k)^t \in \mathcal{X} = \{(y_1, \dots, y_k)^t \in (0, 1)^k : \sum_{i=1}^k y_i < 1\}$ , where  $B(\boldsymbol{\beta}) = [\prod_{i=1}^{k+1} \Gamma(\beta_i)] / \Gamma(\sum_{i=1}^{k+1} \beta_i)$  denotes the beta function evaluated at  $\boldsymbol{\beta}$ . Hence,  $\mathcal{D}(\bullet) = \{\mathcal{D}(\boldsymbol{\beta}) : \boldsymbol{\beta} \in (0, \infty)^{k+1}\}$  and, in particular, by setting  $k = 1$ , the family  $\mathcal{B}(\bullet, \bullet) = \{\mathcal{D}(\beta_1, \beta_2) : \beta_1, \beta_2 \in (0, \infty)\}$  of *beta distributions* are seen to form EFs. The EF of Dirichlet distributions is studied in Sect. 6.2.

Further continuous univariate EFs are given by the family of *log-normal distributions* and the family of *inverse gamma distributions*. Other examples of continuous multivariate distributions, which lead to multiparameter EFs, are multivariate normal distributions, particular multivariate Liouville distributions, and *Wishart distributions* (for the latter, see, e.g., [37, pp. 662/663] and [36, Section 2.4]).

**Remark 2.14** A general class of continuous multivariate distributions is given by so-called *Liouville distributions* with Lebesgue density functions proportional to

$$h \left( \sum_{j=1}^k x_j \right) \prod_{j=1}^k x_j^{a_j-1}, \quad x_1, \dots, x_k > 0, \quad (2.3)$$

where  $h$  is an appropriate non-negative function and  $a_1, \dots, a_k$  are positive parameters (see [37, Chapter 50] and [47, Section 8.5] for details). We denote the Liouville distribution with density function (2.3) by  $\mathcal{L}(h, (a_1, \dots, a_k)^t)$ . Obviously, for fixed  $h$ ,  $\mathcal{L}(h, \bullet) = \{\mathcal{L}(h, \mathbf{a}) : \mathbf{a} \in (0, \infty)^k\}$  then forms an EF. An EF structure

may also arise if  $h$  depends on additional parameters as the following two examples show.

For  $h(y) = y^{\beta-1} \exp\{-y/\alpha\}$ ,  $y > 0$ , with positive parameters  $\alpha$  and  $\beta$ , the resulting Liouville distribution (of the first kind) has density function

$$\frac{\Gamma(a_\bullet)}{\Gamma(a_\bullet + \beta - 1) \left( \prod_{j=1}^k \Gamma(a_j) \right) \alpha^{a_\bullet + \beta - 1}} \left( \sum_{j=1}^k x_j \right)^{\beta-1} \prod_{j=1}^k x_j^{a_j-1} e^{-x_j/\alpha}$$

for  $x_1, \dots, x_k > 0$ , where  $a_\bullet = \sum_{i=1}^k a_i$ . The particular case  $\beta = 1$  yields the joint density function of  $k$  independent gamma-distributed random variables with common scale parameter  $\alpha$  (see Example 2.12).

Choosing  $h(y) = (1 - y)^{\beta-1}$  for  $0 < y < 1$  (zero otherwise) and some  $\beta > 0$ , we obtain the Dirichlet distribution  $\mathcal{D}(a_1, \dots, a_k, \beta)$  defined on the open simplex  $\{(x_1, \dots, x_k)^t \in (0, 1)^k : \sum_{i=1}^k x_i < 1\}$  (see Example 2.13) as a particular Liouville distribution (of the second kind).

The function  $h$  in Definition 2.1 is free of  $\vartheta$  and, indeed, as we will later see, it does not play a role in statistical inference. As a  $(\mathcal{B} - \mathbb{B}^1)$ -measurable non-negative mapping,  $h$  may be considered a  $\mu$ -density of a measure  $\nu$ , i.e., we have  $\nu(A) = \int h \mathbb{1}_A d\mu$ ,  $A \in \mathcal{B}$ , where  $\mathbb{1}_A$  denotes the indicator function of the set  $A$ . For short, we write  $\nu = h\mu$  in that case. Since  $\mu$  is  $\sigma$ -finite and  $h$  is real-valued,  $\nu$  is then  $\sigma$ -finite as well (see, e.g., [6, pp. 104/105]). We may therefore absorb  $h$  into the dominating measure to obtain a shorter representation of the EF.

**Lemma 2.15** *Let  $\mathcal{P}$  be an EF with representation (2.1), and let  $\nu = h\mu$  be the measure with  $\mu$ -density  $h$ . Then, for every  $\vartheta \in \Theta$ , a  $\nu$ -density of  $P_\vartheta$  is given by*

$$g_\vartheta(x) = \frac{dP_\vartheta}{d\nu}(x) = C(\vartheta) \exp \left\{ \sum_{j=1}^k Z_j(\vartheta) T_j(x) \right\}, \quad x \in \mathcal{X}. \quad (2.4)$$

**Remark 2.16** In formula (2.4), we have  $g_\vartheta > 0$  for every  $\vartheta \in \Theta$ , which implies that  $\nu$  and  $P_\vartheta$  are equivalent measures, i.e.,  $\nu(N) = 0$  if and only if  $P_\vartheta(N) = 0$  is valid for every  $N \in \mathcal{B}$ . As a consequence, all distributions in an EF are necessarily equivalent. In particular, the *support* of  $P_\vartheta$ , defined by

$$\text{supp}(P_\vartheta) = \{x \in \mathcal{X} : f_\vartheta(x) > 0\}$$

for the discrete case and by

$$\text{supp}(P_\vartheta) = \{x \in \mathcal{X} : P_\vartheta(V) > 0 \text{ for all open sets } V \subset \mathcal{X} \text{ with } x \in V\}$$

for the continuous case, does not depend on  $\vartheta$ . By fixing some parameter  $\vartheta_0 \in \Theta$ , this allows for a further representation of the EF given by the  $P_{\vartheta_0}$ -densities

$$\frac{dP_{\vartheta}}{dP_{\vartheta_0}}(x) = \frac{C(\vartheta)}{C(\vartheta_0)} \exp \left\{ \sum_{j=1}^k (Z_j(\vartheta) - Z_j(\vartheta_0)) T_j(x) \right\}, \quad x \in \mathcal{X},$$

for  $\vartheta \in \Theta$ .

By Remark 2.16, a necessary condition for a family of distributions to form an EF is that all distributions have the same support. Counterexamples are therefore near at hand. For instance, the continuous uniform distribution on the interval  $[a, b]$ ,  $-\infty < a < b < \infty$ , does not form an EF when  $a$  or  $b$  are supposed to vary. Another counterexample is provided by the two-parameter exponential distribution with varying location parameter. Both distribution families are examples for so-called *truncated EFs*, which have also been studied in the literature. For a recent work on estimation in truncated EFs, we refer to [1].

For the sake of a short notation, we introduce the (column) vectors  $\mathbf{Z} = (Z_1, \dots, Z_k)^t$  and  $\mathbf{T} = (T_1, \dots, T_k)^t$  of parameter functions and statistics, respectively. Then, for  $\vartheta \in \Theta$  and  $x \in \mathcal{X}$ ,  $\mathbf{Z}(\vartheta)^t \mathbf{T}(x) = \sum_{j=1}^k Z_j(\vartheta) T_j(x)$  is the scalar product of  $\mathbf{Z}(\vartheta)$  and  $\mathbf{T}(x)$ .

Moreover, let  $P_{\vartheta}^T$  denote the distribution of  $\mathbf{T}$  under  $P_{\vartheta}$  being defined on the Borel sets  $\mathbb{B}^k$  of  $\mathbb{R}^k$ , i.e.,  $P_{\vartheta}^T(A) = P_{\vartheta}(\mathbf{T}^{-1}(A))$ ,  $A \in \mathbb{B}^k$ . By an integral transformation, it is easily seen that if  $\mathcal{P} = \{P_{\vartheta} : \vartheta \in \Theta\}$  forms an EF so do the image distributions  $P_{\vartheta}^T$ ,  $\vartheta \in \Theta$ , of  $\mathbf{T}$  (see, e.g., [39, pp. 31/32] or [63, p. 149]).

**Lemma 2.17** *Let  $\mathcal{P}$  be an EF with representation (2.4), and let  $\nu^T(A) = \nu(\mathbf{T}^{-1}(A))$ ,  $A \in \mathbb{B}^k$ , denote the image measure of  $\nu$  under  $\mathbf{T}$ . Then, for every  $\vartheta \in \Theta$ , a  $\nu^T$ -density of  $P_{\vartheta}^T$  is given by*

$$g_{\vartheta}^T(\mathbf{t}) = \frac{dP_{\vartheta}^T}{d\nu^T}(\mathbf{t}) = C(\vartheta) \exp \{ \mathbf{Z}(\vartheta)^t \mathbf{t} \}, \quad \mathbf{t} = (t_1, \dots, t_k)^t \in \mathbb{R}^k,$$

and  $\mathcal{P}^T = \{P_{\vartheta}^T : \vartheta \in \Theta\}$  forms an EF.

Note that the measure  $\nu^T$  in Lemma 2.17 is  $\sigma$ -finite, since  $0 < g_{\vartheta}^T < \infty$  and  $\int g_{\vartheta}^T d\nu^T = 1 < \infty$  for  $\vartheta \in \Theta$  (see, e.g., [6, pp. 98/99]).

**Example:**  $\mathcal{N}_k(\bullet, \Sigma_0)$

**Example 2.18 (Multivariate Normal Distribution)** The Lebesgue density of the multivariate normal distribution  $\mathcal{N}_k(\boldsymbol{\mu}, \Sigma)$  with mean vector  $\boldsymbol{\mu} \in \mathbb{R}^k$  and positive

definite covariance matrix  $\Sigma \in \mathbb{R}^{k \times k}$  is given by

$$f_{\mu, \Sigma}(x) = \frac{1}{(2\pi)^{k/2} \sqrt{\det(\Sigma)}} \exp \left\{ -\frac{1}{2} \|x - \mu\|_{\Sigma}^2 \right\}, \quad x \in \mathbb{R}^k,$$

$$\text{where} \quad \|y\|_{\Sigma} = \sqrt{y^t \Sigma^{-1} y}, \quad y \in \mathbb{R}^k,$$

denotes the Mahalanobis norm with weight matrix  $\Sigma^{-1}$ . For fixed  $\Sigma = \Sigma_0$ ,  $\mathcal{P} = \mathcal{N}_k(\bullet, \Sigma_0) = \{\mathcal{N}_k(\mu, \Sigma_0) : \mu \in \mathbb{R}^k\}$  forms an EF with representation (2.1) given by  $\mu = \lambda_{\mathbb{R}^k} = \lambda^k$ , say,

$$C(\mu) = \exp \left\{ -\frac{\|\mu\|_{\Sigma_0}^2}{2} \right\}, \quad Z(\mu) = \mu, \quad \mu \in \mathbb{R}^k,$$

$$\text{and} \quad T(x) = \Sigma_0^{-1} x, \quad h(x) = \frac{\exp \left\{ -\|x\|_{\Sigma_0}^2 / 2 \right\}}{(2\pi)^{k/2} \sqrt{\det(\Sigma_0)}}, \quad x \in \mathbb{R}^k.$$

Now, let  $\nu = h\lambda^k = \mathcal{N}_k(\mathbf{0}, \Sigma_0)$ . For  $A \in \mathbb{B}^k$ , an integral transformation then yields

$$\begin{aligned} \nu^T(A) &= \nu(T^{-1}(A)) = \int_{T^{-1}(A)} h(x) d\lambda^k(x) \\ &= \int_A h(\Sigma_0 u) |\det(\Sigma_0)| d\lambda^k(u) = \int_A f_{\mathbf{0}, \Sigma_0^{-1}}(u) d\lambda^k(u). \end{aligned}$$

Hence,  $\nu^T = \mathcal{N}_k(\mathbf{0}, \Sigma_0^{-1})$ . By Lemma 2.17, the distribution of  $T$  under  $P_{\mu}$  is given by

$$\begin{aligned} P_{\mu}^T(A) &= \int_A C(\mu) e^{Z(\mu)^t t} d\nu^T(t) \\ &= \int_A C(\mu) e^{\mu^t t} f_{\mathbf{0}, \Sigma_0^{-1}}(t) d\lambda^k(t) = \int_A f_{\Sigma_0^{-1}\mu, \Sigma_0^{-1}}(t) d\lambda^k(t), \end{aligned}$$

i.e.,  $P_{\mu}^T = \mathcal{N}_k(\Sigma_0^{-1}\mu, \Sigma_0^{-1})$ , for  $\mu \in \mathbb{R}^k$ . Hence,  $\mathcal{P}^T = \mathcal{N}_k(\bullet, \Sigma_0^{-1})$ , here.

In general, however, the dominating measure  $\nu^T$  of  $\mathcal{P}^T$  in Lemma 2.17 will have a complicated structure. For instance, in case of the EF  $\mathcal{N}(\bullet, \bullet)$  in Example 2.10, we have  $T(x) = (x, x^2)^t$ ,  $x \in \mathbb{R}$ , and, since  $\nu^T$  and  $P_{\theta}^T$  are equivalent by Remark 2.16,  $\text{supp}(\nu^T) = \{(x, x^2)^t : x \in \mathbb{R}\}$ , i.e., the support of  $\nu^T$  is given by the graph of the



standard parabola in the plane. In particular,  $\text{supp}(v^T)$  has no inner points, and  $v^T$  does not have a Lebesgue density.

*Remark 2.19* As the preceding examples show, the concept of the EF is rich and allows for a unifying study of various different distributions. It comprises univariate, one-parameter distributions, such as  $\text{poi}(\bullet)$  and  $\mathcal{E}(\bullet)$  shown in Examples 2.3 and 2.12, as well as multivariate, multiparameter distributions as, for instance,  $\text{m}(n_0, \bullet)$  and  $\mathcal{D}(\bullet)$  in Examples 2.8 and 2.13. Examples for univariate EFs with more than one parameter are provided by  $\text{fn}(\bullet, \bullet)$  and  $\mathcal{N}(\bullet, \bullet)$  in Examples 2.6 and 2.10. Finally, multivariate, one-parameter EFs appear in a natural way when sampling from a univariate, one-parameter EF (see Sect. 3.3); a particular example is the family  $\{\mathcal{N}_k((\mu, \dots, \mu)^t, \mathbf{E}_k) : \mu \in \mathbb{R}\}$ , where  $\mathbf{E}_k$  denotes the  $k$ -dimensional unit matrix.

Many of the aforementioned examples of EFs will be considered again and studied in more detail in the sequel. An overview is provided in Table 2.1.

**Table 2.1** EFs discussed in this book listed by type and number of parameters

Type	Parameters	EF	Discussed in . . .
Discrete	1	$\text{poi}(\bullet)$	Examples 2.3, 2.37, 2.64, 3.2, 3.14, 3.22, 3.31, 4.7, 4.28, 5.6
	1	$\text{b}(n_0, \bullet)$	Examples 2.4, 2.38, 2.57, 3.3, 3.22, 4.8, 4.28
	1	$\text{nb}(r_0, \bullet)$	Examples 2.5, 2.51, 3.22, 3.32, 4.28
	1	$\text{geo}(\bullet)$	Examples 2.5, 2.51, 3.22, 3.32, 4.28
	2	$\text{fn}(\bullet, \bullet)$	Example 2.6
	2	$\text{g}(\bullet, \bullet)$	Examples 2.7, 3.23
	$k (\geq 2)$	$\text{m}(n_0, \bullet)$	Examples 2.8, 2.46, 3.4, 3.24
	$k (\geq 2)$	$\text{nm}(r_0, \bullet)$	Section 6.1
	$k (\geq 2)$	$\text{mps}(h, \bullet)$	Remark 2.9
Continuous	1	$\mathcal{E}(\bullet)$	Examples 2.12, 3.10, 3.22, 4.28, 5.9
	1	$\mathcal{N}(\bullet, \sigma_0^2)$	Examples 2.10, 3.16, 5.7
	1	$\mathcal{N}(\mu_0, \bullet)$	Example 2.10
	2	$\mathcal{N}(\bullet, \bullet)$	Examples 2.10, 2.27, 2.31, 2.39, 2.52, 2.59, 3.5, 3.15, 3.25, 3.33, 4.9, 4.15
	2	$\mathcal{IN}(\bullet, \bullet)$	Examples 2.11, 2.25, 2.41, 2.58
	2	$\mathcal{G}(\bullet, \bullet)$	Examples 2.12, 2.24, 2.40, 2.53, 3.6, 3.23, 3.34, 4.24
	2	$\mathcal{B}(\bullet, \bullet)$	Example 2.13, Section 6.2
	$k (\geq 2)$	$\mathcal{N}_k(\bullet, \Sigma_0)$	Examples 2.18, 2.28, 2.54, 2.65, 3.7, 3.16, 3.26, 3.35, 3.41, 4.10, 4.25, 5.14, 5.18
	$k (\geq 2)$	$\mathcal{D}(\bullet)$	Example 2.13, Section 6.2
	$k (\geq 2)$	$\mathcal{L}(h, \bullet)$	Remark 2.14

## 2.2 Order and Minimal Representation

For both, mathematical and practical reasons, a representation (2.1) of the EF in Definition 2.1 is desired, where the number  $k$  of summands is minimum. On the one hand, this will ensure that the parameter  $\vartheta \in \Theta$  is identifiable meaning that different parameters correspond to different distributions in the EF. On the other hand, the vector  $\mathbf{T}$  of statistics, which plays an important role in statistical inference, should have a dimension as small as possible. First, we define what is meant by a minimal representation of the EF.

**Definition 2.20** Let  $\mathcal{P}$  be an EF with representation (2.1). Then,  $\mathcal{P}$  is said to have *order*  $k$  if the following implication holds true:

If  $\tilde{\mu}$  is a  $\sigma$ -finite measure on  $\mathcal{B}$ ,  $\tilde{k} \in \mathbb{N}$ , and if there are mappings  $\tilde{C}, \tilde{Z}_1, \dots, \tilde{Z}_{\tilde{k}} : \Theta \rightarrow \mathbb{R}$ , and  $\tilde{h}, \tilde{T}_1, \dots, \tilde{T}_{\tilde{k}} : (\mathcal{X}, \mathcal{B}) \rightarrow (\mathbb{R}^1, \mathbb{B}^1)$  with  $\tilde{h} \geq 0$  such that

$$\tilde{f}_{\vartheta}(x) = \frac{dP_{\vartheta}}{d\tilde{\mu}}(x) = \tilde{C}(\vartheta) \exp \left\{ \sum_{j=1}^{\tilde{k}} \tilde{Z}_j(\vartheta) \tilde{T}_j(x) \right\} \tilde{h}(x), \quad x \in \mathcal{X}, \quad (2.5)$$

is a  $\tilde{\mu}$ -density of  $P_{\vartheta}$  for every  $\vartheta \in \Theta$ , then  $\tilde{k} \geq k$ .

In that case, we write  $\text{ord}(\mathcal{P}) = k$ , and representation (2.1) is said to be *minimal*.

Definition 2.20 is not suitable to determine the order of an EF in a given situation. Aiming for necessary and sufficient conditions for an EF to have order  $k$ , we introduce affine independence of mappings.

**Definition 2.21** Let  $\mathcal{Y} \neq \emptyset$  be a set,  $k \in \mathbb{N}$ , and  $u_1, \dots, u_k : \mathcal{Y} \rightarrow \mathbb{R}$  be mappings.

(a)  $u_1, \dots, u_k$  are called *affinely independent* if, for any  $a_0, a_1, \dots, a_k \in \mathbb{R}$ ,

$$a_0 + \sum_{j=1}^k a_j u_j(y) = 0 \quad \text{for all } y \in \mathcal{Y}$$

implies that  $a_0 = a_1 = \dots = a_k = 0$ .

(continued)

**Definition 2.21** (continued)

- (b) Let  $(\mathcal{Y}, C)$  be a measurable space,  $Q$  be a measure on  $C$ , and  $u_1, \dots, u_k$  be  $(C - \mathbb{B}^1)$ -measurable. Then,  $u_1, \dots, u_k$  are called  *$Q$ -affinely independent* if  $\tilde{u}_1, \dots, \tilde{u}_k : N^c \rightarrow \mathbb{R}$  are affinely independent in the sense of (a) for every  $N \in C$  with  $Q(N) = 0$ .

The following lemma connects the concepts of affine independence of mappings and linear independence of vectors in  $\mathbb{R}^k$ . A proof can be found in the web appendix.

**Lemma 2.22** *Let  $\mathcal{Y} \neq \emptyset$  be a set,  $k \in \mathbb{N}$ , and  $u_1, \dots, u_k : \mathcal{Y} \rightarrow \mathbb{R}$  be mappings. Moreover, let  $\mathbf{u} = (u_1, \dots, u_k) : \mathcal{Y} \rightarrow \mathbb{R}^k$  and  $y_0 \in \mathcal{Y}$ . Then, the following assertions are equivalent:*

- (a)  $u_1, \dots, u_k$  are affinely independent.  
 (b) *There exist  $y_1, \dots, y_k \in \mathcal{Y}$  such that  $\mathbf{u}(y_1) - \mathbf{u}(y_0), \dots, \mathbf{u}(y_k) - \mathbf{u}(y_0)$  are linearly independent (in  $\mathbb{R}^k$ ).*

Given representation (2.4) of an EF, it is clear that if  $Z_1, \dots, Z_k$  are not affinely independent or if  $T_1, \dots, T_k$  are not  $\nu$ -affinely independent, the exponent in the densities can be reduced by a summand under respective modifications of  $C$  or  $\nu$ . Hence, affine independence of  $Z_1, \dots, Z_k$  and  $\nu$ -affine independence of  $T_1, \dots, T_k$  is a necessary condition for the EF to have order  $k$ . Indeed, the condition is also sufficient (see, e.g., [4, p. 113]). A detailed proof can be found, for instance, in [63, pp. 145/146].

**Theorem 2.23** *Let  $\mathcal{P}$  be an EF with representation (2.4). Then, the following assertions are equivalent:*

- (a)  $\text{ord}(\mathcal{P}) = k$ ,  
 (b)  $Z_1, \dots, Z_k$  are affinely independent and  $T_1, \dots, T_k$  are  $\nu$ -affinely independent.

Note that, by Remark 2.16,  $\nu$ -affine independence coincides with  $P_\vartheta$ -affine independence for every  $\vartheta \in \Theta$ .

**Examples:**  $\mathcal{G}(\bullet, \bullet)$ ,  $\mathcal{IN}(\bullet, \bullet)$ 

*Example 2.24* Consider the representation of  $\mathcal{G}(\bullet, \bullet)$  in Example 2.12 with  $Z_1(\alpha, \beta) = -1/\alpha$ ,  $Z_2(\alpha, \beta) = \beta$  for  $\alpha, \beta \in (0, \infty)$ , and  $T_1(x) = x$ ,  $T_2(x) = \ln(x)$  for  $x \in (0, \infty)$ . Let

$$a_0 - \frac{a_1}{\alpha} + a_2\beta = 0 \quad \text{for all } \alpha, \beta \in (0, \infty).$$

For  $\alpha \rightarrow \infty$  and  $\beta \rightarrow 0$ , this gives  $a_0 = 0$ . Then, choosing  $(\alpha, \beta) = (1, 1)$  and  $(\alpha, \beta) = (1, 2)$  yields that  $a_2 = 0$  and, thus,  $a_1 = 0$ . Hence,  $Z_1$  and  $Z_2$  are

affinely independent. Now, suppose that  $T_1$  and  $T_2$  are not  $\nu$ -affinely independent. Then, there exists a vector  $(a_0, a_1, a_2) \neq (0, 0, 0)$  with

$$a_0 + a_1x + a_2 \ln(x) = 0 \quad \nu\text{-almost everywhere } (\nu\text{-a.e.}).$$

For  $x \rightarrow \infty$ , the left-hand side tends to either  $-\infty$  or  $\infty$  and is therefore not equal to 0 on the interval  $(M, \infty)$  for some  $M > 0$ . Since  $\nu((M, \infty)) > 0$ , this forms a contradiction. Hence,  $T_1$  and  $T_2$  are  $\nu$ -affinely independent, and we have  $\text{ord}(\mathcal{G}(\bullet, \bullet)) = 2$  by Theorem 2.23. In particular, the representation is minimal.

*Example 2.25* Consider the representation of  $\mathcal{IN}(\bullet, \bullet)$  in Example 2.11 with

$$Z_1(\boldsymbol{\vartheta}) = -\frac{\eta}{2\mu^2}, \quad Z_2(\boldsymbol{\vartheta}) = -\frac{\eta}{2}, \quad \boldsymbol{\vartheta} = (\mu, \eta)^t \in (0, \infty)^2,$$

and  $T_1(x) = x$ ,  $T_2(x) = 1/x$  for  $x \in (0, \infty)$ . Setting  $\boldsymbol{\vartheta}_0 = (\sqrt{2}, 4)^t$ ,  $\boldsymbol{\vartheta}_1 = (1, 2)^t$ , and  $\boldsymbol{\vartheta}_2 = (1, 4)^t$ , we obtain that

$$\mathbf{Z}(\boldsymbol{\vartheta}_1) - \mathbf{Z}(\boldsymbol{\vartheta}_0) = (-1, -1)^t - (-1, -2)^t = (0, 1)^t$$

$$\text{and} \quad \mathbf{Z}(\boldsymbol{\vartheta}_2) - \mathbf{Z}(\boldsymbol{\vartheta}_0) = (-2, -2)^t - (-1, -2)^t = (-1, 0)^t$$

are linearly independent in  $\mathbb{R}^2$ . Hence,  $Z_1$  and  $Z_2$  are affinely independent by Lemma 2.22. Now, suppose that  $T_1$  and  $T_2$  are not  $\nu$ -affinely independent. Then, there exists a vector  $(a_0, a_1, a_2) \neq (0, 0, 0)$  with

$$a_0 + a_1x + \frac{a_2}{x} = 0 \quad \nu\text{-a.e.},$$

or, equivalently,

$$a_1x^2 + a_0x + a_2 = 0 \quad \nu\text{-a.e.}.$$

This quadratic equation has at most two positive solutions  $x_1$  and  $x_2$ , say. Since  $\nu(\{x_1, x_2\}) = 0$ , we have a contradiction. Hence,  $T_1$  and  $T_2$  are  $\nu$ -affinely independent, and we obtain  $\text{ord}(\mathcal{IN}(\bullet, \bullet)) = 2$  from Theorem 2.23. In particular, the representation is minimal.

---

To check for a representation of  $\mathcal{P}$  to be minimal, the following equivalence is often useful. The assertion is obvious from the general rule  $\mathbf{a}'\mathbf{Cov}(\mathbf{U})\mathbf{a} = \text{Var}(\mathbf{a}'\mathbf{U})$  for a random vector  $\mathbf{U}$  and a constant vector  $\mathbf{a}$ , which implies that  $\mathbf{a}'\mathbf{Cov}(\mathbf{U})\mathbf{a} = 0$  if and only if  $\mathbf{a}'\mathbf{U}$  is almost surely (a.s.) constant (see also [63, pp. 145/146]).

**Lemma 2.26** Let  $(\mathcal{Y}, \mathcal{C})$  be a measurable space,  $Q$  be a probability measure on  $\mathcal{C}$ , and  $U = (U_1, \dots, U_k)^t : (\mathcal{Y}, \mathcal{C}) \rightarrow (\mathbb{R}^k, \mathbb{B}^k)$  a random vector with  $E(U_j^2) < \infty$ ,  $1 \leq j \leq k$ . Then, the following assertions are equivalent:

- (a)  $U_1, \dots, U_k$  are  $Q$ -affinely independent.
- (b)  $\text{Cov}(U) > 0$ , i.e., the covariance matrix of  $U$  is positive definite.

**Examples:**  $\mathcal{N}(\bullet, \bullet)$ ,  $\mathcal{N}_k(\bullet, \Sigma_0)$

*Example 2.27* Consider the representation of  $\mathcal{N}(\bullet, \bullet)$  in Example 2.10 with

$$Z_1(\boldsymbol{\vartheta}) = \frac{\mu}{\sigma^2}, \quad Z_2(\boldsymbol{\vartheta}) = -\frac{1}{2\sigma^2}, \quad \boldsymbol{\vartheta} = (\mu, \sigma^2)^t \in \mathbb{R} \times (0, \infty),$$

and  $T_1(x) = x$ ,  $T_2(x) = x^2$ , for  $x \in \mathbb{R}$ . Setting  $\boldsymbol{\vartheta}_0 = (0, 1)^t$ ,  $\boldsymbol{\vartheta}_1 = (0, 1/2)^t$ , and  $\boldsymbol{\vartheta}_2 = (1/2, 1/2)^t$ , we obtain that

$$Z(\boldsymbol{\vartheta}_1) - Z(\boldsymbol{\vartheta}_0) = (0, -1/2)^t$$

$$\text{and} \quad Z(\boldsymbol{\vartheta}_2) - Z(\boldsymbol{\vartheta}_0) = (1, -1/2)^t$$

are linearly independent in  $\mathbb{R}^2$ . Hence,  $Z_1$  and  $Z_2$  are affinely independent by Lemma 2.22. Moreover, the covariance matrix of  $T(X) = (X, X^2)^t$  for  $X \sim Q = \mathcal{N}(0, 1)$  is given by

$$\text{Cov}_Q(T) = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix} > 0.$$

Thus,  $T_1$  and  $T_2$  are  $\nu$ -affinely independent by Lemma 2.26 and Remark 2.16. Applying Theorem 2.23 then yields that  $\text{ord}(\mathcal{N}(\bullet, \bullet)) = 2$ . In particular, the representation is minimal.

*Example 2.28* Consider the representation of  $\mathcal{N}_k(\bullet, \Sigma_0)$  in Example 2.18 with  $Z(\boldsymbol{\mu}) = \boldsymbol{\mu}$ ,  $\boldsymbol{\mu} \in \mathbb{R}^k$ , and  $T(\mathbf{x}) = \Sigma_0^{-1} \mathbf{x}$ ,  $\mathbf{x} \in \mathbb{R}^k$ . If  $a_0 \in \mathbb{R}$  and  $\mathbf{a} \in \mathbb{R}^k$  are such that  $a_0 + \mathbf{a}^t \boldsymbol{\mu} = 0$  for all  $\boldsymbol{\mu} \in \mathbb{R}^k$ , it directly follows that  $a_0 = 0$  and  $\mathbf{a} = \mathbf{0}$ . Hence,  $Z_1, \dots, Z_k$  are affinely independent. Moreover, for every  $\boldsymbol{\mu} \in \mathbb{R}^k$ ,  $\text{Cov}_{\boldsymbol{\mu}}(T) = \Sigma_0^{-1} > 0$  such that  $T_1, \dots, T_k$  are  $\nu$ -affinely independent by Lemma 2.26. Applying Theorem 2.23 then yields that  $\text{ord}(\mathcal{N}_k(\bullet, \Sigma_0)) = k$ . In particular, the representation is minimal.

In a minimal representation of an EF, the vectors  $\mathbf{Z}$  and  $\mathbf{T}$  of parameter functions and statistics, respectively, are uniquely determined up to non-degenerated affine transformations (see [4, p. 112]); a detailed proof is provided in [63, pp. 145–147].

**Theorem 2.29** *Let  $\mathcal{P}$  be an EF with minimal representations (2.1) and (2.5), i.e.,  $k = \text{ord}(\mathcal{P}) = \tilde{k}$ . Then, there exist regular matrices  $\mathbf{L}, \mathbf{M} \in \mathbb{R}^{k \times k}$  and vectors  $\mathbf{l}, \mathbf{m} \in \mathbb{R}^k$  with*

$$\mathbf{Z} = \mathbf{L}\tilde{\mathbf{Z}} + \mathbf{l} \quad \text{and} \quad \mathbf{T} = \mathbf{M}\tilde{\mathbf{T}} + \mathbf{m} \quad v\text{-a.e.},$$

where  $\tilde{\mathbf{Z}} = (\tilde{Z}_1, \dots, \tilde{Z}_k)^t$  and  $\tilde{\mathbf{T}} = (\tilde{T}_1, \dots, \tilde{T}_k)^t$ .

Finding a minimal representation for a given EF is an important step when aiming at statistical inference. Among others, it may ensure that the parameter  $\vartheta \in \Theta$  is identifiable. Alternatively, in the literature, the identifiability property is ascribed to the set  $\Theta$  of parameters or to the parametric family  $\mathcal{P}$  of distributions.

**Lemma 2.30** *Let  $\mathcal{P}$  be an EF with representation (2.4). If  $\mathbf{Z}$  is one-to-one and  $T_1, \dots, T_k$  are  $v$ -affinely independent, then the parameter  $\vartheta \in \Theta$  is identifiable, i.e., the mapping  $\vartheta \mapsto P_\vartheta$ ,  $\vartheta \in \Theta$ , is one-to-one.*

**Proof** Let  $P_{\vartheta_1} = P_{\vartheta_2}$  for  $\vartheta_1, \vartheta_2 \in \Theta$ . Then, we have  $g_{\vartheta_1} = g_{\vartheta_2}$   $v$ -a.e., which is equivalent to

$$(\mathbf{Z}(\vartheta_1) - \mathbf{Z}(\vartheta_2))^t \mathbf{T} = \ln \left( \frac{C(\vartheta_2)}{C(\vartheta_1)} \right) \quad v\text{-a.e.}$$

Hence, we have  $\mathbf{Z}(\vartheta_1) = \mathbf{Z}(\vartheta_2)$  by using that  $T_1, \dots, T_k$  are  $v$ -affinely independent. Since  $\mathbf{Z}$  is one-to-one, it follows that  $\vartheta_1 = \vartheta_2$ .  $\square$

In many applications, the parameters of some EF are supposed to satisfy certain constraints. This assumption leads to subfamilies which, of course, form EFs. The order of these embedded EFs may be smaller than or equal to the order of the original EF. We give a classical example (see [35, pp. 85–87]).

#### Example: Subfamilies of $\mathcal{N}(\bullet, \bullet)$

*Example 2.31* Let  $\mathcal{P} = \mathcal{N}(\bullet, \bullet)$  with representation as in Example 2.10, and recall that  $\text{ord}(\mathcal{P}) = 2$  (see Example 2.27).

- (a) First, we consider the subfamily  $\mathcal{P}_1 = \{\mathcal{N}(\mu, \mu) : \mu \in (0, \infty)\}$  of  $\mathcal{P}$ . Setting  $\sigma^2 = \mu$  in the density of  $\mathcal{N}(\mu, \sigma^2)$ , it is easily seen that  $\mathcal{P}_1$  forms an EF with representation (2.1) given by  $k = 1$ ,

$$C(\mu) = \frac{1}{\sqrt{\mu}} \exp \left\{ -\frac{\mu}{2} \right\}, \quad Z_1(\mu) = -\frac{1}{2\mu}, \quad \mu \in (0, \infty),$$

$$\text{and } T_1(x) = x^2, \quad h(x) = \frac{1}{\sqrt{2\pi}}e^x, \quad x \in \mathbb{R}.$$

Hence,  $\text{ord}(\mathcal{P}_1) = 1$ .

- (b) Now, let  $\mathcal{P}_2 = \{\mathcal{N}(\mu, \mu^2) : \mu \in \mathbb{R} \setminus \{0\}\}$ . Replacing  $\sigma^2$  by  $\mu^2$  in the density of  $\mathcal{N}(\mu, \sigma^2)$ , we arrive at a representation (2.1) of  $\mathcal{P}_2$  given by  $k = 2$ ,

$$C(\mu) = \frac{1}{\sqrt{\mu^2}}, \quad Z_1(\mu) = \frac{1}{\mu}, \quad \text{and} \quad Z_2(\mu) = -\frac{1}{2\mu^2}, \quad \mu \in \mathbb{R} \setminus \{0\},$$

$$\text{and } h(x) = \frac{1}{\sqrt{2\pi}e}, \quad T_1(x) = x, \quad \text{and} \quad T_2(x) = x^2, \quad x \in \mathbb{R}.$$

Since  $\mathbf{Z}(1) - \mathbf{Z}(-1) = (2, 0)^t$  and  $\mathbf{Z}(1/2) - \mathbf{Z}(-1) = (3, -3/2)^t$ ,  $Z_1$  and  $Z_2$  are affinely independent by Lemma 2.22. As shown in Example 2.27,  $T_1$  and  $T_2$  are  $\mathcal{N}(0, 1)$ -affinely independent and thus also  $\mathcal{N}(1, 1)$ -affinely independent. Hence, we have  $\text{ord}(\mathcal{P}_2) = 2$  by Theorem 2.23.

Although both EFs,  $\mathcal{P}_1$  and  $\mathcal{P}_2$ , in Example 2.31 are one-parameter subfamilies of  $\mathcal{N}(\bullet, \bullet)$ , their orders are different. When imposing the constraint for the parameters, the mappings  $Z_1$  and  $Z_2$  in Example 2.27 are no longer affinely independent in case (a) but they are in case (b).  $\mathcal{P}_2$  forms an example for a so-called curved EF. We give a fairly general definition of the concept to keep it short, here (see [50, p. 28]); other definitions in the literature impose additional regularity conditions as, e.g., continuous second derivatives of  $\mathbf{Z}$  on  $\Theta$ , where  $\Theta$  is assumed to be an open subset of  $\mathbb{R}^m$  with  $m < k$ .

**Definition 2.32** Let  $\mathcal{P}$  be an EF with representation (2.1). If  $Z_1, \dots, Z_k$  are affinely independent and  $\text{int}(\mathbf{Z}(\Theta)) = \emptyset$ , i.e., the interior of  $\mathbf{Z}(\Theta)$  is empty, then  $\mathcal{P}$  is called a *curved EF*.

The statistical literature on curved EFs is vast; among others, geometrical aspects, inference, and applications, e.g., in graphical models and networks, have widely been studied. In the sequel, several properties of curved EFs are addressed while the main focus is on EFs with parameter spaces containing at least some inner points. More results on curved EFs can be found, for instance, in [20, 2, 5, 39, 44, 35, 56].

## 2.3 Structural Properties and Natural Parametrization

In Lemma 2.2, we have pointed out that the term  $C(\vartheta)$  in representation (2.1) is just a normalizing constant. Moreover, it depends on  $\vartheta$  only via  $\mathbf{Z}(\vartheta) \in \mathbf{Z}(\Theta) = \{\mathbf{Z}(\vartheta) : \vartheta \in \Theta\}$ . This finding allows for a more natural parametrization of the EF.

**Definition 2.33** Let  $\mathcal{P}$  be an EF with representation (2.1). The *natural parameter space* is defined by

$$\Xi^* = \{\boldsymbol{\zeta} = (\zeta_1, \dots, \zeta_k)^t \in \mathbb{R}^k : 0 < \int \exp \left\{ \sum_{j=1}^k \zeta_j T_j(x) \right\} h(x) d\mu(x) < \infty\}.$$

Note that the natural parameter space depends on representation (2.1) of the EF. Evidently, we have that  $\mathbf{Z}(\Theta) \subset \Xi^*$ . Moreover, by using the Hölder inequality,  $\Xi^*$  is readily seen to be convex; for a detailed proof, see [40, p. 48].

**Theorem 2.34** Let  $\mathcal{P}$  be an EF with representation (2.1). Then,  $\Xi^*$  is a convex set in  $\mathbb{R}^k$ , i.e., if  $\boldsymbol{\zeta}, \boldsymbol{\eta} \in \Xi^*$  and  $\alpha \in [0, 1]$ , then  $\alpha\boldsymbol{\zeta} + (1 - \alpha)\boldsymbol{\eta} \in \Xi^*$ .

The natural parametrization of the EF and related notions are now introduced as follows.

**Definition 2.35** Let  $\mathcal{P}$  be an EF with representation (2.1), and let  $\Xi = \mathbf{Z}(\Theta)$ . The *natural parametrization*  $P_{\boldsymbol{\zeta}}^*$ ,  $\boldsymbol{\zeta} \in \Xi$ , of  $\mathcal{P}$  is defined by the  $\mu$ -densities

$$f_{\boldsymbol{\zeta}}^*(x) = \frac{dP_{\boldsymbol{\zeta}}^*}{d\mu}(x) = C^*(\boldsymbol{\zeta}) \exp \left\{ \sum_{j=1}^k \zeta_j T_j(x) \right\} h(x), \quad x \in \mathcal{X}, \quad (2.6)$$

with normalizing constant  $C^*(\boldsymbol{\zeta})$  for  $\boldsymbol{\zeta} \in \Xi$ . Formula (2.6) is also referred to as *canonical representation* of  $\mathcal{P}$ . Moreover,  $\mathcal{P}$  is called

- full if  $\Xi = \Xi^*$ , and
- regular if  $\Xi = \Xi^*$  and  $\Xi^*$  is open.



*Remark 2.36* Following Lemma 2.15, a  $\nu$ -density of  $P_\zeta^*$  in Definition 2.35 is given by

$$g_\zeta^*(x) = \frac{dP_\zeta^*}{d\nu}(x) = C^*(\zeta) \exp \left\{ \sum_{j=1}^k \zeta_j T_j(x) \right\}, \quad x \in \mathcal{X},$$

for  $\zeta \in \Xi$ ; moreover, by Lemma 2.17, a  $\nu^T$ -density of the image distribution  $(P_\zeta^*)^T$  of  $T = (T_1, \dots, T_k)^T$  is represented by

$$(g_\zeta^*)^T(t) = \frac{d(P_\zeta^*)^T}{d\nu^T}(t) = C^*(\zeta) e^{\zeta^T t}, \quad t \in \mathbb{R}^k, \quad (2.7)$$

for  $\zeta \in \Xi$ . A full EF with densities (2.7), which moreover satisfies that  $\text{int}(\Xi^*) \neq \emptyset$  and that  $\nu^T$  is not concentrated on an affine hyperplane of  $\mathbb{R}^k$ , is referred to as *natural EF* (with *generating measure*  $\nu^T$ ) in the literature (see, e.g., [37, Chapter 54], and the references therein).

The natural parametrization of the EF is more convenient for inferential purposes and leads to various useful statistical properties. If exponential distributions are introduced by the Lebesgue density functions  $\lambda \exp\{-\lambda x\}$ ,  $x > 0$ , for parameter  $\lambda > 0$ , then the respective EF is already given in natural parametrization with  $\lambda \in \Xi = (0, \infty)$ . The natural parametrization for the class  $\mathcal{N}_k(\bullet, \Sigma_0)$  of multivariate normal distributions with fixed covariance matrix  $\Sigma_0$  is already given in Example 2.18. Some other simple examples for EFs in canonical representations are shown in what follows, and the corresponding natural parameter spaces are determined.

**Examples:**  $\text{poi}(\bullet)$ ,  $\text{b}(n_0, \bullet)$ ,  $\mathcal{N}(\bullet, \bullet)$ ,  $\mathcal{G}(\bullet, \bullet)$

*Example 2.37* Consider the representation of  $\mathcal{P} = \text{poi}(\bullet)$  in Example 2.3 with  $k = 1$ ,  $Z_1(a) = \ln(a)$  for  $a \in (0, \infty)$  and  $T_1(x) = x$ ,  $h(x) = 1/x!$  for  $x \in \mathbb{N}_0$ . Since

$$\int e^{\zeta x} \frac{1}{x!} d\mu_{\mathbb{N}_0}(x) = \sum_{x=0}^{\infty} e^{\zeta x} \frac{1}{x!} = \exp\{e^\zeta\}, \quad \zeta \in \mathbb{R},$$

we have  $\Xi^* = \mathbb{R} = Z_1((0, \infty))$ . Hence,  $\mathcal{P}$  is regular, and the canonical representation of  $\mathcal{P}$  is given by the counting densities

$$f_\zeta^*(x) = C^*(\zeta) e^{\zeta x} \frac{1}{x!}, \quad x \in \mathbb{N}_0,$$

with

$$C^*(\zeta) = \exp \{-e^\zeta\}$$

for  $\zeta \in \mathbb{R}$ .

*Example 2.38* Consider the representation of  $\mathcal{P} = \mathfrak{b}(n_0, \bullet)$  in Example 2.4 for fixed  $n_0 \in \mathbb{N}$  with  $k = 1$ ,  $Z_1(p) = \ln(p/(1-p))$  for  $p \in (0, 1)$  and  $T_1(x) = x$ ,  $h(x) = \binom{n_0}{x}$  for  $x \in \{0, 1, \dots, n_0\} = \mathcal{X}$ . Since

$$\int e^{\zeta x} \binom{n_0}{x} d\mu_{\mathcal{X}}(x) = \sum_{x=0}^{n_0} e^{\zeta x} \binom{n_0}{x} = (1 + e^\zeta)^{n_0}, \quad \zeta \in \mathbb{R},$$

we have  $\Xi^* = \mathbb{R} = Z_1((0, 1))$ . Hence,  $\mathcal{P}$  is regular, and the canonical representation of  $\mathcal{P}$  is given by the counting densities

$$f_\zeta^*(x) = C^*(\zeta) e^{\zeta x} \binom{n_0}{x}, \quad x \in \mathcal{X},$$

with

$$C^*(\zeta) = (1 + e^\zeta)^{-n_0}$$

for  $\zeta \in \mathbb{R}$ .

*Example 2.39* Consider the representation of  $\mathcal{P} = \mathcal{N}(\bullet, \bullet)$  in Example 2.10 with  $k = 2$ ,

$$C(\mu, \sigma^2) = \frac{1}{\sqrt{\sigma^2}} \exp \left\{ -\frac{\mu^2}{2\sigma^2} \right\},$$

$Z_1(\mu, \sigma^2) = \mu/\sigma^2$ , and  $Z_2(\mu, \sigma^2) = -1/(2\sigma^2)$  for  $\mu \in \mathbb{R}$ ,  $\sigma^2 \in (0, \infty)$ , and  $h(x) = 1/\sqrt{2\pi}$ ,  $T_1(x) = x$ , and  $T_2(x) = x^2$  for  $x \in \mathbb{R}$ . Here, we have

$$\mathbb{R} \times (-\infty, 0) = \mathbf{Z}(\mathbb{R} \times (0, \infty)) \subset \Xi^*.$$

To establish that  $\Xi^* = \mathbb{R} \times (-\infty, 0)$ , it is sufficient to show that  $(\zeta_1, 0)^t \notin \Xi^*$  for  $\zeta_1 \in \mathbb{R}$ , since  $\Xi^*$  is convex by Theorem 2.34. Indeed, for every  $\zeta_1 \in \mathbb{R}$ , we have

$$\int_{-\infty}^{\infty} e^{\zeta_1 x} \frac{1}{\sqrt{2\pi}} dx = \infty.$$

Hence,  $\mathcal{P}$  is regular, and, by means of the parameter transformation

$$\left( \frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2} \right) = (\zeta_1, \zeta_2) \quad \Leftrightarrow \quad (\mu, \sigma^2) = \left( -\frac{\zeta_1}{2\zeta_2}, -\frac{1}{2\zeta_2} \right),$$

the canonical representation of  $\mathcal{P}$  turns out to be

$$f_{\zeta}^*(x) = C^*(\zeta) e^{\zeta_1 x + \zeta_2 x^2} \frac{1}{\sqrt{2\pi}}, \quad x \in \mathbb{R},$$

with

$$C^*(\zeta) = C\left(-\frac{\zeta_1}{2\zeta_2}, -\frac{1}{2\zeta_2}\right) = \sqrt{-2\zeta_2} \exp\left\{\frac{\zeta_1^2}{4\zeta_2}\right\}$$

for  $\zeta = (\zeta_1, \zeta_2)^t \in \mathbb{R} \times (-\infty, 0)$ .

*Example 2.40* Consider the representation of  $\mathcal{P} = \mathcal{G}(\bullet, \bullet)$  in Example 2.12 with  $k = 2$ ,  $C(\alpha, \beta) = 1/[\Gamma(\beta)\alpha^\beta]$ ,  $Z_1(\alpha, \beta) = -1/\alpha$ ,  $Z_2(\alpha, \beta) = \beta$  for  $\alpha, \beta \in (0, \infty)$ , and  $h(x) = 1/x$ ,  $T_1(x) = x$ ,  $T_2(x) = \ln(x)$  for  $x \in (0, \infty)$ . Here, we have

$$(-\infty, 0) \times (0, \infty) = \mathbf{Z}((0, \infty)^2) \subset \Xi^*.$$

To establish that  $\Xi^* = (-\infty, 0) \times (0, \infty)$ , it is sufficient to show that  $(\zeta_1, 0)^t \notin \Xi^*$  and  $(0, \zeta_2)^t \notin \Xi^*$  for  $\zeta_1 \leq 0$  and  $\zeta_2 > 0$ , since  $\Xi^*$  is convex by Theorem 2.34. Indeed, for  $\zeta_1 \leq 0$  and  $\zeta_2 > 0$ , we have

$$\int_0^\infty e^{\zeta_1 x} \frac{1}{x} dx \geq \int_0^1 e^{\zeta_1 x} \frac{1}{x} dx \geq e^{\zeta_1} \int_0^1 \frac{1}{x} dx = \infty$$

$$\text{and} \quad \int_0^\infty e^{\zeta_2 \ln(x)} \frac{1}{x} dx = \frac{x^{\zeta_2}}{\zeta_2} \Big|_0^\infty = \infty.$$

Hence,  $\mathcal{P}$  is regular, and, by means of the parameter transformation  $\zeta_1 = -1/\alpha$  and  $\zeta_2 = \beta$ , the canonical representation of  $\mathcal{P}$  turns out to be

$$f_{\zeta}^*(x) = C^*(\zeta) e^{\zeta_1 x + \zeta_2 \ln(x)} \frac{1}{x}, \quad x \in (0, \infty),$$

with

$$C^*(\zeta) = C\left(-\frac{1}{\zeta_1}, \zeta_2\right) = \frac{(-\zeta_1)^{\zeta_2}}{\Gamma(\zeta_2)}$$

for  $\zeta = (\zeta_1, \zeta_2)^t \in (-\infty, 0) \times (0, \infty)$ .

The EFs considered in the preceding examples are all full and moreover regular, since the natural parameter space turns out to be open. Indeed, this is the case for most of the common EFs. An example for a full but non-regular EF is provided by the full family of inverse normal distributions (see, e.g., [4, p. 117] or [20, p. 72]).

**Example:  $IN(\bullet, \bullet)$**

*Example 2.41* Consider the representation of  $\mathcal{P} = IN(\bullet, \bullet)$  in Example 2.11 with  $k = 2$ ,

$$C(\mu, \eta) = \sqrt{\eta} \exp\left\{\frac{\eta}{\mu}\right\},$$

$Z_1(\mu, \eta) = -\eta/(2\mu^2)$ , and  $Z_2(\mu, \eta) = -\eta/2$  for  $\mu, \eta \in (0, \infty)$ , and  $h(x) = \sqrt{1/(2\pi x^3)}$ ,  $T_1(x) = x$ , and  $T_2(x) = 1/x$  for  $x \in (0, \infty)$ . Here, we have

$$(-\infty, 0)^2 = \mathbf{Z}((0, \infty)^2) \subset \Xi^*.$$

However, it turns out that  $\Xi^* = (-\infty, 0] \times (-\infty, 0)$ , which can be seen as follows.

First, we have  $(0, \zeta_2)^t \in \Xi^*$  for  $\zeta_2 < 0$ , since by substituting  $z = 1/x$

$$\begin{aligned} \int_0^\infty e^{\zeta_2/x} \sqrt{\frac{1}{2\pi x^3}} dx &= \sqrt{\frac{1}{2\pi}} \int_0^\infty e^{\zeta_2 z} z^{-1/2} dz \\ &= \sqrt{\frac{1}{2\pi}} \frac{\Gamma(1/2)}{(-\zeta_2)^{1/2}} = \sqrt{\frac{1}{-2\zeta_2}} < \infty. \end{aligned}$$

Moreover, it holds that  $(\zeta_1, 0)^t \notin \Xi^*$  for  $\zeta_1 \leq 0$ , since

$$\begin{aligned} \int_0^\infty e^{\zeta_1 x} \sqrt{\frac{1}{2\pi x^3}} dx &\geq \sqrt{\frac{1}{2\pi}} \int_0^1 e^{\zeta_1 x} \sqrt{\frac{1}{x^3}} dx \\ &\geq \sqrt{\frac{1}{2\pi}} e^{\zeta_1} \int_0^1 \frac{1}{x} dx = \infty. \end{aligned}$$

Finally, we have  $(\zeta_1, \zeta_2)^t \notin \Xi^*$  for  $\zeta_1 > 0$  and  $\zeta_2 < 0$ , since by using the series representation of  $\exp\{\zeta_1 x\}$

$$\begin{aligned} \int_0^\infty e^{\zeta_1 x + \zeta_2/x} \sqrt{\frac{1}{2\pi x^3}} dx &\geq \sqrt{\frac{1}{2\pi}} \int_1^\infty e^{\zeta_1 x + \zeta_2/x} \sqrt{\frac{1}{x^3}} dx \\ &\geq \sqrt{\frac{1}{2\pi}} e^{\zeta_2} \int_1^\infty e^{\zeta_1 x} \frac{1}{x^2} dx \\ &\geq \sqrt{\frac{1}{2\pi}} e^{\zeta_2} \int_1^\infty \frac{(\zeta_1 x)^2}{2!} \frac{1}{x^2} dx = \infty. \end{aligned}$$

The convexity of  $\Xi^*$  as pointed out in Theorem 2.34 now ensures that  $\Xi^* = (-\infty, 0] \times (-\infty, 0)$ . In particular,  $\mathcal{P}$  is not full. By using the parameter transformation

$$\left(-\frac{\eta}{2\mu^2}, -\frac{\eta}{2}\right) = (\zeta_1, \zeta_2) \quad \Leftrightarrow \quad (\mu, \eta) = \left(\sqrt{\frac{\zeta_2}{\zeta_1}}, -2\zeta_2\right)$$

the canonical representation of  $\mathcal{P}$  is given by

$$f_\zeta^*(x) = C^*(\zeta) e^{\zeta_1 x + \zeta_2/x} \sqrt{\frac{1}{2\pi x^3}}, \quad x \in (0, \infty),$$

with

$$C^*(\zeta) = C\left(\sqrt{\frac{\zeta_2}{\zeta_1}}, -2\zeta_2\right) = \sqrt{-2\zeta_2} e^{2\sqrt{\zeta_1 \zeta_2}}$$

for  $\zeta = (\zeta_1, \zeta_2)^t \in (-\infty, 0)^2$ . Note that the associated full EF  $\{P_\zeta^* : \zeta \in (-\infty, 0] \times (-\infty, 0)\}$  with  $C^*(0, \zeta_2) = \sqrt{-2\zeta_2}$  for  $\zeta_2 < 0$  is not regular.

Having seen that the natural parameter space of the EF is not open in general, we address a simple sufficient condition in the following lemma guaranteeing that it has at least some inner points. The proof makes use of the convexity property of  $\Xi^*$  and Lemma 2.22 (for details, see [63, p. 150]).

**Lemma 2.42** *Let  $\mathcal{P}$  be an EF with representation (2.1). If  $Z_1, \dots, Z_k$  are affinely independent, then  $\text{int}(\Xi^*) \neq \emptyset$ .*

To address further properties related to the natural parametrization, we introduce a function, which is of high relevance in statistical inference.

**Definition 2.43** Let  $\mathcal{P}$  be an EF with canonical representation (2.6). The mapping  $\kappa : \Xi \rightarrow \mathbb{R}$  defined by  $\kappa = -\ln(C^*)$  is called the *cumulant function* of  $\mathcal{P}$ .

For the motivation of the term cumulant function in Definition 2.43 and for remarks on so-called cumulants, we refer to Sects. 2.4 and 3.1.

The following properties of the cumulant function are immediately obtained from the proof of Theorem 2.34 and Lemma 2.30.

**Theorem 2.44** *Let  $\mathcal{P}$  be a full EF with canonical representation (2.6) and cumulant function  $\kappa = -\ln(C^*)$  defined on  $\Xi^*$ . Then, we have:*

(a)  *$\kappa$  is convex, i.e., if  $\xi, \eta \in \Xi^*$  and  $\alpha \in (0, 1)$ , then*

$$\kappa(\alpha\xi + (1-\alpha)\eta) \leq \alpha\kappa(\xi) + (1-\alpha)\kappa(\eta). \quad (2.8)$$

*Here, equality holds true if and only if  $P_\xi^* = P_\eta^*$ .*

(b) *If  $T_1, \dots, T_k$  are  $\nu$ -affinely independent, then  $\kappa$  is strictly convex on  $\Xi^*$ , i.e., equality in Eq. (2.8) holds true if and only if  $\xi = \eta$ .*

Note that, since  $\kappa$  is convex and finite on  $\Xi^*$ , it is continuous on the interior  $\text{int}(\Xi^*)$  of  $\Xi^*$  (see, e.g., [51, Section 10]). For maximum likelihood estimation in a subsequent section, it will be convenient to extend  $\kappa$  in a natural way to a function on  $\mathbb{R}^k$ . The following property then directly follows from Fatou's lemma (see [20, pp. 19/20]). Fatou's lemma along with its proof can be found, e.g., in [17, p. 209].

**Lemma 2.45** *Let  $\mathcal{P}$  be a full EF with canonical representation (2.6) and cumulant function  $\kappa = -\ln(C^*)$  on  $\Xi^*$ . Moreover, let  $\tilde{\kappa} : \mathbb{R}^k \rightarrow \mathbb{R} \cup \{\infty\}$  be the extension of  $\kappa$  defined by  $\tilde{\kappa}(\xi) = \infty$  for  $\xi \in \mathbb{R}^k \setminus \Xi^*$ . Then,  $\tilde{\kappa}$  is lower semi-continuous, i.e.,*

$$\liminf_{\eta \rightarrow \xi} \tilde{\kappa}(\eta) \geq \tilde{\kappa}(\xi) \quad \text{for every } \xi \in \mathbb{R}^k.$$

*Moreover,  $\tilde{\kappa}$  is continuous on  $\text{int}(\Xi^*)$ .*

When aiming for inference with a given EF, it is advisable to find a minimal representation, first, before transforming it to a canonical representation. Otherwise, statistical problems may arise as it is demonstrated in the following example, where we discuss the structure of the multinomial EF (see [20, pp. 4–6]).

**Example:  $m(n_0, \bullet)$**

*Example 2.46* Consider the representation of  $\mathcal{P} = m(n_0, \bullet)$  in Example 2.8 for fixed  $n_0 \in \mathbb{N}$  with  $C(\mathbf{p}) = 1$ ,

$$Z_j(\mathbf{p}) = \ln(p_j), \quad \mathbf{p} \in \Theta = \{(q_1, \dots, q_k)^t \in (0, 1)^k : \sum_{i=1}^k q_i = 1\},$$

and  $h(\mathbf{x}) = n_0! / \prod_{i=1}^k x_i!$ ,

$$T_j(\mathbf{x}) = x_j, \quad \mathbf{x} \in \mathcal{X} = \{(y_1, \dots, y_k)^t \in \mathbb{N}_0^k : \sum_{i=1}^k y_i = n_0\},$$

for  $j \in \{1, \dots, k\}$ . For  $\boldsymbol{\zeta} = (\zeta_1, \dots, \zeta_k)^t \in \mathbb{R}^k$ , it follows by the multinomial theorem that

$$\begin{aligned} \int \exp \left\{ \sum_{j=1}^k \zeta_j x_j \right\} h(\mathbf{x}) d\mu_{\mathcal{X}}(\mathbf{x}) &= \sum_{\mathbf{x} \in \mathcal{X}} \frac{n_0!}{\prod_{i=1}^k x_i!} \prod_{j=1}^k (e^{\zeta_j})^{x_j} \\ &= \left( \sum_{j=1}^k e^{\zeta_j} \right)^{n_0}. \end{aligned}$$

Hence, we have  $\Xi^* = \mathbb{R}^k$ , and, since  $\mathbf{Z}(\Theta) \subset (-\infty, 0)^k$ ,  $\mathcal{P}$  is not full. The full EF is given by  $\mathcal{P}^* = \{P_{\boldsymbol{\zeta}}^* : \boldsymbol{\zeta} \in \mathbb{R}^k\}$ , where  $P_{\boldsymbol{\zeta}}^*$  has counting density

$$f_{\boldsymbol{\zeta}}^*(\mathbf{x}) = \left( \sum_{j=1}^k e^{\zeta_j} \right)^{-n_0} \exp \left\{ \sum_{j=1}^k \zeta_j x_j \right\} h(\mathbf{x}), \quad \mathbf{x} \in \mathcal{X},$$

for  $\boldsymbol{\zeta} \in \mathbb{R}^k$ . Nevertheless, it holds that  $\mathcal{P} = \mathcal{P}^*$ , which can be seen as follows. First, for  $\boldsymbol{\zeta} \in \mathbb{R}^k$ ,  $\mathbf{1}^{(k)} = (1, \dots, 1)^t \in \mathbb{R}^k$ , and  $a \in \mathbb{R}$ , we have

$$f_{\boldsymbol{\zeta}+a\mathbf{1}^{(k)}}^*(\mathbf{x}) = \left( \sum_{j=1}^k e^{\zeta_j+a} \right)^{-n_0} \exp \left\{ \sum_{j=1}^k (\zeta_j + a) x_j \right\} h(\mathbf{x})$$

$$\begin{aligned}
&= e^{-n_0 a} \exp \left\{ a \sum_{j=1}^k x_j \right\} f_{\zeta}^*(\mathbf{x}) \\
&= f_{\zeta}^*(\mathbf{x}), \quad \mathbf{x} \in \mathcal{X}.
\end{aligned}$$

Hence,  $P_{\zeta}^* = P_{\zeta+a\mathbf{1}^{(k)}}^*$ . Now, let  $\zeta \in \mathbb{R}^k$ . We have to show that there exists some  $\mathbf{p} \in \Theta$  with  $\mathbf{m}(n_0, \mathbf{p}) = P_{\zeta}^*$ . To this end, let  $\mathbf{p} = (p_1, \dots, p_k)^t$  with

$$p_i = \exp \left\{ \zeta_i - \ln \left( \sum_{j=1}^k e^{\zeta_j} \right) \right\}, \quad 1 \leq i \leq k.$$

Obviously,  $p_i > 0$  for  $1 \leq i \leq k$ , and  $\sum_{j=1}^k p_j = 1$ , such that  $\mathbf{p} \in \Theta$ . Moreover, since  $f_{n_0, \mathbf{p}} = f_{\mathbf{Z}(\mathbf{p})}^*$  by definition, it follows for  $a = -\ln(\sum_{j=1}^k \exp\{\zeta_j\})$  that

$$\mathbf{m}(n_0, \mathbf{p}) = P_{\mathbf{Z}(\mathbf{p})}^* = P_{\zeta+a\mathbf{1}^{(k)}}^* = P_{\zeta}^*.$$

Hence,  $\mathcal{P} = \mathcal{P}^*$  is shown. In particular, this yields that the parameter  $\zeta \in \Xi^*$  is not identifiable.

This problem occurs, since the representation of  $\mathcal{P}$  in Example 2.8 is not minimal. Indeed, we have  $\text{ord}(\mathcal{P}) = k - 1$ , which is shown in the following. First, we rewrite

$$\begin{aligned}
f_{n_0, \mathbf{p}}(\mathbf{x}) &= \exp \left\{ \sum_{j=1}^{k-1} \ln(p_j) x_j + \ln(p_k) \left( n_0 - \sum_{j=1}^{k-1} x_j \right) \right\} h(\mathbf{x}) \\
&= p_k^{n_0} \exp \left\{ \sum_{j=1}^{k-1} (\ln(p_j) - \ln(p_k)) x_j \right\} h(\mathbf{x}), \quad \mathbf{x} \in \mathcal{X}. \quad (2.9)
\end{aligned}$$

This is another representation of  $\mathcal{P}$  with parameter functions  $\tilde{Z}_j(\mathbf{p}) = \ln(p_j) - \ln(p_k)$  for  $\mathbf{p} \in \Theta$  and statistics  $T_j(\mathbf{x}) = x_j$  for  $\mathbf{x} \in \mathcal{X}$ ,  $1 \leq j \leq k - 1$ . Hence,



$ord(\mathcal{P}) \leq k - 1$ . Since  $\emptyset$  is the only set  $N \subset \mathcal{X}$  with  $v(N) = 0$ , we show that  $T_1, \dots, T_{k-1}$  are affinely independent. Let  $a_0, a_1, \dots, a_{k-1} \in \mathbb{R}$  with

$$0 = a_0 + \sum_{j=1}^{k-1} a_j T_j(\mathbf{x}) = a_0 + \sum_{j=1}^{k-1} a_j x_j \quad \text{for all } \mathbf{x} \in \mathcal{X}. \quad (2.10)$$

Moreover, we denote by  $\mathbf{e}_j^{(k)}$  the  $j$ -th unit vector in  $\mathbb{R}^k$ . Then, setting  $\mathbf{x} = n_0 \mathbf{e}_k^{(k)} \in \mathcal{X}$  in Eq. (2.10) gives  $a_0 = 0$ , and the choice  $\mathbf{x} = n_0 \mathbf{e}_j^{(k)} \in \mathcal{X}$  leads to  $a_j = 0$ ,  $1 \leq j \leq k - 1$ . To show that  $\tilde{Z}_1, \dots, \tilde{Z}_{k-1}$  are affinely independent, let  $\mathbf{p}^{(0)} = \mathbf{1}^{(k)}/k \in \Theta$  and  $\mathbf{p}^{(j)} = (\mathbf{1}^{(k)} + \mathbf{e}_j^{(k)})/(k+1) \in \Theta$ ,  $1 \leq j \leq k - 1$ . By setting  $\tilde{\mathbf{Z}} = (\tilde{Z}_1, \dots, \tilde{Z}_{k-1})^t$ , we have

$$\tilde{\mathbf{Z}}(\mathbf{p}^{(j)}) - \tilde{\mathbf{Z}}(\mathbf{p}^{(0)}) = \tilde{\mathbf{Z}}(\mathbf{p}^{(j)}) = \ln(2) \mathbf{e}_j^{(k-1)} \in \mathbb{R}^{k-1}, \quad 1 \leq j \leq k - 1.$$

From Lemma 2.22, we obtain that  $\tilde{Z}_1, \dots, \tilde{Z}_{k-1}$  are affinely independent, and applying Theorem 2.23 gives  $ord(\mathcal{P}) = k - 1$ . In particular, representation (2.9) is minimal. Now, by using the multinomial theorem again,  $\Xi^* = \mathbb{R}^{k-1} = \tilde{\mathbf{Z}}(\Theta)$ , such that  $\mathcal{P}$  is full and regular. Finally, by setting  $\zeta_j = \ln(p_j) - \ln(p_k)$ ,  $1 \leq j \leq k - 1$ , in density (2.9) and by noticing that then

$$p_k = 1 - \sum_{j=1}^{k-1} p_j = 1 - p_k \sum_{j=1}^{k-1} e^{\zeta_j} \quad \text{and thus} \quad p_k = \left( 1 + \sum_{j=1}^{k-1} e^{\zeta_j} \right)^{-1},$$

a minimal canonical representation of  $\mathcal{P}$  is given by the counting densities

$$\tilde{f}_{\zeta}^*(\mathbf{x}) = \left( 1 + \sum_{j=1}^{k-1} e^{\zeta_j} \right)^{-n_0} \exp \left\{ \sum_{j=1}^{k-1} \zeta_j x_j \right\} h(\mathbf{x}), \quad \mathbf{x} \in \mathcal{X},$$

for  $\zeta \in \mathbb{R}^{k-1}$ .

## 2.4 Analytical Properties and Mean Value Parametrization

We now turn to useful analytical properties of the EF in natural parametrization. Essentially, they are consequences of the following lemma, which can be shown by using the Lebesgue dominated-convergence theorem; a detailed proof is provided, e.g., in [40, pp. 49/50].

**Lemma 2.47** *Let  $\mathcal{P}$  be an EF with canonical representation (2.6), and let  $A \subset \Xi$  be an open set. Moreover, let the function  $\varphi: (X, \mathcal{B}) \rightarrow (\mathbb{R}^1, \mathbb{B}^1)$  be  $P_\xi^*$ -integrable for all  $\xi = (\xi_1, \dots, \xi_k)^t \in A$ . Then, the function  $\delta: A \rightarrow \mathbb{R}$  defined by*

$$\delta(\xi) = \int \varphi(x) \exp \left\{ \sum_{j=1}^k \xi_j T_j(x) \right\} h(x) d\mu(x), \quad \xi \in A,$$

*is infinitely often differentiable, and the partial derivatives can be obtained by differentiating under the integral sign, i.e.,*

$$\frac{\partial^{l_1}}{\partial \xi_1^{l_1}} \dots \frac{\partial^{l_k}}{\partial \xi_k^{l_k}} \delta(\xi) = \int \varphi(x) T_1^{l_1}(x) \dots T_k^{l_k}(x) \exp \left\{ \sum_{j=1}^k \xi_j T_j(x) \right\} h(x) d\mu(x) \quad (2.11)$$

for  $\xi \in A$  and  $l_1, \dots, l_k \in \mathbb{N}_0$ .

**Remark 2.48** Lemma 2.47 can be extended to functions  $\delta$  with complex arguments: If  $\delta$  is defined on the complex set  $\{z = \xi + i\eta \in \mathbb{C}^k : \xi \in A\}$ , then  $\delta$  is a holomorphic function of every variable  $z_j$  for fixed other components  $z_i, i \neq j$ . As in Lemma 2.47, its (complex) derivatives can be obtained by differentiating under the integral sign. For a proof, see, e.g., [40, pp. 49/50].

The usefulness of Lemma 2.47 is that it yields existence and simple representations for the moments of  $T$ . Setting  $\varphi \equiv 1$ , which is  $P_\xi^*$ -integrable for all  $\xi \in \Xi^*$  as a constant function, and multiplying both sides of Eq. (2.11) by  $C^*(\xi)$ , we directly obtain the following theorem by noticing that  $E_\xi[T_1^{l_1} \dots T_k^{l_k}] = \int T_1^{l_1} \dots T_k^{l_k} f_\xi^* d\mu$ .

**Theorem 2.49** *Let  $\mathcal{P}$  be an EF with canonical representation (2.6). Then, for  $\xi \in \text{int}(\Xi)$ ,  $T = (T_1, \dots, T_k)^t$  has finite multivariate moments of any order with respect to  $P_\xi^*$ , which are given by*

$$E_\xi[T_1^{l_1} \dots T_k^{l_k}] = C^*(\xi) \frac{\partial^{l_1}}{\partial \xi_1^{l_1}} \dots \frac{\partial^{l_k}}{\partial \xi_k^{l_k}} \int \exp \left\{ \sum_{j=1}^k \xi_j T_j \right\} h d\mu$$

for  $l_1, \dots, l_k \in \mathbb{N}_0$ .

As another important finding, the mean and covariance matrix of  $T$  can be obtained as derivatives of the cumulant function  $\kappa = -\ln(C^*)$  introduced in Definition 2.43; the proof gives insight into the convenient structure of the canonical representation.

**Corollary 2.50** *Let  $\mathcal{P}$  be an EF with canonical representation (2.6). Then, the functions  $C^*$  and  $\kappa = -\ln(C^*)$  are infinitely often differentiable at every  $\xi \in \text{int}(\Xi)$ , where*

$$\begin{aligned}\nabla\kappa(\xi) &= E_\xi[T], \\ \mathbf{H}_\kappa(\xi) &= \mathbf{Cov}_\xi(T).\end{aligned}$$

Here,  $\nabla\kappa(\xi)$  and  $\mathbf{H}_\kappa(\xi)$  denote the gradient and the Hessian matrix of  $\kappa$  evaluated at  $\xi$ .

**Proof** By Lemma 2.47 with  $\varphi \equiv 1$ , the function  $1/C^*$  and thus  $C^*$  and  $\kappa = -\ln(C^*)$  are infinitely often differentiable at  $\xi \in \text{int}(\Xi)$ . Moreover, with  $v = h\mu$ , we obtain that

$$\frac{\partial}{\partial \xi_i} \kappa(\xi) = \frac{\partial}{\partial \xi_i} \ln \left( \int e^{\xi^T T} dv \right) = \frac{\int T_i e^{\xi^T T} dv}{\int e^{\xi^T T} dv} = E_\xi[T_i]$$

for  $1 \leq i \leq k$ . Applying Lemma 2.47 with  $\varphi = T_i$ , which is admissible by Theorem 2.49, then yields

$$\begin{aligned}\frac{\partial^2}{\partial \xi_i \partial \xi_j} \kappa(\xi) &= \frac{\partial}{\partial \xi_j} \left( \frac{\int T_i e^{\xi^T T} dv}{\int e^{\xi^T T} dv} \right) \\ &= \frac{(\int T_i T_j e^{\xi^T T} dv)(\int e^{\xi^T T} dv) - (\int T_i e^{\xi^T T} dv)(\int T_j e^{\xi^T T} dv)}{(\int e^{\xi^T T} dv)^2} \\ &= E_\xi[T_i T_j] - E_\xi[T_i] E_\xi[T_j] = \text{Cov}_\xi(T_i, T_j)\end{aligned}$$

for  $1 \leq i, j \leq k$ . □

Moments in EFs with canonical representation are easily determined by the respective normalizing constant  $C^*(\xi)$ .

**Examples:**  $\text{nb}(r_0, \bullet)$ ,  $\mathcal{N}(\bullet, \bullet)$ ,  $\mathcal{G}(\bullet, \bullet)$ ,  $\mathcal{N}_k(\bullet, \Sigma_0)$

*Example 2.51* From Example 2.5, it is readily seen that the EF  $\mathcal{P} = \text{nb}(r_0, \bullet)$  is regular with canonical representation

$$f_\xi^*(x) = C^*(\xi) e^{\xi x} \binom{x + r_0 - 1}{x}, \quad x \in \mathbb{N}_0,$$

with  $C^*(\zeta) = (1 - \exp\{\zeta\})^{r_0}$  for  $\zeta \in \Xi^* = (-\infty, 0)$ . Hence,

$$\kappa(\zeta) = -r_0 \ln(1 - e^\zeta), \quad \zeta \in (-\infty, 0),$$

and, by using Corollary 2.50, the mean and variance of the negative binomial distribution  $P_\zeta^*$  are given by

$$\kappa'(\zeta) = \frac{r_0 \exp\{\zeta\}}{1 - \exp\{\zeta\}} \quad \text{and} \quad \kappa''(\zeta) = \frac{r_0 \exp\{\zeta\}}{(1 - \exp\{\zeta\})^2}.$$

In the original parametrization, this yields the expressions  $r_0(1-p)/p$  and  $r_0(1-p)/p^2$  for the mean and variance of  $P_p = \text{nb}(r_0, p)$ .

*Example 2.52* In Example 2.39, the canonical representation of  $\mathcal{P} = \mathcal{N}(\bullet, \bullet)$  yields

$$C^*(\xi) = \sqrt{-2\xi_2} \exp\left\{\frac{\xi_1^2}{4\xi_2}\right\}, \quad \xi = (\xi_1, \xi_2)^t \in \mathbb{R} \times (-\infty, 0),$$

and  $\mathbf{T}(x) = (T_1(x), T_2(x))^t = (x, x^2)^t$ ,  $x \in \mathbb{R}$ . Hence, by Corollary 2.50 with  $\kappa(\xi) = -\ln(-2\xi_2)/2 - \xi_1^2/(4\xi_2)$ ,

$$E_\xi[T_1] = \frac{\partial}{\partial \xi_1} \kappa(\xi) = -\frac{\xi_1}{2\xi_2}$$

$$\text{and} \quad E_\xi[T_2] = \frac{\partial}{\partial \xi_2} \kappa(\xi) = -\frac{1}{2\xi_2} + \frac{\xi_1^2}{4\xi_2^2}$$

for  $\xi = (\xi_1, \xi_2)^t \in \mathbb{R} \times (-\infty, 0)$ . In terms of the common parametrization of the univariate normal distribution (see Example 2.10), we find for a random variable  $X \sim \mathcal{N}(\mu, \sigma^2)$  with parameters  $(\mu, \sigma^2) = (-\xi_1/(2\xi_2), -1/(2\xi_2))^t$  that the means of  $X = T_1(X)$  and  $X^2 = T_2(X)$  are given by  $\mu$  and  $\sigma^2 + \mu^2$ ; the variance of  $X$  is then equal to  $\sigma^2$ .

*Example 2.53* For  $\mathcal{P} = \mathcal{G}(\bullet, \bullet)$ , we conclude from Example 2.40 that

$$\kappa(\xi) = \ln(\Gamma(\xi_2)) - \xi_2 \ln(-\xi_1), \quad \xi = (\xi_1, \xi_2)^t \in (-\infty, 0) \times (0, \infty).$$

Hence, by using Corollary 2.50, we have for  $\mathbf{T}(x) = (x, \ln(x))^t$ ,  $x \in (0, \infty)$ , that

$$E_\xi[\mathbf{T}] = \nabla \kappa(\xi) = \left( -\frac{\xi_2}{\xi_1}, \psi(\xi_2) - \ln(-\xi_1) \right)^t, \quad \xi \in (-\infty, 0) \times (0, \infty),$$

where  $\psi(y) = [\ln(\Gamma(y))]' = \Gamma'(y)/\Gamma(y)$ ,  $y \in (0, \infty)$ , denotes the digamma function (see [22, p. 258]). Moreover, it follows that

$$\mathbf{Cov}_\zeta(T) = \mathbf{H}_\kappa(\zeta) = \begin{pmatrix} \frac{\zeta_2}{\zeta_1^2} & -\frac{1}{\zeta_1} \\ -\frac{1}{\zeta_1} & \psi_1(\zeta_2) \end{pmatrix}, \quad \zeta \in (-\infty, 0) \times (0, \infty),$$

where  $\psi_1 = \psi'$  denotes the trigamma function (see [22, p. 260]). In particular, the mean and variance of the gamma distribution  $P_\zeta^*$  are given by  $-\zeta_2/\zeta_1$  and  $\zeta_2/\zeta_1^2$ .

*Example 2.54* For the EF  $\mathcal{P} = \mathcal{N}_k(\bullet, \Sigma_0)$  of multivariate normal distributions as in Example 2.18, which is already given in natural parametrization ( $\zeta = \mu$ ,  $C^* = C$ ), we have

$$C(\mu) = \exp \left\{ -\frac{\|\mu\|_{\Sigma_0}^2}{2} \right\} = \exp \left\{ -\frac{1}{2} \mu^t \Sigma_0^{-1} \mu \right\}, \quad \mu \in \mathbb{R}^k.$$

Then, by Corollary 2.50 with  $\kappa(\mu) = \mu^t \Sigma_0^{-1} \mu/2$ ,

$$E_\mu[T] = \nabla \kappa(\mu) = \Sigma_0^{-1} \mu,$$

$$\mathbf{Cov}_\mu(T) = \mathbf{H}_\kappa(\mu) = \Sigma_0^{-1}, \quad \mu \in \mathbb{R}^k,$$

where  $T(x) = \Sigma_0^{-1} x$ ,  $x \in \mathbb{R}^k$ . Hence, for a random vector  $X \sim \mathcal{N}_k(\mu, \Sigma_0)$ , we find with the identity  $X = \Sigma_0 T(X)$  that the mean and covariance matrix of  $X$  are given by  $\Sigma_0 E_\mu[T] = \mu$  and  $\Sigma_0 \mathbf{Cov}_\mu(T) \Sigma_0 = \Sigma_0$ .

Lemma 2.47 and Corollary 2.50 may be applied to establish the following result. It is important for the derivation of optimal statistical tests.

**Corollary 2.55** *Under the assumptions of Lemma 2.47, the mapping*

$$\beta_\varphi : A \rightarrow \mathbb{R} : \zeta \mapsto E_\zeta[\varphi]$$

*is infinitely often differentiable with gradient*

$$\nabla \beta_\varphi(\zeta) = E_\zeta[\varphi T] - E_\zeta[\varphi] E_\zeta[T] = \mathbf{Cov}_\zeta(\varphi, T), \quad \zeta \in A.$$

**Proof** Since

$$\beta_\varphi(\xi) = C^*(\xi) \int \varphi e^{\xi' T} h d\mu, \quad \xi \in A,$$

$\beta_\varphi$  is infinitely often differentiable by Lemma 2.47 and Corollary 2.50. Moreover, for  $1 \leq i \leq k$ , we obtain that

$$\frac{\partial}{\partial \xi_i} C^*(\xi) = \frac{\partial}{\partial \xi_i} \left[ \left( \int e^{\xi' T} h d\mu \right)^{-1} \right] = - \frac{\int T_i e^{\xi' T} h d\mu}{\left( \int e^{\xi' T} h d\mu \right)^2} = -C^*(\xi) E_\xi[T_i]$$

and, hence,

$$\begin{aligned} \frac{\partial}{\partial \xi_i} \beta_\varphi(\xi) &= \left( \frac{\partial}{\partial \xi_i} C^*(\xi) \right) \int \varphi e^{\xi' T} h d\mu + C^*(\xi) \int \varphi T_i e^{\xi' T} h d\mu \\ &= E_\xi[\varphi T_i] - E_\xi[\varphi] E_\xi[T_i]. \end{aligned}$$

□

There is an important function, which describes the relationship between the natural parameter space  $\Xi^*$  and the expectation space of the vector  $T$  of statistics. Its properties allow for another parametrization of a regular EF, and they are also useful in point estimation.

**Theorem 2.56** *Let  $\mathcal{P}$  be a full EF with minimal canonical representation (2.6). Then, the so-called mean value function*

$$\pi : \text{int}(\Xi^*) \rightarrow \pi(\text{int}(\Xi^*)) : \quad \xi \mapsto E_\xi[T]$$

*has the following properties:*

- (a)  $\pi(\xi) = \nabla \kappa(\xi)$  for every  $\xi \in \text{int}(\Xi^*)$ .
- (b)  $\pi$  is continuously differentiable with Jacobian matrix

$$\mathbf{D}_\pi(\xi) = \mathbf{H}_\kappa(\xi) = \mathbf{Cov}_\xi(T) > 0, \quad \xi \in \text{int}(\Xi^*).$$

- (c)  $\pi$  is bijective and has a continuously differentiable inverse function  $\pi^{-1} : \pi(\text{int}(\Xi^*)) \rightarrow \text{int}(\Xi^*)$ .

**Proof** Assertions (a) and (b) directly follow from Corollary 2.50 together with Theorem 2.23 and Lemma 2.26. For statement (c), let  $\xi, \eta \in \text{int}(\Xi^*)$  with  $\xi \neq \eta$ . By Theorem 2.23 and 2.44(b),  $\kappa$  is strictly convex on  $\Xi^*$ . Hence, by using part (a),

$$(\eta - \xi)^t \pi(\xi) < \kappa(\eta) - \kappa(\xi),$$

and, for symmetry reasons,

$$(\zeta - \eta)^t \pi(\eta) < \kappa(\zeta) - \kappa(\eta)$$

(see, for instance, [19, pp. 69/70]). Summation of both inequalities leads to

$$(\zeta - \eta)^t (\pi(\eta) - \pi(\zeta)) < 0.$$

This yields  $\pi(\zeta) \neq \pi(\eta)$ , and  $\pi$  is shown to be one-to-one. By definition,  $\pi$  is onto and thus bijective. The remaining part follows by assertion (b) and the inverse function theorem (see, e.g., [52, p. 221]).  $\square$

**Examples:**  $\mathfrak{b}(n_0, \bullet)$ ,  $\mathcal{IN}(\bullet, \bullet)$ ,  $\mathcal{N}(\bullet, \bullet)$

*Example 2.57* For  $\mathcal{P} = \mathfrak{b}(n_0, \bullet)$ , we obtain from Example 2.38 that

$$\kappa(\zeta) = n_0 \ln(1 + e^\zeta), \quad \zeta \in \mathbb{R}.$$

Hence, the mean value function is

$$\pi : \mathbb{R} \rightarrow (0, n_0) : \zeta \mapsto \frac{n_0 \exp\{\zeta\}}{1 + \exp\{\zeta\}}$$

with inverse function

$$\pi^{-1} : (0, n_0) \rightarrow \mathbb{R} : t \mapsto \ln\left(\frac{t}{n_0 - t}\right).$$

*Example 2.58* For  $\mathcal{P} = \mathcal{IN}(\bullet, \bullet)$ , we conclude from Example 2.41 that

$$\kappa(\zeta) = -\frac{\ln(-2\zeta_2)}{2} - 2\sqrt{\zeta_1\zeta_2}, \quad \zeta = (\zeta_1, \zeta_2)^t \in (-\infty, 0)^2.$$

Hence, the mean value function is

$$\pi : (-\infty, 0)^2 \rightarrow \pi((-\infty, 0)^2) : \zeta \mapsto \left( \sqrt{\frac{\zeta_2}{\zeta_1}}, -\frac{1}{2\zeta_2} + \sqrt{\frac{\zeta_1}{\zeta_2}} \right)^t$$

with inverse function

$$\pi^{-1} : \pi((-\infty, 0)^2) \rightarrow (-\infty, 0)^2 : t \mapsto \left( \frac{1}{2t_1(1 - t_1t_2)}, \frac{t_1}{2(1 - t_1t_2)} \right)^t.$$

*Example 2.59* For  $\mathcal{P} = \mathcal{N}(\bullet, \bullet)$ , we obtain from Example 2.52 that the mean value function is given by

$$\pi : \mathbb{R} \times (-\infty, 0) \rightarrow \pi(\mathbb{R} \times (-\infty, 0)) : \quad \xi \mapsto \left( -\frac{\xi_1}{2\xi_2}, \frac{\xi_1^2 - 2\xi_2}{4\xi_2^2} \right)^t$$

with inverse function

$$\pi^{-1} : \pi(\mathbb{R} \times (-\infty, 0)) \rightarrow \mathbb{R} \times (-\infty, 0) : \quad t \mapsto \left( \frac{t_1}{t_2 - t_1^2}, \frac{1}{2(t_1^2 - t_2)} \right)^t.$$

Moreover, we have

$$\pi(\mathbb{R} \times (-\infty, 0)) = \{(x, y)^t \in \mathbb{R}^2 : x^2 < y\}.$$

In Theorem 2.56, the image of the mapping  $\pi$  can be stated explicitly provided that the EF is steep (see [4, p. 117] and [20, p. 71] for the definition and an illustration). The concept of steepness is originally introduced in a more general context for convex functions. Here, for simplicity, we give a definition directly for the full EF with minimal canonical representation.

**Definition 2.60** Let  $\mathcal{P}$  be a full EF with minimal canonical representation (2.6).  $\mathcal{P}$  is called *steep* if, for all  $\eta \in \Xi^* \setminus \text{int}(\Xi^*)$  and  $\zeta \in \text{int}(\Xi^*)$ , we have

$$\lim_{\alpha \nearrow 1} \frac{\partial}{\partial \alpha} \kappa(\zeta^{(\alpha)}) = \lim_{\alpha \nearrow 1} (\eta - \zeta)^t \nabla \kappa(\zeta^{(\alpha)}) = \infty, \quad (2.12)$$

where  $\zeta^{(\alpha)} = \zeta + \alpha(\eta - \zeta)$ ,  $\alpha \in (0, 1)$ .

Roughly speaking, Eq. (2.12) ensures that the cumulant function of the EF has infinite slope when moving towards boundary points in  $\Xi^*$  on straight lines. A regular EF trivially meets the condition, since such boundary points do not exist in that case.



**Lemma 2.61** *Let  $\mathcal{P}$  be an EF with minimal canonical representation (2.6). Then, the following implications are true:*

$$\mathcal{P} \text{ is regular} \quad \Rightarrow \quad \mathcal{P} \text{ is steep} \quad \Rightarrow \quad \mathcal{P} \text{ is full} .$$

As pointed out in Example 2.41, the full EF of inverse normal distributions is not regular, but it can be shown to be steep (see [4, p. 117] or [20, p. 72]). An example for a full EF which is not steep is presented in [20, p. 86].

*Remark 2.62* In Theorem 2.56, we have

$$\pi(\text{int}(\Xi^*)) \subset \text{int}(M) , \quad (2.13)$$

where  $M$  denotes the convex support of  $\nu^T$ , i.e., the closed convex hull of the support of  $\nu^T$  (see, e.g., [37, p. 668]). If  $\mathcal{P}$  is steep, then equality holds true in formula (2.13); for a proof, see, e.g., [20, pp. 74/75]. Recall that the support of  $\nu^T$  coincides with the support of  $(P_\zeta^*)^T$  for every  $\zeta \in \Xi^*$ , since these measures are all equivalent; this finding may ease to derive  $M$ .

The properties of the mean value function  $\pi$  in Theorem 2.56 allow for the following parametrization of a regular EF.

**Definition 2.63** Let  $\mathcal{P}$  be a regular EF with minimal canonical representation (2.6). Then,

$$\tilde{P}_m = P_{\pi^{-1}(m)}^* , \quad m \in \pi(\Xi^*) = \text{int}(M)$$

is called the *mean value parametrization* of  $\mathcal{P}$ , where  $M$  is as defined in Remark 2.62.

Some examples illustrate the mean value parametrization of an EF.

**Examples:**  $\text{poi}(\bullet)$ ,  $\mathcal{N}_k(\bullet, \Sigma_0)$

*Example 2.64* According to Example 2.37,  $\mathcal{P} = \text{poi}(\bullet)$  is regular with canonical representation given by the counting densities

$$f_\zeta^*(x) = C^*(\zeta) e^{\zeta x} \frac{1}{x!} , \quad x \in \mathbb{N}_0 ,$$

with  $C^*(\zeta) = \exp\{-e^\zeta\}$  for  $\zeta \in \Xi^* = \mathbb{R}$ . Hence, the cumulant function is

$$\kappa(\zeta) = e^\zeta, \quad \zeta \in \mathbb{R},$$

and we conclude by Corollary 2.50 that the mean and variance of the statistic  $T(x) = x$ ,  $x \in \mathbb{N}_0$ , under  $P_\zeta^*$  are given by

$$E_\zeta[T] = \kappa'(\zeta) = e^\zeta,$$

$$\text{Var}_\zeta(T) = \kappa''(\zeta) = e^\zeta.$$

Here, the mean value function  $\pi$  is

$$\pi : \mathbb{R} \rightarrow (0, \infty) : \zeta \mapsto e^\zeta$$

with inverse function

$$\pi^{-1} : (0, \infty) \rightarrow \mathbb{R} : t \mapsto \ln(t).$$

To obtain the mean value parametrization, we set  $\zeta = \pi^{-1}(m) = \ln(m)$  in  $f_\zeta^*$  and arrive at the counting densities

$$\tilde{f}_m(x) = e^{-m} m^x \frac{1}{x!}, \quad x \in \mathbb{N}_0,$$

for  $m \in (0, \infty)$ , which is the common representation of  $\text{poi}(\bullet)$  as introduced in Example 2.3. In this parametrization, mean and variance of  $T$  under  $\tilde{P}_m$  are both equal to  $m$ .

*Example 2.65* For  $\mathcal{P} = \mathcal{N}_k(\bullet, \Sigma_0)$ , we obtain from Example 2.54 that the mean value function  $\pi$  and its inverse function  $\pi^{-1}$  are given by

$$\pi : \mathbb{R}^k \rightarrow \mathbb{R}^k : \mu \mapsto \Sigma_0^{-1} \mu$$

$$\text{and} \quad \pi^{-1} : \mathbb{R}^k \rightarrow \mathbb{R}^k : t \mapsto \Sigma_0 t.$$

To find the mean value parametrization, we set  $\mu = \pi^{-1}(m) = \Sigma_0 m$  in  $f_{\mu, \Sigma_0}$  (see Example 2.18) and obtain the Lebesgue densities

$$\tilde{f}_m(x) = \exp \left\{ -||m||_{\Sigma_0^{-1}/2}^2 \right\} e^{m^t x} h(x), \quad x \in \mathbb{R}^k,$$

for  $\mathbf{m} \in \mathbb{R}^k$ . In this parametrization, mean vector and covariance matrix of  $\mathbf{T}$  under  $\tilde{P}_{\mathbf{m}}$  are given by  $\mathbf{m}$  and  $\Sigma_0^{-1}$ , respectively. Note that the parametrizations coincide if  $\Sigma_0 = \mathbf{E}_k$ , i.e., if  $\Sigma_0$  is the  $k$ -dimensional unit matrix.

Related to the mean value parametrization is the so-called variance function, which leads to a wide range of publications with structural considerations of EFs.

*Remark 2.66* An EF  $\mathcal{P} = \{P_{\xi}^* : \xi \in \text{int}(\Xi^*)\}$  with minimal canonical representation (2.6) and  $\mathbf{T}(\mathbf{x}) = \mathbf{x}$ ,  $\mathbf{x} \in \mathcal{X}$ , i.e., with  $\nu$ -densities

$$\frac{dP_{\xi}^*}{d\nu}(\mathbf{x}) = C^*(\xi) e^{\xi^t \mathbf{x}} = e^{\xi^t \mathbf{x} - \kappa(\xi)}, \quad \mathbf{x} \in \mathbb{R}^k, \quad (2.14)$$

for  $\xi \in \text{int}(\Xi^*)$ , forms a natural EF with generating measure  $\nu$  as introduced in [37, Chapter 54] and Remark 2.36. Here, minimality of representation (2.14) ensures that  $\text{int}(\Xi^*) \neq \emptyset$  and that the support of  $\nu$  is not contained in an affine hyperplane of  $\mathbb{R}^k$ . If  $\mathcal{P}$  is moreover steep, the *variance function* of  $\mathcal{P}$  is defined by

$$V_{\mathcal{P}}(\mathbf{m}) = \text{Cov}_{\tilde{P}_{\mathbf{m}}}(X), \quad \mathbf{m} \in \text{int}(M),$$

i.e.,  $V_{\mathcal{P}}$  maps the mean  $\mathbf{m}$  of the random vector  $X$  to the covariance matrix of  $X$  under  $\tilde{P}_{\mathbf{m}}$ , where  $\text{int}(M)$  is the domain of the means as defined in Remark 2.62 (see, e.g., [37, pp. 669/670]). By Theorem 2.56, we have the identity

$$V_{\mathcal{P}}(\mathbf{m}) = \mathbf{H}_{\kappa}(\pi^{-1}(\mathbf{m})), \quad \mathbf{m} \in \text{int}(M).$$

In Example 2.64,  $\mathcal{P}$  forms a natural EF with linear variance function  $V_{\mathcal{P}}(\mathbf{m}) = \mathbf{m}$ ,  $\mathbf{m} \in (0, \infty)$ . For  $\Sigma_0 = \mathbf{E}_k$ , the EF  $\mathcal{P}$  in Example 2.65 is natural and has constant variance function  $V_{\mathcal{P}}(\mathbf{m}) = \mathbf{E}_k$ ,  $\mathbf{m} \in \mathbb{R}^k$ . One motivation for considering variance functions is that it enables to characterize, classify, and construct natural EFs. The interested reader is referred to, e.g., [41], [37, pp. 669/670 and pp. 678–691], and the references therein. A recent overview on the topic is provided in [3].

## Chapter 3

# Distributional and Statistical Properties



In the third chapter, some further distributional and structural properties of an EF are examined, and basic results for statistical inference in EFs are provided.

### 3.1 Generating Functions

At the beginning of Sect. 2.2, we have mentioned that the vector  $\mathbf{T} = (T_1, \dots, T_k)^t$  of statistics is fundamental in statistical inference. Theorem 2.49 already shows how to derive the moments of  $\mathbf{T}$  in the EF with natural parametrization. To find the distribution of  $\mathbf{T}$  under  $P_{\boldsymbol{\zeta}}^*$  and for several other applications, it is useful to have simple representations of the *moment generating function*  $m_{\mathbf{T}}(\mathbf{a}) = E_{\boldsymbol{\zeta}}[\exp\{\mathbf{a}^t \mathbf{T}\}]$ ,  $\mathbf{a} \in \mathbb{R}^k$ , and the *characteristic function*  $\varphi_{\mathbf{T}}(\mathbf{a}) = E_{\boldsymbol{\zeta}}[\exp\{i\mathbf{a}^t \mathbf{T}\}]$ ,  $\mathbf{a} \in \mathbb{R}^k$ , available. These are stated in the following lemma.

**Lemma 3.1** *Let  $\mathcal{P}$  be an EF with canonical representation (2.6), and let  $\boldsymbol{\zeta} \in \Xi$ .*

(a) *If  $A \subset \mathbb{R}^k$  is such that  $\boldsymbol{\zeta} + A = \{\boldsymbol{\zeta} + \mathbf{a} : \mathbf{a} \in A\} \subset \Xi$ , then*

$$m_{\mathbf{T}}(\mathbf{a}) = \frac{C^*(\boldsymbol{\zeta})}{C^*(\boldsymbol{\zeta} + \mathbf{a})}, \quad \mathbf{a} \in A,$$

*where integration has been carried out with respect to  $P_{\boldsymbol{\zeta}}^*$ .*

(b) *If the function  $C^*$  is extended to the complex set  $\{\boldsymbol{\zeta} + i\mathbf{a} \in \mathbb{C}^k : \mathbf{a} \in \mathbb{R}^k\}$  in a canonical way, then*

$$\varphi_{\mathbf{T}}(\mathbf{a}) = \frac{C^*(\boldsymbol{\zeta})}{C^*(\boldsymbol{\zeta} + i\mathbf{a})}, \quad \mathbf{a} \in \mathbb{R}^k,$$

*where integration has been carried out with respect to  $P_{\boldsymbol{\zeta}}^*$ .*

**Proof**

(a) Let  $\zeta \in \Xi$  and  $\zeta + A \subset \Xi$ . Obviously, we have for  $a \in A$

$$\begin{aligned} m_T(a) &= E_\zeta[e^{a^t T}] = \int e^{a^t T(x)} f_\zeta^*(x) d\mu(x) \\ &= \frac{C^*(\zeta)}{C^*(\zeta + a)} \int C^*(\zeta + a) e^{(\zeta + a)^t T(x)} h(x) d\mu(x) \\ &= \frac{C^*(\zeta)}{C^*(\zeta + a)}. \end{aligned}$$

(b) The representation is obtained by proceeding as in the proof of statement (a).  $\square$

**Examples:**  $\text{poi}(\bullet)$ ,  $\text{b}(n_0, \bullet)$ ,  $\text{m}(n_0, \bullet)$ ,  $\mathcal{N}(\bullet, \bullet)$ ,  $\mathcal{G}(\bullet, \bullet)$ ,  $\mathcal{N}_k(\bullet, \Sigma_0)$

*Example 3.2* For  $\mathcal{P} = \text{poi}(\bullet)$ , we have by Example 2.37 that  $C^*(\zeta) = \exp\{-e^\zeta\}$ ,  $\zeta \in \mathbb{R}$ , and  $T(x) = x$ ,  $x \in \mathcal{X} = \mathbb{N}_0$ . Application of Lemma 3.1(a) then yields

$$m_T(b) = \exp\left\{e^\zeta(e^b - 1)\right\}, \quad b \in \mathbb{R},$$

where integration is with respect to  $P_\zeta^*$ . The transformation  $\zeta = \ln(a)$  then gives the moment generating function of  $T$  in the original parametrization (see Example 2.3), i.e.,

$$m_T(b) = \exp\left\{a(e^b - 1)\right\}, \quad b \in \mathbb{R},$$

where integration is with respect to  $P_a = \text{poi}(a)$ .

*Example 3.3* For  $\mathcal{P} = \text{b}(n_0, \bullet)$ , we have by Example 2.38 that  $C^*(\zeta) = (1 + \exp\{\zeta\})^{-n_0}$ ,  $\zeta \in \mathbb{R}$ , and  $T(x) = x$ ,  $x \in \mathcal{X} = \{0, 1, \dots, n_0\}$ . Application of Lemma 3.1(a) then yields

$$m_T(a) = \left( \frac{1 + \exp\{\zeta + a\}}{1 + \exp\{\zeta\}} \right)^{n_0}, \quad a \in \mathbb{R},$$

where integration is with respect to  $P_\zeta^*$ . The transformation  $\zeta = \ln(p/(1 - p))$  then gives the moment generating function of  $T$  in the original parametrization

(see Example 2.4), i.e.,

$$m_T(a) = (1 - p + pe^a)^{n_0}, \quad a \in \mathbb{R},$$

where integration is with respect to  $P_p = \mathbf{b}(n_0, p)$ .

*Example 3.4* For  $\mathcal{P} = \mathbf{m}(n_0, \bullet)$  with minimal canonical representation  $\tilde{f}_\zeta^*$ ,  $\zeta \in \mathbb{R}^{k-1}$ , as shown in Example 2.46, we have  $C^*(\zeta) = (1 + \sum_{j=1}^{k-1} \exp\{\zeta_j\})^{-n_0}$  for  $\zeta \in \mathbb{R}^{k-1}$  and  $\mathbf{T}(\mathbf{x}) = (x_1, \dots, x_{k-1})^t$  for  $\mathbf{x} \in \mathcal{X} = \{(y_1, \dots, y_k)^t \in \mathbb{N}_0^k : \sum_{i=1}^k y_i = n_0\}$ . Application of Lemma 3.1(a) then yields

$$m_T(\mathbf{a}) = \left( \frac{1 + \sum_{j=1}^{k-1} \exp\{\zeta_j + a_j\}}{1 + \sum_{j=1}^{k-1} \exp\{\zeta_j\}} \right)^{n_0}, \quad \mathbf{a} = (a_1, \dots, a_{k-1})^t \in \mathbb{R}^{k-1},$$

where integration is with respect to  $\tilde{P}_\zeta^*$  with counting density  $\tilde{f}_\zeta^*$ . For  $k = 2$ , we are in the case of Example 3.3.

*Example 3.5* In Example 2.39, the canonical representation of  $\mathcal{P} = \mathcal{N}(\bullet, \bullet)$  yields

$$C^*(\zeta) = \sqrt{-2\zeta_2} \exp \left\{ \frac{\zeta_1^2}{4\zeta_2} \right\}, \quad \zeta = (\zeta_1, \zeta_2)^t \in \mathbb{R} \times (-\infty, 0),$$

and  $\mathbf{T}(x) = (T_1(x), T_2(x))^t = (x, x^2)^t$ ,  $x \in \mathbb{R}$ . Then, by Lemma 3.1(b), the characteristic function of  $T_1$  is

$$\begin{aligned} \varphi_{T_1}(a_1) &= \varphi_T((a_1, 0)^t) = \frac{C^*(\zeta)}{C^*(\zeta + i(a_1, 0)^t)} \\ &= \exp \left\{ \frac{\zeta_1^2}{4\zeta_2} - \frac{(\zeta_1 + ia_1)^2}{4\zeta_2} \right\} = \exp \left\{ -\frac{i\zeta_1 a_1}{2\zeta_2} + \frac{a_1^2}{4\zeta_2} \right\}, \quad a_1 \in \mathbb{R}, \end{aligned}$$

where integration is with respect to  $P_\zeta^*$ . In terms of the common parametrization of the univariate normal distribution (see Example 2.10), we find for the characteristic function of a random variable  $X = T_1(X) \sim \mathcal{N}(\mu, \sigma^2)$  with  $(\zeta_1, \zeta_2) = (\mu/\sigma^2, -1/(2\sigma^2))$  the well known expression

$$\varphi_{T_1}(t) = \exp \left\{ i\mu t - \frac{\sigma^2 t^2}{2} \right\}, \quad t \in \mathbb{R},$$

where integration is with respect to  $P_{\mu, \sigma^2} = \mathcal{N}(\mu, \sigma^2)$ .

*Example 3.6* For  $\mathcal{P} = \mathcal{G}(\bullet, \bullet)$ , we have by Example 2.40 that  $C^*(\boldsymbol{\zeta}) = (-\zeta_1)^{\zeta_2} / \Gamma(\zeta_2)$ ,  $\boldsymbol{\zeta} = (\zeta_1, \zeta_2)^t \in (-\infty, 0) \times (0, \infty)$ , and  $\mathbf{T}(x) = (x, \ln(x))^t$ ,  $x \in (0, \infty)$ . Application of Lemma 3.1(a) then yields

$$m_T(\mathbf{a}) = \frac{\Gamma(\zeta_2 + a_2)}{\Gamma(\zeta_2)} \frac{(-\zeta_1)^{\zeta_2}}{(-\zeta_1 - a_1)^{\zeta_2 + a_2}}$$

for  $\mathbf{a} = (a_1, a_2)^t \in (-\infty, -\zeta_1) \times (-\zeta_2, \infty)$ , where integration is with respect to  $P_{\boldsymbol{\zeta}}^*$ . The transformation  $\zeta_1 = -1/\alpha$ ,  $\zeta_2 = \beta$  then gives the moment generating function of  $\mathbf{T}$  in the original parametrization (see Example 2.12), i.e.,

$$m_T(\mathbf{a}) = \frac{\Gamma(\beta + a_2)}{\Gamma(\beta)} \frac{\alpha^{a_2}}{(1 - \alpha a_1)^{\beta + a_2}}, \quad \mathbf{a} \in (-\infty, 1/\alpha) \times (-\beta, \infty),$$

where integration is with respect to  $P_{\alpha, \beta} = \mathcal{G}(\alpha, \beta)$ .

*Example 3.7* For the EF  $\mathcal{P} = \mathcal{N}_k(\bullet, \boldsymbol{\Sigma}_0)$  of multivariate normal distributions as in Example 2.18, we have

$$C(\boldsymbol{\mu}) = \exp \left\{ -\frac{\|\boldsymbol{\mu}\|_{\boldsymbol{\Sigma}_0}^2}{2} \right\} = \exp \left\{ -\frac{1}{2} \boldsymbol{\mu}^t \boldsymbol{\Sigma}_0^{-1} \boldsymbol{\mu} \right\}, \quad \boldsymbol{\mu} \in \mathbb{R}^k,$$

and  $\mathbf{T}(\mathbf{x}) = \boldsymbol{\Sigma}_0^{-1} \mathbf{x}$ ,  $\mathbf{x} \in \mathbb{R}^k$ . Hence, by Lemma 3.1(b) with  $\boldsymbol{\zeta} = \boldsymbol{\mu}$  and  $C^* = C$  (cf. Example 2.54), the characteristic function of  $\mathbf{T}$  is given by

$$\begin{aligned} \varphi_T(\mathbf{a}) &= \exp \left\{ -\frac{1}{2} \boldsymbol{\mu}^t \boldsymbol{\Sigma}_0^{-1} \boldsymbol{\mu} + \frac{1}{2} (\boldsymbol{\mu} + i\mathbf{a})^t \boldsymbol{\Sigma}_0^{-1} (\boldsymbol{\mu} + i\mathbf{a}) \right\} \\ &= \exp \left\{ i \boldsymbol{\mu}^t \boldsymbol{\Sigma}_0^{-1} \mathbf{a} - \frac{1}{2} \mathbf{a}^t \boldsymbol{\Sigma}_0^{-1} \mathbf{a} \right\}, \quad \mathbf{a} \in \mathbb{R}^k, \end{aligned}$$

where integration is with respect to  $P_{\boldsymbol{\mu}, \boldsymbol{\Sigma}_0} = \mathcal{N}_k(\boldsymbol{\mu}, \boldsymbol{\Sigma}_0)$ . This yields the well known expression for the characteristic function of a random vector  $\mathbf{X} = \boldsymbol{\Sigma}_0 \mathbf{T}(\mathbf{X}) \sim \mathcal{N}_k(\boldsymbol{\mu}, \boldsymbol{\Sigma}_0)$  as

$$\varphi_{\boldsymbol{\Sigma}_0 \mathbf{T}}(\mathbf{t}) = \varphi_T(\boldsymbol{\Sigma}_0 \mathbf{t}) = \exp \left\{ i \boldsymbol{\mu}^t \mathbf{t} - \frac{1}{2} \mathbf{t}^t \boldsymbol{\Sigma}_0 \mathbf{t} \right\}, \quad \mathbf{t} \in \mathbb{R}^k.$$

It is known in general that if the moment generating function  $m_T$  of a random vector  $T$  with values in  $\mathbb{R}^k$  is finite in a neighborhood of  $\mathbf{0} \in \mathbb{R}^k$ , then all partial derivatives of  $m_T$  exist finitely, and moments of  $T$  can be obtained by calculating respective partial derivatives at  $\mathbf{0}$ . The logarithm of a moment generating function is called a *cumulant generating function*, and its partial derivatives at  $\mathbf{0}$  are termed *cumulants* of  $T$ , which are applied, for instance, in Edgeworth expansions (see, e.g., [36] and [35]). For further details on moments and cumulants as well as their interrelation, we refer to the literature.

**Remark 3.8** In the situation of Lemma 3.1(a) with  $k = 1$  (one-parameter case), let  $K_T = \ln(m_T)$  denote the cumulant generating function of  $T$ . Then, we find the relation

$$K_T(a) = \ln \left( \frac{C^*(\zeta)}{C^*(\zeta + a)} \right) = \kappa(\zeta + a) - \kappa(\zeta), \quad a \in A,$$

to the cumulant function  $\kappa = -\ln(C^*)$  of the EF (see Definition 2.43). If  $0 \in \text{int}(A)$ , the  $n$ -th cumulant  $\mathcal{K}_T^{(n)}$  of  $T$  is defined as the  $n$ -th derivative of  $K_T$  at 0, i.e., we have

$$\mathcal{K}_T^{(n)} = \frac{\partial^n}{\partial a^n} K_T(a)|_{a=0} = \frac{\partial^n}{\partial \zeta^n} \kappa(\zeta).$$

In particular, we obtain the first two cumulants  $\mathcal{K}_T^{(1)}$  and  $\mathcal{K}_T^{(2)}$  of  $T$ , which are known to give the mean and variance of  $T$  under  $P_\zeta^*$ , as respective derivatives of  $\kappa$  (as already shown in Corollary 2.50).

## 3.2 Marginal and Conditional Distributions

It is natural to ask whether the EF structure is preserved by considering marginal and conditional distributions. To begin with, the restriction of the sample space to a subset  $\mathcal{Y} \subset X$  yields an EF, again (see, e.g., [44, p. 37]).

**Lemma 3.9** *Let  $\mathcal{P}$  be an EF with representation (2.1) and  $v = h\mu$  as in Lemma 2.15. Moreover, let  $\mathcal{Y} \in \mathcal{B}$  with  $v(\mathcal{Y}) > 0$ , and let  $Q_\vartheta(\cdot) = P_\vartheta(\cdot|\mathcal{Y})$  for  $\vartheta \in \Theta$ . Then, the family  $\{Q_\vartheta : \vartheta \in \Theta\}$  of conditional distributions forms an EF with  $\mu$ -density*

$$\frac{dQ_\vartheta}{d\mu}(y) = C_{\mathcal{Y}}(\vartheta) \exp \left\{ \sum_{j=1}^k Z_j(\vartheta) T_j(y) \right\} h(y), \quad y \in \mathcal{Y},$$

and normalizing constant  $C_{\mathcal{Y}}(\vartheta)$  for  $\vartheta \in \Theta$ .



**Proof** Let  $\vartheta \in \Theta$ . Since  $\nu(\mathcal{Y}) > 0$ , we have  $P_{\vartheta}(\mathcal{Y}) > 0$  by Remark 2.16. Then, for  $A \in \mathcal{B} \cap \mathcal{Y}$ , we find

$$Q_{\vartheta}(A) = \frac{P_{\vartheta}(A \cap \mathcal{Y})}{P_{\vartheta}(\mathcal{Y})} = \frac{\int_A f_{\vartheta} d\mu}{\int_{\mathcal{Y}} f_{\vartheta} d\mu},$$

and thus

$$\frac{dQ_{\vartheta}}{d\mu} = \frac{f_{\vartheta}}{\int_{\mathcal{Y}} f_{\vartheta} d\mu},$$

where  $\int_{\mathcal{Y}} f_{\vartheta} d\mu = C(\vartheta)/C_{\mathcal{Y}}(\vartheta)$ , say. □

### Example: $\mathcal{E}(\bullet)$

*Example 3.10* Let  $\mathcal{P} = \mathcal{E}(\bullet)$ ,  $P_{\alpha} = \mathcal{E}(\alpha)$  for  $\alpha \in (0, \infty)$ , and  $\mathcal{Y} = [b, \infty)$  for some fixed constant  $b \in (0, \infty)$ . Then, for  $\alpha \in (0, \infty)$ ,  $P_{\alpha}(\cdot|\mathcal{Y})$  has Lebesgue density

$$C_{\mathcal{Y}}(\alpha) \exp\left\{-\frac{y}{\alpha}\right\}, \quad y \in [b, \infty),$$

where

$$C_{\mathcal{Y}}(\alpha) = \left(\int_b^{\infty} \exp\left\{-\frac{y}{\alpha}\right\} dy\right)^{-1} = \frac{\exp\{b/\alpha\}}{\alpha}, \quad \alpha \in (0, \infty).$$

Hence,  $\{P_{\alpha}(\cdot|\mathcal{Y}) : \alpha \in (0, \infty)\}$  is the EF of exponential distributions with location parameter  $b$ .

The conditional EF in Lemma 3.9 is not to be mixed up with the truncated EF, which does not form an EF in the proper sense (see Sect. 2.1).

For the EF  $\mathcal{P}^T$  of distributions of  $\mathbf{T}$ , the following results for marginal and conditional distributions can be shown; for the proof, see, e.g., [35, pp. 255–257].

**Theorem 3.11** *Let  $\mathcal{P}$  be an EF with canonical representation (2.6) and  $k \geq 2$ . Moreover, let  $\mathbf{U} = (T_1, \dots, T_r)^t$  and  $\mathbf{V} = (T_{r+1}, \dots, T_k)^t$  for some fixed  $r \in \{1, \dots, k-1\}$ .*

(a) *Let  $\zeta_1, \dots, \zeta_r$  be fixed. Then, there exists a  $\sigma$ -finite measure  $Q_{\zeta_1, \dots, \zeta_r}$  on  $\mathbb{B}^{k-r}$  depending on  $\zeta_1, \dots, \zeta_r$ , such that the distributions of  $\mathbf{V}$  form an EF with*

*densities*

$$\frac{d(P_\xi^*)^V}{dQ_{\zeta_1, \dots, \zeta_r}}(t_{r+1}, \dots, t_k) = C(\xi) \exp \left\{ \sum_{j=r+1}^k \zeta_j t_j \right\}, \quad t_{r+1}, \dots, t_k \in \mathbb{R}. \quad (3.1)$$

(b) *There exists a  $\sigma$ -finite measure  $\tilde{Q}_v$  on  $\mathbb{B}^r$ , such that the conditional distributions of  $U$  given  $V = v$  form an EF with densities*

$$\frac{d(P_\xi^*)^{U|V=v}}{d\tilde{Q}_v}(t_1, \dots, t_r) = C_v(\zeta_1, \dots, \zeta_r) \exp \left\{ \sum_{j=1}^r \zeta_j t_j \right\}, \quad t_1, \dots, t_r \in \mathbb{R},$$

where  $C_v(\zeta_1, \dots, \zeta_r)$  denotes a normalizing constant.  $\tilde{Q}_v$  and  $C_v(\zeta_1, \dots, \zeta_r)$  are independent of  $\zeta_{r+1}, \dots, \zeta_k$ .

We shall highlight the findings in Theorem 3.11 from an inferential perspective. Suppose that we are only interested in estimating the parameters  $\zeta_j$ ,  $j \in J$ , for some index set  $J \subsetneq \{1, \dots, k\}$ . If the parameters  $\zeta_j$ ,  $j \notin J$ , are assumed to be known, inference may be carried out in the marginal model based on the distribution of  $(T_j)_{j \in J}$  forming an EF by varying  $\zeta_j$ ,  $j \in J$ . However, if  $\zeta_j$ ,  $j \notin J$ , are unknown, the dominating measure and normalizing constant in the marginal density will be unknown as well, such that inference based on the marginal model will not be feasible or at least gets more involved. In contrast, conditioning  $(T_j)_{j \in J}$  on  $(T_j)_{j \notin J}$  will eliminate the unknown nuisance parameters  $\zeta_j$ ,  $j \notin J$ , while maintaining the EF structure of the model.

*Remark 3.12* If the density in formula (3.1) does not depend on  $\zeta_1, \dots, \zeta_r$ ,  $V$  is called a *cut* for  $\mathcal{P}$  (see, e.g., [37, pp. 673–676], and the references cited therein). Necessary and sufficient conditions for a statistic to be a cut for a (natural) EF can be found in [4, Chapter 10]. For cuts and the related topic of *mixed parametrizations*, the reader is also referred to [44, pp. 240–245] and [20, pp. 78–81].

### 3.3 Product Measures

We are aiming for statistical inference with EFs in later sections, where we assume to have a sample of independent random variables. Hence, we shall now focus on the corresponding family of product measures. From the respective product densities it becomes clear that the EF structure is preserved for the product case. In what follows, let  $\bigotimes_{i=1}^n \mathcal{B}_i$  denote the product  $\sigma$ -algebra of the  $\sigma$ -algebras  $\mathcal{B}_1, \dots, \mathcal{B}_n$  and  $\bigotimes_{i=1}^n Q_i$  denote the product measure of the  $\sigma$ -finite measures  $Q_1, \dots, Q_n$  for  $n \in \mathbb{N}$ . Moreover, for  $n \in \mathbb{N}$ , we use the notation  $\mathcal{B}^n = \bigotimes_{i=1}^n \mathcal{B}$  and  $Q^{(n)} = \bigotimes_{i=1}^n Q$  for the  $n$ -fold product  $\sigma$ -algebra of a  $\sigma$ -algebra  $\mathcal{B}$  and the  $n$ -fold product measure of a  $\sigma$ -finite measure  $Q$ , respectively.

**Theorem 3.13**

- (a) For  $1 \leq i \leq n$ , let  $(X_i, \mathcal{B}_i)$  be a measurable space and  $\mathcal{P}_i = \{P_{\vartheta_i; i} : \vartheta_i \in \Theta_i\}$  be an EF of distributions on  $\mathcal{B}_i$  dominated by a  $\sigma$ -finite measure  $\mu_i$ . Then, the class

$$\left\{ \bigotimes_{i=1}^n P_{\vartheta_i; i} : \vartheta_i \in \Theta_i, 1 \leq i \leq n \right\}$$

of product measures on  $\bigotimes_{i=1}^n \mathcal{B}_i$  forms an EF dominated by  $\bigotimes_{i=1}^n \mu_i$ .

- (b) Let  $\mathcal{P}$  be an EF with representation (2.1). For  $n \in \mathbb{N}$ , let  $P_{\vartheta}^{(n)} = \bigotimes_{i=1}^n P_{\vartheta}$  for  $P_{\vartheta} \in \mathcal{P}$ . Then, the class

$$\mathcal{P}^{(n)} = \left\{ P_{\vartheta}^{(n)} : \vartheta \in \Theta \right\}$$

of product measures on  $\mathcal{B}^n$  forms an EF dominated by  $\mu^{(n)}$ , where, for  $\vartheta \in \Theta$ , a  $\mu^{(n)}$ -density of  $P_{\vartheta}^{(n)}$  is given by

$$f_{\vartheta}^{(n)}(\tilde{\mathbf{x}}) = C(\vartheta)^n \exp \left\{ \sum_{j=1}^k Z_j(\vartheta) T_j^{(n)}(\tilde{\mathbf{x}}) \right\} \left( \prod_{i=1}^n h(x^{(i)}) \right)$$

with statistics

$$T_j^{(n)}(\tilde{\mathbf{x}}) = \sum_{i=1}^n T_j(x^{(i)}), \quad 1 \leq j \leq k, \quad (3.2)$$

for  $\tilde{\mathbf{x}} = (x^{(1)}, \dots, x^{(n)}) \in \mathcal{X}^n$ .

Consequently, the joint distribution of independent random variables  $X_1, \dots, X_n$  forms an EF if the distribution of  $X_i$ ,  $1 \leq i \leq n$ , forms an EF. In Theorem 3.13, case (a) corresponds to a sample of independent not necessarily identically distributed (inid) random variables; case (b) then refers to the situation of an independent and identically distributed (iid) sample.

**Examples:**  $\text{poi}(\bullet)$ ,  $\mathcal{N}(\bullet, \bullet)$ ,  $\mathcal{N}_k(\bullet, \Sigma_0)$

*Example 3.14* Let  $X_1, \dots, X_n$  be independent random variables with  $X_i \sim \text{poi}(a_i)$  for  $1 \leq i \leq n$  as in Example 2.3. Then, the distribution of  $\tilde{\mathbf{X}} = (X_1, \dots, X_n)$  forms a multiparameter EF with representation (2.1) given by  $\mu = \mu_{\mathbb{N}_0}^{(n)}$ ,  $k = n$ ,

$$C(\mathbf{a}) = \exp \left\{ - \sum_{i=1}^n a_i \right\}, \quad Z_j(\mathbf{a}) = \ln(a_j), \quad 1 \leq j \leq n,$$

for  $\mathbf{a} = (a_1, \dots, a_n)^t \in (0, \infty)^n$ , and

$$T_j(\tilde{\mathbf{x}}) = x_j, \quad 1 \leq j \leq n, \quad h(\tilde{\mathbf{x}}) = \prod_{i=1}^n \frac{1}{x_i!}$$

for  $\tilde{\mathbf{x}} = (x_1, \dots, x_n) \in \mathbb{N}_0^n$ . Clearly, this EF has order  $n$  and is regular. On the other hand, by assuming that  $a_1 = \dots = a_n$ , we are in the iid situation with an EF of order 1. Obviously, we can generate EFs with one parameter up to  $n$  parameters, here. Now, suppose that the parameters  $a_1, \dots, a_n$  are connected via a log-linear link function, i.e., we set  $\ln(a_j) = \zeta_1 + \zeta_2 v_j$ ,  $1 \leq j \leq n$ , where  $\zeta_1$  and  $\zeta_2$  are real-valued parameters and  $v_1, \dots, v_n$  are fixed real numbers, which are not all identical. By inserting, the distribution of  $\tilde{X}$  turns out to form a regular EF of order 2 with canonical representation (2.6), where

$$C^*(\boldsymbol{\zeta}) = \exp \left\{ - \sum_{i=1}^n e^{\zeta_1 + \zeta_2 v_i} \right\}, \quad \boldsymbol{\zeta} = (\zeta_1, \zeta_2)^t \in \mathbb{R}^2,$$

and  $\mathbf{T}(\tilde{\mathbf{x}}) = (\sum_{i=1}^n x_i, \sum_{i=1}^n v_i x_i)^t$  for  $\tilde{\mathbf{x}} = (x_1, \dots, x_n) \in \mathbb{N}_0^n$  (see [44, p. 35]).

*Example 3.15* Let  $X_1, \dots, X_{n_1}$  be an iid sample from  $\mathcal{N}(\mu_1, \sigma_1^2)$  and  $Y_1, \dots, Y_{n_2}$  be an iid sample from  $\mathcal{N}(\mu_2, \sigma_2^2)$  as introduced in Example 2.10. Moreover, let the samples be independent. Then, the distribution of  $(X_1, \dots, X_{n_1}, Y_1, \dots, Y_{n_2})$  forms an EF with representation (2.1) given by  $\mu = \lambda_{\mathbb{R}}^{n_1+n_2}$ ,  $k = 4$ ,

$$C(\boldsymbol{\vartheta}) = (\sigma_1^2)^{-n_1/2} (\sigma_2^2)^{-n_2/2} \exp \left\{ - \frac{n_1 \mu_1^2}{2\sigma_1^2} - \frac{n_2 \mu_2^2}{2\sigma_2^2} \right\},$$

$$\mathbf{Z}(\boldsymbol{\vartheta}) = \left( \frac{\mu_1}{\sigma_1^2}, \frac{\mu_2}{\sigma_2^2}, -\frac{1}{2\sigma_1^2}, -\frac{1}{2\sigma_2^2} \right)^t,$$

for  $\boldsymbol{\vartheta} = (\mu_1, \mu_2, \sigma_1^2, \sigma_2^2)^t \in \mathbb{R}^2 \times (0, \infty)^2$ , and

$$\mathbf{T}(\tilde{\mathbf{x}}) = \left( \sum_{i=1}^{n_1} x_i, \sum_{j=1}^{n_2} y_j, \sum_{i=1}^{n_1} x_i^2, \sum_{j=1}^{n_2} y_j^2 \right)^t, \quad h(\tilde{\mathbf{x}}) = \left( \frac{1}{\sqrt{2\pi}} \right)^{n_1+n_2},$$

for  $\tilde{\mathbf{x}} = (x_1, \dots, x_{n_1}, y_1, \dots, y_{n_2}) \in \mathbb{R}^{n_1+n_2}$ . Clearly, the EF has order 4 and is regular. Now, suppose that  $\sigma_1^2 = \sigma_2^2 = \sigma^2$ , say, i.e., we have a common-scale

model. By inserting, the distribution of  $\tilde{X}$  turns out to form an EF with mappings

$$C(\tilde{\boldsymbol{\theta}}) = (\sigma^2)^{-(n_1+n_2)/2} \exp \left\{ -\frac{n_1\mu_1^2 + n_2\mu_2^2}{2\sigma^2} \right\},$$

$$\mathbf{Z}(\tilde{\boldsymbol{\theta}}) = \left( \frac{\mu_1}{\sigma^2}, \frac{\mu_2}{\sigma^2}, -\frac{1}{2\sigma^2} \right)^t,$$

for  $\tilde{\boldsymbol{\theta}} = (\mu_1, \mu_2, \sigma^2)^t \in \mathbb{R}^2 \times (0, \infty)$ , and

$$\mathbf{T}(\tilde{\mathbf{x}}) = \left( \sum_{i=1}^{n_1} x_i, \sum_{j=1}^{n_2} y_j, \sum_{i=1}^{n_1} x_i^2 + \sum_{j=1}^{n_2} y_j^2 \right)^t,$$

for  $\tilde{\mathbf{x}} = (x_1, \dots, x_{n_1}, y_1, \dots, y_{n_2}) \in \mathbb{R}^{n_1+n_2}$ . This is a regular EF of order 3. In contrast, the common-location assumption  $\mu_1 = \mu_2$  leads to a curved EF of order 4 (see also [35, pp. 87/88]).

*Example 3.16* Let  $X_1, \dots, X_n$  be independent random variables with  $X_i \sim \mathcal{N}(\mu_i, \sigma_0^2)$ ,  $i \in \{1, \dots, n\}$ , for some fixed variance  $\sigma_0^2 > 0$ . Then, by Theorem 3.13(a), the joint distributions of  $X_1, \dots, X_n$  form an EF of distributions on  $(\mathbb{R}^n, \mathbb{B}^n)$  with parameters  $\mu_1, \dots, \mu_n \in \mathbb{R}$ .

Now, let  $\mathbf{X}^{(1)}, \dots, \mathbf{X}^{(m)}$  be iid random vectors following a multivariate normal distribution with mean vector  $\boldsymbol{\mu} = (\mu_1, \dots, \mu_n)^t \in \mathbb{R}^n$  and fixed positive definite covariance matrix  $\boldsymbol{\Sigma}_0 \in \mathbb{R}^{n \times n}$ , i.e.,  $\mathbf{X}^{(i)} \sim \mathcal{N}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma}_0)$ ,  $i \in \{1, \dots, m\}$ . Then, by Theorem 3.13(b), the joint distributions of  $\mathbf{X}^{(1)}, \dots, \mathbf{X}^{(m)}$  form an EF of distributions on  $(\mathbb{R}^{mn}, \mathbb{B}^{mn})$  with parameters  $\mu_1, \dots, \mu_n \in \mathbb{R}$ .

Based on Theorem 3.13, the latter examples show the flexibility of constructing EFs. Suppose that, in the situation of Example 3.16, we want to have a 3-parameter EF of distributions on  $(\mathbb{R}^6, \mathbb{B}^6)$ . This is achieved, for instance, by choosing  $n = 6$  in the independent case while assuming  $\mu_2 = \mu_3$  and  $\mu_4 = \mu_5 = \mu_6$ , or by setting  $n = 3$  and  $m = 2$  in the iid case.

Given the iid situation, several properties of the EF are transferred to the product EF. In what follows, let  $\mathbf{T}^{(n)} = (T_1^{(n)}, \dots, T_k^{(n)})^t$  with  $T_j^{(n)}$ ,  $1 \leq j \leq k$ , as defined in formula (3.2), be the vector of statistics for the  $n$ -fold product case,  $n \in \mathbb{N}$ .

**Corollary 3.17** *Let  $\mathcal{P}$  be an EF with representation (2.1), and let  $\mathcal{P}^{(n)}$ ,  $n \in \mathbb{N}$ , denote the corresponding EF of  $n$ -fold product measures. Then,  $\text{ord}(\mathcal{P}^{(n)}) = \text{ord}(\mathcal{P})$ . Moreover, if  $\mathcal{P}$  is full (regular, steep), so is  $\mathcal{P}^{(n)}$ .*

**Proof** Considering  $T^{(n)}$  as function of iid random variables  $X_1, \dots, X_n \sim P_\vartheta$ , it follows that

$$\mathbf{Cov}_{P_\vartheta^{(n)}}(T^{(n)}) = n \mathbf{Cov}_{P_\vartheta}(T).$$

Hence,  $\mathcal{P}$  and  $\mathcal{P}^{(n)}$  have the same order by applying Lemma 2.26 and Theorem 2.23. Moreover, by Fubini's theorem, we have for  $\zeta \in \mathbb{R}^k$  that

$$\int e^{\zeta^t T^{(n)}} h_n d\mu^{(n)} = \left( \int e^{\zeta^t T} h d\mu \right)^n,$$

where  $h_n(\tilde{x}) = \prod_{i=1}^n h(x^{(i)})$  for  $\tilde{x} = (x^{(1)}, \dots, x^{(n)}) \in \mathcal{X}^n$ , such that the natural parameter spaces of  $\mathcal{P}^{(n)}$  and  $\mathcal{P}$  coincide. Trivially,  $\mathcal{P}^{(n)}$  is then full (regular) if  $\mathcal{P}$  is full (regular). Finally, since the cumulant function of  $\mathcal{P}^{(n)}$  is given by  $n\kappa$ ,  $\mathcal{P}^{(n)}$  is steep if  $\mathcal{P}$  is steep.  $\square$

Note that Lemma 2.17 and Theorem 3.13(b) with  $T(x) = x$ ,  $x \in \mathcal{X}$ , imply that the distribution of a sum of iid random variables forms an EF if this is the case for the distribution of the summands (for more details, see [37, pp. 670–673]). In other words, the EF structure is preserved for the family of  $n$ -fold convolutions. Here, the cumulant function for the  $n$ -fold convolution family is given by  $n\kappa$ ,  $n \in \mathbb{N}$ . For the steep natural EF  $\mathcal{P}$  with  $\nu$ -densities (2.14), this finding suggests the construction of a more general family of distributions. Let  $\Lambda \subset (0, \infty)$  denote the set of all parameters  $\lambda$ , for which  $\lambda\kappa$  is the cumulant function of a natural EF generated by some measure  $\nu_\lambda$  on  $\mathbb{B}^k$ . Then,

$$\mathcal{D} = \{P_{\zeta, \lambda}^* : \zeta \in \text{int}(\Xi^*), \lambda \in \Lambda\}$$

is called the *exponential dispersion model* associated with  $\mathcal{P}$ , where a  $\nu_\lambda$ -density of  $P_{\zeta, \lambda}^*$  is given by

$$\frac{dP_{\zeta, \lambda}^*}{d\nu_\lambda}(x) = e^{\zeta^t x - \lambda\kappa(\zeta)}, \quad x \in \mathbb{R}^k,$$

for  $\zeta \in \text{int}(\Xi^*)$  and  $\lambda \in \Lambda$ .  $\mathcal{D}$  is the union of the natural EFs  $\{P_{\zeta, \lambda}^* : \zeta \in \text{int}(\Xi^*)\}$  with variance functions  $V_\lambda(\mathbf{m}) = \lambda V_\varphi(\mathbf{m}/\lambda)$ ,  $\mathbf{m} \in \text{int}(\lambda M)$ , for  $\lambda \in \Lambda$  (see Remark 2.66). The parameter  $1/\lambda$  is called the *dispersion parameter*, and  $\Lambda$  is also referred to as the Jørgensen set in the literature. For an extensive account on the model, see [29]. Applications of exponential dispersion models can be found, for instance, in [3] and the references cited therein. For their particular use in generalized linear models and actuarial risk theory, see, e.g., [61, 23, 56, 30].

### 3.4 Sufficiency and Completeness

The concepts of sufficiency and completeness of a statistic for a family of distributions is fundamental in various aspects of statistical inference as, for example, the derivation of optimal point estimators and statistical tests. Roughly speaking, a sufficient statistic contains and summarizes all information in the data about the underlying true distribution, and it comprises no redundant information if it is complete, too. In general, sufficiency is introduced via conditional expectations and probabilities (see, e.g., [40, pp. 36–44]). Here, for short, we introduce sufficiency by means of the form of the density functions in the family, which is known as the Neyman criterion (see, e.g., [40, pp. 44–46]).

**Definition 3.18** Let  $(\mathcal{X}, \mathcal{B})$  be a measurable space and  $\mathcal{P} = \{P_\vartheta : \vartheta \in \Theta\}$  be a family of distributions on  $\mathcal{B}$  dominated by a  $\sigma$ -finite measure  $\mu$ . Moreover, let  $T : (\mathcal{X}, \mathcal{B}) \rightarrow (\mathcal{Y}, \mathcal{C})$  be a statistic.

- (a)  $T$  is called *sufficient* for  $\vartheta \in \Theta$  if there exist a function  $h : (\mathcal{X}, \mathcal{B}) \rightarrow (\mathbb{R}^1, \mathbb{B}^1)$  and functions  $q_\vartheta : (\mathcal{Y}, \mathcal{C}) \rightarrow (\mathbb{R}^1, \mathbb{B}^1)$ ,  $\vartheta \in \Theta$ , such that a  $\mu$ -density of  $P_\vartheta$  is of the form

$$\frac{dP_\vartheta}{d\mu}(x) = q_\vartheta(T(x)) h(x), \quad x \in \mathcal{X}, \quad (3.3)$$

for all  $\vartheta \in \Theta$ .

- (b)  $T$  is called *minimal sufficient* for  $\vartheta \in \Theta$  if  $T$  is sufficient for  $\vartheta \in \Theta$  and factorizes over any other sufficient statistic, i.e., for every statistic  $\tilde{T} : (\mathcal{X}, \mathcal{B}) \rightarrow (\tilde{\mathcal{Y}}, \tilde{\mathcal{C}})$  being sufficient for  $\vartheta \in \Theta$ , there exist a function  $q : (\tilde{\mathcal{Y}}, \tilde{\mathcal{C}}) \rightarrow (\mathcal{Y}, \mathcal{C})$  and a set  $N \in \mathcal{B}$  with  $P_\vartheta(N) = 0$  for all  $\vartheta \in \Theta$ , such that

$$T = q \circ \tilde{T} \quad \text{on } N^c.$$

A sufficient statistic allows for reduction of the data without a loss of information meaning that  $T(x)$  contains all information in  $x$  about  $\vartheta$ . The main issue in formula (3.3) is that the term  $q_\vartheta(T(x))$  depends on the observation  $x$  only through  $T(x)$ . A minimal sufficient statistic is as simple as possible in the sense of Definition 3.18(b).

**Definition 3.19** Let  $(X, \mathcal{B})$  be a measurable space and  $\mathcal{P} = \{P_\vartheta : \vartheta \in \Theta\}$  be a family of distributions on  $\mathcal{B}$ .

- (a)  $\mathcal{P}$  is called *(boundedly) complete* if, for every (bounded) function  $r : (X, \mathcal{B}) \rightarrow (\mathbb{R}^1, \mathbb{B}^1)$ , the following implication holds:

$$\int r dP_\vartheta = 0 \quad \text{for all } \vartheta \in \Theta$$

$\Rightarrow$  there exists a set  $N \in \mathcal{B}$  with  $P_\vartheta(N) = 0$  for all  $\vartheta \in \Theta$  and  $r(x) = 0$  for all  $x \in N^c$ .

- (b) A statistic  $T : (X, \mathcal{B}) \rightarrow (\mathcal{Y}, \mathcal{C})$  is called *(boundedly) complete* for  $\vartheta \in \Theta$  if  $\mathcal{P}^T = \{P_\vartheta^T : \vartheta \in \Theta\}$  is (boundedly) complete.

The completeness of a statistic refers to the richness of the underlying family  $\mathcal{P}$  of distributions. By definition, every complete statistic is boundedly complete. In an EF, a (minimal) sufficient and complete statistic is easily found, which means an important advantage in statistical inference.

**Theorem 3.20** Let  $\mathcal{P}$  be an EF with representation (2.1). Then, we have the following properties:

- (a) The vector  $T = (T_1, \dots, T_k)^t$  is sufficient for  $\vartheta \in \Theta$ .
- (b) If  $Z_1, \dots, Z_k$  are affinely independent, then  $T$  is minimal sufficient for  $\vartheta \in \Theta$ .
- (c) If  $\text{int}(\mathbf{Z}(\Theta)) \neq \emptyset$ , then  $T$  is complete for  $\vartheta \in \Theta$ .

Statement (a) in Theorem 3.20 directly follows by setting  $q_\vartheta(\mathbf{t}) = C(\vartheta) \exp\{\mathbf{Z}(\vartheta)^t \mathbf{t}\}$ ,  $\mathbf{t} \in \mathbb{R}^k$ , in Eq. (3.3), and it holds true in any EF without imposing further conditions on  $Z_1, \dots, Z_k$ . A proof of statement (b) can be found, for instance, in [39, p. 39] and [50, pp. 25/26]. Statement (c) can be shown by using characteristic functions and the uniqueness theorem (see, e.g., [40, pp. 116/117], [50, pp. 26/27], or [20, p. 43]).

The property of an EF to be full along with a minimal representation (see Definitions 2.35 and 2.20) has direct applications for the statistic  $T$ , namely, it is minimal sufficient and complete.

**Corollary 3.21** Let  $\mathcal{P}$  be a full EF with minimal representation (2.1). Then,  $T = (T_1, \dots, T_k)^t$  is minimal sufficient and complete for  $\vartheta$ .

**Proof** The assertion follows from Theorem 3.20, Lemma 2.42, and Theorem 2.23.  $\square$



Note that the dimension of the vector  $\mathbf{T}^{(n)}$  of statistics related to the product case is the same for all  $n \in \mathbb{N}$  (see Theorem 3.13(b)). Hence, in the iid case, sampling from an EF does not increase the dimension of the sufficient statistic.

**Examples:**  $\text{poi}(\bullet)$ ,  $\text{b}(n_0, \bullet)$ ,  $\text{nb}(r_0, \bullet)$ ,  $\text{m}(n_0, \bullet)$ ,  $\mathcal{E}(\bullet)$ ,  $\text{g}(\bullet, \bullet)$ ,  $\mathcal{G}(\bullet, \bullet)$ ,  $\mathcal{N}(\bullet, \bullet)$ ,  $\mathcal{N}_k(\bullet, \Sigma_0)$

*Example 3.22* Let  $\mathcal{P}$  be any of the EFs  $\text{poi}(\bullet)$ ,  $\text{b}(n_0, \bullet)$ ,  $\text{nb}(r_0, \bullet)$ , or  $\mathcal{E}(\bullet)$  introduced in Examples 2.3, 2.4, 2.5, and 2.12. Moreover, let  $X_1, \dots, X_n$  be an iid sample from  $P_{\vartheta} \in \mathcal{P}$ . Then, by Theorem 3.20,

$$\mathbf{T}^{(n)}(\tilde{\mathbf{x}}) = \sum_{i=1}^n x_i, \quad \tilde{\mathbf{x}} = (x_1, \dots, x_n) \in \mathcal{X}^n,$$

is minimal sufficient and complete for  $\vartheta$ .

*Example 3.23* Let  $\mathcal{P}$  be the EF  $\text{g}(\bullet, \bullet)$  or  $\mathcal{G}(\bullet, \bullet)$  defined in Examples 2.7 and 2.12. Moreover, let  $X_1, \dots, X_n$  be an iid sample from  $P_{\vartheta} \in \mathcal{P}$ . Then, by Theorem 3.20,

$$\mathbf{T}^{(n)}(\tilde{\mathbf{x}}) = \left( \sum_{i=1}^n x_i, \sum_{i=1}^n \ln(x_i) \right)^t, \quad \tilde{\mathbf{x}} = (x_1, \dots, x_n) \in \mathcal{X}^n,$$

is minimal sufficient and complete for  $\vartheta$ .

*Example 3.24* Let  $\mathbf{X}^{(1)}, \dots, \mathbf{X}^{(n)}$  be an iid sample from  $\text{m}(n_0, \mathbf{p}) \in \text{m}(n_0, \bullet)$  as introduced in Example 2.8. Then, by Theorem 3.20 and Example 2.46,

$$\tilde{\mathbf{T}}^{(n)}(\tilde{\mathbf{x}}) = \sum_{i=1}^n (x_1^{(i)}, \dots, x_{k-1}^{(i)})^t, \quad \tilde{\mathbf{x}} = (\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}) \in \mathcal{X}^n,$$

$$\mathbf{x}^{(i)} = (x_1^{(i)}, \dots, x_k^{(i)})^t \in \mathcal{X}, \quad 1 \leq i \leq n,$$

is minimal sufficient and complete for  $\mathbf{p}$ .

*Example 3.25* Let  $X_1, \dots, X_n$  be an iid sample from  $\mathcal{N}(\mu, \sigma^2) \in \mathcal{N}(\bullet, \bullet)$  as introduced in Example 2.10. Then, by Theorem 3.20 and Example 2.27,

$$\mathbf{T}^{(n)}(\tilde{\mathbf{x}}) = \left( \sum_{i=1}^n x_i, \sum_{i=1}^n x_i^2 \right)^t, \quad \tilde{\mathbf{x}} = (x_1, \dots, x_n) \in \mathbb{R}^n,$$

is minimal sufficient and complete for  $(\mu, \sigma^2)^t$ .

*Example 3.26* Let  $\mathbf{X}^{(1)}, \dots, \mathbf{X}^{(n)}$  be an iid sample from  $\mathcal{N}_k(\boldsymbol{\mu}, \boldsymbol{\Sigma}_0) \in \mathcal{N}_k(\bullet, \boldsymbol{\Sigma}_0)$  as introduced in Example 2.18. Then, by Theorem 3.20 and Example 2.28,

$$\mathbf{T}^{(n)}(\tilde{\mathbf{x}}) = \sum_{i=1}^n \mathbf{x}^{(i)}, \quad \tilde{\mathbf{x}} = (\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}) \in (\mathbb{R}^k)^n,$$

is minimal sufficient and complete for  $\boldsymbol{\mu}$ .

*Remark 3.27* For the curved EF introduced in Definition 2.32, statements (a) and (b) of Theorem 3.20 are true. Completeness of  $\mathbf{T}$ , however, may or may not be the case. In [50, pp. 28–31], examples for complete and non-complete curved EFs are shown and sufficient conditions are given for  $\mathbf{T}$  to be not complete or not even boundedly complete.

### 3.5 Score Statistic and Information Matrix

The Fisher information matrix is a measure for the amount of information a random variable/vector contains about the true underlying parameter. It appears, for instance, in the context of point estimation and allows for defining efficient and asymptotically efficient estimators.

**Definition 3.28** Let  $(X, \mathcal{B})$  be a measurable space and  $\mathcal{P} = \{P_{\vartheta} = f_{\vartheta} \mu : \vartheta \in \Theta\}$ ,  $\Theta \subset \mathbb{R}^k$  open, be a family of distributions on  $\mathcal{B}$  dominated by a  $\sigma$ -finite measure  $\mu$ . If all appearing derivatives (with respect to  $\vartheta$ ) and integrals

(continued)

**Definition 3.28** (continued)  
(with respect to  $P_{\vartheta}$ ) exist, then

$$\mathbf{I}(\vartheta) = E_{P_{\vartheta}}[U_{\vartheta} U_{\vartheta}^t] \quad \text{with} \quad U_{\vartheta} = \nabla \ln(f_{\vartheta})$$

is called the *Fisher information matrix* of  $\mathcal{P}$  at  $\vartheta \in \Theta$ . The random vector  $U_{\vartheta}$  is referred to as the *score statistic* at  $\vartheta \in \Theta$ .

In an EF with natural parametrization, differentiability of the densities with respect to  $\zeta$  is guaranteed by Corollary 2.50, and the score statistic and Fisher information matrix are readily obtained.

**Theorem 3.29** *Let  $\mathcal{P}$  be an EF with canonical representation (2.6), and let  $\zeta \in \text{int}(\Xi)$ . Then,*

$$U_{\zeta} = T - E_{\zeta}[T],$$

$$\mathbf{I}(\zeta) = \text{Cov}_{\zeta}(T).$$

*In particular, all entries in  $\mathbf{I}(\zeta)$  are finite, i.e.,  $[\mathbf{I}(\zeta)]_{i,j} < \infty$ ,  $1 \leq i, j \leq k$ . If  $\text{ord}(\mathcal{P}) = k$ , then  $\mathbf{I}(\zeta) > 0$ .*

**Proof** Let  $\zeta \in \text{int}(\Xi)$ . By Corollary 2.50,

$$U_{\zeta} = \nabla \ln(\exp\{\zeta^t T - \kappa(\zeta)\} h) = T - E_{\zeta}[T].$$

Hence,  $\mathbf{I}(\zeta) = \text{Cov}_{\zeta}(T)$ . Moreover, from Theorem 2.49, all entries in  $\mathbf{I}(\zeta)$  are finite. If  $\text{ord}(\mathcal{P}) = k$ , then  $\mathbf{I}(\zeta) > 0$  by Theorem 2.23 and Lemma 2.26.  $\square$

Moreover, the following simple formula for the Fisher information matrix results from Corollary 2.50.

**Corollary 3.30** *Let  $\mathcal{P}$  be an EF with canonical representation (2.6). Then,*

$$\mathbf{I}(\zeta) = \mathbf{H}_{\kappa}(\zeta), \quad \zeta \in \text{int}(\Xi).$$

For one-parameter EFs with canonical representation and  $\zeta \in \text{int}(\Xi)$ , the Fisher information is simply given by the second derivative of  $\kappa$  at  $\zeta$ .

**Examples:**  $\text{poi}(\bullet)$ ,  $\text{nb}(r_0, \bullet)$ ,  $\mathcal{N}(\bullet, \bullet)$ ,  $\mathcal{G}(\bullet, \bullet)$ ,  $\mathcal{N}_k(\bullet, \Sigma_0)$

*Example 3.31* For  $\mathcal{P} = \text{poi}(\bullet)$ , Corollary 3.30 and Example 2.64 directly yield  $I(\zeta) = e^\zeta$  for  $\zeta \in \mathbb{R}$ .

*Example 3.32* For  $\mathcal{P} = \text{nb}(r_0, \bullet)$ , Corollary 3.30 and Example 2.51 directly yield

$$I(\zeta) = \frac{r_0 \exp\{\zeta\}}{(1 - \exp\{\zeta\})^2}, \quad \zeta \in (-\infty, 0).$$

*Example 3.33* For  $\mathcal{P} = \mathcal{N}(\bullet, \bullet)$ , Corollary 3.30 and Example 2.52 yield

$$\mathbf{I}(\zeta) = \begin{pmatrix} -\frac{1}{2\zeta_2} & \frac{\zeta_1}{2\zeta_2^2} \\ \frac{\zeta_1}{2\zeta_2^2} & \frac{1}{2\zeta_2^2} - \frac{\zeta_1^2}{2\zeta_2^3} \end{pmatrix} = \frac{1}{2\zeta_2^3} \begin{pmatrix} -\zeta_2^2 & \zeta_1\zeta_2 \\ \zeta_1\zeta_2 & \zeta_2 - \zeta_1^2 \end{pmatrix}$$

for  $\zeta = (\zeta_1, \zeta_2)^t \in \mathbb{R} \times (-\infty, 0)$ .

*Example 3.34* For  $\mathcal{P} = \mathcal{G}(\bullet, \bullet)$ , Corollary 3.30 and Example 2.53 directly yield

$$\mathbf{I}(\zeta) = \begin{pmatrix} \frac{\zeta_2}{\zeta_1^2} & -\frac{1}{\zeta_1} \\ -\frac{1}{\zeta_1} & \psi_1(\zeta_2) \end{pmatrix}, \quad \zeta = (\zeta_1, \zeta_2)^t \in (-\infty, 0) \times (0, \infty),$$

where  $\psi_1 = \psi'$  denotes the trigamma function.

*Example 3.35* For  $\mathcal{P} = \mathcal{N}_k(\bullet, \Sigma_0)$ , Corollary 3.30 and Example 2.54 directly yield

$$\mathbf{I}(\mu) = \Sigma_0^{-1}, \quad \mu \in \mathbb{R}^k.$$

The Fisher information matrix generally depends on the parametrization of the distribution family. In the situation of Definition 3.28, let  $\mathbf{I}(\vartheta)$  exist for  $\vartheta \in \Theta$ . Moreover, let  $\Delta \subset \mathbb{R}^q$ ,  $q \leq k$ , be open, and let  $\Psi : \Delta \rightarrow \Theta$  be a one-to-one and continuously differentiable function with Jacobian matrix  $\mathbf{D}_\Psi(\eta) \in \mathbb{R}^{k \times q}$  of full rank for all  $\eta \in \Delta$ . Then, the score statistic and Fisher information matrix at  $\eta \in \Delta$

of the family  $\{\tilde{P}_\eta : \eta \in \Delta\} \subset \mathcal{P}$  with parametrization  $\tilde{P}_\eta = P_{\Psi(\eta)}$ ,  $\eta \in \Delta$ , are given by

$$\begin{aligned}\tilde{\mathbf{U}}_\eta &= \mathbf{D}_\Psi(\eta)^t \mathbf{U}_{\Psi(\eta)}, \\ \tilde{\mathbf{I}}(\eta) &= \mathbf{D}_\Psi(\eta)^t \mathbf{I}(\Psi(\eta)) \mathbf{D}_\Psi(\eta).\end{aligned}\tag{3.4}$$

### Example

*Example 3.36* Let  $\mathcal{P}$  be a regular EF with minimal canonical representation (2.6), and let  $\pi : \Xi^* \rightarrow \pi(\Xi^*)$  be the mean value function as defined in Theorem 2.56 with Jacobian matrix  $\mathbf{D}_\pi(\zeta) = \mathbf{Cov}_\zeta(T)$  for  $\zeta \in \Xi^*$ . Then, by using Theorem 3.29, formula (3.4) with  $\Psi = \pi^{-1}$ , and  $\mathbf{D}_{\pi^{-1}}(\mathbf{m}) = [\mathbf{D}_\pi(\pi^{-1}(\mathbf{m}))]^{-1}$ , the Fisher information matrix of  $\mathcal{P}$  corresponding to the mean value parametrization  $\tilde{P}_\mathbf{m} = P_{\pi^{-1}(\mathbf{m})}$ ,  $\mathbf{m} \in \pi(\Xi^*)$ , is given by

$$\begin{aligned}\tilde{\mathbf{I}}(\mathbf{m}) &= \mathbf{D}_{\pi^{-1}}(\mathbf{m})^t \mathbf{I}(\pi^{-1}(\mathbf{m})) \mathbf{D}_{\pi^{-1}}(\mathbf{m}) \\ &= [\mathbf{Cov}_{\pi^{-1}(\mathbf{m})}(T)]^{-1} \mathbf{Cov}_{\pi^{-1}(\mathbf{m})}(T) [\mathbf{Cov}_{\pi^{-1}(\mathbf{m})}(T)]^{-1} \\ &= [\mathbf{Cov}_{\pi^{-1}(\mathbf{m})}(T)]^{-1} \\ &= [\mathbf{Cov}_{\tilde{P}_\mathbf{m}}(T)]^{-1}, \quad \mathbf{m} \in \pi(\Xi^*).\end{aligned}$$

## 3.6 Divergence and Distance Measures

*Entropy measures* as, for instance, the Shannon entropy defined by  $H(P) = -E_P[\ln(f)]$  for a distribution  $P$  with  $\mu$ -density  $f$ , measure the disorder and uncertainty of distributions. Closely related are *divergence measures*, which quantify the ‘distance’ of two distributions in some family  $\mathcal{P}$  in a certain sense (see, e.g., [57] and [48]). In statistical inference, they serve to construct statistical tests or confidence regions for underlying parameters. If the distributions in  $\mathcal{P}$  are dominated by a  $\sigma$ -finite measure  $\mu$ , various divergence measures have been introduced by means of the corresponding densities. Some important examples are the *Kullback-Leibler divergence*

$$D_{KL}(P_1, P_2) = \int f_1 \ln \left( \frac{f_1}{f_2} \right) d\mu,$$

the *Jeffrey distance*  $D_J(P_1, P_2) = D_{KL}(P_1, P_2) + D_{KL}(P_2, P_1)$ , the *Rényi divergence*

$$D_{R_q}(P_1, P_2) = \frac{1}{q(q-1)} \ln \left( \int f_1^q f_2^{1-q} d\mu \right), \quad q \in \mathbb{R} \setminus \{0, 1\},$$

the *Cressie-Read divergence*

$$D_{CR_q}(P_1, P_2) = \frac{1}{q(q-1)} \int f_1 \left( \left( \frac{f_1}{f_2} \right)^{q-1} - 1 \right) d\mu, \quad q \in \mathbb{R} \setminus \{0, 1\},$$

and the *Hellinger metric*

$$D_H(P_1, P_2) = \left( \int (\sqrt{f_1} - \sqrt{f_2})^2 d\mu \right)^{1/2}$$

for  $P_i = f_i \mu \in \mathcal{P}$ ,  $i = 1, 2$ , provided that the respective expressions are defined and the integrals finitely exist. If  $\mathcal{P} = \{P_{\vartheta} : \vartheta \in \Theta\}$  is a parametric family of distributions, it is convenient to use the parameters as arguments and briefly write, e.g.,  $D_{KL}(\vartheta_1, \vartheta_2)$  for  $D_{KL}(P_{\vartheta_1}, P_{\vartheta_2})$ ,  $\vartheta_1, \vartheta_2 \in \Theta$ ; note that then  $D_{KL}(\vartheta_1, \vartheta_2) = 0$  implies  $\vartheta_1 = \vartheta_2$  if the parameter  $\vartheta \in \Theta$  is identifiable (see Lemma 2.30).

The EF structure admits simple formulas for various divergence measures as indicated in Lemma 3.37 and Corollary 3.39. Note that the Kullback-Leibler divergence, which can be considered as an expected log-likelihood ratio, is a starting point for many statistical procedures. Under common assumptions, we obtain a simple representation in EFs (see [20, pp. 174–178]). This representation is used in [33] in a two-class classification procedure within an EF, where classes are defined by means of left-sided Kullback-Leibler balls.

**Lemma 3.37** *Let  $\mathcal{P}$  be a full EF with minimal canonical representation (2.6). Then, for  $\zeta, \eta \in \text{int}(\Xi^*)$ , we have*

$$D_{KL}(\zeta, \eta) = \kappa(\eta) - \kappa(\zeta) + (\zeta - \eta)^t \pi(\zeta),$$

$$D_J(\zeta, \eta) = (\zeta - \eta)^t (\pi(\zeta) - \pi(\eta)).$$

**Proof** For  $\zeta, \eta \in \text{int}(\Xi^*)$ , we have by Theorem 2.56

$$\begin{aligned} D_{KL}(\zeta, \eta) &= \int \left( \ln(f_{\zeta}^*) - \ln(f_{\eta}^*) \right) f_{\zeta}^* d\mu \\ &= \int ((\zeta - \eta)^t \mathbf{T} - \kappa(\zeta) + \kappa(\eta)) f_{\zeta}^* d\mu \end{aligned}$$

$$\begin{aligned}
&= \kappa(\boldsymbol{\eta}) - \kappa(\boldsymbol{\zeta}) + (\boldsymbol{\zeta} - \boldsymbol{\eta})^t E_{\boldsymbol{\zeta}}[\mathbf{T}] \\
&= \kappa(\boldsymbol{\eta}) - \kappa(\boldsymbol{\zeta}) + (\boldsymbol{\zeta} - \boldsymbol{\eta})^t \boldsymbol{\pi}(\boldsymbol{\zeta}).
\end{aligned}$$

From this, the representation of  $D_J$  is obvious.  $\square$

The Rényi divergence, Cressie-Read divergence, Hellinger metric, and other divergences are functions of the quantity  $A_q(P_1, P_2) = \int f_1^q f_2^{1-q} d\mu$ , which has a nice representation in EFs for  $q \in (0, 1)$  (see, e.g., [43, 60]).

**Lemma 3.38** *Let  $\mathcal{P}$  be a full EF with minimal canonical representation (2.6). Then, for  $\boldsymbol{\zeta}, \boldsymbol{\eta} \in \Xi^*$  and  $q \in (0, 1)$ , we have*

$$A_q(\boldsymbol{\zeta}, \boldsymbol{\eta}) = \exp \{ \kappa(q\boldsymbol{\zeta} + (1-q)\boldsymbol{\eta}) - (q\kappa(\boldsymbol{\zeta}) + (1-q)\kappa(\boldsymbol{\eta})) \}.$$

**Proof** Let  $\boldsymbol{\zeta}, \boldsymbol{\eta} \in \Xi^*$  and  $q \in (0, 1)$ . Then,

$$\begin{aligned}
A_q(\boldsymbol{\zeta}, \boldsymbol{\eta}) &= \int (f_{\boldsymbol{\zeta}}^*)^q (f_{\boldsymbol{\eta}}^*)^{1-q} d\mu \\
&= \int \exp \{ (q\boldsymbol{\zeta} + (1-q)\boldsymbol{\eta})^t \mathbf{T} - (q\kappa(\boldsymbol{\zeta}) + (1-q)\kappa(\boldsymbol{\eta})) \} h d\mu \\
&= \exp \{ \kappa(q\boldsymbol{\zeta} + (1-q)\boldsymbol{\eta}) - (q\kappa(\boldsymbol{\zeta}) + (1-q)\kappa(\boldsymbol{\eta})) \}.
\end{aligned}$$

$\square$

Recall that the cumulant function  $\kappa$  is convex (see Theorem 2.44). The logarithm of  $A_q(\boldsymbol{\zeta}, \boldsymbol{\eta})$  in Lemma 3.38 is just the difference of  $\kappa$  evaluated at a convex combination of  $\boldsymbol{\zeta}$  and  $\boldsymbol{\eta}$  and the respective convex combination of  $\kappa(\boldsymbol{\zeta})$  and  $\kappa(\boldsymbol{\eta})$ .

**Corollary 3.39** *Let  $\mathcal{P}$  be a full EF with minimal canonical representation (2.6). Then, for  $\boldsymbol{\zeta}, \boldsymbol{\eta} \in \Xi^*$  and  $q \in (0, 1)$ , we have*

$$\begin{aligned}
D_{R_q}(\boldsymbol{\zeta}, \boldsymbol{\eta}) &= \frac{\kappa(q\boldsymbol{\zeta} + (1-q)\boldsymbol{\eta}) - (q\kappa(\boldsymbol{\zeta}) + (1-q)\kappa(\boldsymbol{\eta}))}{q(q-1)}, \\
D_{CR_q}(\boldsymbol{\zeta}, \boldsymbol{\eta}) &= \frac{\exp \{ \kappa(q\boldsymbol{\zeta} + (1-q)\boldsymbol{\eta}) - (q\kappa(\boldsymbol{\zeta}) + (1-q)\kappa(\boldsymbol{\eta})) \} - 1}{q(q-1)}, \\
D_H(\boldsymbol{\zeta}, \boldsymbol{\eta}) &= \left( 2 - 2 \exp \left\{ \kappa \left( \frac{\boldsymbol{\zeta} + \boldsymbol{\eta}}{2} \right) - \frac{\kappa(\boldsymbol{\zeta}) + \kappa(\boldsymbol{\eta})}{2} \right\} \right)^{1/2}.
\end{aligned}$$

**Proof** The assertion immediately follows from Lemma 3.38, since  $D_{R_q} = (\ln A_q)/[q(q-1)]$ ,  $D_{CR_q} = (A_q - 1)/[q(q-1)]$ , and  $D_H = (2 - 2A_{1/2})^{1/2}$ .  $\square$

*Remark 3.40* Every divergence measure  $D$ , say, of  $D_{KL}$ ,  $D_J$ ,  $D_{R_q}$ ,  $D_{CR_q}$ , and  $D_H$  in Lemma 3.37 and Corollary 3.39 satisfies the properties that, for any  $\xi, \eta \in \text{int}(\Xi^*)$ ,  $D(\xi, \eta) \geq 0$  and

$$D(\xi, \eta) = 0 \quad \Leftrightarrow \quad \xi = \eta.$$

Moreover,  $D_J$  and  $D_H$  are symmetric. Beyond that,  $D_H$  meets the triangle inequality and therefore defines a metric on  $\text{int}(\Xi^*) \times \text{int}(\Xi^*)$ .

**Example:**  $\mathcal{N}_k(\bullet, \Sigma_0)$

*Example 3.41* In the minimal canonical representation of  $\mathcal{N}_k(\bullet, \Sigma_0)$  in Example 2.65, the cumulant function and mean value function are given by  $\kappa(\mu) = \|\mu\|_{\Sigma_0}^2/2$  and  $\pi(\mu) = \Sigma_0^{-1}\mu$  for  $\mu \in \mathbb{R}^k$ . Hence, by Lemma 3.37,

$$D_{KL}(\mu, \eta) = \frac{\|\mu - \eta\|_{\Sigma_0}^2}{2},$$

$$D_J(\mu, \eta) = \|\mu - \eta\|_{\Sigma_0}^2,$$

for  $\mu, \eta \in \mathbb{R}^k$ . Moreover, we have for  $q \in (0, 1)$  that

$$\kappa(q\mu + (1-q)\eta) - (q\kappa(\mu) + (1-q)\kappa(\eta)) = \frac{q(q-1)\|\mu - \eta\|_{\Sigma_0}^2}{2},$$

which, by using Corollary 3.39 gives

$$D_{R_q}(\mu, \eta) = \frac{\|\mu - \eta\|_{\Sigma_0}^2}{2},$$

$$D_{CR_q}(\mu, \eta) = \frac{1}{q(q-1)} \left( \exp \left\{ \frac{q(q-1)\|\mu - \eta\|_{\Sigma_0}^2}{2} \right\} - 1 \right),$$

and, by setting  $q = 1/2$ ,

$$D_H(\mu, \eta) = \left( 2 - 2 \exp \left\{ -\frac{\|\mu - \eta\|_{\Sigma_0}^2}{8} \right\} \right)^{1/2},$$

for  $\mu, \eta \in \mathbb{R}^k$ . Hence,  $D_{KL}(\mu, \eta)$ ,  $D_J(\mu, \eta)$ ,  $D_{R_q}(\mu, \eta)$ ,  $D_{CR_q}(\mu, \eta)$ , and  $D_H(\mu, \eta)$  each admit a representation as  $h(\|\mu - \eta\|_{\Sigma_0}^2)$  for some increasing function  $h : [0, \infty) \rightarrow [0, \infty)$ .



The above representations of divergences can be applied to develop statistical procedures such as confidence regions (see, e.g., [60]). By using the EF structure, simple formulas for other related quantities may be derived as well. In the full EF with minimal canonical representation (2.6), the *weighted Matusita affinity*

$$\rho_n(\zeta^{(1)}, \dots, \zeta^{(n)}) = \int \left( \prod_{i=1}^n f_{\zeta^{(i)}}^* \right)^{w_i} d\mu, \quad \zeta^{(1)}, \dots, \zeta^{(n)} \in \Xi^*,$$

with positive weights  $w_1, \dots, w_n$  satisfying  $\sum_{i=1}^n w_i = 1$  and  $n \geq 2$  is given by

$$\rho_n(\zeta^{(1)}, \dots, \zeta^{(n)}) = \exp \left\{ \kappa \left( \sum_{i=1}^n w_i \zeta^{(i)} \right) - \sum_{i=1}^n w_i \kappa(\zeta^{(i)}) \right\}$$

(cf. Lemma 3.38); it can be applied, e.g., to construct a homogeneity test with null hypothesis  $\zeta^{(1)} = \dots = \zeta^{(n)}$  or to decide in a discriminant problem (see [25, 42, 34]).

## Chapter 4

# Parameter Estimation



In this chapter, parameter estimation is discussed when the underlying parametric model forms an EF. For this, we focus on the natural parametrization of the EF according to Definition 2.35, i.e.,  $\mathcal{P} = \{P_{\zeta}^* : \zeta \in \Xi\}$  with  $\mu$ -densities

$$\begin{aligned} f_{\zeta}^*(x) &= \frac{dP_{\zeta}^*}{d\mu}(x) = C^*(\zeta) \exp \left\{ \sum_{j=1}^k \zeta_j T_j(x) \right\} h(x) \\ &= e^{\zeta^t T(x) - \kappa(\zeta)} h(x), \quad x \in X, \end{aligned} \quad (4.1)$$

for  $\zeta \in \Xi$ , where  $\emptyset \neq \Xi \subset \Xi^*$ . To estimate  $\zeta$  or, more generally, a parameter  $\gamma(\zeta)$ ,  $\zeta \in \Xi$ , we will assume to have observations  $x^{(i)}$  considered realizations of iid random variables  $X^{(i)}$  with distribution  $P_{\zeta}^* \in \mathcal{P}$  for  $i \in I \subset \mathbb{N}$ . Every measurable function of  $X^{(i)}$ ,  $i \in I$ , with values in  $\Xi$  is then called an *estimator* of  $\zeta$  (in  $\Xi$  based on  $X^{(i)}$ ,  $i \in I$ ); an estimator of  $\gamma(\zeta)$ ,  $\zeta \in \Xi$ , or, for short, of  $\gamma$  is a measurable function of  $X^{(i)}$ ,  $i \in I$ , with values in  $\gamma(\Xi)$ . Here, all random variables  $X^{(i)}$ ,  $i \in I$ , are formally defined on the probability space  $(\Omega, \mathcal{A}, \mathbb{P}_{\zeta})$ , i.e., we have

$$\mathbb{P}_{\zeta}^{X^{(i)}} = P_{\zeta}^*, \quad i \in I.$$

Moreover, the notation  $\mathbb{E}_{\zeta}[\cdot]$ ,  $\mathbb{V}\text{ar}_{\zeta}(\cdot)$ , and  $\mathbb{C}\text{ov}_{\zeta}(\cdot)$  for the mean, variance, and covariance (matrix) is used when integration is with respect to  $\mathbb{P}_{\zeta}$ .

## 4.1 Maximum Likelihood Estimation

The maximum likelihood method, where, given some observation  $x$  of a random variable  $X$  with density function  $f_\zeta^*$ , we want to assign a global maximum point of  $f_\zeta^*(x)$ , as a function of  $\zeta$ , to an estimate of  $\zeta$ , is closely connected to the study of EFs. Under mild conditions and in the case of Lebesgue density functions, the unique existence of the maximum likelihood estimator can be guaranteed. For counting densities, existence is an issue, while uniqueness is ensured.

**Definition 4.1** Let  $x \in \mathcal{X}$  be a realization of  $X \sim P_\zeta^*$ . Every parameter in  $\Xi$  that maximizes the *likelihood function*

$$L(\zeta; x) = f_\zeta^*(x), \quad \zeta \in \Xi,$$

is called a *maximum likelihood estimate* (ML estimate) of  $\zeta$  in  $\Xi$  based on  $x$ .

Equivalently and usually analytically simpler, one may maximize the *log-likelihood function*

$$\ell = \ln(L)$$

to find the ML estimate(s). From formula (4.1), the log-likelihood function in our model is

$$\ell(\zeta; x) = \zeta^t \mathbf{T}(x) - \kappa(\zeta) + \ln(h(x)), \quad \zeta \in \Xi.$$

Hence, the aim is to maximize the expression

$$\zeta^t \mathbf{T}(x) - \kappa(\zeta)$$

with respect to  $\zeta$  in  $\Xi$ . For this, it will be convenient to consider the natural extension  $\tilde{\kappa}$  of  $\kappa$  with domain  $\mathbb{R}^k$  as introduced in Lemma 2.45 and the respective extension  $\tilde{\ell}$  of  $\ell$  to  $\mathbb{R}^k$  by setting  $\tilde{\ell}(\zeta; x) = -\infty$  for  $\zeta \in \mathbb{R}^k \setminus \Xi^*$ . For a simple notation, these extensions are again denoted by  $\kappa$  and  $\ell$  in the following. Theorem 2.44, Lemma 2.45, and Theorem 2.56 directly yield some useful properties of  $\ell$ .

**Lemma 4.2** Let  $\mathcal{P}$  be a full EF with minimal canonical representation (4.1). Then, we have for every  $x \in \mathcal{X}$ :

(a)  $\ell(\cdot; x)$  is strictly concave on  $\Xi^*$ .

(b)  $\ell(\cdot; x)$  is upper semi-continuous, i.e.,

$$\limsup_{\eta \rightarrow \zeta} \ell(\eta; x) \leq \ell(\zeta; x) \quad \text{for every } \zeta \in \mathbb{R}^k.$$

(c)  $\ell(\cdot; x)$  is continuously differentiable on  $\text{int}(\Xi^*)$  with derivative

$$\nabla \ell(\zeta; x) = \mathbf{T}(x) - \pi(\zeta), \quad \zeta \in \text{int}(\Xi^*),$$

where  $\pi$  is the mean value function introduced in Theorem 2.56.

Note that in Lemma 4.2 the mean value function  $\pi : \text{int}(\Xi^*) \rightarrow \pi(\text{int}(\Xi^*))$ , defined by  $\pi(\zeta) = E_{\zeta}[T]$  for  $\zeta \in \text{int}(\Xi^*)$ , is bijective with  $\pi(\text{int}(\Xi^*)) \subset \text{int}(M)$ , where  $M$  denotes the convex support of  $\nu^T$  (see Theorem 2.56 and Remark 2.62). If  $\mathcal{P}$  is moreover steep, we have

$$\pi(\text{int}(\Xi^*)) = \text{int}(M).$$

Now, we have the following fundamental theorem for steep EFs (see [4, p. 151]).

**Theorem 4.3** *Let  $\mathcal{P}$  be a steep EF with minimal canonical representation (4.1), and let  $x \in X$  be a given realization of the random variable  $X$  with distribution  $P_{\zeta}^* \in \mathcal{P}$ . Then, the ML estimate of  $\zeta$  in  $\Xi^*$  based on  $x$  exists if and only if*

$$\mathbf{T}(x) \in \pi(\text{int}(\Xi^*)) \quad (= \text{int}(M)).$$

If it exists, it is uniquely determined and given by  $\pi^{-1}(\mathbf{T}(x))$ .

**Proof** First, we show that the maximum of  $\ell = \ell(\cdot; x)$  cannot be attained at a boundary point of  $\Xi^*$ . Let  $\eta \in \Xi^* \setminus \text{int}(\Xi^*)$ ,  $\zeta \in \text{int}(\Xi^*)$ , and  $\zeta^{(\alpha)} = \zeta + \alpha(\eta - \zeta) \in \text{int}(\Xi^*)$ ,  $\alpha \in [0, 1)$ . By Lemma 4.2,

$$\begin{aligned} \liminf_{\alpha \nearrow 1} \ell(\zeta^{(\alpha)}) &\geq \liminf_{\alpha \nearrow 1} [\ell(\zeta) + \alpha(\ell(\eta) - \ell(\zeta))] \\ &= \ell(\eta) \geq \limsup_{\alpha \nearrow 1} \ell(\zeta^{(\alpha)}), \end{aligned}$$

i.e.,  $\lim_{\alpha \nearrow 1} \ell(\zeta^{(\alpha)}) = \ell(\eta)$ . Moreover, since  $\mathcal{P}$  is steep,

$$\lim_{\alpha \nearrow 1} \frac{\partial}{\partial \alpha} \ell(\zeta^{(\alpha)}) = \lim_{\alpha \nearrow 1} \left[ (\eta - \zeta)^t \mathbf{T}(x) - \frac{\partial}{\partial \alpha} \kappa(\zeta^{(\alpha)}) \right] = -\infty,$$

such that there exists some  $\tilde{\alpha} \in (0, 1)$  with  $\ell(\zeta^{(\tilde{\alpha})}) > \ell(\eta)$ . Hence, if  $\ell$  has a maximum, it is attained on  $\text{int}(\Xi^*)$  and necessary a solution of the likelihood equation

$$\mathbf{0} = \nabla \ell(\zeta) = \mathbf{T}(x) - \pi(\zeta).$$

In case of  $\mathbf{T}(x) \in \pi(\text{int}(\Xi^*))$ , this equation has the only solution  $\pi^{-1}(\mathbf{T}(x))$ , which is then also the unique ML estimate of  $\zeta$  in  $\Xi^*$  based on  $x$ . Otherwise, an ML estimate in  $\Xi^*$  based on  $x$  does not exist.  $\square$

*Remark 4.4* In Theorem 4.3, the ML estimate of  $\zeta$  in  $\Xi^*$  based on a single observation of  $X \sim P_\zeta^*$  exists with probability

$$P_\zeta^*(\mathbf{T} \in \pi(\text{int}(\Xi^*))) = P_\zeta^*(\mathbf{T} \in \text{int}(M)).$$

Since  $\nu^{\mathbf{T}}(M^c) = 0$  as well as  $\nu^{\mathbf{T}}$  and  $(P_\zeta^*)^{\mathbf{T}}$  are equivalent measures by Remark 2.16 and Lemma 2.17, it follows that

$$P_\zeta^*(\mathbf{T} \notin M) = 0 \quad \text{for } \zeta \in \Xi^*.$$

Hence, the ML estimate of  $\zeta$  in  $\Xi^*$  exists with  $P_\zeta^*$ -probability 1 if and only if  $\mathbf{T}$  lies on the boundary of  $M$  with  $P_\zeta^*$ -probability 0. In particular, this is the case if the distribution of  $\mathbf{T}$  has a Lebesgue density, since the boundary of any convex set in  $\mathbb{R}^k$  has Lebesgue measure 0.

Based on an iid sample from  $P_\zeta^*$ , we arrive at the following version of Theorem 4.3.

**Corollary 4.5** *Let  $\mathcal{P}$  be a steep EF with minimal canonical representation (4.1). Moreover, let  $X^{(1)}, \dots, X^{(n)}$  be an iid sample from the distribution  $P_\zeta^* \in \mathcal{P}$ , and let  $x^{(1)}, \dots, x^{(n)}$  denote realizations of  $X^{(1)}, \dots, X^{(n)}$ . Then, the ML estimate of  $\zeta$  in  $\Xi^*$  based on  $x^{(1)}, \dots, x^{(n)}$  exists if and only if*

$$\frac{1}{n} \sum_{i=1}^n \mathbf{T}(x^{(i)}) \in \pi(\text{int}(\Xi^*)) \quad (= \text{int}(M)). \quad (4.2)$$

*If the ML estimate exists, it is uniquely determined and given by*

$$\pi^{-1} \left( \frac{1}{n} \sum_{i=1}^n \mathbf{T}(x^{(i)}) \right).$$

**Proof** By using Theorem 3.13 and Corollary 3.17, we may apply Theorem 4.3 to the EF  $\mathcal{P}^{(n)}$  of distributions of  $(X^{(1)}, \dots, X^{(n)})$ . By doing so,  $\mathbf{T}$  and  $\pi$  have to be replaced by  $\mathbf{T}^{(n)}$  and  $n\pi$ . The assertion then follows from

$$\mathbf{T}^{(n)}(x^{(1)}, \dots, x^{(n)}) \in (n\pi)(\text{int}(\Xi^*)) \quad \Leftrightarrow \quad \frac{1}{n} \sum_{i=1}^n \mathbf{T}(x^{(i)}) \in \pi(\text{int}(\Xi^*)).$$

$\square$

**Notation 4.6** If, in the situation of Corollary 4.5, condition (4.2) is true with  $(P_\zeta^*)^{(n)}$ -probability 1, the *maximum likelihood estimator* (MLE) of  $\zeta$  in  $\Xi^*$  based on  $X^{(1)}, \dots, X^{(n)}$ , which then uniquely exists  $\mathbb{P}_\zeta$ -a.s., is denoted by

$$\hat{\zeta}^{(n)} = \pi^{-1} \left( \frac{1}{n} \sum_{i=1}^n T(X^{(i)}) \right). \quad (4.3)$$

**Examples:**  $\text{poi}(\bullet)$ ,  $\text{b}(n_0, \bullet)$ ,  $\mathcal{N}(\bullet, \bullet)$ ,  $\mathcal{N}_k(\bullet, \Sigma_0)$

*Example 4.7* From Example 2.64 for  $\mathcal{P} = \text{poi}(\bullet)$ , we have  $T(x) = x$  for  $x \in \mathbb{N}_0$  and  $\pi(\zeta) = \exp\{\zeta\}$  for  $\zeta \in \Xi^* = \mathbb{R}$ . Here,  $\pi(\text{int}(\Xi^*)) = (0, \infty)$ . Now, let  $X_1, \dots, X_n$  be an iid sample from  $P_\zeta^*$ , and let  $x_1, \dots, x_n \in \mathbb{N}_0$  denote realizations of  $X_1, \dots, X_n$ . Then,

$$\frac{1}{n} \sum_{i=1}^n x_i \notin (0, \infty) \quad \Leftrightarrow \quad x_1 = \dots = x_n = 0.$$

Hence, by Corollary 4.5, an ML estimate of  $\zeta$  based on  $x_1, \dots, x_n \in \mathbb{N}_0$  does not exist if  $x_1 = \dots = x_n = 0$ . In any other case,

$$\ln \left( \frac{1}{n} \sum_{i=1}^n x_i \right)$$

is the unique ML estimate. The probability for the non-existence of the ML estimate is given by

$$\mathbb{P}_\zeta(X_1 = 0, \dots, X_n = 0) = \exp \{ -ne^\zeta \}, \quad \zeta \in \mathbb{R},$$

and converges to 0 for  $n \rightarrow \infty$ .

*Example 4.8* From Example 2.57 for  $\mathcal{P} = \text{b}(n_0, \bullet)$ , we have  $T(x) = x$  for  $x \in \{0, 1, \dots, n_0\}$ ,  $\Xi^* = \mathbb{R}$ , and  $\pi(\Xi^*) = (0, n_0)$ . Now, let  $X_1, \dots, X_n$  be an iid sample from  $P_\zeta^*$ , and let  $x_1, \dots, x_n \in \{0, 1, \dots, n_0\}$  denote realizations of

$X_1, \dots, X_n$ . Then,

$$\frac{1}{n} \sum_{i=1}^n x_i \notin (0, n_0) \quad \Leftrightarrow \quad (x_1 = \dots = x_n = 0 \quad \vee \quad x_1 = \dots = x_n = n_0)$$

Hence, by Corollary 4.5, an ML estimate of  $\zeta$  based on  $x_1, \dots, x_n \in \{0, 1, \dots, n_0\}$  does not exist if  $x_1 = \dots = x_n = 0$  or  $x_1 = \dots = x_n = n_0$ . In any other case,

$$\pi^{-1} \left( \frac{1}{n} \sum_{i=1}^n x_i \right) = \ln \left( \frac{\sum_{i=1}^n x_i}{n_0 n - \sum_{i=1}^n x_i} \right)$$

is the unique ML estimate. The probability for the non-existence of the ML estimate is given by

$$\begin{aligned} \mathbb{P}_\zeta(\{X_1 = \dots = X_n = 0\} \cup \{X_1 = \dots = X_n = n_0\}) \\ = \left( \frac{1}{1 + \exp\{\zeta\}} \right)^{n_0 n} + \left( \frac{\exp\{\zeta\}}{1 + \exp\{\zeta\}} \right)^{n_0 n} \end{aligned}$$

and converges to 0 for  $n \rightarrow \infty$ .

*Example 4.9* From Example 2.59 for  $\mathcal{P} = \mathcal{N}(\bullet, \bullet)$ , we have  $T(x) = (x, x^2)^t$  for  $x \in \mathbb{R}$ ,  $\Xi^* = \mathbb{R} \times (-\infty, 0)$ , and

$$\pi(\Xi^*) = \{(x, y)^t \in \mathbb{R}^2 : x^2 < y\}.$$

By Corollary 4.5, the probability for the existence of the ML estimate of  $\zeta$  in  $\Xi^*$  based on  $n$  observations is given by

$$(P_\zeta^*)^{(n)} \left( \left\{ (x_1, \dots, x_n) \in \mathbb{R}^n : \frac{1}{n} \sum_{i=1}^n T(x_i) \in \pi(\Xi^*) \right\} \right) = \begin{cases} 0, & n = 1, \\ 1, & n \geq 2, \end{cases}$$

since  $(x_1, x_1^2)^t \notin \pi(\Xi^*)$  for all  $x_1 \in \mathbb{R}$ , and

$$\left( \frac{1}{n} \sum_{i=1}^n x_i \right)^2 \leq \frac{1}{n} \sum_{i=1}^n x_i^2$$

for all  $x_1, \dots, x_n \in \mathbb{R}$  and  $n \geq 2$  (the empirical variance is non-negative) with equality if and only if  $x_1 = \dots = x_n$ . Hence, for  $n \geq 2$ , the MLE of  $\zeta$  based on

an iid sample  $X_1, \dots, X_n$  from  $P_\zeta^*$  uniquely exists ( $\mathbb{P}_\zeta$ -a.s.) and is given by

$$\hat{\zeta}^{(n)} = \pi^{-1} \left( \frac{1}{n} \sum_{i=1}^n X_i, \frac{1}{n} \sum_{i=1}^n X_i^2 \right)$$

with inverse function  $\pi^{-1}$  as stated in Example 2.59.

*Example 4.10* From Example 2.65 for  $\mathcal{P} = \mathcal{N}_k(\bullet, \Sigma_0)$ , we have  $T(x) = \Sigma_0^{-1}x$  for  $x \in \mathbb{R}^k$ ,  $\Xi^* = \mathbb{R}^k$ , and  $\pi(\mu) = \Sigma_0^{-1}\mu$  for  $\mu \in \mathbb{R}^k$ . Here,  $\pi(\Xi^*) = \mathbb{R}^k$  such that condition (4.2) is trivially met. Hence, by Corollary 4.5, the MLE of  $\mu$  based on an iid sample  $X^{(1)}, \dots, X^{(n)}$  from  $P_\mu$  uniquely exists and is given by

$$\hat{\mu}^{(n)} = \pi^{-1} \left( \frac{1}{n} \sum_{i=1}^n T(X^{(i)}) \right) = \Sigma_0 \left( \frac{1}{n} \sum_{i=1}^n \Sigma_0^{-1} X^{(i)} \right) = \frac{1}{n} \sum_{i=1}^n X^{(i)}.$$

## 4.2 Constrained Maximum Likelihood Estimation

We now consider maximum likelihood estimation in non-full EFs. Any such non-full family is represented by a subfamily  $\{P_\zeta^* : \zeta \in \Xi\}$  of the full family  $\{P_\zeta^* : \zeta \in \Xi^*\}$  for some proper subset  $\Xi$  of  $\Xi^*$  (with  $\Xi \neq \Xi^*$ ). Thus, the aim is now to find all parameters within the *constrained* set  $\Xi$  at which the likelihood function has a global maximum on  $\Xi$ .

For every  $x \in \mathcal{X}$  and every compact set  $\Xi \subset \text{int}(\Xi^*)$ , it is clear that there exists an ML estimate of  $\zeta$  in  $\Xi$  based on  $x$ , since  $\ell(\cdot; x)$  is continuous on  $\Xi$  by Lemma 4.2 and thus attains its maximum on  $\Xi$ . On the other hand, the strict concavity of  $\ell(\cdot; x)$  on  $\Xi^*$  ensures uniqueness of the ML estimate in every convex set  $\Xi \subset \Xi^*$  provided that it exists. Moreover, in the steep EF and if  $T(x) \in \pi(\Xi)$ , then  $\pi^{-1}(T(x))$  must be the ML estimate in  $\Xi$ , since it already maximizes  $\ell(\cdot; x)$  over  $\Xi^*$ . In [20, pp. 152–155], under different assumptions on the constrained set  $\Xi$ , sufficient and necessary conditions for the existence and uniqueness of ML estimates in non-full EFs are stated. The findings are presented in this section. First, as a preliminary result, we address the following lemma.

**Lemma 4.11** *Let  $\mathcal{P}$  be a full EF with minimal canonical representation (4.1). Moreover, let  $x \in \mathcal{X}$  and  $\emptyset \neq A \subset \mathbb{R}^k$  be closed. If*

$$\ell(\zeta) = \ell(\zeta; x) \rightarrow -\infty \quad \text{for } \|\zeta\|_2 \rightarrow \infty, \quad (4.4)$$

*then there exists some  $\eta \in A$  with  $\ell(\eta) = \sup_{\zeta \in A} \ell(\zeta)$ .*



**Proof** Let  $m = \sup_{\xi \in A} \ell(\xi)$ , and  $\xi_n \in A$ ,  $n \in \mathbb{N}$ , be a sequence with  $\ell(\xi_n) \rightarrow m$  for  $n \rightarrow \infty$ . If  $m = -\infty$ , the assertion is trivially met. For  $m > -\infty$ , it holds by assumption that  $\xi_n$ ,  $n \in \mathbb{N}$ , is bounded. Hence, there exists a convergent subsequence  $\xi_{n_k}$ ,  $k \in \mathbb{N}$ , with  $\xi_{n_k} \rightarrow \eta \in A$  for  $k \rightarrow \infty$ . By Lemma 4.2,  $\ell$  is upper semi-continuous, which yields

$$m \geq \ell(\eta) \geq \limsup_{k \rightarrow \infty} \ell(\xi_{n_k}) = m,$$

i.e.,  $\ell(\eta) = m$ . □

To obtain a sufficient condition for property (4.4) to hold true, we introduce the half spaces

$$H^+(\mathbf{y}, a) = \{\mathbf{z} \in \mathbb{R}^k : \mathbf{y}^t \mathbf{z} > a\},$$

$$H^-(\mathbf{y}, a) = \{\mathbf{z} \in \mathbb{R}^k : \mathbf{y}^t \mathbf{z} < a\},$$

for  $\mathbf{y} \in \mathbb{R}^k$  and  $a \in \mathbb{R}$ .

**Lemma 4.12** *Let  $\mathcal{P}$  be a full EF with minimal canonical representation (4.1), and let  $\mathbf{0} \in \text{int}(M)$ . Then,  $\kappa(\xi) \rightarrow \infty$  for  $\|\xi\|_2 \rightarrow \infty$ .*

**Proof** First, we show that there exist some  $a > 0$  and  $c > 0$  such that

$$v^T(H^+(\mathbf{y}, a)) > c \quad \text{for all } \mathbf{y} \in B = \{\mathbf{z} \in \mathbb{R}^k : \|\mathbf{z}\|_2 = 1\}.$$

Suppose the contrary assertion is true. Then, there exist a sequence  $a_n$ ,  $n \in \mathbb{N}$ , with  $\lim_{n \rightarrow \infty} a_n = 0$ , and a sequence  $\mathbf{y}_n \in B$ ,  $n \in \mathbb{N}$ , with

$$\lim_{n \rightarrow \infty} v^T(H^+(\mathbf{y}_n, a_n)) = 0.$$

Since  $B$  is compact, there exist  $\mathbf{y} \in B$  and a convergent subsequence  $\mathbf{y}_{n_k}$ ,  $k \in \mathbb{N}$ , with  $\mathbf{y}_{n_k} \rightarrow \mathbf{y}$  for  $k \rightarrow \infty$ . Note that, for every  $\mathbf{z} \in \mathbb{R}^k$  with  $\mathbf{y}^t \mathbf{z} > 0$ , there exists an index  $k_0 \in \mathbb{N}$  such that  $\mathbf{y}_{n_k}^t \mathbf{z} > a_{n_k}$  for all  $k \geq k_0$ . Together with Fatou's lemma (see, e.g., [17, p. 209]), this yields

$$\begin{aligned} v^T(H^+(\mathbf{y}, 0)) &= \int \mathbb{1}_{H^+(\mathbf{y}, 0)} dv^T \leq \int \liminf_{k \rightarrow \infty} \mathbb{1}_{H^+(\mathbf{y}_{n_k}, a_{n_k})} dv^T \\ &\leq \liminf_{k \rightarrow \infty} \int \mathbb{1}_{H^+(\mathbf{y}_{n_k}, a_{n_k})} dv^T = \lim_{k \rightarrow \infty} v^T(H^+(\mathbf{y}_{n_k}, a_{n_k})) = 0. \end{aligned}$$

Hence, we have  $\text{supp}(v^T) \subset H^+(\mathbf{y}, 0)^c$  and thus  $M \subset H^+(\mathbf{y}, 0)^c$ , since  $H^+(\mathbf{y}, 0)^c$  is convex and closed. By assumption,  $\mathbf{0} \in \text{int}(M)$ , which gives the contradiction  $\mathbf{0} \in \text{int}(H^+(\mathbf{y}, 0)^c) = H^-(\mathbf{y}, 0)$ .

Now, for  $\zeta \in \mathbb{R}^k \setminus \{\mathbf{0}\}$ ,

$$\begin{aligned} \kappa(\zeta) &= \ln \left( \int e^{\zeta^t t} dv^T(t) \right) = \ln \left( \int \exp \left\{ \|\zeta\|_2 \left( \frac{\zeta}{\|\zeta\|_2} \right)^t t \right\} dv^T(t) \right) \\ &\geq \ln \left( \int_{H^+(\zeta/\|\zeta\|_2, a)} \exp \left\{ \|\zeta\|_2 \left( \frac{\zeta}{\|\zeta\|_2} \right)^t t \right\} dv^T(t) \right) \\ &\geq \ln \left( e^{\|\zeta\|_2 a} c \right) = a \|\zeta\|_2 + \ln(c). \end{aligned}$$

For  $\|\zeta\|_2 \rightarrow \infty$ , this yields the assertion.  $\square$

For a relatively closed subset  $\Xi$  of  $\Xi^*$ , we have the following theorem (see [20, pp. 152/153]).

**Theorem 4.13** *Let  $\mathcal{P}$  be a steep EF with minimal canonical representation (4.1), and let  $x \in X$  be a realization of the random variable  $X$  with distribution  $P_\zeta^* \in \mathcal{P}$ . Moreover, let  $\Xi \neq \emptyset$  be relatively closed in  $\Xi^*$ , i.e.,  $\Xi = B \cap \Xi^*$  for a closed set  $B \subset \mathbb{R}^k$ . Then, an ML estimate of  $\zeta$  in  $\Xi$  based on  $x$  exists if one of the following conditions is met:*

- (a)  $T(x) \in \text{int}(M)$ ,
- (b)  $T(x) \in M \setminus \text{int}(M)$  and there exist  $t_1, \dots, t_m \in \text{int}(M)$  and  $a_1, \dots, a_m \in \mathbb{R}$  with

$$\Xi \subset \bigcup_{i=1}^m H^-(T(x) - t_i, a_i). \quad (4.5)$$

**Proof** Let  $t = T(x)$ . First, let (a) be true, i.e.,  $t \in \text{int}(M)$ . Since for  $\zeta \in \Xi^*$

$$f_\zeta^* = e^{\zeta^t T - \kappa(\zeta)} h = e^{\zeta^t (T-t) - (\kappa(\zeta) - \zeta^t t)} h$$

we have another representation of  $\mathcal{P}$  with statistic  $\check{T} = T - t$  and cumulant function  $\check{\kappa}(\zeta) = \kappa(\zeta) - \zeta^t t$ ,  $\zeta \in \mathbb{R}^k$ . The corresponding convex support  $\check{M}$  of  $v^{\check{T}}$  is then

$$\check{M} = M - t$$

and satisfies  $\mathbf{0} \in \text{int}(\check{M})$  by assumption. Hence, by Lemma 4.12, we have

$$\ell(\zeta; x) = \ln(h(x)) - \check{\kappa}(\zeta) \rightarrow -\infty$$

for  $\|\zeta\|_2 \rightarrow \infty$ . Applying Lemma 4.11 then yields that there exists some  $\eta \in B$  with  $\ell(\eta; x) = \sup_{\zeta \in B} \ell(\zeta; x)$ . Necessarily holds that  $\eta \in \Xi = B \cap \Xi^*$ , since  $\eta \notin \Xi^*$  would imply that  $\ell(\eta; x) = -\infty$ , which forms a contradiction to  $\emptyset \neq \Xi \subset \Xi^*$ .

Now, let assertion (b) be true. Since  $\mathbf{t}_1, \dots, \mathbf{t}_m \in \text{int}(M)$ , we have  $\kappa_i(\boldsymbol{\zeta}) = \kappa(\boldsymbol{\zeta}) - \boldsymbol{\zeta}^t \mathbf{t}_i \rightarrow \infty$  for  $\|\boldsymbol{\zeta}\|_2 \rightarrow \infty$  and  $1 \leq i \leq m$  by the same arguments as above. Moreover, for  $\boldsymbol{\zeta} \in \Xi$ , there exists an index  $j \in \{1, \dots, m\}$  with

$$\begin{aligned} \ell(\boldsymbol{\zeta}; x) &= \ln(h(x)) + \boldsymbol{\zeta}^t (\mathbf{T}(x) - \mathbf{t}_j) - \kappa_j(\boldsymbol{\zeta}) \\ &\leq \ln(h(x)) + a_j - \kappa_j(\boldsymbol{\zeta}). \end{aligned}$$

Thus,

$$\ell(\boldsymbol{\zeta}; x) \leq \sup_{1 \leq i \leq m} \{\ln(h(x)) + a_i - \kappa_i(\boldsymbol{\zeta})\} \rightarrow -\infty$$

for  $\|\boldsymbol{\zeta}\|_2 \rightarrow \infty$ . Again, applying Lemma 4.11 completes the proof.  $\square$

Theorem 4.13 may be useful to ensure existence of ML estimates in curved EFs as indicated in the following remark.

*Remark 4.14* Let  $\mathcal{P}$  be a steep EF with minimal canonical representation (4.1). Moreover, let  $\Theta \subset \mathbb{R}^m$  for some  $m < k$ ,

$$\mathbf{Z} = (Z_1, \dots, Z_k)^t : \Theta \rightarrow \Xi^*$$

be a one-to-one mapping with  $\text{int}(\mathbf{Z}(\Theta)) = \emptyset$ , and  $Z_1, \dots, Z_k$  be affinely independent. Then, the non-full subfamily

$$\mathcal{P}_{\mathbf{Z}} = \{P_{\mathbf{Z}(\vartheta)}^* : \vartheta \in \Theta\}$$

of  $\mathcal{P}$  forms a curved EF according to Definition 2.32. Now, let  $x$  be a realization of  $X \sim P_{\mathbf{Z}(\vartheta)}^* \in \mathcal{P}_{\mathbf{Z}}$ . If  $\mathbf{Z}(\Theta)$  is relatively closed in  $\Xi^*$ , Theorem 4.13 is applicable and may be used to ensure existence of ML estimates of  $\vartheta$  in  $\Theta$  based on  $x$ . For example, an ML estimate of  $\vartheta$  in  $\Theta$  exists with probability 1 if the distribution of  $\mathbf{T}$  has a Lebesgue density (see Remark 4.4). Under the additional assumptions that  $\Theta$  is open,  $\mathbf{Z}(\Theta) \subset \text{int}(\Xi^*)$ , and  $\mathbf{Z}$  is differentiable, any ML estimate of  $\vartheta$  in  $\Theta$  based on  $x$  is necessarily a solution of the likelihood equation

$$\mathbf{D}_{\mathbf{Z}}(\vartheta)^t [\mathbf{T}(x) - \pi(\mathbf{Z}(\vartheta))] = \mathbf{0} \quad (4.6)$$

with respect to  $\vartheta \in \Theta$  (see Theorem 2.56). Note, however, that the likelihood function may have several local maxima and thus Eq. (4.6) more than one solution in  $\Theta$  (see [5, p. 69]).

Uniqueness of the ML estimate may also be guaranteed by imposing further assumptions on both  $\mathbf{Z}$  and  $\mathbf{T}$ . For instance, in case of Lebesgue density functions, the MLE of  $\vartheta$  in  $\Theta$  is unique with probability 1 if  $X \subset \mathbb{R}^n$  for some  $n \geq k$  and  $X$  is open,  $\Theta$  is open with  $\mathbf{Z}(\Theta) \subset \text{int}(\Xi^*)$ ,  $\mathbf{T}$  is continuously differentiable and  $\mathbf{Z}$  is twice continuously differentiable with Jacobian matrix  $\mathbf{D}_{\mathbf{T}}(x)$  of full rank  $k$  at every

$x \in \mathcal{X}$  and Jacobian matrix  $\mathbf{D}_{\mathbf{Z}}(\vartheta)$  of full rank  $m$  at every  $\vartheta \in \Theta$ , respectively (see [49] and [38]).

**Example: Curved EF in  $\mathcal{N}(\bullet, \bullet)$**

*Example 4.15* We consider the curved EF  $\mathcal{P}_2 = \{\mathcal{N}(\mu, \mu^2) : \mu \in \mathbb{R} \setminus \{0\}\}$  in Example 2.31(b) forming a subfamily of  $\mathcal{N}(\bullet, \bullet)$  with  $\text{ord}(\mathcal{P}_2) = 2$ . By using the canonical representation of  $\mathcal{N}(\bullet, \bullet)$  introduced in Example 2.39, we have  $\mathcal{P}_2 = \{P_{\mathbf{Z}(\mu)}^* : \mu \in \Theta\}$  with  $\Theta = \mathbb{R} \setminus \{0\}$  and  $\mathbf{Z} = (Z_1, Z_2)^t : \Theta \rightarrow \Xi^* = \mathbb{R} \times (-\infty, 0)$ , where

$$Z_1(\mu) = \frac{1}{\mu} \quad \text{and} \quad Z_2(\mu) = -\frac{1}{2\mu^2} \quad \text{for } \mu \in \Theta.$$

The corresponding statistic is given by  $\mathbf{T}(x) = (x, x^2)^t$ ,  $x \in \mathbb{R}$ . Obviously,  $\mathbf{Z}$  is one-to-one, and

$$\mathbf{Z}(\Theta) = \{(\eta, -\eta^2/2)^t : \eta \in \mathbb{R} \setminus \{0\}\}$$

is relatively closed in  $\Xi^*$ , since  $\mathbf{Z}(\Theta) = B \cap \Xi^*$  for  $B = \{(\eta, -\eta^2/2)^t : \eta \in \mathbb{R}\}$ , which is closed in  $\mathbb{R}^2$ . From Remark 2.62 and Example 2.59, we have that

$$\text{int}(M) = \pi(\Xi^*) = \{(x, y)^t \in \mathbb{R}^2 : x^2 < y\}.$$

Now, let us assume to have realizations  $x_1, \dots, x_n$  of an iid sample  $X_1, \dots, X_n$  of size  $n \geq 2$  from  $P_{\mathbf{Z}(\mu)}^*$  for some  $\mu \in \Theta$ . Since

$$\frac{1}{n} \sum_{i=1}^n \mathbf{T}(x_i) = \left( \frac{1}{n} \sum_{i=1}^n x_i, \frac{1}{n} \sum_{i=1}^n x_i^2 \right)^t \in \text{int}(M)$$

if  $x_1, \dots, x_n$  are not all identical (see Example 4.9), Theorem 4.13 applied to the product EF  $\mathcal{P}^{(n)}$  with statistic  $\mathbf{T}^{(n)}(\mathbf{x}) = \sum_{i=1}^n \mathbf{T}(x_i)$  for  $\mathbf{x} = (x_1, \dots, x_n) \in \mathcal{X}^n = \mathbb{R}^n$  ensures that an MLE of  $\mu$  in  $\Theta$  based on  $X_1, \dots, X_n$  exists with probability 1. Moreover, the MLE is uniquely determined with probability 1 by Remark 4.14. To see this, we formally replace the sample space  $\mathcal{X}^n = \mathbb{R}^n$  by  $\mathcal{Y} = \mathbb{R}^n \setminus N$ , where  $N = \{(x, \dots, x) \in \mathbb{R}^n : x \in \mathbb{R}\}$  has probability 0. Then,  $\mathcal{Y}$  is open and  $\mathbf{T}^{(n)}$  is continuously differentiable on  $\mathcal{Y}$  with Jacobian matrix

$$\mathbf{D}_{\mathbf{T}^{(n)}}(\mathbf{y}) = \begin{pmatrix} \mathbf{1} \\ 2\mathbf{y} \end{pmatrix}, \quad \mathbf{y} \in \mathcal{Y},$$

with  $\mathbf{1} = (1, \dots, 1) \in \mathbb{R}^n$ , which has full rank  $k = 2 \leq n$  at every  $\mathbf{y} \in \mathcal{Y}$ . Since  $\Theta = \mathbb{R} \setminus \{0\}$  is open and  $\mathbf{Z}$  is twice continuously differentiable with Jacobian matrix

$$\mathbf{D}_{\mathbf{Z}}(\mu) = \left( -\frac{1}{\mu^2}, \frac{1}{\mu^3} \right)^t$$

of rank  $m = 1$  at every  $\mu \in \Theta$ , all conditions in Remark 4.14 for the uniqueness of the MLE are met (see also [49, Ex. 3]).

If  $\Xi$  in Theorem 4.13 is moreover convex, the ML estimate of  $\zeta$  in  $\Xi$  based on  $x$  is unique provided that it exists. Furthermore, we have the following theorem (see [20, pp. 153–155]).

**Theorem 4.16** *In the situation of Theorem 4.13, let  $\Xi$  be moreover convex and  $\Xi \cap \text{int}(\Xi^*) \neq \emptyset$ . Then, there exists a unique ML estimate of  $\zeta$  in  $\Xi$  based on  $x$  if and only if one of the following conditions is met:*

- (a)  $\mathbf{T}(x) \in \text{int}(M)$ ,
- (b)  $\mathbf{T}(x) \in M \setminus \text{int}(M)$  and there exist  $\mathbf{t}_1 \in \text{int}(M)$  and  $a_1 \in \mathbb{R}$  with

$$\Xi \subset H^-(\mathbf{T}(x) - \mathbf{t}_1, a_1).$$

If an ML estimate of  $\zeta$  in  $\Xi$  based on  $x$  exists, it is the unique parameter  $\zeta^* \in \Xi \cap \text{int}(\Xi^*)$  satisfying

$$(\mathbf{T}(x) - \pi(\zeta^*))^t (\zeta^* - \eta) \geq 0 \quad \text{for all } \eta \in \Xi. \quad (4.7)$$

**Proof** First, let us show that if inequality (4.7) has a solution in  $\Xi \cap \text{int}(\Xi^*)$ , it is uniquely determined. For this, let  $\zeta_1^*, \zeta_2^* \in \Xi \cap \text{int}(\Xi^*)$  be solutions of inequality (4.7). Then, in particular,

$$(\mathbf{T}(x) - \pi(\zeta_1^*))^t (\zeta_1^* - \zeta_2^*) \geq 0,$$

$$(\mathbf{T}(x) - \pi(\zeta_2^*))^t (\zeta_2^* - \zeta_1^*) \geq 0.$$

Summation of both inequalities and applying Lemma 3.37 yields

$$0 \leq (\pi(\zeta_2^*) - \pi(\zeta_1^*))^t (\zeta_1^* - \zeta_2^*) = -D_J(\zeta_1^*, \zeta_2^*)$$

By Remark 3.40, it follows that  $D_J(\zeta_1^*, \zeta_2^*) = 0$  and  $\zeta_1^* = \zeta_2^*$ .

Next, let us show that  $\zeta^*$  is an ML estimate of  $\zeta$  in  $\Xi$  based on  $x$  if and only if  $\zeta^* \in \Xi \cap \text{int}(\Xi^*)$  and  $\zeta^*$  satisfies inequality (4.7).

First, if  $\zeta^* \notin \Xi \cap \text{int}(\Xi^*)$ ,  $\zeta$  cannot be an ML estimate of  $\zeta$  in  $\Xi$  based on  $x$ , since the steepness of  $\mathcal{P}$  ensures the existence of some  $\eta \in \Xi \cap \text{int}(\Xi^*)$  with  $\ell(\eta; x) > \ell(\zeta; x)$ , as shown in the proof of Theorem 4.3. Now, let  $\zeta^* \in \Xi \cap \text{int}(\Xi^*)$  and suppose that there exists some  $\eta \in \Xi$  with

$$(\mathbf{T}(x) - \pi(\zeta^*))^t (\zeta^* - \eta) < 0.$$

Since  $\Xi$  is convex by assumption,  $\eta^{(\alpha)} = \zeta^* + \alpha(\eta - \zeta^*) \in \Xi$ ,  $\alpha \in [0, 1]$ , and we obtain from Lemma 4.2

$$\left. \frac{\partial}{\partial \alpha} \ell(\eta^{(\alpha)}; x) \right|_{\alpha=0} = (\mathbf{T}(x) - \pi(\zeta^*))^t (\eta - \zeta^*) > 0.$$

Hence, there exists  $\tilde{\alpha} \in (0, 1)$  with  $\ell(\eta^{(\tilde{\alpha})}; x) > \ell(\zeta^*; x)$ , which implies that  $\zeta^*$  is not an ML estimate of  $\zeta$  in  $\Xi$  based on  $x$ .

Now, let  $\zeta^* \in \Xi \cap \text{int}(\Xi^*)$  satisfy inequality (4.7). Then, by using Lemma 3.37, it holds for  $\eta \in \Xi \cap \text{int}(\Xi^*)$  that

$$\begin{aligned} \ell(\zeta^*; x) - \ell(\eta; x) &= (\zeta^* - \eta)^t \mathbf{T}(x) - \kappa(\zeta^*) + \kappa(\eta) \\ &= (\zeta^* - \eta)^t (\mathbf{T}(x) - \pi(\zeta^*)) + D_{KL}(\zeta^*, \eta) \\ &\geq 0, \end{aligned}$$

since  $\zeta^*$  satisfies inequality (4.7) and  $D_{KL}$  is non-negative by Remark 3.40. Moreover, by the same arguments as in the proof of Theorem 4.3, the steepness of  $\mathcal{P}$  ensures that for every  $\tilde{\eta} \in \Xi \cap (\Xi^* \setminus \text{int}(\Xi^*))$  there exists some  $\eta \in \Xi \cap \text{int}(\Xi^*)$  with  $\ell(\eta; x) > \ell(\tilde{\eta}; x)$ . Hence,  $\zeta^*$  is an ML estimate of  $\zeta$  in  $\Xi$  based on  $x$ .

Finally, we prove the equivalence assertion. If either statement (a) or (b) is met, then existence of an ML estimate of  $\zeta$  in  $\Xi$  based on  $x$  follows from Theorem 4.13. Since  $\ell(\cdot; x)$  is strictly concave on  $\Xi$  by Lemma 4.2, this ML estimate is uniquely determined. For the other implication, let  $\zeta^*$  denote the unique ML estimate of  $\zeta$  in  $\Xi$  based on  $x$ . Then, as already shown above,  $\zeta^* \in \Xi \cap \text{int}(\Xi^*)$  and  $\zeta^*$  satisfies inequality (4.7), which is equivalent to

$$\eta^t (\mathbf{T}(x) - \pi(\zeta^*)) \leq (\zeta^*)^t (\mathbf{T}(x) - \pi(\zeta^*)) \quad \text{for all } \eta \in \Xi.$$

Hence, assertion (b) is true with  $t_1 = \pi(\zeta^*) \in \text{int}(M)$  and some arbitrary  $a_1 > (\zeta^*)^t (\mathbf{T}(x) - \pi(\zeta^*))$ .  $\square$

For more details on maximum likelihood estimation in EFs including, in particular, geometrical interpretations and illustrations of formulas (4.5) and (4.7), see [20, Chapter 5].

So far in this chapter, we have considered maximum likelihood estimation for the natural parameter of the EF. For the sake of completeness, we finally address a result and a comment for other parametrizations.

**Lemma 4.17** *Let  $\mathcal{P}$  be an EF with minimal canonical representation (4.1). Moreover, let  $\Delta \subset \mathbb{R}^k$ ,  $\Psi : \Delta \rightarrow \Xi$  be a bijective mapping, and  $\tilde{P}_\eta = P_{\Psi(\eta)}^*$ ,  $\eta \in \Delta$ , denote the corresponding parametrization of  $\mathcal{P}$  with  $\mu$ -density  $\tilde{f}_\eta = f_{\Psi(\eta)}^*$  of  $\tilde{P}_\eta$  for  $\eta \in \Delta$ . Let  $x$  be a realization of the random variable  $X$  with distribution in  $\mathcal{P}$ , and let  $\check{\zeta}$  denote an ML estimate of  $\zeta$  in  $\Xi$  based on  $x$ . Then,  $\check{\eta} = \Psi^{-1}(\check{\zeta})$  is an ML estimate of  $\eta$  in  $\Delta$  based on  $x$ . Moreover, if  $\check{\zeta}$  is uniquely determined, so is  $\check{\eta}$ .*

**Proof** Since  $\Psi$  is bijective, we have

$$\begin{aligned} \tilde{f}_{\Psi^{-1}(\check{\zeta})}(x) \geq \tilde{f}_{\check{\eta}}(x) \quad \text{for all } \eta \in \Delta & \Leftrightarrow f_{\check{\zeta}}^*(x) \geq f_{\Psi(\eta)}^*(x) \quad \text{for all } \eta \in \Delta \\ & \Leftrightarrow f_{\check{\zeta}}^*(x) \geq f_{\check{\zeta}}^*(x) \quad \text{for all } \check{\zeta} \in \Xi, \end{aligned}$$

which is true by assumption. Hence,  $\check{\eta} = \Psi^{-1}(\check{\zeta})$  is an ML estimate of  $\eta$  in  $\Delta$  based on  $x$ . If  $\check{\zeta}$  is uniquely determined, then the above equivalences hold true with strict inequalities when replacing  $\Delta$  by  $\Delta \setminus \{\check{\eta}\}$  and  $\Xi$  by  $\Xi \setminus \{\check{\zeta}\}$ , and  $\check{\eta}$  is uniquely determined as well.  $\square$

Simply put, Lemma 4.17 states that an ML estimate for a (bijectively) transformed parameter can be obtained by transforming an ML estimate for the original parameter in the same way. This property holds true in parametric models, in general.

From Lemma 4.17, respective statements for the iid case and for the existence and uniqueness of the MLE of  $\eta$  can readily be deduced. As an important example for minimal regular EFs, these results apply to the mean value parametrization  $\tilde{P}_m$ ,  $m \in \text{int}(M)$ , by setting  $\Psi = \pi^{-1}$  (see Theorem 2.56). For instance, provided that the MLE  $\hat{\zeta}^{(n)}$  of  $\zeta$  in  $\Xi^*$  based on an iid sample  $X^{(1)}, \dots, X^{(n)}$  from  $P_\zeta^*$  exists, the MLE of  $m$  in  $\text{int}(M)$  based on  $X^{(1)}, \dots, X^{(n)}$  is given by

$$\hat{m}^{(n)} = \pi(\hat{\zeta}^{(n)}) = \frac{1}{n} \sum_{i=1}^n T(X^{(i)}),$$

where we still use the representation of  $\hat{\zeta}^{(n)}$  in Definition 4.6.

### 4.3 Efficient Estimation

We now turn to properties of point estimators and their connections to EFs. First of all, unbiasedness of an estimator means that, in expectation, we estimate correctly.

**Definition 4.18** Let  $\gamma : \Xi \rightarrow \mathbb{R}^q$  be a function. An estimator  $\hat{\gamma} : (\Omega, \mathcal{A}, \mathbb{P}_\zeta) \rightarrow (\mathbb{R}^q, \mathbb{B}^q)$  is called *unbiased for  $\gamma$*  if

$$\mathbb{E}_\zeta[\hat{\gamma}] = \gamma(\zeta) \quad \text{for every } \zeta \in \Xi.$$

In particular, an estimator  $\hat{\zeta}$  is called *unbiased for  $\zeta$*  if

$$\mathbb{E}_\zeta[\hat{\zeta}] = \zeta \quad \text{for every } \zeta \in \Xi.$$

*Remark 4.19* If under the assumptions of Corollary 4.5 the MLE  $\hat{\zeta}^{(n)}$  in formula (4.3) exists with probability 1, it is a measurable function of the sufficient and complete statistic  $\mathbf{T}^{(n)}$  (see Corollary 3.21 and Theorem 3.13). For the one-parameter EF, the Lehmann-Scheffé theorem then yields that any estimator  $c\hat{\zeta}^{(n)}$  being unbiased for  $\zeta$  for some constant  $c \in \mathbb{R}$  has minimum variance among all unbiased estimators for  $\zeta$ . In that case,  $c\hat{\zeta}^{(n)}$  is called a *uniformly minimum variance unbiased estimator* (UMVUE) for  $\zeta$ .

Estimators with such a minimality property may also be defined in the multivariate case. The *Cramér-Rao inequality* provides a lower bound for the covariance matrix of any unbiased estimator.

**Theorem 4.20** Let  $\mathcal{P}$  be an EF with minimal canonical representation (4.1),  $\Xi \subset \Xi^*$  be open, and  $\gamma : \Xi \rightarrow \mathbb{R}^q$  be a differentiable function. Moreover, let  $X$  be a random variable with distribution  $P_\zeta^*$  for some  $\zeta \in \Xi$ . Then, for every unbiased estimator  $\hat{\gamma}$  for  $\gamma$  based on  $X$ , the Cramér-Rao inequality

$$\mathbb{C}\text{ov}_\zeta(\hat{\gamma}) \geq \mathbf{D}_\gamma(\zeta) [\mathbf{I}(\zeta)]^{-1} \mathbf{D}_\gamma(\zeta)^t \quad (4.8)$$

holds true in the sense of the Löwner order for matrices defined by  $\mathbf{A} \geq \mathbf{B}$  if and only if  $\mathbf{A} - \mathbf{B} \geq 0$ , i.e.,  $\mathbf{A} - \mathbf{B}$  is positive semidefinite. In particular, any unbiased estimator  $\hat{\zeta}$  for  $\zeta$  based on  $X$  satisfies that

$$\mathbb{C}\text{ov}_\zeta(\hat{\zeta}) \geq [\mathbf{I}(\zeta)]^{-1}.$$

Here,  $\mathbf{D}_\gamma(\zeta) \in \mathbb{R}^{q \times k}$  denotes the Jacobian matrix of  $\gamma$  and  $\mathbf{I}(\zeta)$  the Fisher information matrix of  $\mathcal{P}$  at  $\zeta$ .

**Proof** Let  $\zeta \in \Xi$ . Note that then  $\zeta \in \text{int}(\Xi^*)$ , since  $\Xi$  is open by assumption. Recall that by Theorems 3.29, 2.23, and Lemma 2.26

$$U_\zeta = \mathbf{T} - E_\zeta[\mathbf{T}] \quad \text{and} \quad \mathbf{I}(\zeta) = \text{Cov}_\zeta(\mathbf{T}) > 0.$$



In particular, we have

$$E_{\zeta}[U_{\zeta}] = \mathbf{0} \quad \text{and} \quad \text{Cov}_{\zeta}(U_{\zeta}) = \mathbf{I}(\zeta). \quad (4.9)$$

Since  $\hat{\gamma} = (\hat{\gamma}_1, \dots, \hat{\gamma}_q)^t$  is unbiased for  $\gamma = (\gamma_1, \dots, \gamma_q)^t$ , Corollary 2.55 then yields for  $1 \leq j \leq q$  that

$$\begin{aligned} \nabla \gamma_j(\zeta) &= \nabla \mathbb{E}_{\zeta}[\hat{\gamma}_j] = \mathbb{C}\text{ov}_{\zeta}(\hat{\gamma}_j, \mathbf{T}(X)) \\ &= \mathbb{C}\text{ov}_{\zeta}(\hat{\gamma}_j, U_{\zeta}(X)) = \mathbb{E}_{\zeta}[\hat{\gamma}_j U_{\zeta}(X)], \end{aligned}$$

i.e.,

$$\mathbf{D}_{\gamma}(\zeta) = \mathbb{E}_{\zeta}[\hat{\gamma} U_{\zeta}(X)^t].$$

Now, for  $\mathbf{y} \in \mathbb{R}^q$ , it follows by inserting and by using formula (4.9) that

$$\begin{aligned} \mathbf{y}^t \mathbf{D}_{\gamma}(\zeta) [\mathbf{I}(\zeta)]^{-1} \mathbf{D}_{\gamma}(\zeta)^t \mathbf{y} &= \mathbb{E}_{\zeta}[\mathbf{y}^t \hat{\gamma} U_{\zeta}(X)^t [\mathbf{I}(\zeta)]^{-1} \mathbf{D}_{\gamma}(\zeta)^t \mathbf{y}] \\ &= \mathbb{E}_{\zeta}[\mathbf{y}^t (\hat{\gamma} - \gamma(\zeta)) [U_{\zeta}(X)^t [\mathbf{I}(\zeta)]^{-1} \mathbf{D}_{\gamma}(\zeta)^t \mathbf{y}]] \\ &= \mathbb{C}\text{ov}_{\zeta}(\mathbf{y}^t (\hat{\gamma} - \gamma(\zeta)), U_{\zeta}(X)^t [\mathbf{I}(\zeta)]^{-1} \mathbf{D}_{\gamma}(\zeta)^t \mathbf{y}). \end{aligned}$$

From the unbiasedness of  $\hat{\gamma}$  and formula (4.9), the last expression is seen to be the covariance of two random variables with mean 0. Applying the Cauchy-Schwarz inequality then leads to

$$\begin{aligned} &[\mathbf{y}^t \mathbf{D}_{\gamma}(\zeta) [\mathbf{I}(\zeta)]^{-1} \mathbf{D}_{\gamma}(\zeta)^t \mathbf{y}]^2 \\ &\leq \mathbb{V}\text{ar}_{\zeta}(\mathbf{y}^t (\hat{\gamma} - \gamma(\zeta))) \mathbb{V}\text{ar}_{\zeta}(U_{\zeta}(X)^t [\mathbf{I}(\zeta)]^{-1} \mathbf{D}_{\gamma}(\zeta)^t \mathbf{y}) \\ &= [\mathbf{y}^t \mathbb{C}\text{ov}_{\zeta}(\hat{\gamma}) \mathbf{y}] [\mathbf{y}^t \mathbf{D}_{\gamma}(\zeta) [\mathbf{I}(\zeta)]^{-1} \mathbb{C}\text{ov}_{\zeta}(U_{\zeta}(X)) [\mathbf{I}(\zeta)]^{-1} \mathbf{D}_{\gamma}(\zeta)^t \mathbf{y}] \\ &= [\mathbf{y}^t \mathbb{C}\text{ov}_{\zeta}(\hat{\gamma}) \mathbf{y}] [\mathbf{y}^t \mathbf{D}_{\gamma}(\zeta) [\mathbf{I}(\zeta)]^{-1} \mathbf{D}_{\gamma}(\zeta)^t \mathbf{y}]. \end{aligned}$$

Hence,

$$\mathbf{y}^t \mathbf{D}_{\gamma}(\zeta) [\mathbf{I}(\zeta)]^{-1} \mathbf{D}_{\gamma}(\zeta)^t \mathbf{y} \leq \mathbf{y}^t \mathbb{C}\text{ov}_{\zeta}(\hat{\gamma}) \mathbf{y},$$

which yields the assertion.  $\square$

Note that in the proof of Theorem 4.20 we use properties of the EF but not the explicit structure of the densities. Under respective assumptions such as those in formula (4.9), the Cramér-Rao inequality can be shown to hold true in more general models.

Efficient estimators are now introduced as follows, where we implicitly assume that the Cramér-Rao inequality is valid.

**Definition 4.21** Let  $\Xi \subset \Xi^*$  be open,  $\gamma : \Xi \rightarrow \mathbb{R}^q$  be a differentiable function, and  $\hat{\gamma} : (\Omega, \mathcal{A}, \mathbb{P}_\xi) \rightarrow (\mathbb{R}^q, \mathbb{B}^q)$  be an unbiased estimator for  $\gamma$ . Then,  $\hat{\gamma}$  is called *efficient for  $\gamma$*  if its covariance matrix attains the lower bound of the Cramér-Rao inequality in formula (4.8) for every  $\xi \in \Xi$ , i.e.,

$$\mathbb{C}_{\text{ov}}_\xi(\hat{\gamma}) = \mathbf{D}_\gamma(\xi) [\mathbf{I}(\xi)]^{-1} \mathbf{D}_\gamma(\xi)^t, \quad \xi \in \Xi.$$

In particular, any unbiased estimator  $\hat{\xi}$  for  $\xi$  with  $\mathbb{C}_{\text{ov}}_\xi(\hat{\xi}) = [\mathbf{I}(\xi)]^{-1}$ ,  $\xi \in \Xi$ , is called *efficient for  $\xi$* .

An efficient estimator in the sense of Definition 4.21 simultaneously solves different optimization problems.

**Remark 4.22** Let  $\mathbf{A} = [a_{ij}]_{1 \leq i, j \leq k}$ ,  $\mathbf{B} = [b_{ij}]_{1 \leq i, j \leq k} \in \mathbb{R}^{k \times k}$  be two matrices with  $\mathbf{A}, \mathbf{B} \geq \mathbf{0}$  and  $\mathbf{A} \geq \mathbf{B}$  in the sense of the Löwner order. Then,

- (a)  $a_{ii} \geq b_{ii}$ ,  $1 \leq i \leq k$ ,
- (b)  $\text{trace}(\mathbf{A}) \geq \text{trace}(\mathbf{B})$ ,
- (c)  $\det(\mathbf{A}) \geq \det(\mathbf{B})$ .

Note that (a) is obvious from  $\mathbf{e}^t \mathbf{A} \mathbf{e} \geq \mathbf{e}^t \mathbf{B} \mathbf{e}$  for any unit vector  $\mathbf{e} \in \mathbb{R}^k$ , and (b) directly follows from (a); assertion (c) can be obtained from Weyl's Monotonicity Theorem stating that the  $j$ -th largest eigenvalue of  $\mathbf{A}$  is not smaller than that of  $\mathbf{B}$  for  $1 \leq j \leq k$  if  $\mathbf{A}, \mathbf{B} \geq \mathbf{0}$  and  $\mathbf{A} - \mathbf{B} \geq \mathbf{0}$  (see [15, p. 63]). An efficient estimator therefore minimizes the marginal variances (a), the *total variation* (b), and the *generalized variance* (c) among all unbiased estimators. Here, the total variation and generalized variance of a random vector are defined as the trace and determinant of its covariance matrix, respectively (see, e.g., [45, p. 13 and pp. 30/31]).

**Remark 4.23** For  $k = 1$ , an efficient estimator is always a UMVUE. The reverse implication is not true in general, since the lower bound of the Cramér-Rao inequality need not to be attained by the UMVUE. Indeed, it can be shown that efficient estimators essentially only exist in EFs (see, e.g., [16, pp. 182/183]). Analogously, for  $k > 1$ , efficiency of an estimator ensures that it has minimum covariance matrix in the sense of the Löwner order among all unbiased estimators, but not vice versa.

**Examples:**  $\mathcal{G}(\bullet, \bullet)$ ,  $\mathcal{N}_k(\bullet, \Sigma_0)$

*Example 4.24* Let  $\mathcal{P} = \mathcal{G}(\bullet, \bullet)$  and  $X_1, \dots, X_n$  be an iid sample from  $P_\zeta^* \in \mathcal{P}$  (see Example 2.40). Evidently,  $\sum_{i=1}^n X_i/n$  is an unbiased estimator for the mean  $\mathbb{E}_\zeta[X_1]$ , which is given by  $\pi_1(\zeta) = -\zeta_2/\zeta_1$  in the natural parametrization as shown in Example 2.53. From Example 3.34, the Fisher information matrix of  $\mathcal{P}^{(n)}$  is seen to be

$$n\mathbf{I}(\zeta) = n \begin{pmatrix} \frac{\zeta_2}{\zeta_1^2} & -\frac{1}{\zeta_1} \\ -\frac{1}{\zeta_1} & \psi_1(\zeta_2) \end{pmatrix}, \quad \zeta = (\zeta_1, \zeta_2)^t \in (-\infty, 0) \times (0, \infty).$$

Hence, by setting  $\gamma = \pi_1$ , the lower bound of the Cramér-Rao inequality in formula (4.8) is obtained as

$$\begin{aligned} \left( \frac{\zeta_2}{\zeta_1^2}, -\frac{1}{\zeta_1} \right) \frac{1}{n \det(\mathbf{I}(\zeta))} \begin{pmatrix} \psi_1(\zeta_2) & \frac{1}{\zeta_1} \\ \frac{1}{\zeta_1} & \frac{\zeta_2}{\zeta_1^2} \end{pmatrix} \begin{pmatrix} \frac{\zeta_2}{\zeta_1^2} \\ -\frac{1}{\zeta_1} \end{pmatrix} \\ = \frac{\zeta_1^2}{n(\zeta_2 \Psi_1(\zeta_2) - 1)} \left( \frac{\zeta_2}{\zeta_1^2}, -\frac{1}{\zeta_1} \right) \begin{pmatrix} \frac{\zeta_2 \Psi_1(\zeta_2) - 1}{\zeta_1^2} \\ 0 \end{pmatrix} = \frac{\zeta_2}{n \zeta_1^2} \end{aligned}$$

for  $\zeta = (\zeta_1, \zeta_2)^t \in (-\infty, 0) \times (0, \infty)$ . From Example 2.53, we have  $\mathbb{V}_{\text{ar}_\zeta}(X_1) = \zeta_2/\zeta_1^2$  such that the variance of  $\sum_{i=1}^n X_i/n$  attains the lower bound of the Cramér-Rao inequality. Thus,  $\sum_{i=1}^n X_i/n$  is an efficient estimator (and a UMVUE) for  $\pi_1$ .

*Example 4.25* Let  $\mathcal{P} = \mathcal{N}_k(\bullet, \Sigma_0)$  and  $\mathbf{X}^{(1)}, \dots, \mathbf{X}^{(n)}$  be an iid sample from  $P_\mu \in \mathcal{P}$ . By Example 3.35, the Fisher information matrix of  $\mathcal{P}^{(n)}$  is given by  $n\Sigma_0^{-1}$  at every  $\mu \in \mathbb{R}^k$ . Accordingly, by Theorem 4.20, the covariance matrix of any unbiased estimator for  $\mu$  is bounded from below by  $\Sigma_0/n$ . This is the covariance matrix of the MLE  $\hat{\mu}^{(n)} = \sum_{i=1}^n \mathbf{X}^{(i)}/n$  derived in Example 4.10, which is unbiased and therefore also efficient for  $\mu$ .

Example 4.24 provides an efficient estimator for the mean of a gamma distribution. This finding is a particular case of a general result for EFs.

**Theorem 4.26** *Let  $\mathcal{P}$  be an EF with minimal canonical representation (4.1),  $\Xi \subset \Xi^*$  be open, and  $X$  be a random variable with distribution in  $\mathcal{P}$ . Then,  $\mathbf{T}(X)$  is an*

efficient estimator for  $\pi(\zeta)$ ,  $\zeta \in \Xi$ , where  $\pi$  is the mean value function as defined in Theorem 2.56.

**Proof** Trivially,  $T(X)$  is an unbiased estimator for  $\pi(\zeta) = \mathbb{E}_\zeta[T(X)]$ ,  $\zeta \in \Xi$ . Moreover, from Theorems 2.56 and 3.29, we obtain that

$$\begin{aligned} \mathbf{D}_\pi(\zeta) [\mathbf{I}(\zeta)]^{-1} \mathbf{D}_\pi(\zeta)^t &= \mathbf{Cov}_\zeta(T) [\mathbf{Cov}_\zeta(T)]^{-1} \mathbf{Cov}_\zeta(T) \\ &= \mathbb{C}_{\text{ov}_\zeta}(T(X)), \quad \zeta \in \Xi. \end{aligned}$$

□

Based on an iid sample, we arrive at the following version of Theorem 4.26.

**Corollary 4.27** Let  $\mathcal{P}$  be an EF with minimal canonical representation (4.1),  $\Xi \subset \Xi^*$  be open, and  $\pi$  be the corresponding mean value function as defined in Theorem 2.56. Moreover, let  $X^{(1)}, \dots, X^{(n)}$  be an iid sample from a distribution in  $\mathcal{P}$ . Then,  $\sum_{i=1}^n T(X^{(i)})/n$  is an efficient estimator for  $\pi(\zeta)$ ,  $\zeta \in \Xi$ .

**Proof** Since  $n\mathbf{I}(\zeta)$  is the Fisher information matrix of the product EF  $\mathcal{P}^{(n)}$  at  $\zeta$ , the lower bound of the Cramér-Rao inequality is here

$$\begin{aligned} \mathbf{D}_\pi(\zeta) [n\mathbf{I}(\zeta)]^{-1} \mathbf{D}_\pi(\zeta)^t &= \frac{\mathbb{C}_{\text{ov}_\zeta}(T(X^{(1)}))}{n} \\ &= \mathbb{C}_{\text{ov}_\zeta}\left(\frac{1}{n} \sum_{i=1}^n T(X^{(i)})\right), \quad \zeta \in \Xi, \end{aligned}$$

by using the same arguments as in the proof of Theorem 4.26 and that  $X^{(1)}, \dots, X^{(n)}$  are iid. Evidently,  $\sum_{i=1}^n T(X^{(i)})/n$  is unbiased for  $\pi(\zeta)$ ,  $\zeta \in \Xi$ , and the assertion follows. □

In particular, Theorem 4.26 and Corollary 4.27 apply to (minimal) regular EFs, in which the mean value parameter can therefore always be estimated efficiently.

**Examples:**  $\text{poi}(\bullet)$ ,  $\text{b}(n_0, \bullet)$ ,  $\text{nb}(r_0, \bullet)$ ,  $\mathcal{E}(\bullet)$

*Example 4.28* Let  $\mathcal{P}$  be any of the EFs  $\text{poi}(\bullet)$ ,  $\text{b}(n_0, \bullet)$ ,  $\text{nb}(r_0, \bullet)$ , or  $\mathcal{E}(\bullet)$  in Examples 2.37, 2.38, 2.51, and 2.12. Moreover, let  $X_1, \dots, X_n$  be an iid sample from  $P_\zeta^* \in \mathcal{P}$ . Then, by Corollary 4.27,  $\sum_{i=1}^n X_i/n$  is an efficient estimator (and thus a UMVUE) for the mean of  $P_\zeta^*$ .

## 4.4 Asymptotic Properties

In the examples of Sect. 4.1, the probability that the ML estimate of  $\zeta$  in  $\Xi^*$  based on  $n$  observations exists is equal to 1 or at least converges to 1 when the sample size tends to infinity. This finding holds true in general for minimal regular EFs. It is a consequence of the following theorem.

**Theorem 4.29** *Let  $\mathcal{P}$  be an EF with minimal canonical representation (4.1),  $\Xi \subset \Xi^*$  be open, and  $X^{(1)}, X^{(2)}, \dots$  be a sequence of iid random variables with distribution  $P_\zeta^* \in \mathcal{P}$  for some  $\zeta \in \Xi$ . Then,*

$$\lim_{n \rightarrow \infty} \mathbb{P}_\zeta \left( \frac{1}{n} \sum_{i=1}^n T(X^{(i)}) \in \pi(\Xi) \right) = 1,$$

i.e., the probability that the ML estimate of  $\zeta$  in  $\Xi$  based on realizations  $x^{(1)}, \dots, x^{(n)}$  of  $X^{(1)}, \dots, X^{(n)}$  exists converges to 1 for  $n \rightarrow \infty$ .

**Proof** Let  $\zeta \in \Xi$ . By Theorem 2.49, all moments of  $T$  with respect to  $P_\zeta^*$  are finite. The strong law of large numbers together with the definition of  $\pi$  in Theorem 2.56 then yields that

$$\frac{1}{n} \sum_{i=1}^n T(X^{(i)}) \xrightarrow{n \rightarrow \infty} \mathbb{E}_\zeta[T(X^{(1)})] = \pi(\zeta) \quad \mathbb{P}_\zeta\text{-a.s.},$$

which implies that  $\sum_{i=1}^n T(X^{(i)})/n \rightarrow \pi(\zeta)$  in  $\mathbb{P}_\zeta$ -probability (see, e.g., [55, p. 51]). Moreover, by Theorem 2.56,  $\pi^{-1}$  is continuous, and we conclude that  $\pi(\Xi)$  is open as the pre-image  $(\pi^{-1})^{-1}(\Xi) = \{\mathbf{m} \in \pi(\text{int}(\Xi^*)) : \pi^{-1}(\mathbf{m}) \in \Xi\}$  of the open set  $\Xi$  under  $\pi^{-1}$ . Hence, there exist some  $\varepsilon > 0$  and an Euclidean ball  $B_\varepsilon(\pi(\zeta)) = \{\boldsymbol{\eta} \in \mathbb{R}^k : \|\boldsymbol{\eta} - \pi(\zeta)\|_2 < \varepsilon\} \subset \pi(\Xi)$ . Now, it follows that

$$\begin{aligned} \mathbb{P}_\zeta \left( \frac{1}{n} \sum_{i=1}^n T(X^{(i)}) \in \pi(\Xi) \right) &\geq \mathbb{P}_\zeta \left( \frac{1}{n} \sum_{i=1}^n T(X^{(i)}) \in B_\varepsilon(\pi(\zeta)) \right) \\ &= 1 - \mathbb{P}_\zeta \left( \left\| \frac{1}{n} \sum_{i=1}^n T(X^{(i)}) - \pi(\zeta) \right\|_2 \geq \varepsilon \right) \\ &\xrightarrow{n \rightarrow \infty} 1. \end{aligned}$$

Finally, since  $\Xi$  is open, Lemma 4.2 yields that an ML estimate of  $\zeta$  in  $\Xi$  based on realizations  $x^{(1)}, \dots, x^{(n)}$  of  $X^{(1)}, \dots, X^{(n)}$  exists (and is then uniquely determined) if and only if the corresponding likelihood equation  $\pi(\cdot) = \sum_{i=1}^n T(x^{(i)})/n$  has a solution in  $\Xi$ . This completes the proof.  $\square$

In Sect. 4.3, we have shown that the parameter  $\mathbf{m} = \pi(\boldsymbol{\zeta})$  of the mean value parametrization can be estimated efficiently in minimal regular EFs. Based on a single random variable  $X \sim \tilde{P}_{\mathbf{m}} = P_{\pi^{-1}(\mathbf{m})}^*$ , for example,  $\mathbf{T}(X)$  was proven to form an efficient estimator for  $\mathbf{m}$ . In contrast, provided that it exists in  $\Xi^*$  with probability 1, the MLE  $\pi^{-1}(\mathbf{T}(X))$  of the natural parameter does not have to be efficient, not even unbiased for  $\boldsymbol{\zeta}$  in general. Asymptotically, however, it possesses prominent properties, which are introduced in what follows.

First, the consistency of an estimator, more precisely of a sequence of estimators, is the property that the estimator approaches the true parameter when the sample size increases.

**Definition 4.30** Let  $\gamma : \Xi \rightarrow \mathbb{R}^q$  be a function and  $\hat{\gamma}^{(n)} : (\Omega, \mathcal{A}, \mathbb{P}_{\boldsymbol{\zeta}}) \rightarrow (\mathbb{R}^q, \mathbb{B}^q)$  be an estimator of  $\gamma$  for every  $n \in \mathbb{N}$ . Then, the sequence  $\hat{\gamma}^{(n)}$ ,  $n \in \mathbb{N}$ , is called *(strongly) consistent for  $\gamma$*  if

$$\hat{\gamma}^{(n)} \xrightarrow{n \rightarrow \infty} \gamma(\boldsymbol{\zeta})$$

in  $\mathbb{P}_{\boldsymbol{\zeta}}$ -probability ( $\mathbb{P}_{\boldsymbol{\zeta}}$ -a.s.) for every  $\boldsymbol{\zeta} \in \Xi$ .

In particular, a sequence  $\hat{\boldsymbol{\zeta}}^{(n)}$ ,  $n \in \mathbb{N}$ , of estimators is called *(strongly) consistent for  $\boldsymbol{\zeta}$*  if

$$\hat{\boldsymbol{\zeta}}^{(n)} \xrightarrow{n \rightarrow \infty} \boldsymbol{\zeta}$$

in  $\mathbb{P}_{\boldsymbol{\zeta}}$ -probability ( $\mathbb{P}_{\boldsymbol{\zeta}}$ -a.s.) for every  $\boldsymbol{\zeta} \in \Xi$ .

Note that strong consistency implies consistency by definition.

**Theorem 4.31** Let  $\mathcal{P}$  be a regular EF with minimal canonical representation (4.1) and  $X^{(1)}, X^{(2)}, \dots$  be a sequence of iid random variables with distribution  $P_{\boldsymbol{\zeta}}^* \in \mathcal{P}$ . Moreover, let the MLE  $\hat{\boldsymbol{\zeta}}^{(n)}$  of  $\boldsymbol{\zeta}$  in  $\Xi^*$  based on  $X^{(1)}, \dots, X^{(n)}$  exist for every  $n \in \mathbb{N}$ . Then, the sequence  $\hat{\boldsymbol{\zeta}}^{(n)}$ ,  $n \in \mathbb{N}$ , is strongly consistent for  $\boldsymbol{\zeta}$ .

**Proof** Let  $\boldsymbol{\zeta} \in \Xi^*$ . As in the proof of Theorem 4.29, we have

$$\frac{1}{n} \sum_{i=1}^n \mathbf{T}(X^{(i)}) \longrightarrow \pi(\boldsymbol{\zeta}) \quad \mathbb{P}_{\boldsymbol{\zeta}}\text{-a.s.}$$

Since  $\pi^{-1}$  is continuous by Theorem 2.56, we obtain from formula (4.3) that

$$\hat{\xi}^{(n)} = \pi^{-1} \left( \frac{1}{n} \sum_{i=1}^n T(X^{(i)}) \right) \xrightarrow{n \rightarrow \infty} \pi^{-1}(\pi(\xi)) = \xi \quad \mathbb{P}_\xi\text{-a.s.}$$

□

Another main concept for sequences of estimators is asymptotic efficiency, which is motivated by Definition 4.21.

**Definition 4.32** Let  $\Xi \subset \Xi^*$  be open,  $\gamma : \Xi \rightarrow \mathbb{R}^q$  be a differentiable function with Jacobian matrix  $\mathbf{D}_\gamma(\xi) \in \mathbb{R}^{q \times k}$  of rank  $q$  at every  $\xi \in \Xi$ , and  $\hat{\gamma}^{(n)} : (\Omega, \mathcal{A}, \mathbb{P}_\xi) \rightarrow (\mathbb{R}^q, \mathbb{B}^q)$  be an estimator of  $\gamma$  for every  $n \in \mathbb{N}$ . Then, the sequence  $\hat{\gamma}^{(n)}$ ,  $n \in \mathbb{N}$ , is called *asymptotically efficient* for  $\gamma$  if

$$\sqrt{n}(\hat{\gamma}^{(n)} - \gamma(\xi)) \xrightarrow{d} \mathcal{N}_q(\mathbf{0}, \mathbf{D}_\gamma(\xi)[\mathbf{I}(\xi)]^{-1}\mathbf{D}_\gamma(\xi)^t) \quad \text{for } \xi \in \Xi,$$

i.e.,  $\sqrt{n}(\hat{\gamma}^{(n)} - \gamma(\xi))$  is asymptotically multivariate normal distributed with mean vector  $\mathbf{0}$  and covariance matrix  $\mathbf{D}_\gamma(\xi)[\mathbf{I}(\xi)]^{-1}\mathbf{D}_\gamma(\xi)^t$ .

In particular, a sequence  $\hat{\xi}^{(n)}$ ,  $n \in \mathbb{N}$ , of estimators is called *asymptotically efficient* for  $\xi$  if

$$\sqrt{n}(\hat{\xi}^{(n)} - \xi) \xrightarrow{d} \mathcal{N}_k(\mathbf{0}, [\mathbf{I}(\xi)]^{-1}) \quad \text{for } \xi \in \Xi.$$

Here,  $\mathbf{I}(\xi)$  denotes the Fisher information matrix of  $\mathcal{P}$  at  $\xi$ , which is assumed to be regular for all  $\xi \in \Xi$ .

**Remark 4.33** We shall address some motivational aspects for Definition 4.32 (see also [39, pp. 437–443]). Suppose that  $\hat{\xi}^{(n)}$ ,  $n \in \mathbb{N}$ , is asymptotically efficient for  $\xi$  according to Definition 4.32. If we further assume that the covariance matrix of  $\sqrt{n}(\hat{\xi}^{(n)} - \xi)$  converges to that of the asymptotic distribution for  $n \rightarrow \infty$ , this yields for sufficiently large  $n$  that

$$[\mathbf{I}(\xi)]^{-1} \approx \mathbb{C}_{\mathbb{O}\mathbb{W}_\xi}(\sqrt{n}(\hat{\xi}^{(n)} - \xi)) = n \mathbb{C}_{\mathbb{O}\mathbb{W}_\xi}(\hat{\xi}^{(n)}), \quad \xi \in \Xi.$$

That is, the covariance matrix of  $\hat{\xi}^{(n)}$  approximately attains the lower bound  $[\mathbf{I}(\xi)]^{-1}/n$  of the Cramér-Rao inequality for  $\mathcal{P}^{(n)}$  related to an iid sample of size  $n$  (see Sect. 4.3). Note, however, that  $\hat{\xi}^{(n)}$  is not necessarily unbiased for  $\xi$ .

In a minimal regular EF, it can moreover be shown that for every sequence  $\tilde{\xi}^{(n)}$ ,  $n \in \mathbb{N}$ , of estimators for  $\xi$  with

$$\sqrt{n}(\tilde{\xi}^{(n)} - \xi) \xrightarrow{d} \mathcal{N}_k(\mathbf{0}, \Sigma(\xi)) \quad \text{for every } \xi \in \Xi^*,$$

we necessarily have

$$\Sigma(\xi) \geq [\mathbf{I}(\xi)]^{-1} \quad \lambda_{\Xi^*}\text{-a.e.}$$

in the sense of the Löwner order (see [39, pp. 437–443] and the references cited therein). If  $\Sigma$  and  $\mathbf{I}^{-1}$  are continuous on  $\Xi^*$ , then the inequality holds even everywhere on  $\Xi^*$ . Hence, an asymptotically efficient sequence of estimators has minimum asymptotic covariance matrix within a certain class of (sequences of) estimators.

Note that asymptotic efficiency implies consistency. This can be seen by rewriting

$$\hat{\xi}^{(n)} - \xi = \frac{1}{\sqrt{n}} \sqrt{n}(\hat{\xi}^{(n)} - \xi),$$

which converges to  $\mathbf{0}$  in distribution by applying the multivariate Slutsky theorem (see, e.g., [53, p. 130]). Since the limit is a constant, this convergence holds true even in  $\mathbb{P}_\xi$ -probability (see, for instance, [55, p. 51]).

**Theorem 4.34** *Under the assumptions of Theorem 4.31, the sequence  $\hat{\xi}^{(n)}$ ,  $n \in \mathbb{N}$ , is asymptotically efficient for  $\xi$ .*

**Proof** Let  $\xi \in \Xi^*$ . First, recall that by Theorems 2.56 and 3.29, it holds that

$$\mathbb{E}_\xi[\mathbf{T}(X^{(1)})] = \pi(\xi), \quad \mathbb{C}_{\mathbb{D}\mathbb{W}_\xi}(\mathbf{T}(X^{(1)})) = \mathbf{I}(\xi).$$

The multivariate central limit theorem (see, e.g., [55, Section 1.5]) then yields that

$$\sqrt{n} \left( \frac{1}{n} \sum_{i=1}^n \mathbf{T}(X^{(i)}) - \pi(\xi) \right) \xrightarrow{d} \mathbf{N} \quad (4.10)$$

for a random vector  $\mathbf{N} \sim \mathcal{N}_k(\mathbf{0}, \mathbf{I}(\xi))$ .

Now, to ease notation, let  $\rho = (\rho_1, \dots, \rho_k)$  denote the inverse function of  $\pi$ , i.e., we have  $\rho = \pi^{-1}$ . Then, formula (4.3) along with the mean value theorem for real-valued functions of several variables (e.g., in [52, p. 174]) applied to every of



the mappings  $\rho_j$ ,  $1 \leq j \leq k$ , yields

$$\begin{aligned} \sqrt{n}(\hat{\xi}^{(n)} - \xi) &= \sqrt{n} \left( \rho \left( \frac{1}{n} \sum_{i=1}^n T(X^{(i)}) \right) - \rho(\pi(\xi)) \right) \\ &= \mathbf{G}_n \sqrt{n} \left( \frac{1}{n} \sum_{i=1}^n T(X^{(i)}) - \pi(\xi) \right), \end{aligned}$$

where  $\mathbf{G}_n$  is a random matrix with rows

$$[\nabla \rho_j(V_j^{(n)})]^t, \quad 1 \leq j \leq k,$$

and the random vectors  $V_1^{(n)}, \dots, V_k^{(n)}$  are convex combinations of  $\sum_{i=1}^n T(X^{(i)})/n$  and  $\pi(\xi)$  for  $n \in \mathbb{N}$ . For  $n \rightarrow \infty$ , the strong law of large numbers ensures that  $\sum_{i=1}^n T(X^{(i)})/n \rightarrow \pi(\xi) \mathbb{P}_\xi$ -a.s. and with it  $V_j^{(n)} \rightarrow \pi(\xi) \mathbb{P}_\xi$ -a.s. for  $1 \leq j \leq k$ . From Corollary 2.56 and Theorem 3.29, we conclude

$$\mathbf{G}_n \xrightarrow{n \rightarrow \infty} \mathbf{D}_\rho(\pi(\xi)) = [\mathbf{D}_\pi(\xi)]^{-1} = [\mathbf{I}(\xi)]^{-1} \quad \mathbb{P}_\xi\text{-a.s.}$$

By the multivariate Slutsky theorem (see, e.g., [53, p. 130]), it finally follows that

$$\sqrt{n}(\hat{\xi}^{(n)} - \xi) \xrightarrow{d} [\mathbf{I}(\xi)]^{-1} N,$$

and, since  $[\mathbf{I}(\xi)]^{-1} N \sim \mathcal{N}_k(\mathbf{0}, [\mathbf{I}(\xi)]^{-1})$ , the proof is completed.  $\square$

Note that if  $\mathcal{P}$  is not assumed to be regular, Theorems 4.31 and 4.34 remain valid when replacing  $\Xi^*$  by some open set  $\Xi \subset \Xi^*$ .

*Remark 4.35* The assumption in Theorems 4.31 and 4.34 saying that the MLE  $\hat{\xi}^{(n)}$  of  $\xi$  in  $\Xi^*$  based on  $X^{(1)}, \dots, X^{(n)}$  exists for  $n \in \mathbb{N}$  may be dropped in a certain sense. Evidently, we may weaken the assumption from all  $n \in \mathbb{N}$  to eventually all  $n \in \mathbb{N}$ , i.e., for all  $n \geq n_0$  for some  $n_0 \in \mathbb{N}$ , without affecting the findings. If even the latter property is not met, we may proceed as follows to obtain a strongly consistent and asymptotically efficient estimator for  $\xi$ .

Let  $t \in \pi(\Xi^*)$  be arbitrary. Moreover, for  $n \in \mathbb{N}$ , we introduce the set

$$A_n = \left\{ \frac{1}{n} \sum_{i=1}^n T(X^{(i)}) \in \pi(\Xi^*) \right\} \in \mathcal{A}$$

and the estimator

$$\tilde{\xi}^{(n)} = \pi^{-1}(W_n) \quad \text{with} \quad W_n = \left( \frac{1}{n} \sum_{i=1}^n T(X^{(i)}) \right) \mathbb{1}_{A_n} + t \mathbb{1}_{A_n^c}.$$

Now, for  $\xi \in \Xi^*$ , we have that  $\sum_{i=1}^n T(X^{(i)})/n \rightarrow \pi(\xi) \mathbb{P}_\xi$ -a.s. for  $n \rightarrow \infty$  by the strong law of large numbers. Since  $\pi(\Xi^*) = \text{int}(M)$  is open, this implies that  $\mathbb{1}_{A_n} \rightarrow 1 \mathbb{P}_\xi$ -a.s. and  $\mathbb{1}_{A_n^c} \rightarrow 0 \mathbb{P}_\xi$ -a.s. for  $n \rightarrow \infty$ . Hence,

$$W_n \xrightarrow{n \rightarrow \infty} \pi(\xi) \quad \mathbb{P}_\xi\text{-a.s.}, \quad (4.11)$$

and, since  $\pi^{-1}$  is continuous by Theorem 2.56, it follows that  $\tilde{\xi}^{(n)} \rightarrow \xi \mathbb{P}_\xi$ -a.s. for  $n \rightarrow \infty$ , i.e., the sequence  $\tilde{\xi}^{(n)}, n \in \mathbb{N}$ , is strongly consistent for  $\xi$ .

To establish that the sequence  $\tilde{\xi}^{(n)}, n \in \mathbb{N}$ , is also asymptotically efficient for  $\xi$ , note first that for  $n \in \mathbb{N}$

$$W_n = \frac{1}{n} \sum_{i=1}^n T(X^{(i)}) + \left( t - \frac{1}{n} \sum_{i=1}^n T(X^{(i)}) \right) \mathbb{1}_{A_n^c}$$

and thus

$$\sqrt{n}(W_n - \pi(\xi)) = \sqrt{n} \left( \frac{1}{n} \sum_{i=1}^n T(X^{(i)}) - \pi(\xi) \right) + \sqrt{n} \left( t - \frac{1}{n} \sum_{i=1}^n T(X^{(i)}) \right) \mathbb{1}_{A_n^c}.$$

Since  $\mathbb{1}_{A_n^c} \rightarrow 0 \mathbb{P}_\xi$ -a.s. for  $n \rightarrow \infty$ , there exists  $\mathbb{P}_\xi$ -a.s. an index  $m \in \mathbb{N}$  with  $\mathbb{1}_{A_n^c} = 0$  for all  $n \geq m$ . Hence,

$$\sqrt{n} \left( t - \frac{1}{n} \sum_{i=1}^n T(X^{(i)}) \right) \mathbb{1}_{A_n^c} \xrightarrow{n \rightarrow \infty} \mathbf{0} \quad \mathbb{P}_\xi\text{-a.s.}$$

Formula (4.10) together with the multivariate Slutsky theorem then yields that

$$\sqrt{n} (W_n - \pi(\xi)) \xrightarrow{d} \mathcal{N}_k(\mathbf{0}, \mathbf{I}(\xi)). \quad (4.12)$$

As in the proof of Theorem 4.34, it now follows with  $\rho = \pi^{-1}$  that

$$\begin{aligned} \sqrt{n} (\tilde{\xi}^{(n)} - \xi) &= \sqrt{n} (\rho(W_n) - \rho(\pi(\xi))) \\ &= \tilde{\mathbf{G}}_n \sqrt{n} (W_n - \pi(\xi)), \end{aligned}$$

where  $\tilde{\mathbf{G}}_n$  is a random matrix with rows

$$[\nabla \rho_j(\tilde{V}_j^{(n)})]^t, \quad 1 \leq j \leq k,$$

and the random vectors  $\tilde{V}_1^{(n)}, \dots, \tilde{V}_k^{(n)}$  are convex combinations of  $W_n$  and  $\pi(\xi)$  for  $n \in \mathbb{N}$ . From formula (4.11), we obtain for  $n \rightarrow \infty$  that  $\tilde{V}_j^{(n)} \rightarrow \pi(\xi) \mathbb{P}_\xi$ -a.s.

for  $1 \leq j \leq k$  and thus  $\check{\mathbf{G}}_n \rightarrow [\mathbf{I}(\boldsymbol{\xi})]^{-1} \mathbb{P}_{\boldsymbol{\xi}}$ -a.s. as in the proof of Theorem 4.34. The latter finding together with formula (4.12) then yields that

$$\sqrt{n} (\tilde{\boldsymbol{\xi}}^{(n)} - \boldsymbol{\xi}) \xrightarrow{d} \mathcal{N}_k(\mathbf{0}, [\mathbf{I}(\boldsymbol{\xi})]^{-1})$$

by using the multivariate Slutsky theorem, again. Thus, asymptotic efficiency of  $\tilde{\boldsymbol{\xi}}^{(n)}$ ,  $n \in \mathbb{N}$ , is shown.

Asymptotic properties for curved EFs are stated in the following remark.

**Remark 4.36** Let  $\mathcal{P}_{\mathbf{Z}} \subset \mathcal{P}$  with  $\mathbf{Z} : \Theta \rightarrow \Xi^*$  be a curved EF as introduced in Remark 4.14, where we assume that  $\mathbf{Z}(\Theta) \subset \text{int}(\Xi^*)$  and that  $\mathbf{Z}(\Theta)$  is relatively closed in  $\Xi^*$ . Moreover, suppose that the MLE  $\hat{\vartheta}_n$  of  $\vartheta$  in  $\Theta \subset \mathbb{R}^m$  based on a sample of size  $n$  uniquely exists for eventually all  $n \in \mathbb{N}$ . Then the sequence  $\hat{\vartheta}_n$ ,  $n \in \mathbb{N}$ , is strongly consistent provided that  $\mathbf{Z}$  has a continuous inverse  $\mathbf{Z}^{-1} : \mathbf{Z}(\Theta) \rightarrow \Theta$ . If, in addition,  $\Theta$  is open and  $\mathbf{Z}$  is continuously differentiable with Jacobian matrix  $\mathbf{D}_{\mathbf{Z}}(\vartheta)$  of full rank  $m$  at every  $\vartheta \in \Theta$ , then  $\sqrt{n}(\hat{\vartheta}_n - \vartheta)$  converges in distribution to  $\mathcal{N}_m(\mathbf{0}, [\mathbf{D}_{\mathbf{Z}}(\vartheta)^t \mathbf{I}(\mathbf{Z}(\vartheta)) \mathbf{D}_{\mathbf{Z}}(\vartheta)]^{-1})$ . Here,  $\mathbf{I}(\mathbf{Z}(\vartheta))$  denotes the Fisher information matrix of  $\mathcal{P}$  at  $\mathbf{Z}(\vartheta)$ . For a detailed discussion along with proofs of the above results, see, e.g., [14].

Having proposed strongly consistent and asymptotically efficient sequences of estimators for the natural parameter  $\boldsymbol{\xi}$  of the EF, we finally state sufficient conditions under which these properties transfer to a sequence of estimators for  $\gamma(\boldsymbol{\xi})$ .

**Lemma 4.37** *Let  $\mathcal{P}$  be an EF with minimal canonical representation (4.1),  $\Xi \subset \Xi^*$  be open, and  $\gamma : \Xi \rightarrow \mathbb{R}^q$  be a mapping. Moreover, let  $\tilde{\boldsymbol{\xi}}^{(n)}$ ,  $n \in \mathbb{N}$ , be a sequence of estimators for  $\boldsymbol{\xi}$ . Then, the following assertions hold true:*

- (a) *If  $\tilde{\boldsymbol{\xi}}^{(n)}$ ,  $n \in \mathbb{N}$ , is strongly consistent for  $\boldsymbol{\xi}$  and  $\gamma$  is continuous, then the sequence  $\gamma(\tilde{\boldsymbol{\xi}}^{(n)})$ ,  $n \in \mathbb{N}$ , is strongly consistent for  $\gamma$ .*
- (b) *If  $\tilde{\boldsymbol{\xi}}^{(n)}$ ,  $n \in \mathbb{N}$ , is asymptotically efficient for  $\boldsymbol{\xi}$  and  $\gamma$  is continuously differentiable with Jacobian matrix  $\mathbf{D}_{\gamma}(\boldsymbol{\xi}) \in \mathbb{R}^{q \times k}$  of rank  $q$  at every  $\boldsymbol{\xi} \in \Xi$ , then the sequence  $\gamma(\tilde{\boldsymbol{\xi}}^{(n)})$ ,  $n \in \mathbb{N}$ , is asymptotically efficient for  $\gamma$ .*

**Proof** Assertion (a) is clear, since a.s.-convergence is preserved under continuous mappings. To show assertion (b), let  $\gamma$  be continuously differentiable and  $\boldsymbol{\xi} \in \Xi$ . Then, applying the mean value theorem for real-valued functions of several variables (e.g., in [52, p. 174]) to  $\gamma_j$  for  $1 \leq j \leq q$  gives

$$\sqrt{n} (\gamma(\tilde{\boldsymbol{\xi}}^{(n)}) - \gamma(\boldsymbol{\xi})) = \check{\mathbf{G}}_n \sqrt{n} (\tilde{\boldsymbol{\xi}}^{(n)} - \boldsymbol{\xi}),$$

where  $\check{\mathbf{G}}_n$  is a random matrix with rows

$$[\nabla \gamma_j(\check{\mathbf{V}}_j^{(n)})]^t, \quad 1 \leq j \leq q,$$

and the random vectors  $\check{\mathbf{V}}_1^{(n)}, \dots, \check{\mathbf{V}}_q^{(n)}$  are convex combinations of  $\tilde{\boldsymbol{\zeta}}^{(n)}$  and  $\boldsymbol{\zeta}$  for  $n \in \mathbb{N}$ . Since  $\tilde{\boldsymbol{\zeta}}^{(n)}$ ,  $n \in \mathbb{N}$ , is assumed to be asymptotically efficient, it is also consistent for  $\boldsymbol{\zeta}$ , i.e., we have  $\tilde{\boldsymbol{\zeta}}^{(n)} \rightarrow \boldsymbol{\zeta}$  in  $\mathbb{P}_{\boldsymbol{\zeta}}$ -probability for  $n \rightarrow \infty$ . Consequently, for  $1 \leq j \leq q$ , we also have  $\check{\mathbf{V}}_j^{(n)} \rightarrow \boldsymbol{\zeta}$  in  $\mathbb{P}_{\boldsymbol{\zeta}}$ -probability for  $n \rightarrow \infty$ . Moreover, for  $1 \leq j \leq q$ ,  $\nabla \gamma_j$  is continuous by assumption, which implies that

$$\check{\mathbf{G}}_n \xrightarrow{n \rightarrow \infty} \mathbf{D}_{\gamma}(\boldsymbol{\zeta}) \quad \text{in } \mathbb{P}_{\boldsymbol{\zeta}}\text{-probability}$$

by using that convergence in probability is preserved under continuous mappings (see, e.g., [55, p. 59]). Since  $\tilde{\boldsymbol{\zeta}}^{(n)}$ ,  $n \in \mathbb{N}$ , is asymptotically efficient for  $\boldsymbol{\zeta}$ , applying the multivariate Slutsky theorem finally gives

$$\sqrt{n}(\gamma(\tilde{\boldsymbol{\zeta}}^{(n)}) - \gamma(\boldsymbol{\zeta})) \xrightarrow{d} \mathcal{N}_q(\mathbf{0}, \mathbf{D}_{\gamma}(\boldsymbol{\zeta})[\mathbf{I}(\boldsymbol{\zeta})]^{-1}\mathbf{D}_{\gamma}(\boldsymbol{\zeta})^t),$$

i.e.,  $\gamma(\tilde{\boldsymbol{\zeta}}^{(n)})$ ,  $n \in \mathbb{N}$ , is asymptotically efficient for  $\gamma$ . □

# Chapter 5

## Hypotheses Testing



As a basic concept of statistical inference, a statistical test allows for a data-based decision between two contrary hypotheses about some population, where two error probabilities with opposing trends are fundamental.

In a one-parameter EF, uniformly most powerful tests are easily found for particular, but important hypotheses. Moreover, uniformly most powerful unbiased tests can be established in other cases and in multiparameter EFs as well via conditioning, which reduces the problem to the one-parameter case. Finally, statistical tests for several parameters simultaneously are considered with a focus on likelihood-ratio tests.

For a rigorous introduction in hypotheses testing, we refer to the rich textbook literature on Mathematical Statistics. In order to become familiar with the topic, a basic knowledge of statistical testing with some applications is helpful.

### 5.1 One-Sided Test Problems

**Definition 5.1** Let  $\mathcal{P} = \{P_{\zeta}^* : \zeta \in \Xi\}$  be a family of distributions on the measurable space  $(\mathcal{X}, \mathcal{B})$ .

- (a) Every  $(\mathcal{B} - \mathbb{B}^1)$ -measurable mapping  $\varphi : \mathcal{X} \rightarrow [0, 1]$  is called a *(statistical) test*, and the function  $\beta_{\varphi}$  defined by

(continued)

**Definition 5.1** (continued)

$$\beta_\varphi(\zeta) = E_\zeta[\varphi] = \int \varphi dP_\zeta^*, \quad \zeta \in \Xi,$$

is its *power function*.

If  $\varphi$  only takes on the values 0 and 1, then it is called *non-randomized*.

- (b) For non-empty, disjoint subsets  $\Xi_0$  and  $\Xi_1$  of  $\Xi$ , the test problem with *null hypothesis*  $H_0 : \zeta \in \Xi_0$  and *alternative hypothesis*  $H_1 : \zeta \in \Xi_1$  is denoted by

$$H_0 : \zeta \in \Xi_0 \quad \leftrightarrow \quad H_1 : \zeta \in \Xi_1. \quad (5.1)$$

A test  $\varphi$  is then said to have (*significance*) *level*  $\alpha \in (0, 1)$  for test problem (5.1) if

$$\beta_\varphi(\zeta) \leq \alpha \quad \text{for all } \zeta \in \Xi_0,$$

and a test  $\varphi$  with level  $\alpha \in (0, 1)$  for test problem (5.1) is called *unbiased* if

$$\beta_\varphi(\zeta) \geq \alpha \quad \text{for all } \zeta \in \Xi_1.$$

Roughly speaking,  $\varphi(x)$  is interpreted as the probability to decide for the alternative hypothesis  $H_1$  if  $x \in X$  has been observed.  $\varphi(x) = 0$  therefore corresponds to a decision for  $H_0$ , whereas  $\varphi(x) = 1$  yields a decision for  $H_1$ . If there exists some  $x \in X$  for which  $\varphi(x) \in (0, 1)$ , then  $\varphi$  is called a *randomized test*, in the case of which the decision is indeed random and has to be implemented by some additional rule; one may think of flipping a fair coin in case of  $\varphi(x) = 1/2$  as a simple example. Allowing for randomization of the test decision is useful in the derivation of optimal tests; in practice, however, they do not play an important role and are usually avoided by adjusting the level of the test.

For  $\zeta \in \Xi_0$ ,  $\beta_\varphi(\zeta)$  is called the *error probability of the first kind* of  $\varphi$  at  $\zeta$ ; for a non-randomized test, it is the probability of deciding in favor of  $H_1$  and the true parameter is in  $\Xi_0$ . Accordingly, for  $\zeta \in \Xi_1$ ,  $1 - \beta_\varphi(\zeta)$  is called the *error probability of the second kind* of  $\varphi$  at  $\zeta$ ; for a non-randomized test, it is the probability of deciding in favor of  $H_0$  and the true parameter is in  $\Xi_1$ . In the latter case,  $\beta_\varphi(\zeta)$  is also said to be the *power* of  $\varphi$  at  $\zeta$ . Usually, it is not possible to simultaneously minimize the error probabilities of the first and second kind. The derivation of optimal tests is therefore based on a non-symmetric procedure. One minimizes the error probability of the second kind—and thus maximizes power—in a restricted family of tests, for example, among all tests with the same level.

**Definition 5.2** Let  $\Phi$  be a set of statistical tests. A test  $\varphi \in \Phi$  is called *uniformly most powerful (UMP) in  $\Phi$  for test problem (5.1)* if

$$\beta_{\varphi}(\zeta) = \sup_{\psi \in \Phi} \beta_{\psi}(\zeta) \quad \text{for all } \zeta \in \Xi_1.$$

For the particular case

$$\Phi = \{\psi : \psi \text{ is a test with level } \alpha \text{ for test problem (5.1)}\},$$

$\varphi$  is called a *UMP (uniformly most powerful) level- $\alpha$  test*, and for

$$\Phi = \{\psi : \psi \text{ is an unbiased test with level } \alpha \text{ for test problem (5.1)}\},$$

$\varphi$  is called a *UMPU (uniformly most powerful unbiased) level- $\alpha$  test for test problem (5.1)*.

Note that a UMP level- $\alpha$  test is necessarily unbiased, since its power is not smaller than that of the constant level- $\alpha$  test  $\psi \equiv \alpha$  at every  $\zeta \in \Xi_1$ . As a consequence, every UMP level- $\alpha$  test is also a UMPU level- $\alpha$  test.

For one- and multiparameter regular EFs, UMP or UMPU level- $\alpha$  tests can be provided for various test problems concerning a single parameter. These optimal tests are readily deduced from the (generalized) Neyman-Pearson lemma, which is presented in what follows. For the proof, see, e.g., [63, Section 2.4.].

**Lemma 5.3** Let  $\mu$  be a  $\sigma$ -finite measure and  $f_0, f_1, \dots, f_n$  be real-valued,  $\mu$ -integrable functions on the measurable space  $(X, \mathcal{B})$  for some  $n \in \mathbb{N}$ . Moreover, assume that  $(\alpha_0, \alpha_1, \dots, \alpha_{n-1}) \in \text{int}(Q)$  with

$$Q = \left\{ (q_0, q_1, \dots, q_{n-1}) \in \mathbb{R}^n : \text{there exists a test } \psi \text{ with} \right. \\ \left. \int \psi f_i d\mu = q_i \text{ for all } i = 0, 1, \dots, n-1 \right\}.$$

Then, the following properties hold true:

(a) A test  $\varphi$  maximizes  $\int \psi f_n d\mu$  among all tests  $\psi$  in

$$\Phi = \{\psi : \psi \text{ is test with } \int \psi f_i d\mu = \alpha_i, i = 0, 1, \dots, n-1\}$$

if and only if there exist constants  $k_0, k_1, \dots, k_{n-1} \in \mathbb{R}$  with

$$\varphi(x) = \begin{cases} 1, & f_n(x) > \sum_{i=0}^{n-1} k_i f_i(x), \\ 0, & f_n(x) < \sum_{i=0}^{n-1} k_i f_i(x), \end{cases} \quad \mu\text{-a.e.} \quad (5.2)$$

(b) There exists a maximizing test  $\varphi$  according to (a).

(c) If  $\alpha_i = \alpha \int f_i d\mu$  for  $i = 0, 1, \dots, n-1$  and some  $\alpha \in (0, 1)$ , then a test  $\varphi$  satisfying formula (5.2) with  $k_0, k_1, \dots, k_{n-1} \geq 0$  maximizes  $\int \psi f_n d\mu$  even among all tests  $\psi$  in

$$\tilde{\Phi} = \{\psi : \psi \text{ is test with } \int \psi f_i d\mu \leq \alpha_i, i = 0, 1, \dots, n-1\}.$$

First, we focus on optimal tests in one-parameter EFs, where UMP level- $\alpha$  tests for one-sided test problems are derived by applying Lemma 5.3. In what follows, as commonly used, a hypothesis  $\zeta \in (-\infty, \zeta_0] \cap \Xi^*$ , say, is denoted by  $\zeta \leq \zeta_0$ , for short.

**Theorem 5.4** Let  $\mathcal{P} = \{P_\zeta^* = f_\zeta^* \mu : \zeta \in \Xi^*\}$  be a regular EF with  $\mu$ -densities

$$f_\zeta^*(x) = C^*(\zeta) e^{\zeta T(x)} h(x), \quad x \in X,$$

for  $\zeta \in \Xi^* \subset \mathbb{R}$ . Moreover, let  $\zeta_0 \in \Xi^*$  and  $\alpha \in (0, 1)$ . Then, we have:

(a) For the one-sided test problem

$$H_0 : \zeta \leq \zeta_0 \quad \leftrightarrow \quad H_1 : \zeta > \zeta_0, \quad (5.3)$$

a UMP level- $\alpha$  test is given by

$$\varphi(x) = \mathbb{1}_{(c, \infty)}(T(x)) + \gamma \mathbb{1}_{\{c\}}(T(x)), \quad x \in X.$$

(b) For the one-sided test problem

$$H_0 : \zeta \geq \zeta_0 \quad \leftrightarrow \quad H_1 : \zeta < \zeta_0,$$

a UMP level- $\alpha$  test is given by

$$\varphi(x) = \mathbb{1}_{(-\infty, c)}(T(x)) + \gamma \mathbb{1}_{\{c\}}(T(x)), \quad x \in X.$$

In (a) and (b), respectively, the constants  $c \in \mathbb{R}$  and  $\gamma \in [0, 1]$  are each determined by the equation

$$E_{\zeta_0}[\varphi] = \alpha.$$



**Proof** (a) Let  $\varphi$  be as stated. Without loss of generality, we may assume that  $h(x) > 0$  for all  $x \in \mathcal{X}$ , since  $P_\zeta^*(h = 0) = 0$  for  $\zeta \in \Xi^*$ . We first show that, for every fixed  $\zeta_1 \in \Xi^*$  with  $\zeta_1 > \zeta_0$ ,  $\varphi$  is a UMP level- $\alpha$  test for the test problem

$$\tilde{H}_0 : \zeta = \zeta_0 \quad \leftrightarrow \quad \tilde{H}_1 : \zeta = \zeta_1 \quad (5.4)$$

with two *simple* hypotheses ( $\Xi_0 = \{\zeta_0\}$  and  $\Xi_1 = \{\zeta_1\}$  are singletons).

Obviously, the likelihood ratio

$$\frac{f_{\zeta_1}^*(x)}{f_{\zeta_0}^*(x)} = \frac{C^*(\zeta_1)}{C^*(\zeta_0)} e^{(\zeta_1 - \zeta_0) T(x)}$$

is a strictly increasing function in  $T(x)$ .  $\varphi$  can therefore equivalently be represented as

$$\varphi(x) = \begin{cases} 1, & f_{\zeta_1}^*(x) > k_0 f_{\zeta_0}^*(x), \\ \gamma, & f_{\zeta_1}^*(x) = k_0 f_{\zeta_0}^*(x), \\ 0, & f_{\zeta_1}^*(x) < k_0 f_{\zeta_0}^*(x), \end{cases}$$

for some constant  $k_0 \in \mathbb{R}$  such that  $E_{\zeta_0}[\varphi] = \alpha$ . Necessarily, it holds that  $k_0 \geq 0$ . Applying Lemma 5.3(c) with  $n = 1$ ,  $f_i = f_{\zeta_i}^*$ ,  $i = 0, 1$ , and  $\alpha_0 = \alpha$  ( $Q = [0, 1]$ ),  $\varphi$  then maximizes

$$\beta_\psi(\zeta_1) = E_{\zeta_1}[\psi] = \int \psi f_{\zeta_1}^* d\mu$$

among all tests  $\psi$  with

$$\beta_\psi(\zeta_0) = E_{\zeta_0}[\psi] = \int \psi f_{\zeta_0}^* d\mu \leq \alpha.$$

By Definitions 5.1 and 5.2, this means that  $\varphi$  is a UMP level- $\alpha$  test for test problem (5.4).

Moreover, since  $\varphi$  does not depend on the special value of  $\zeta_1 > \zeta_0$ , it is clear from Definition 5.2 that  $\varphi$  is even a UMP level- $\alpha$  test for the test problem

$$\tilde{H}_0 : \zeta = \zeta_0 \quad \leftrightarrow \quad H_1 : \zeta > \zeta_0. \quad (5.5)$$

Now, note that every level- $\alpha$  test for test problem (5.3) is a level- $\alpha$  test for test problem (5.5), too. Hence, what is left to verify for  $\varphi$  to be a UMP level- $\alpha$  test for test problem (5.3) is that  $\varphi$  has level  $\alpha$  for test problem (5.3). For this, it is sufficient to show that  $\beta_\varphi$  is non-decreasing, since then  $\beta_\varphi(\zeta) \leq \beta_\varphi(\zeta_0) = \alpha$  for  $\zeta \leq \zeta_0$  by assumption.

By Corollary 2.55,  $\beta_\varphi$  is differentiable with derivative

$$\beta'_\varphi(\zeta) = E_\zeta[(T - E_\zeta[T]) \varphi], \quad \zeta \in \Xi^*.$$

Hence, we have for all  $\zeta \in \Xi^*$  with  $c - E_\zeta[T] \geq 0$  that

$$\beta'_\varphi(\zeta) = \int_{(c, \infty)} (t - E_\zeta[T]) d(P_\zeta^*)^T(t) + \gamma(c - E_\zeta[T]) P_\zeta^*(T = c) \geq 0,$$

and for all  $\zeta \in \Xi^*$  with  $c - E_\zeta[T] < 0$

$$\begin{aligned} \beta'_\varphi(\zeta) &\geq \int_{[c, \infty)} (t - E_\zeta[T]) d(P_\zeta^*)^T(t) \\ &= \int_{(-\infty, \infty)} (t - E_\zeta[T]) d(P_\zeta^*)^T(t) - \int_{(-\infty, c)} (t - E_\zeta[T]) d(P_\zeta^*)^T(t) \\ &= - \int_{(-\infty, c)} (t - E_\zeta[T]) d(P_\zeta^*)^T(t) \geq 0. \end{aligned}$$

The proof of assertion (b) is analogue to that of (a) and therefore omitted.  $\square$

Obviously, the UMP tests in Theorem 5.4 can be chosen as non-randomized tests if  $T$  has a continuous distribution.

*Remark 5.5* The UMP tests in Theorem 5.4 have even strictly monotone power functions on  $\{\zeta \in \Xi^* : 0 < \beta_\varphi(\zeta) < 1\}$ , which can be seen as follows. Let  $\varphi$  be the UMP test in Theorem 5.4(a), and let  $\zeta_1, \zeta_2 \in \Xi^*$  with  $\zeta_1 < \zeta_2$  and  $\alpha_0 = \beta_\varphi(\zeta_1) \in (0, 1)$ . Then, by inspecting the proof of Theorem 5.4(a) and by Lemma 5.3(a),  $\varphi$  also maximizes  $E_{\zeta_2}[\psi]$  among all tests  $\psi$  with  $E_{\zeta_1}[\psi] = \alpha_0$ . In particular, by setting  $\psi \equiv \alpha_0$ , it follows that  $E_{\zeta_2}[\varphi] \geq \alpha_0$ . Here, if equality would be true, then  $\psi \equiv \alpha_0$  would also be a solution of this maximizing problem and Lemma 5.3(a) then would ensure that there exists a constant  $k_0 \in \mathbb{R}$  with  $f_{\zeta_2}^* = k_0 f_{\zeta_1}^* \mu$ -a.e., which would yield  $k_0 = 1$  and thus the contradiction  $P_{\zeta_2}^* = P_{\zeta_1}^*$ . Hence, we have  $\beta_\varphi(\zeta_2) = E_{\zeta_2}[\varphi] > \alpha_0 = \beta_\varphi(\zeta_1)$ . By the same arguments, the assertion can be shown for the UMP test in Theorem 5.4(b).

**Examples:**  $\text{poi}(\bullet), \mathcal{N}(\bullet, \sigma_0^2)$

*Example 5.6* Let  $X_1, \dots, X_n$  be an iid sample from  $P_a = \text{poi}(a)$  for some  $a \in (0, \infty)$ , and consider the test problem

$$H_0 : a \leq a_0 \quad \Leftrightarrow \quad H_1 : a > a_0,$$

for some fixed  $a_0 \in (0, \infty)$ . By transforming to the canonical representation of  $\text{poi}(\bullet)$  via  $P_\zeta^* = P_{\exp\{\zeta\}}$ ,  $\zeta \in \mathbb{R}$ , according to Example 2.37, the test problem becomes

$$H_0 : \zeta \leq \zeta_0 \quad \leftrightarrow \quad H_1 : \zeta > \zeta_0$$

with  $\zeta_0 = \ln(a_0)$ . Applying Theorem 5.4(a) to the corresponding product family  $\mathcal{P}^{(n)}$  then yields that

$$\varphi(\mathbf{x}) = \mathbb{1}_{(c, \infty)}(T_n(\mathbf{x})) + \gamma \mathbb{1}_{\{c\}}(T_n(\mathbf{x})), \quad \mathbf{x} = (x_1, \dots, x_n) \in \mathbb{N}_0^n,$$

forms a UMP level- $\alpha$  test for this test problem, where  $T_n(\mathbf{x}) = \sum_{i=1}^n x_i$ ,  $\mathbf{x} \in \mathbb{N}_0^n$ , and the constants  $c \in \mathbb{R}$  and  $\gamma \in [0, 1]$  are determined by the equation

$$\begin{aligned} \alpha &= (P_{\zeta_0}^*)^{(n)}(T_n > c) + \gamma (P_{\zeta_0}^*)^{(n)}(T_n = c) \\ &= P_{a_0}^{(n)}(T_n > c) + \gamma P_{a_0}^{(n)}(T_n = c). \end{aligned}$$

By using Example 3.2, the moment generating function of  $T_n$  is given by

$$m_{T_n}(b) = m_T^n(b) = \exp \left\{ na(e^b - 1) \right\}, \quad b \in \mathbb{R},$$

where integration is with respect to  $P_a^{(n)}$ . This yields  $T_n \sim P_{na} = \text{poi}(na)$ . Hence, since  $0 \leq \gamma \leq 1$ ,  $c$  must be the minimum integer  $m \in \mathbb{N}_0$  such that

$$\alpha \geq P_{a_0}^{(n)}(T_n > m) = e^{-na_0} \sum_{j=m+1}^{\infty} \frac{(na_0)^j}{j!},$$

and  $\gamma$  is then already specified as  $\gamma = [\alpha - P_{a_0}^{(n)}(T_n > c)] / P_{a_0}^{(n)}(T_n = c)$ .

*Example 5.7* Let  $X_1, \dots, X_n$  be an iid sample from  $\mathcal{N}(\mu, \sigma_0^2)$  for some  $\mu \in \mathbb{R}$  and fixed  $\sigma_0^2 \in (0, \infty)$ , and consider the test problem

$$H_0 : \mu \geq \mu_0 \quad \leftrightarrow \quad H_1 : \mu < \mu_0, \quad (5.6)$$

for some fixed  $\mu_0 \in \mathbb{R}$ . Obviously, the product family  $\mathcal{P}^{(n)}$  of  $\mathcal{P} = \mathcal{N}(\bullet, \sigma_0^2) = \{\mathcal{N}(\mu, \sigma_0^2) : \mu \in \mathbb{R}\}$  forms an EF with canonical representation

$$f_\mu^{(n)}(\mathbf{x}) = C(\mu) e^{\mu T_n(\mathbf{x})} h(\mathbf{x}), \quad \mathbf{x} = (x_1, \dots, x_n) \in \mathbb{R}^n,$$

where  $C(\mu) = \exp\{-n\mu^2/(2\sigma_0^2)\}$ ,  $\mu \in \mathbb{R}$ , and

$$T_n(\mathbf{x}) = \frac{1}{\sigma_0^2} \sum_{i=1}^n x_i \quad \text{and} \quad h(\mathbf{x}) = \left( \frac{1}{2\pi\sigma_0^2} \right)^{n/2} \exp \left\{ -\frac{1}{2\sigma_0^2} \sum_{i=1}^n x_i^2 \right\}$$

for  $\mathbf{x} = (x_1, \dots, x_n) \in \mathbb{R}^n$  (see Example 2.10). Applying Theorem 5.4(b) to the EF  $\mathcal{P}^{(n)}$  then yields that

$$\varphi(\mathbf{x}) = \mathbb{1}_{(-\infty, c)}(T_n(\mathbf{x})), \quad \mathbf{x} = (x_1, \dots, x_n) \in \mathbb{R}^n,$$

forms a UMP level- $\alpha$  test for test problem (5.6), where the constant  $c \in \mathbb{R}$  is determined by

$$P_{\mu_0}^{(n)}(T_n < c) = \alpha.$$

Here, since  $T_n$  is continuously distributed, we have set  $\gamma = 0$  and thus chosen a non-randomized test. Since  $T_n \sim \mathcal{N}(n\mu_0/\sigma_0^2, n/\sigma_0^2)$  under  $P_{\mu_0}^{(n)}$ , we conclude that  $c$  fulfils

$$\frac{c - n\mu_0/\sigma_0^2}{\sqrt{n/\sigma_0^2}} = u_\alpha,$$

where  $u_\alpha$  denotes the  $\alpha$ -quantile of the standard normal distribution  $\mathcal{N}(0, 1)$ . Equivalently, we have

$$c = \sqrt{\frac{n}{\sigma_0^2}} u_\alpha + \frac{n\mu_0}{\sigma_0^2},$$

and thus, in terms of the arithmetic mean of  $x_1, \dots, x_n$ , a representation of  $\varphi$  as

$$\varphi(\mathbf{x}) = \mathbb{1}_{\left(-\infty, \mu_0 + u_\alpha \sqrt{\sigma_0^2/n}\right)} \left( \frac{1}{n} \sum_{i=1}^n x_i \right), \quad \mathbf{x} = (x_1, \dots, x_n) \in \mathbb{R}^n.$$

This test is commonly known as Gauss test.

## 5.2 Two-Sided Test Problems

For *two-sided* test problems, a UMP level- $\alpha$  test does not exist, in general; however, optimal tests for these test problems can be established for one-parameter EFs by farther restricting the set  $\Phi$  of admissible tests, i.e., from all level- $\alpha$  tests to all unbiased level- $\alpha$  tests for the test problem.

**Theorem 5.8** *Under the assumptions of Theorem 5.4 a UMPU level- $\alpha$  test for the two-sided test problem*

$$H_0 : \zeta = \zeta_0 \quad \leftrightarrow \quad H_1 : \zeta \neq \zeta_0 \quad (5.7)$$

is given by

$$\varphi(x) = \begin{cases} 1, & T(x) < c_1 \text{ or } T(x) > c_2, \\ \gamma_i, & T(x) = c_i, \quad i = 1, 2, \\ 0, & c_1 < T(x) < c_2, \end{cases} \quad \text{for } x \in X, \quad (5.8)$$

where the constants  $c_1, c_2 \in \mathbb{R}$  and  $\gamma_1, \gamma_2 \in [0, 1]$  are determined by the equations

$$E_{\zeta_0}[\varphi] = \alpha \quad \text{and} \quad E_{\zeta_0}[\varphi T] = \alpha E_{\zeta_0}[T].$$

**Proof** We first show that every unbiased test with level  $\alpha$  for test problem (5.7) is contained in the set

$$\Phi = \{\psi : \psi \text{ is test with } E_{\zeta_0}[\psi] = \alpha \text{ and } E_{\zeta_0}[\psi T] = \alpha E_{\zeta_0}[T]\}.$$

Let  $\psi$  be an unbiased test with level  $\alpha$  for test problem (5.7). From Corollary 2.55, we have that the power function  $\beta_\psi(\zeta) = E_\zeta[\psi]$ ,  $\zeta \in \Xi^*$ , of  $\psi$  is continuously differentiable with derivative

$$\beta'_\psi(\zeta) = E_\zeta[\psi T] - E_\zeta[\psi]E_\zeta[T], \quad \zeta \in \Xi^*.$$

Since unbiasedness of  $\psi$  for test problem (5.7) yields that  $\beta_\psi(\zeta_0) \leq \alpha$  and  $\beta_\psi(\zeta) \geq \alpha$  for  $\zeta \in \Xi^* \setminus \{\zeta_0\}$ , it necessarily holds that  $\beta_\psi(\zeta_0) = \alpha$  by using the continuity of  $\beta_\psi$ . Moreover, unbiasedness of  $\psi$  implies that  $\beta_\psi$  has a global minimum at  $\zeta_0$ . Since  $\psi$  is differentiable, it follows that

$$0 = \beta'_\psi(\zeta_0) = E_{\zeta_0}[\psi T] - E_{\zeta_0}[\psi]E_{\zeta_0}[T] = E_{\zeta_0}[\psi T] - \alpha E_{\zeta_0}[T].$$

Hence,  $\psi \in \Phi$  is shown.

Next, note that every UMP test in  $\Phi$  for test problem (5.7) has level  $\alpha$  and must be unbiased for test problem (5.7), as a power comparison with the constant test

$\psi \equiv \alpha$  in  $\Phi$  shows. Hence, it also forms a UMPU level- $\alpha$  test for this test problem. We therefore aim to find a UMP test in  $\Phi$  for test problem (5.7). For this, we derive a UMP test in  $\Phi$  for test problem (5.4) with two simple hypotheses, where  $\zeta_1 \in \Xi^*$  with  $\zeta_1 \neq \zeta_0$  is fixed; this will lead to the test  $\varphi$  as stated and, since  $\varphi$  does not depend on the special value of  $\zeta_1$ , it is UMP in  $\Phi$  even for test problem (5.7) by Definition 5.2.

Let  $\zeta_1 \in \Xi^*$  with  $\zeta_1 \neq \zeta_0$  be fixed. As in the proof of Theorem 5.4, let  $h > 0$  without loss of generality. We apply Lemma 5.3 with  $n = 2$ ,  $f_0 = f_{\zeta_0}^*$ ,  $f_1 = T f_{\zeta_0}^*$  and  $f_2 = f_{\zeta_1}^*$ ,  $\alpha_0 = \alpha$  and  $\alpha_1 = \alpha E_{\zeta_0}[T]$ . Here, the condition that  $(\alpha, \alpha E_{\zeta_0}[T]) \in \text{int}(Q)$  can be verified as follows.

First, by choosing  $\psi \equiv \alpha + \varepsilon$  and  $\psi \equiv \alpha - \varepsilon$  for a sufficiently small  $\varepsilon > 0$ , we obtain that

$$(\alpha + \varepsilon, (\alpha + \varepsilon)E_{\zeta_0}[T]) \in Q \quad \text{and} \quad (\alpha - \varepsilon, (\alpha - \varepsilon)E_{\zeta_0}[T]) \in Q. \quad (5.9)$$

By Lemma 5.3(b) with  $n = 1$ ,  $f_0 = f_{\zeta_0}^*$ ,  $f_1 = T f_{\zeta_0}^*$ , and  $\alpha_0 = \alpha$ , there exists a test  $\varphi_1$ , which maximizes  $E_{\zeta_0}[\psi T]$  among all tests  $\psi$  with  $E_{\zeta_0}[\psi] = \alpha$ . In particular, the choice  $\psi \equiv \alpha$  yields that  $E_{\zeta_0}[\varphi_1 T] \geq \alpha E_{\zeta_0}[T]$ . If this would be an equality,  $\psi \equiv \alpha$  would also be a solution of the maximizing problem, and Lemma 5.3(a) then would ensure the existence of a constant  $k_0 \in \mathbb{R}$  with  $T = k_0$   $\mu$ -a.e., which cannot be the case. Hence,

$$E_{\zeta_0}[\varphi_1] = \alpha \quad \text{and} \quad E_{\zeta_0}[\varphi_1 T] > \alpha E_{\zeta_0}[T]. \quad (5.10)$$

By the same arguments, there exists a test  $\varphi_2$  with  $E_{\zeta_0}[\varphi_2] = 1 - \alpha$  and  $E_{\zeta_0}[\varphi_2 T] > (1 - \alpha)E_{\zeta_0}[T]$ . The test  $1 - \varphi_2$  therefore meets

$$E_{\zeta_0}[1 - \varphi_2] = \alpha \quad \text{and} \quad E_{\zeta_0}[(1 - \varphi_2)T] < \alpha E_{\zeta_0}[T]. \quad (5.11)$$

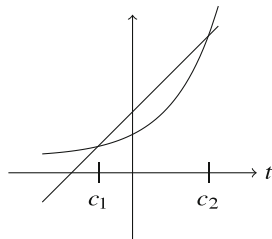
Equations (5.9)–(5.11) together with the convexity of  $Q$  then guarantee that  $(\alpha, \alpha E_{\zeta_0}[T]) \in \text{int}(Q)$ .

Now, applying Lemma 5.3(a, b) yields that there exists a UMP test  $\varphi$  in  $\Phi$  for test problem (5.4) with

$$\begin{aligned} \varphi(x) &= \begin{cases} 1, & f_{\zeta_1}^*(x) > (k_0 + k_1 T(x)) f_{\zeta_0}^*(x), \\ 0, & f_{\zeta_1}^*(x) < (k_0 + k_1 T(x)) f_{\zeta_0}^*(x), \end{cases} \\ &= \begin{cases} 1, & \frac{C^*(\zeta_1)}{C^*(\zeta_0)} \exp\{(\zeta_1 - \zeta_0)T(x)\} > k_0 + k_1 T(x), \\ 0, & \frac{C^*(\zeta_1)}{C^*(\zeta_0)} \exp\{(\zeta_1 - \zeta_0)T(x)\} < k_0 + k_1 T(x), \end{cases} \quad \mu\text{-a.e.}, \end{aligned}$$

with constants  $k_0, k_1 \in \mathbb{R}$  such that  $\varphi \in \Phi$ .

**Fig. 5.1** Illustration of the solutions  $c_1$  and  $c_2$  of Eq. (5.12) in  $t$  being intersections of a line and a strict convex function



Finally, we consider all solutions of the equation

$$\frac{C^*(\zeta_1)}{C^*(\zeta_0)} e^{(\zeta_1 - \zeta_0)t} = k_0 + k_1 t \quad (5.12)$$

with respect to  $t \in \mathbb{R}$ . If Eq. (5.12) would have no solution in  $t$ , then it would follow that  $\varphi \equiv 0$  or  $\varphi \equiv 1$  in contradiction to  $E_{\zeta_0}[\varphi] = \alpha$ . On the other hand, since the left-hand side of Eq. (5.12) is strictly convex as a function of  $t$ , there exist at most two solutions of Eq. (5.12) in  $t$ .  $\varphi$  therefore admits a representation as in formula (5.8) with constants  $c_1, c_2 \in \mathbb{R}$  and  $\gamma_1, \gamma_2 \in [0, 1]$  such that  $\varphi \in \Phi$  ( $c_1 = c_2$  and  $\gamma_1 = \gamma_2$  if Eq. (5.12) has only one solution). See Fig. 5.1 for an illustration.  $\square$

### Example: $\mathcal{E}(\bullet)$

*Example 5.9* Let  $X_1, \dots, X_n$  be an iid sample from  $\mathcal{E}(\lambda) \in \mathcal{P} = \mathcal{E}(\bullet)$  for some  $\lambda \in (0, \infty)$ , and consider the test problem

$$H_0 : \lambda = \lambda_0 \quad \leftrightarrow \quad H_1 : \lambda \neq \lambda_0,$$

for some fixed  $\lambda_0 \in \mathbb{R}$ . By transforming to the canonical representation of  $\mathcal{E}(\bullet)$  via  $P_\zeta^* = \mathcal{E}(-1/\zeta)$ ,  $\zeta \in (-\infty, 0)$ , according to Example 2.12, the test problem becomes

$$H_0 : \zeta = \zeta_0 \quad \leftrightarrow \quad H_1 : \zeta \neq \zeta_0$$

with  $\zeta_0 = -1/\lambda_0$ . Applying Theorem 5.8 to the corresponding product family  $\mathcal{P}^{(n)}$  then yields that

$$\varphi(\mathbf{x}) = 1 - \mathbb{1}_{[c_1, c_2]}(T_n(\mathbf{x})), \quad \mathbf{x} = (x_1, \dots, x_n) \in (0, \infty)^n,$$

forms a UMPU level- $\alpha$  test for this test problem, where  $T_n(\mathbf{x}) = \sum_{i=1}^n x_i$ ,  $\mathbf{x} \in (0, \infty)^n$ , and the constants  $c_1, c_2 \in \mathbb{R}$  are determined by the equations

$$\int [1 - \mathbb{1}_{[c_1, c_2]}(T_n)] d(P_{\zeta_0}^*)^{(n)} = \alpha$$

and

$$\int [1 - \mathbb{1}_{[c_1, c_2]}(T_n)] T_n d(P_{\zeta_0}^*)^{(n)} = \alpha \int T_n d(P_{\zeta_0}^*)^{(n)}.$$

Here, since  $T_n$  is continuously distributed, we have set  $\gamma_i = 0$ ,  $i = 1, 2$ . Moreover, we have that  $T_n \sim \mathcal{G}(-1/\zeta_0, n) = \mathcal{G}(\lambda_0, n)$  with corresponding density  $f_{\lambda_0, n}$  and mean  $n\lambda_0$  under  $(P_{\zeta_0}^*)^{(n)}$ . Hence, we may rewrite the above equations as

$$\int_{c_1}^{c_2} f_{\lambda_0, n}(t) dt = 1 - \alpha \quad \text{and} \quad \int_{c_1}^{c_2} t f_{\lambda_0, n}(t) dt = n\lambda_0(1 - \alpha).$$

It can be shown analytically that these equations have a unique solution in  $(c_1, c_2) \in \mathbb{R}^2$ , which does not depend on  $\lambda_0$ ; for a given sample size  $n$  and a desired significance level  $\alpha$ , this solution can be obtained numerically, for instance, via the Newton-Raphson method (see, e.g., [7]).

Besides one- and two-sided test problems, choosing a bounded interval as (either null or alternative) hypothesis may also be of interest in applications, which are known as *hypotheses of (non)-equivalence* in the literature (see [62] for an extensive account on the topic). For one-parameter EFs, optimal tests for these test problems can be derived in a similar way as in the proofs of Theorems 5.4 and 5.8. They are stated in the following theorem; a proof can be found, for instance, in [63, Section 2.4].

**Theorem 5.10** *Let the assumptions of Theorem 5.4 be given, and let  $\zeta_1, \zeta_2 \in \Xi^*$  with  $\zeta_1 < \zeta_2$ . Then, the following assertions are true:*

(a) *For the test problem*

$$H_0 : \zeta \in [\zeta_1, \zeta_2] \quad \leftrightarrow \quad H_1 : \zeta \notin [\zeta_1, \zeta_2],$$

*a UMPU level- $\alpha$ -test is given by  $\varphi$  in formula (5.8).*

(b) *For the test problem*

$$H_0 : \zeta \notin (\zeta_1, \zeta_2) \quad \leftrightarrow \quad H_1 : \zeta \in (\zeta_1, \zeta_2),$$



a UMP level- $\alpha$  test is given by

$$\varphi(x) = \begin{cases} 1, & c_1 < T(x) < c_2, \\ \gamma_i, & T(x) = c_i, \ i = 1, 2, \\ 0, & T(x) < c_1 \text{ or } T(x) > c_2, \end{cases} \quad \text{for } x \in \mathcal{X}.$$

In (a) and (b), respectively, the constants  $c_1, c_2 \in \mathbb{R}$  and  $\gamma_1, \gamma_2 \in [0, 1]$  are each determined by the equations

$$E_{\zeta_1}[\varphi] = \alpha \quad \text{and} \quad E_{\zeta_2}[\varphi] = \alpha.$$

**Remark 5.11** Note that, since every UMP level- $\alpha$  test is unbiased, the tests in Theorems 5.4 and 5.10(b) are also UMPU level- $\alpha$  tests for the given test problem.

### 5.3 Testing for a Selected Parameter

In multiparameter EFs ( $k > 1$ ), UMPU level- $\alpha$  tests for test problems concerning a single parameter  $\zeta_1$ , say, may be derived as well. We illustrate the key arguments by considering the regular EF  $\mathcal{P} = \{P_{\zeta}^* : \zeta = (\zeta_1, \zeta_2)^t \in \Xi^*\}$  with minimal canonical representation

$$f_{\zeta}^*(x) = C^*(\zeta) \exp\{\zeta_1 T_1(x) + \zeta_2 T_2(x)\} h(x), \quad x \in \mathcal{X}, \quad (5.13)$$

and the test problem (5.1) with  $\Xi_0 = \{\zeta \in \Xi^* : \zeta_1 \leq \zeta_0\}$  and  $\Xi_1 = \{\zeta \in \Xi^* : \zeta_1 > \zeta_0\}$  or, for short,

$$H_0 : \zeta_1 \leq \zeta_0 \quad \leftrightarrow \quad H_1 : \zeta_1 > \zeta_0, \quad (5.14)$$

where  $\zeta_0$  is fixed. Since  $\mathbf{T} = (T_1, T_2)^t$  is sufficient for  $\mathcal{P}$  by Corollary 3.21, we may restrict ourselves to consider tests  $\psi = \psi(\mathbf{T})$  that depend on  $x$  only through  $\mathbf{T}(x)$ .

Let  $B = \{\zeta \in \Xi^* : \zeta_1 = \zeta_0\}$  denote the common boundary of  $\Xi_0$  and  $\Xi_1$ . Since the power function of any test is continuous by Corollary 2.55, every unbiased test  $\psi$  with level  $\alpha$  for test problem (5.14) is necessarily  $\alpha$ -similar on  $B$ , i.e.,

$$E_{\zeta}[\psi(\mathbf{T})] = \alpha \quad \text{for all } \zeta \in B. \quad (5.15)$$

Hence, if we find a UMP test among all tests with level  $\alpha$  for test problem (5.14) satisfying condition (5.15), then this is also a UMPU level- $\alpha$  test for test problem (5.14), where unbiasedness of the test is obtained by a power comparison with the constant test  $\psi \equiv \alpha$ .

Next, note that the power function of every test  $\psi$  at  $\zeta \in \Xi^*$  can be represented as

$$\beta_\psi(\zeta) = E_\zeta[\psi(T)] = \int \int \psi(t_1, t_2) d(P_\zeta^*)^{T_1|T_2=t_2}(t_1) d(P_\zeta^*)^{T_2}(t_2). \quad (5.16)$$

A crucial point is now that  $T_2$  is a sufficient and (boundedly) complete statistic for the EF  $\{P_\zeta^* : \zeta \in B\}$ , which follows by setting  $\zeta_1 = \zeta_0$  in formula (5.13) and then applying Corollary 3.21. The tests satisfying condition (5.15) are therefore exactly the tests having so-called *Neyman structure* with respect to  $T_2$ , i.e., the tests  $\psi$  that meet

$$\int \psi(t_1, t_2) d(P_\zeta^*)^{T_1|T_2=t_2}(t_1) = \alpha \quad (P_\zeta^*)^{T_2}\text{-a.s.}, \quad \zeta \in B$$

(see, for instance, [40, p. 118/119] and [55, pp. 405/406]).

This suggests to consider the inner integral in representation (5.16) separately for every fixed  $t_2$ , which reduces the maximizing problem to the one-parameter case, since the conditional distribution of  $T_1$  given  $T_2 = t_2$  does not depend on  $\zeta_2$  and forms a one-parameter EF in  $\zeta_1$  (see Theorem 3.11).

Detailed proofs of the following theorems can be found, e.g., in [40, pp. 119–122] and [55, pp. 406–408]. Without loss of generality, the results are formulated in terms of the first parameter.

**Theorem 5.12** *Let  $\mathcal{P} = \{P_\zeta^* = f_\zeta^* \mu : \zeta \in \Xi^*\}$  be a regular EF with minimal canonical representation of the  $\mu$ -densities given by*

$$f_\zeta^*(x) = C^*(\zeta) \exp\{\zeta_1 U(x) + (\zeta_2, \dots, \zeta_k) V(x)\} h(x), \quad x \in \mathcal{X},$$

for  $\zeta = (\zeta_1, \dots, \zeta_k)^t \in \Xi^* \subset \mathbb{R}^k$ , where  $U : (\mathcal{X}, \mathcal{B}) \rightarrow (\mathbb{R}^1, \mathbb{B}^1)$  is a random variable and  $V : (\mathcal{X}, \mathcal{B}) \rightarrow (\mathbb{R}^{k-1}, \mathbb{B}^{k-1})$  is a column random vector. Moreover, let  $\eta = (\eta_1, \dots, \eta_k)^t \in \Xi^*$  and  $\alpha \in (0, 1)$ . Then, we have:

(a) *For the one-sided test problem*

$$H_0 : \zeta_1 \leq \eta_1 \quad \leftrightarrow \quad H_1 : \zeta_1 > \eta_1,$$

a UMPU level- $\alpha$  test is given by  $\varphi = \psi \circ (U, V)$  with mapping

$$\psi(u, v) = \mathbb{1}_{(c(v), \infty)}(u) + \gamma(v) \mathbb{1}_{\{c(v)\}}(u), \quad u \in \mathbb{R}, \quad v \in \mathbb{R}^{k-1}.$$

(b) *For the one-sided test problem*

$$H_0 : \zeta_1 \geq \eta_1 \quad \leftrightarrow \quad H_1 : \zeta_1 < \eta_1$$

a UMPU level- $\alpha$  test is given by  $\varphi = \psi \circ (U, V)$  with mapping

$$\psi(u, v) = \mathbb{1}_{(-\infty, c(v))}(u) + \gamma(v) \mathbb{1}_{\{c(v)\}}(u), \quad u \in \mathbb{R}, v \in \mathbb{R}^{k-1}.$$

In (a) and (b), respectively, the mappings  $c, \gamma : (\mathbb{R}^{k-1}, \mathbb{B}^{k-1}) \rightarrow (\mathbb{R}^1, \mathbb{B}^1)$  with  $0 \leq \gamma \leq 1$  are each determined by the equations

$$E_\eta[\varphi|V = v] = \alpha, \quad v \in \mathbb{R}^{k-1}. \quad (5.17)$$

Note that the expected value in formula (5.17) is independent of  $\eta_2, \dots, \eta_k$ , since

$$E_\eta[\varphi|V = v] = E_\eta[\psi(U, v)|V = v] = \int \psi(u, v) d(P_\eta^*)^{U|V=v}(u),$$

and the conditional distribution of  $U$  given  $V = v$  under  $P_\eta^*$  does only depend on  $\eta_1$ .

**Theorem 5.13** *Under the assumptions of Theorem 5.12 a UMPU level- $\alpha$  test for the two-sided test problem*

$$H_0 : \zeta_1 = \eta_1 \quad \leftrightarrow \quad H_1 : \zeta_1 \neq \eta_1$$

is given by  $\varphi = \psi \circ (U, V)$  with mapping

$$\psi(u, v) = \begin{cases} 1, & u < c_1(v) \text{ or } u > c_2(v), \\ \gamma_i(v), & u = c_i(v), i = 1, 2, \\ 0, & c_1(v) < u < c_2(v), \end{cases} \quad u \in \mathbb{R}, v \in \mathbb{R}^{k-1}.$$

Here, the mappings  $c_i, \gamma_i : (\mathbb{R}^{k-1}, \mathbb{B}^{k-1}) \rightarrow (\mathbb{R}^1, \mathbb{B}^1)$  with  $0 \leq \gamma_i \leq 1, i = 1, 2$ , are determined by the equations

$$E_\eta[\varphi|V = v] = \alpha \quad \text{and} \quad E_\eta[\varphi U|V = v] = \alpha E_\eta[U|V = v] \quad (5.18)$$

for  $v \in \mathbb{R}^{k-1}$ .

The expected values in formula (5.18) are all independent of  $\eta_2, \dots, \eta_k$ .

Respective statements hold true for the test problems in Theorem 5.10 with interval hypotheses.

**Example:**  $\mathcal{N}_2(\bullet, \Sigma_0)$

*Example 5.14* Let  $\mathcal{P} = \mathcal{N}_2(\bullet, \Sigma_0) = \{\mathcal{N}_2(\mu, \Sigma_0) : \mu \in \mathbb{R}^2\}$  be the EF of bivariate normal distributions according to Example 2.18 with  $k = 2$  and fixed

covariance matrix

$$\Sigma_0 = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix},$$

where  $\sigma_1, \sigma_2 > 0$  and  $\sigma_{12} \in \mathbb{R}$  with  $\sigma_{12}^2 < \sigma_1^2 \sigma_2^2$ . We assume to have an iid sample  $\mathbf{X}^{(1)}, \dots, \mathbf{X}^{(n)}$  from  $P_\mu$  and consider the test problem

$$H_0 : \mu_1 \leq \eta_1 \quad \leftrightarrow \quad H_1 : \mu_1 > \eta_1 \quad (5.19)$$

for fixed  $\eta_1 \in \mathbb{R}$ . From Example 2.18, we know that  $P_\mu^T = \mathcal{N}_2(\Sigma_0^{-1}\mu, \Sigma_0^{-1})$  for  $\mathbf{T}(\mathbf{x}) = \Sigma_0^{-1}\mathbf{x}$ ,  $\mathbf{x} \in \mathbb{R}^2$ . For the product family  $\mathcal{P}^{(n)}$ , it therefore follows that  $\mathbf{T}^{(n)}(\tilde{\mathbf{x}}) = \sum_{i=1}^n \mathbf{T}(\mathbf{x}^{(i)})$ ,  $\tilde{\mathbf{x}} = (\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}) \in (\mathbb{R}^2)^n$  has a  $\mathcal{N}_2(n\Sigma_0^{-1}\mu, n\Sigma_0^{-1})$ -distribution under  $P_\mu^{(n)}$ , where

$$n\Sigma_0^{-1} = \frac{n}{\sigma_1^2\sigma_2^2 - \sigma_{12}^2} \begin{pmatrix} \sigma_2^2 & -\sigma_{12} \\ -\sigma_{12} & \sigma_1^2 \end{pmatrix} = \begin{pmatrix} \tau_1^2 & \tau_{12} \\ \tau_{12} & \tau_2^2 \end{pmatrix}, \quad \text{say.} \quad (5.20)$$

The conditional distribution of  $T_1^{(n)}$  given  $T_2^{(n)} = v$  under  $P_\mu^{(n)}$  is a univariate normal distribution with mean  $\mu(v)$  and variance  $\sigma^2(v)$  given by

$$\begin{aligned} \mu(v) &= \tau_1^2\mu_1 + \tau_{12}\mu_2 + \frac{\tau_{12}}{\tau_2^2} \left( \frac{\tau_{12}}{\tau_1\tau_2} \right) [v - (\tau_{12}\mu_1 + \tau_2^2\mu_2)] \\ &= \tau_1^2\mu_1 + \tau_{12}\mu_2 + \frac{\tau_{12}}{\tau_2^2} (v - \tau_{12}\mu_1 - \tau_2^2\mu_2) \\ &= \frac{1}{\tau_2^2} (\tau_1^2\tau_2^2 - \tau_{12}^2) \mu_1 + \frac{\tau_{12}}{\tau_2^2} v = \frac{n\mu_1 - \sigma_{12}v}{\sigma_1^2} \\ \text{and} \quad \sigma^2(v) &= \left( 1 - \frac{\tau_{12}^2}{\tau_1^2\tau_2^2} \right) \tau_1^2 = \frac{1}{\tau_2^2} (\tau_1^2\tau_2^2 - \tau_{12}^2) = \frac{n}{\sigma_1^2}, \end{aligned}$$

where the latter is free of  $v$ . In the derivation, we have used the identity

$$\tau_1^2\tau_2^2 - \tau_{12}^2 = \frac{n^2}{\sigma_1^2\sigma_2^2 - \sigma_{12}^2},$$

which is obtained by considering determinants in formula (5.20), as well as the equations  $\tau_{12}/\tau_2^2 = -\sigma_{12}/\sigma_1^2$  and  $\tau_2^2 = n\sigma_1^2/(\sigma_1^2\sigma_2^2 - \sigma_{12}^2)$ . Note that the

conditional distribution does not depend on  $\mu_2$ . Denoting by  $c_{1-\alpha}(v)$  the  $(1-\alpha)$ -quantile of this distribution, it holds for  $v \in \mathbb{R}$  that

$$u_{1-\alpha} = \frac{c_{1-\alpha}(v) - \mu(v)}{\sigma(v)}$$

and thus 
$$c_{1-\alpha}(v) = u_{1-\alpha} \sigma(v) + \mu(v) = u_{1-\alpha} \sqrt{\frac{n}{\sigma_1^2}} + \frac{n\mu_1 - \sigma_{12}v}{\sigma_1^2},$$

where  $u_{1-\alpha}$  is the  $(1-\alpha)$ -quantile of the standard normal distribution  $\mathcal{N}(0, 1)$ . Hence, by Theorem 5.12(a), a UMPU level- $\alpha$  test for test problem (5.19) is given by

$$\varphi(\tilde{\mathbf{x}}) = \mathbb{1}_{(c(T_2^{(n)}(\tilde{\mathbf{x}})), \infty)}(T_1^{(n)}(\tilde{\mathbf{x}})), \quad \tilde{\mathbf{x}} \in (\mathbb{R}^2)^n,$$

with 
$$c(v) = u_{1-\alpha} \sqrt{\frac{n}{\sigma_1^2}} + \frac{n\eta_1 - \sigma_{12}v}{\sigma_1^2}, \quad v \in \mathbb{R}.$$

## 5.4 Testing for Several Parameters

In the preceding sections, we have proposed optimal tests for test problems concerning a single parameter of a one- or multiparameter EF. Here, we consider tests for test problem (5.1) with parameter set  $\Xi_0 \subset \mathbb{R}^k$  making assumptions on more than one component of  $\boldsymbol{\zeta}$ , where we focus on likelihood-ratio tests. On the one hand, these tests enable to check conjectures on several parameters simultaneously as, for instance, a simple null hypothesis by choosing  $\Xi_0 = \{\boldsymbol{\eta}\}$  for some fixed  $\boldsymbol{\eta} \in \Xi$ . On the other hand, they may be used to test for relations between the parameters; a homogeneity test, for example, is obtained by setting  $\Xi_0 = \{\boldsymbol{\zeta} \in \Xi : \zeta_1 = \dots = \zeta_k\}$ .

**Definition 5.15** Let  $\mu$  be a  $\sigma$ -finite measure and  $\mathcal{P} = \{f_{\boldsymbol{\zeta}}^* \mu : \boldsymbol{\zeta} \in \Xi\}$ ,  $\Xi \subset \mathbb{R}^k$ , be a family of probability distributions on a measurable space  $(\mathcal{X}, \mathcal{B})$ . Moreover, let  $\emptyset \neq \Xi_0 \subsetneq \Xi$  and  $\alpha \in (0, 1)$ . For the test problem

$$H_0 : \boldsymbol{\zeta} \in \Xi_0 \quad \leftrightarrow \quad H_1 : \boldsymbol{\zeta} \in \Xi \setminus \Xi_0, \quad (5.21)$$

(continued)

**Definition 5.15** (continued)  
the *likelihood-ratio (LR) test* with level  $\alpha$  is defined as

$$\varphi_{\Lambda}(x) = \mathbb{1}_{(c, \infty)}(\Lambda(x)) + \gamma \mathbb{1}_{[c]}(\Lambda(x)), \quad x \in \mathcal{X},$$

provided that the test statistic

$$\Lambda(x) = -2 \ln \left( \frac{\sup_{\zeta \in \Xi_0} f_{\zeta}^*(x)}{\sup_{\zeta \in \Xi} f_{\zeta}^*(x)} \right), \quad x \in \mathcal{X},$$

is well-defined  $\mu$ -a.e.. Here, the constants  $c \geq 0$  and  $\gamma \in [0, 1]$  are determined by the equation

$$\sup_{\zeta \in \Xi_0} E_{\zeta}[\varphi_{\Lambda}] = \alpha.$$

The LR test compares the maximum ‘probability’  $f_{\zeta}^*(x)$  of observing  $x$  under  $\Xi_0$  with the one under  $\Xi$ ; the null hypothesis is then rejected if the ratio of both values is (too) small (see the representations of  $\Lambda$  and  $\varphi_{\Lambda}$  in Definition 5.15). The numerator and denominator in the expression of  $\Lambda$  are nothing but the density  $f_{\zeta}^*(x)$ , where the parameter is chosen as the ML estimate of  $\zeta$  in  $\Xi_0$  and in  $\Xi$ , respectively.

In minimal regular EFs,  $\Lambda$  is closely related to the Kullback-Leibler divergence.

**Theorem 5.16** *Let  $\mathcal{P}$  be a regular EF with minimal canonical representation (4.1), and let  $x$  be a realization of the random variable  $X$  with distribution  $P_{\zeta}^* \in \mathcal{P}$ . If the ML estimate  $\hat{\zeta}(x)$  of  $\zeta$  in  $\Xi^*$  and the ML estimate  $\hat{\zeta}_0(x)$  of  $\zeta$  in  $\Xi_0$  based on  $x$  exist, then the test statistic of the LR test for test problem (5.21) with  $\Xi = \Xi^*$  is given by*

$$\Lambda(x) = 2 D_{KL}(\hat{\zeta}(x), \Xi_0) = 2 D_{KL}(\hat{\zeta}(x), \hat{\zeta}_0(x)), \quad (5.22)$$

$$\text{where} \quad D_{KL}(\zeta, \Xi_0) = \inf_{\eta \in \Xi_0} D_{KL}(\zeta, \eta), \quad \zeta \in \Xi^*.$$

**Proof** Without loss of generality, we assume that  $h(x) > 0$ . Then, since the logarithm is strictly increasing, we have

$$\begin{aligned} \Lambda(x) &= 2 [\ln(f_{\hat{\zeta}(x)}^*(x)) - \ln(\sup_{\zeta \in \Xi_0} f_{\zeta}^*(x))] \\ &= 2 \left( \hat{\zeta}(x)^T T(x) - \kappa(\hat{\zeta}(x)) - \sup_{\zeta \in \Xi_0} \{\zeta^T T(x) - \kappa(\zeta)\} \right) \end{aligned}$$

$$\begin{aligned}
&= 2 \inf_{\boldsymbol{\zeta} \in \Xi_0} \left( \hat{\boldsymbol{\zeta}}(x)^t \mathbf{T}(x) - \kappa(\hat{\boldsymbol{\zeta}}(x)) - (\boldsymbol{\zeta}^t \mathbf{T}(x) - \kappa(\boldsymbol{\zeta})) \right) \\
&= 2 \inf_{\boldsymbol{\zeta} \in \Xi_0} \left( \kappa(\boldsymbol{\zeta}) - \kappa(\hat{\boldsymbol{\zeta}}(x)) + (\hat{\boldsymbol{\zeta}}(x) - \boldsymbol{\zeta})^t \mathbf{T}(x) \right).
\end{aligned}$$

Now, by Theorem 4.3,  $\mathbf{T}(x) = \pi(\hat{\boldsymbol{\zeta}}(x))$ , and the first equality in formula (5.22) follows from Lemma 3.37. Since the supremum of  $\boldsymbol{\zeta}^t \mathbf{T}(x) - \kappa(\boldsymbol{\zeta})$ ,  $\boldsymbol{\zeta} \in \Xi_0$ , is attained at  $\hat{\boldsymbol{\zeta}}_0(x)$ , the equation  $\Lambda(x) = 2D_{KL}(\hat{\boldsymbol{\zeta}}(x), \hat{\boldsymbol{\zeta}}_0(x))$  is also valid.  $\square$

In particular, for fixed  $\boldsymbol{\eta} \in \Xi^*$  and the test problem

$$H_0 : \boldsymbol{\zeta} = \boldsymbol{\eta} \quad \Leftrightarrow \quad H_1 : \boldsymbol{\zeta} \neq \boldsymbol{\eta}, \quad (5.23)$$

with a simple null hypothesis, Theorem 5.16 yields that  $\Lambda = 2D_{KL}(\hat{\boldsymbol{\zeta}}, \boldsymbol{\eta})$  provided that the ML estimate  $\hat{\boldsymbol{\zeta}}(x)$  of  $\boldsymbol{\zeta}$  exists for  $x \in \mathcal{X}$ .

*Remark 5.17* The EF structure may also be useful for deriving other well known test statistics for multivariate test problems. For test problem (5.21), if all appearing quantities exist, the *Rao score statistic* and the *Wald statistic* are defined as

$$R = \mathbf{U}_{\hat{\boldsymbol{\zeta}}_0}^t \mathbf{I}(\hat{\boldsymbol{\zeta}}_0)^{-1} \mathbf{U}_{\hat{\boldsymbol{\zeta}}_0}$$

$$\text{and} \quad W = g(\hat{\boldsymbol{\zeta}})^t [\mathbf{D}_g(\hat{\boldsymbol{\zeta}}) \mathbf{I}(\hat{\boldsymbol{\zeta}})^{-1} \mathbf{D}_g(\hat{\boldsymbol{\zeta}})^t]^{-1} g(\hat{\boldsymbol{\zeta}}),$$

where  $\mathbf{U}_{\boldsymbol{\zeta}}$  is the score statistic and  $\mathbf{I}(\boldsymbol{\zeta})$  is the Fisher information matrix at  $\boldsymbol{\zeta}$ , and  $g : \Xi \rightarrow \mathbb{R}^q$ ,  $q \leq k$ , is a differentiable function with Jacobian matrix  $\mathbf{D}_g(\boldsymbol{\zeta})$  of full rank  $q$  at every  $\boldsymbol{\zeta} \in \Xi$ , which satisfies that  $g(\boldsymbol{\zeta}) = \mathbf{0}$  if and only if  $\boldsymbol{\zeta} \in \Xi_0$  (see, e.g., [53, pp. 239/240]). Note that  $R$  does only depend on the MLE  $\hat{\boldsymbol{\zeta}}_0$  of  $\boldsymbol{\zeta}$  in  $\Xi_0$ , whereas  $W$  only depends on the MLE  $\hat{\boldsymbol{\zeta}}$  of  $\boldsymbol{\zeta}$  in  $\Xi$ . The *Rao score test* and the *Wald test* then reject  $H_0$  for (too) large values of  $R$  and  $W$ , respectively. These decision rules are based on the fact that, if  $\boldsymbol{\zeta} \in \Xi_0$  is true, then, on the one hand,  $\mathbf{U}_{\hat{\boldsymbol{\zeta}}_0}$  will tend to be close to  $\mathbf{U}_{\hat{\boldsymbol{\zeta}}} = \mathbf{0}$  and, on the other hand,  $g(\hat{\boldsymbol{\zeta}})$  is expected to be small as an estimator of  $g(\boldsymbol{\zeta}) = \mathbf{0}$ . Now, if the underlying model forms a minimal regular EF, we may utilize the representations for  $\mathbf{U}_{\boldsymbol{\zeta}}$  and  $\mathbf{I}(\boldsymbol{\zeta})$  as derived in Theorem 3.29 and make also use of the results in Sects. 4.1 and 4.2 to find the MLEs  $\hat{\boldsymbol{\zeta}}$  and  $\hat{\boldsymbol{\zeta}}_0$ .

#### Example: $\mathcal{N}_k(\bullet, \boldsymbol{\Sigma}_0)$

*Example 5.18* Let  $\mathbf{X}^{(1)}, \dots, \mathbf{X}^{(n)}$  be an iid sample from  $\mathcal{N}_k(\boldsymbol{\mu}, \boldsymbol{\Sigma}_0)$  for some  $\boldsymbol{\mu} \in \mathbb{R}^k$  and fixed  $\boldsymbol{\Sigma}_0 \in \mathbb{R}^{k \times k}$  with  $\boldsymbol{\Sigma}_0 > 0$ , and consider the test problem

$$H_0 : \boldsymbol{\mu} = \boldsymbol{\eta} \quad \Leftrightarrow \quad H_1 : \boldsymbol{\mu} \neq \boldsymbol{\eta}, \quad (5.24)$$

for some fixed  $\eta \in \mathbb{R}^k$ . From Example 4.10, the ML estimate of  $\mu$  based on realizations  $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}$  of  $\mathbf{X}^{(1)}, \dots, \mathbf{X}^{(n)}$  is given by  $\hat{\mu}^{(n)}(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}) = \sum_{i=1}^n \mathbf{x}^{(i)} / n$ . Hence, by using Theorem 5.16 and the representation of  $D_{KL}$  derived in Example 3.41, the test statistic of the LR test for test problem (5.24) is given by

$$\Lambda_n = 2n D_{KL}(\hat{\mu}^{(n)}, \eta) = n \|\hat{\mu}^{(n)} - \eta\|_{\Sigma_0}^2.$$

Here, note that  $nD_{KL}$  is the Kullback-Leibler divergence corresponding to the EF  $\mathcal{P}^{(n)}$  of product measures. Under  $H_0$ , we have

$$\Sigma_0^{-1/2} \sqrt{n} (\hat{\mu}^{(n)} - \eta) \sim \mathcal{N}_k(\mathbf{0}, \mathbf{E}_k),$$

where  $\mathbf{E}_k$  is the  $k$ -dimensional unit matrix and  $\Sigma_0^{-1/2} \in \mathbb{R}^{k \times k}$  is a matrix satisfying  $\Sigma_0^{-1/2} > 0$  and  $\Sigma_0^{-1/2} \Sigma_0^{-1/2} = \Sigma_0^{-1}$ , the existence of which is ensured by the spectral theorem. Thus, under  $H_0$ ,  $\Lambda_n \sim \chi^2(k)$ , i.e.,  $\Lambda_n$  has a chi-square distribution with  $k$  degrees of freedom. The LR test with level  $\alpha$  is therefore given by

$$\varphi(\tilde{\mathbf{x}}) = \mathbb{1}_{(\chi_{1-\alpha}^2(k), \infty)} \left( n \|\hat{\mu}(\tilde{\mathbf{x}}) - \eta\|_{\Sigma_0}^2 \right), \quad \tilde{\mathbf{x}} = (\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}) \in (\mathbb{R}^k)^n,$$

where  $\chi_{1-\alpha}^2(k)$  denotes the  $(1 - \alpha)$ -quantile of  $\chi^2(k)$  (see, e.g., [45, p. 124]).

Now, let us derive the Rao score statistic and the Wald statistic for test problem (5.24) as introduced in Remark 5.17, where we choose  $g(\mu) = \mu - \eta$ ,  $\mu \in \mathbb{R}^k$ . First, note that by Example 2.18 and Theorem 3.13(b) the statistic  $\mathbf{T}^{(n)}$  for the product EF  $\mathcal{P}^{(n)}$  can be represented as  $\mathbf{T}^{(n)} = n \Sigma_0^{-1} \hat{\mu}^{(n)}$ . According to Theorem 3.29 and Example 3.35, the score statistic and the Fisher information matrix belonging to  $\mathcal{P}^{(n)}$  are therefore given by  $\mathbf{U}_{\mu}^{(n)} = n \Sigma_0^{-1} (\hat{\mu}^{(n)} - \mu)$  and  $\mathbf{I}^{(n)}(\mu) = n \Sigma_0^{-1}$  for  $\mu \in \mathbb{R}^k$ . Hence, the Rao score statistic is

$$R_n = [\mathbf{U}_{\eta}^{(n)}]^t [\mathbf{I}^{(n)}(\eta)]^{-1} \mathbf{U}_{\eta}^{(n)} = n \|\hat{\mu}^{(n)} - \eta\|_{\Sigma_0}^2.$$

Moreover, we have  $\mathbf{D}_g(\mu) = \mathbf{E}_k$ ,  $\mu \in \mathbb{R}^k$ , and the Wald statistic is

$$W_n = g(\hat{\mu}^{(n)})^t \mathbf{I}^{(n)}(\hat{\mu}^{(n)}) g(\hat{\mu}^{(n)}) = n \|\hat{\mu}^{(n)} - \eta\|_{\Sigma_0}^2.$$

Hence, in the actual case, all three test statistics  $\Lambda_n$ ,  $R_n$ , and  $W_n$  coincide and thus yield the same test.



In general, for finite sample size  $n \in \mathbb{N}$ , the distribution of  $\Lambda_n$  under the null hypothesis is not explicitly available. However, for sufficiently large  $n$ , we may employ the asymptotic distribution of  $\Lambda_n$  under  $H_0$  instead, which is easily derived by using the properties of the EF. Choosing the  $(1 - \alpha)$ -quantile of the asymptotic distribution as critical value for  $\Lambda_n$ , the actual level of the test  $\varphi_{\Lambda_n}$  will converge to the nominal level  $\alpha$ , when  $n$  tends to infinity.

For a simple null hypothesis, the asymptotic distribution of  $\Lambda_n$  can be obtained from a Taylor series expansion of the log-likelihood function and by using (asymptotic) properties of the MLE of  $\zeta$  (see, e.g., [54, pp. 151–156]).

**Theorem 5.19** *Let  $\mathcal{P}$  be a regular EF with minimal canonical representation (4.1), and let  $X^{(1)}, X^{(2)}, \dots$  be a sequence of iid random variables with distribution  $P_\zeta^* \in \mathcal{P}$ . Moreover, for eventually all  $n \in \mathbb{N}$ , let the MLE  $\hat{\zeta}^{(n)}$  of  $\zeta$  in  $\Xi^*$  based on  $X^{(1)}, \dots, X^{(n)}$  exist and  $\Lambda_n = \Lambda_n(X^{(1)}, \dots, X^{(n)})$  denote the test statistic of the LR test for test problem (5.23) based on  $X^{(1)}, \dots, X^{(n)}$ . Then, under  $H_0$ ,*

$$\Lambda_n \xrightarrow{d} \chi^2(k)$$

for  $n \rightarrow \infty$ , i.e.,  $\Lambda_n$  is asymptotically chi-square distributed with  $k$  degrees of freedom if  $n$  tends to infinity.

**Proof** Without loss of generality, let  $\hat{\zeta}^{(n)}$  exist for all  $n \in \mathbb{N}$ . Let  $\ell_n(\zeta) = \ln(\prod_{i=1}^n f_\zeta^*(X^{(i)}))$ ,  $\zeta \in \Xi^*$ , denote the log-likelihood function based on  $X^{(1)}, \dots, X^{(n)}$  for  $n \in \mathbb{N}$ . Moreover, suppose that  $H_0$  is true, i.e.,  $\zeta = \eta$ . Since  $\ell_n$  has a maximum at  $\hat{\zeta}^{(n)}$  and  $\Xi^*$  is open, it necessarily holds that  $\nabla \ell_n(\hat{\zeta}^{(n)}) = 0$  for all  $n \in \mathbb{N}$ . Hence, the Taylor series expansion of  $\ell_n$  about  $\hat{\zeta}^{(n)}$  at  $\zeta$  is given by

$$\ell_n(\zeta) = \ell_n(\hat{\zeta}^{(n)}) + \frac{1}{2} (\hat{\zeta}^{(n)} - \zeta)^t \mathbf{H}_{\ell_n}(\tilde{\zeta}^{(n)}) (\hat{\zeta}^{(n)} - \zeta),$$

where  $\tilde{\zeta}^{(n)}$  is a convex combination of  $\zeta$  and  $\hat{\zeta}^{(n)}$ ,  $n \in \mathbb{N}$ . For the test statistic  $\Lambda_n = \Lambda_n(X^{(1)}, \dots, X^{(n)})$  of the LR test based on  $X^{(1)}, \dots, X^{(n)}$ , we thus obtain

$$\begin{aligned} \Lambda_n &= -2(\ell_n(\zeta) - \ell_n(\hat{\zeta}^{(n)})) \\ &= \sqrt{n}(\hat{\zeta}^{(n)} - \zeta)^t \left[ -\frac{1}{n} \mathbf{H}_{\ell_n}(\tilde{\zeta}^{(n)}) \right] \sqrt{n}(\hat{\zeta}^{(n)} - \zeta). \end{aligned}$$

From Corollary 3.30 and Theorem 2.56, we have that

$$-\frac{1}{n} \mathbf{H}_{\ell_n}(\zeta) = \mathbf{H}_\kappa(\zeta) = \mathbf{I}(\zeta) > 0, \quad \zeta \in \Xi^*,$$

which allows for the representation

$$\Lambda_n = \|\mathbf{I}(\tilde{\boldsymbol{\zeta}}^{(n)})^{1/2} \sqrt{n}(\hat{\boldsymbol{\zeta}}^{(n)} - \boldsymbol{\zeta})\|_2^2, \quad (5.25)$$

with a matrix  $\mathbf{I}(\tilde{\boldsymbol{\zeta}}^{(n)})^{1/2} > 0$  satisfying  $\mathbf{I}(\tilde{\boldsymbol{\zeta}}^{(n)})^{1/2} \mathbf{I}(\tilde{\boldsymbol{\zeta}}^{(n)})^{1/2} = \mathbf{I}(\tilde{\boldsymbol{\zeta}}^{(n)})$  for  $n \in \mathbb{N}$ ; here, existence of such a matrix is guaranteed by the spectral theorem. From Theorem 4.31, we have that  $\hat{\boldsymbol{\zeta}}^{(n)} \rightarrow \boldsymbol{\zeta}$   $\mathbb{P}_{\boldsymbol{\zeta}}$ -a.s. and thus also  $\tilde{\boldsymbol{\zeta}}^{(n)} \rightarrow \boldsymbol{\zeta}$   $\mathbb{P}_{\boldsymbol{\zeta}}$ -a.s. for  $n \rightarrow \infty$ . Since the mapping  $\mathbf{I}(\cdot)^{1/2}$  is continuous on  $\Xi^*$ , it follows that  $\mathbf{I}(\tilde{\boldsymbol{\zeta}}^{(n)})^{1/2} \rightarrow \mathbf{I}(\boldsymbol{\zeta})^{1/2}$   $\mathbb{P}_{\boldsymbol{\zeta}}$ -a.s. for  $n \rightarrow \infty$ . Moreover, by Theorem 4.34,

$$\sqrt{n}(\hat{\boldsymbol{\zeta}}^{(n)} - \boldsymbol{\zeta}) \xrightarrow{d} N \sim \mathcal{N}_k(\mathbf{0}, [\mathbf{I}(\boldsymbol{\zeta})]^{-1}),$$

and we conclude from the multivariate Slutsky theorem that

$$\mathbf{I}(\tilde{\boldsymbol{\zeta}}^{(n)})^{1/2} \sqrt{n}(\hat{\boldsymbol{\zeta}}^{(n)} - \boldsymbol{\zeta}) \xrightarrow{d} \mathbf{I}(\boldsymbol{\zeta})^{1/2} N \sim \mathcal{N}_k(\mathbf{0}, \mathbf{E}_k),$$

where  $\mathbf{E}_k \in \mathbb{R}^{k \times k}$  denotes the  $k$ -dimensional unit matrix. In virtue of the continuous mapping  $\|\cdot\|_2^2$ , the assertion now follows from formula (5.25) together with the continuous mapping theorem (see, e.g., [18, Chapter 1, Section 5]).  $\square$

Theorem 5.19 is a special case of the following, more general result; its proof can be found, for instance, in [53, p. 240] and [54, pp. 156–160].

**Theorem 5.20** *In the situation of Theorem 5.19, let the test problem be replaced by test problem (5.21) with  $\Xi = \Xi^*$  and*

$$\Xi_0 = \{\boldsymbol{\zeta} \in \Xi^* : g(\boldsymbol{\zeta}) = \mathbf{0}\},$$

where  $g : \Xi^* \rightarrow \mathbb{R}^q$ ,  $q \leq k$ , is a differentiable function with Jacobian matrix  $\mathbf{D}_g(\boldsymbol{\zeta})$  of full rank  $q$  at every  $\boldsymbol{\zeta} \in \Xi^*$ . If the MLE  $\hat{\boldsymbol{\zeta}}_0^{(n)}$  of  $\boldsymbol{\zeta}$  in  $\Xi_0$  based on  $X^{(1)}, \dots, X^{(n)}$  exists for eventually all  $n \in \mathbb{N}$ , then

$$\Lambda_n \xrightarrow{d} \chi^2(q)$$

under  $H_0$  for  $n \rightarrow \infty$ .

The function  $g$  in Theorem 5.20 characterizes the null hypothesis in the same way as in Remark 5.17 for the Wald test. Setting  $g(\boldsymbol{\zeta}) = \boldsymbol{\zeta} - \boldsymbol{\eta}$ ,  $\boldsymbol{\zeta} \in \Xi^*$ , in Theorem 5.20 gives the result of Theorem 5.19. The choice

$$g(\boldsymbol{\zeta}) = (\zeta_{i_1} - \eta_{i_1}, \dots, \zeta_{i_q} - \eta_{i_q})^t, \quad \boldsymbol{\zeta} \in \Xi^*,$$

for fixed  $\boldsymbol{\eta} = (\eta_1, \dots, \eta_k)^t \in \Xi^*$  and  $i_1, \dots, i_q \in \{1, \dots, k\}$  with  $i_1 < \dots < i_q$  corresponds to the null hypothesis  $H_0 : \zeta_{i_j} = \eta_{i_j}$  for all  $1 \leq j \leq q$ , which may

therefore be used to check only for particular components of  $\xi$ . Moreover, we may test for homogeneity of the parameters, i.e., for  $H_0 : \zeta_1 = \cdots = \zeta_k$ , by defining, e.g.,

$$g(\xi) = (\zeta_2 - \zeta_1, \dots, \zeta_k - \zeta_{k-1})^t, \quad \xi \in \Xi^*,$$

in the case of which  $\Lambda_n \xrightarrow{d} \chi^2(k-1)$  under  $H_0$  for  $n \rightarrow \infty$ .

*Remark 5.21* Let the iid setup and the test problem in Theorem 5.20 be given. If the respective estimators  $\hat{\xi}_0^{(n)}$  and  $\hat{\xi}^{(n)}$  exist for eventually all  $n \in \mathbb{N}$ , the Rao score statistic  $R_n$  and the Wald statistic  $W_n$  (defined on  $\mathcal{X}^n$  according to Remark 5.17) are asymptotically  $\chi^2(q)$ -distributed under  $H_0$  for  $n \rightarrow \infty$ . For the proofs, see, e.g., [53, p. 240] and [54, pp. 156–160].

For further reading on testing hypotheses in EFs, we refer to, for example, [63, 20, 40, 42, 34].

# Chapter 6

## Exemplary Multivariate Applications



In this chapter, we demonstrate the usefulness of the proposed results for EFs by means of three multivariate examples of different kind. First, for the discrete case, the family of negative multinomial distributions is considered. Then, we turn to the continuous case and study properties of Dirichlet distributions. Finally, the results are applied to joint distributions of generalized order statistics, which form a unifying approach to various models of ordered random variables. In the joint density function of generalized order statistics, an absolutely continuous baseline distribution function can be arbitrarily chosen.

### 6.1 Negative Multinomial Distribution

The negative multinomial distribution is a natural multivariate extension of the negative binomial distribution (see, for example, [28, Chapter 36]). Suppose that we have an experiment with  $k + 1$  possible outcomes  $O_1, \dots, O_{k+1}$ , which is independently repeated until a given number  $r$  of a particular outcome  $O_{k+1}$ , say, is observed. Then, the negative multinomial distribution describes the probabilities of observing exactly  $x_i$  times outcome  $O_i$  for  $1 \leq i \leq k$  up to the  $r$ -th observation of  $O_{k+1}$ .

**Definition 6.1** The *negative multinomial distribution*  $\text{nm}(r, \mathbf{p})$  with  $r \in \mathbb{N}$  and  $\mathbf{p} = (p_1, \dots, p_k)^t \in (0, 1)^k$ ,  $k \in \mathbb{N}$ , satisfying  $\sum_{j=1}^k p_j < 1$  is defined

(continued)

**Definition 6.1** (continued)

as the probability distribution on  $\mathbb{N}_0^k$  with counting density

$$f_{\mathbf{p}}(\mathbf{x}) = \frac{(r - 1 + \sum_{j=1}^k x_j)!}{(r - 1)! \prod_{j=1}^k x_j!} \left( \prod_{j=1}^k p_j^{x_j} \right) \left( 1 - \sum_{j=1}^k p_j \right)^r, \quad (6.1)$$

for  $\mathbf{x} = (x_1, \dots, x_k)^t \in \mathbb{N}_0^k$ .

The negative multinomial distribution is a member of the more general family of multivariate power series distributions introduced in Remark 2.9.

Rewriting density function (6.1) as

$$f_{\mathbf{p}}(\mathbf{x}) = C(\mathbf{p}) \exp \left\{ \sum_{j=1}^k Z_j(\mathbf{p}) T_j(\mathbf{x}) \right\} h(\mathbf{x}), \quad \mathbf{x} = (x_1, \dots, x_k)^t \in \mathbb{N}_0^k,$$

with  $Z_j(\mathbf{p}) = \ln(p_j)$ ,  $1 \leq j \leq k$ , and

$$C(\mathbf{p}) = \left( 1 - \sum_{j=1}^k p_j \right)^r,$$

and  $T_j(\mathbf{x}) = x_j$ ,  $1 \leq j \leq k$ , and

$$h(\mathbf{x}) = \frac{(r - 1 + \sum_{j=1}^k x_j)!}{(r - 1)! \prod_{j=1}^k x_j!}, \quad \mathbf{x} = (x_1, \dots, x_k)^t \in \mathbb{N}_0^k, \quad (6.2)$$

we directly find that, for fixed  $r = r_0 \in \mathbb{N}$ ,  $\text{nm}(r_0, \bullet) = \{\text{nm}(r_0, \mathbf{p}) : \mathbf{p} \in \Theta\}$  forms an EF with parameter space

$$\Theta = \left\{ (p_1, \dots, p_k)^t \in (0, 1)^k : \sum_{j=1}^k p_j < 1 \right\}.$$

Aiming for inferential results for the negative multinomial distribution, we turn to a minimal canonical representation of  $\text{nm}(r_0, \bullet)$ .

**Theorem 6.2** *Let  $\mathcal{P} = \text{nm}(r_0, \bullet)$  for some fixed  $r_0 \in \mathbb{N}$  as introduced above. Then, we find:*

(a)  $\text{ord}(\mathcal{P}) = k$ .

(b) The natural parameter space of  $\mathcal{P}$  is given by

$$\Xi^* = \left\{ \boldsymbol{\zeta} = (\zeta_1, \dots, \zeta_k)^t \in (-\infty, 0)^k : \sum_{j=1}^k e^{\zeta_j} < 1 \right\}.$$

(c)  $\mathcal{P}$  is regular with minimal canonical representation given by the counting densities

$$f_{\boldsymbol{\zeta}}^*(\mathbf{x}) = \frac{dP_{\boldsymbol{\zeta}}^*}{d\mu_{\mathbb{N}_0^k}}(\mathbf{x}) = e^{\boldsymbol{\zeta}^t \mathbf{T}(\mathbf{x}) - \kappa(\boldsymbol{\zeta})} h(\mathbf{x}), \quad \mathbf{x} \in \mathbb{N}_0^k, \quad (6.3)$$

for  $\boldsymbol{\zeta} \in \Xi^*$ , where  $\mathbf{T}(\mathbf{x}) = \mathbf{x}$ ,  $\mathbf{x} \in \mathbb{N}_0^k$ ,  $h$  is as in formula (6.2), and the cumulant function is given by

$$\kappa(\boldsymbol{\zeta}) = -r_0 \ln \left( 1 - \sum_{j=1}^k e^{\zeta_j} \right), \quad \boldsymbol{\zeta} = (\zeta_1, \dots, \zeta_k)^t \in \Xi^*.$$

### Proof

- (a) Let  $\mathbf{Z} = (Z_1, \dots, Z_k)^t$  and  $\mathbf{T} = (T_1, \dots, T_k)^t$  with  $Z_j(\mathbf{p}) = \ln(p_j)$ ,  $\mathbf{p} \in \Theta$ , and  $T_j(\mathbf{x}) = x_j$ ,  $\mathbf{x} \in \mathbb{N}_0^k$ , for  $1 \leq j \leq k$ . Moreover, let  $\mathbf{e}_j$  denote the  $j$ -th unit vector in  $\mathbb{R}^k$ ,  $1 \leq j \leq k$ . Since  $\mathbf{T}(\mathbf{e}_j) - \mathbf{T}(\mathbf{0}) = \mathbf{e}_j$ ,  $1 \leq j \leq k$ , are linearly independent in  $\mathbb{R}^k$ ,  $T_1, \dots, T_k$  are affinely independent by Lemma 2.22. Setting  $\mathbf{p}^{(0)} = (\exp\{-k\}, \dots, \exp\{-k\})^t$  and  $\mathbf{p}^{(j)} = (p_1^{(j)}, \dots, p_k^{(j)})^t$  with  $p_i^{(j)} = \exp\{-k\}$ ,  $i \neq j$ , and  $p_j^{(j)} = \exp\{-k-1\}$  for  $1 \leq i, j \leq k$ , we find that  $\mathbf{Z}(\mathbf{p}^{(j)}) - \mathbf{Z}(\mathbf{p}^{(0)}) = -\mathbf{e}_j$ ,  $1 \leq j \leq k$ , are linearly independent in  $\mathbb{R}^k$ . Again, by Lemma 2.22,  $Z_1, \dots, Z_k$  are affinely independent. The assertion then follows from Theorem 2.23.
- (b) Let  $\boldsymbol{\zeta} = (\zeta_1, \dots, \zeta_k)^t \in \mathbb{R}^k$ . Then,

$$\begin{aligned} \sum_{\mathbf{x} \in \mathbb{N}_0^k} \exp \left\{ \sum_{j=1}^k \zeta_j x_j \right\} h(\mathbf{x}) &= \sum_{n=0}^{\infty} \sum_{\mathbf{x} \in \mathbb{N}_0^k : \sum_{j=1}^k x_j = n} \exp \left\{ \sum_{j=1}^k \zeta_j x_j \right\} h(\mathbf{x}) \\ &= \sum_{n=0}^{\infty} \frac{(n+r_0-1)!}{(r_0-1)! n!} \sum_{\mathbf{x} \in \mathbb{N}_0^k : \sum_{j=1}^k x_j = n} \frac{n!}{\prod_{j=1}^k x_j!} \prod_{j=1}^k (e^{\zeta_j})^{x_j} \\ &= \sum_{n=0}^{\infty} \binom{n+r_0-1}{r_0-1} \left( \sum_{j=1}^k e^{\zeta_j} \right)^n, \end{aligned}$$

where the last equation follows from the multinomial theorem. By D'Alembert's criterion, this series converges if  $\sum_{j=1}^k \exp\{\zeta_j\} < 1$  and diverges if  $\sum_{j=1}^k \exp\{\zeta_j\} \geq 1$ . This gives  $\Xi^*$  as stated.

- (c) From assertion (b), we have that  $\mathbf{Z}(\Theta) = \Xi^*$  and  $\Xi^*$  is open, such that  $\mathcal{P}$  is regular. By setting  $\zeta_j = \ln(p_j)$ ,  $1 \leq j \leq k$ , in density function  $f_{\mathbf{p}}$ , the canonical representation is obtained, which is minimal according to statement (a).  $\square$

Utilizing this finding, the mean vector and covariance matrix of the negative multinomial distribution are directly obtained.

**Theorem 6.3** *Let  $\mathcal{P} = \text{nm}(r_0, \bullet)$  for some fixed  $r_0 \in \mathbb{N}$  with minimal canonical representation (6.3). Then, the mean vector and the covariance matrix of  $\mathbf{X} \sim P_{\boldsymbol{\zeta}}^*$  are given by*

$$\mathbb{E}_{\boldsymbol{\zeta}}(\mathbf{X}) = \left( \frac{r_0 e^{\zeta_1}}{1 - \sum_{j=1}^k e^{\zeta_j}}, \dots, \frac{r_0 e^{\zeta_k}}{1 - \sum_{j=1}^k e^{\zeta_j}} \right)^t \quad (6.4)$$

and

$$\mathbb{C}_{\text{ov}}_{\boldsymbol{\zeta}}(X_i, X_j) = \frac{r_0}{1 - \sum_{l=1}^k e^{\zeta_l}} \left( \frac{e^{\zeta_i + \zeta_j}}{1 - \sum_{l=1}^k e^{\zeta_l}} + \delta_{ij} e^{\zeta_i} \right), \quad 1 \leq i, j \leq k, \quad (6.5)$$

where  $\delta_{ij}$  is the Kronecker delta, i.e.,  $\delta_{ij} = 1$  for  $i = j$ , and  $\delta_{ij} = 0$  for  $i \neq j$ .

**Proof** Applying Corollary 2.50, we have  $\mathbb{E}_{\boldsymbol{\zeta}}(\mathbf{X}) = E_{\boldsymbol{\zeta}}[\mathbf{T}] = \nabla \kappa(\boldsymbol{\zeta})$  and  $\mathbb{C}_{\text{ov}}_{\boldsymbol{\zeta}}(\mathbf{X}) = \text{Cov}_{\boldsymbol{\zeta}}(\mathbf{T}) = \mathbf{H}_{\kappa}(\boldsymbol{\zeta})$  for  $\boldsymbol{\zeta} \in \Xi^*$ , which leads to the stated formulas.  $\square$

In terms of the original parameters  $p_1, \dots, p_k$ , the mean vector and covariance matrix of  $\text{nm}(r_0, \mathbf{p})$  can be obtained by setting  $\zeta_j = \ln(p_j)$ ,  $1 \leq j \leq k$ , in formulas (6.4) and (6.5).

Noticing that one-dimensional marginals are negative binomial distributions (see [28, Chapter 36]), the expressions in Theorem 6.3 for one-dimensional marginal expectations and variances are known from Example 2.51. More precisely, the parameters of the  $j$ -th marginal distribution are  $r_0$  and  $(1 - \sum_{i=1}^n p_i)/(1 - \sum_{i=1, i \neq j}^n p_i)$ .

As two by-products of Theorem 6.3, the mean value function of  $\mathcal{P}$  turns out to be

$$\pi(\boldsymbol{\zeta}) = \left( \frac{r_0 e^{\zeta_1}}{1 - \sum_{j=1}^k e^{\zeta_j}}, \dots, \frac{r_0 e^{\zeta_k}}{1 - \sum_{j=1}^k e^{\zeta_j}} \right)^t, \quad \boldsymbol{\zeta} \in \Xi^*, \quad (6.6)$$

and the Fisher information matrix of  $\mathcal{P}$  at  $\zeta \in \Xi^*$  is given by

$$[\mathbf{I}(\zeta)]_{i,j} = \frac{r_0}{1 - \sum_{l=1}^k e^{\zeta_l}} \left( \frac{e^{\zeta_i + \zeta_j}}{1 - \sum_{l=1}^k e^{\zeta_l}} + \delta_{ij} e^{\zeta_i} \right), \quad 1 \leq i, j \leq k \quad (6.7)$$

(see Theorem 2.56 and Theorem 3.29).

Let us now develop inferential results for the negative multinomial distribution. For simplicity, we restrict ourselves to the case of a single observation from  $P_\zeta^*$ . First, maximum likelihood estimation of the natural parameters  $\zeta_1, \dots, \zeta_k$  is considered.

**Theorem 6.4** *Let  $\mathcal{P} = \text{nm}(r_0, \bullet)$  for some fixed  $r_0 \in \mathbb{N}$  with minimal canonical representation (6.3), and let  $\mathbf{x} \in \mathbb{N}_0^k$  be a realization of  $\mathbf{X} \sim P_\zeta^*$  for some  $\zeta \in \Xi^*$ . Then, an ML estimate of  $\zeta$  in  $\Xi^*$  based on  $\mathbf{x}$  exists if and only if  $\mathbf{x} \in (0, \infty)^k$ . If it exists, it is uniquely determined and given by*

$$\left( \ln \left( \frac{x_1}{r_0 + \sum_{i=1}^k x_i} \right), \dots, \ln \left( \frac{x_k}{r_0 + \sum_{i=1}^k x_i} \right) \right)^t.$$

**Proof** From Theorem 4.3, we obtain that an ML estimate of  $\zeta$  in  $\Xi^*$  based on  $\mathbf{x}$  exists if and only if  $\mathbf{x} \in \text{int}(M)$ , where  $M = [0, \infty)^k$  is the convex support of  $\mathbf{X}$ . Hence, the first statement is clear. Now, let us compute the ML estimate  $\pi^{-1}(\mathbf{x})$  of  $\zeta$  for  $\mathbf{x} \in (0, \infty)^k$ , where  $\pi^{-1}$  is the inverse function of  $\pi : \Xi^* \rightarrow (0, \infty)^k$  in formula (6.6). Suppose that  $\pi(\zeta) = \mathbf{x}$  for some  $\zeta \in \Xi^*$ . Summing up all  $k$  equations of this system then gives

$$\frac{r_0 \sum_{j=1}^k e^{\zeta_j}}{1 - \sum_{i=1}^k e^{\zeta_i}} = \sum_{j=1}^k x_j,$$

which is equivalent to

$$\sum_{i=1}^k e^{\zeta_i} = \frac{\sum_{j=1}^k x_j}{r_0 + \sum_{j=1}^k x_j}.$$

Thus, for  $1 \leq j \leq r$ , since  $r_0 \exp\{\zeta_j\} / (1 - \sum_{i=1}^k \exp\{\zeta_i\}) = x_j$ , we find

$$\zeta_j = \ln \left( \frac{x_j}{r_0} \left( 1 - \sum_{i=1}^k e^{\zeta_i} \right) \right) = \ln \left( \frac{x_j}{r_0 + \sum_{i=1}^k x_i} \right).$$

□



In particular, Theorem 6.4 yields that the probability of non-existence of an ML estimate based on a single observation of  $\mathbf{X}$  is positive, since

$$\mathbb{P}_\zeta(\mathbf{X} \notin (0, \infty)^k) \geq \mathbb{P}_\zeta(\mathbf{X} = \mathbf{0}) = \left(1 - \sum_{j=1}^k e^{\zeta_j}\right)^{r_0} > 0.$$

If an estimator for the mean vector of  $\text{nm}(r_0, \mathbf{p})$  is desired, then one may choose  $\mathbf{X}$  itself.

**Theorem 6.5** *Under the assumptions of Theorem 6.4,  $\mathbf{X}$  is an efficient estimator of  $\pi$ , i.e., its covariance matrix attains the lower bound of the Cramér-Rao inequality given by*

$$[\mathbf{I}(\zeta)^{-1}]_{i,j} = \frac{1 - \sum_{l=1}^k e^{\zeta_l}}{r_0} (e^{-\zeta_i} \delta_{ij} - 1), \quad 1 \leq i, j \leq k.$$

**Proof**  $\mathbf{X}$  is an efficient estimator of  $\pi$  by Theorem 4.26. To find the inverse of  $\mathbf{I}(\zeta)$  in formula (6.7), note that

$$\mathbf{I}(\zeta) = \frac{r_0}{a} \left( \mathbf{P} + \frac{\mathbf{p}\mathbf{p}^t}{a} \right)$$

for the diagonal matrix  $\mathbf{P}$  with positive diagonal entries  $p_j = e^{\zeta_j}$ ,  $1 \leq j \leq k$ , positive number  $a = 1 - \sum_{j=1}^k p_j$ , and vector  $\mathbf{p} = (p_1, \dots, p_k)^t$ . A well known matrix identity (see, e.g., [45, p. 458]) then yields that

$$\begin{aligned} \mathbf{I}(\zeta)^{-1} &= \frac{a}{r_0} \left( \mathbf{P}^{-1} - \mathbf{P}^{-1} \mathbf{p} \left( a + \mathbf{p}^t \mathbf{P}^{-1} \mathbf{p} \right)^{-1} \mathbf{p}^t \mathbf{P}^{-1} \right) \\ &= \frac{a}{r_0} \left( \mathbf{P}^{-1} - (1, \dots, 1)^t (1, \dots, 1) \right), \end{aligned}$$

noticing that  $\mathbf{p}^t \mathbf{P}^{-1} \mathbf{p} = \sum_{j=1}^k p_j$ . Inserting for  $p_1, \dots, p_k$  and  $a$  completes the proof.  $\square$

Next, suppose we are interested to statistically confirm that the probability  $p_1$  of outcome  $O_1$  in a single experiment exceeds a certain threshold  $p_0 \in (0, 1)$ , say. The corresponding test problem is

$$H_0 : p_1 \leq p_0 \quad \leftrightarrow \quad H_1 : p_1 > p_0 \quad (6.8)$$

or, equivalently and in terms of the natural parameter  $\zeta_1$ ,

$$H_0 : \zeta_1 \leq \zeta_0 \quad \leftrightarrow \quad H_1 : \zeta_1 > \zeta_0 \quad (6.9)$$

with  $\zeta_0 = \ln(p_0)$ . An optimal test for these test problems can be obtained from Theorem 5.12(a) by using that the conditional distribution of  $X_1$  given  $(X_2, \dots, X_k) = \mathbf{y}$  is a negative binomial distribution with respective parameters.

**Theorem 6.6** *Let the assumptions of Theorem 6.4 be given, and let  $\alpha \in (0, 1)$ . Then, for test problems (6.8) and (6.9),*

$$\varphi(\mathbf{x}) = \mathbb{1}_{(c(\mathbf{y}), \infty)}(x_1) + \gamma(\mathbf{y})\mathbb{1}_{\{c(\mathbf{y})\}}(x_1), \quad \mathbf{x} = (x_1, \mathbf{y}) \in \mathbb{N}_0^k,$$

*forms a UMPU level- $\alpha$  test, where, for  $\mathbf{y} = (x_2, \dots, x_k) \in \mathbb{N}_0^{k-1}$ ,  $c(\mathbf{y})$  is the minimum number  $m \in \mathbb{N}_0$  satisfying*

$$\sum_{i=m+1}^{\infty} g_{\mathbf{y}}(i) \leq \alpha,$$

*and*

$$\gamma(\mathbf{y}) = \frac{\alpha - \sum_{i=c(\mathbf{y})+1}^{\infty} g_{\mathbf{y}}(i)}{g_{\mathbf{y}}(c(\mathbf{y}))},$$

*where  $g_{\mathbf{y}}$  denotes the counting density of  $\text{nb}(r_0 + \sum_{i=2}^k x_i, 1 - e^{\zeta_0})$  defined by Example 2.5.*

Let us now consider the multivariate test problem

$$H_0 : \mathbf{p} = \mathbf{q} \quad \leftrightarrow \quad H_1 : \mathbf{p} \neq \mathbf{q} \quad (6.10)$$

with a simple null hypothesis and fixed  $\mathbf{q} = (q_1, \dots, q_k)^t \in \Theta$ , or, equivalently and in terms of the natural parameter  $\boldsymbol{\zeta}$ ,

$$H_0 : \boldsymbol{\zeta} = \boldsymbol{\eta} \quad \leftrightarrow \quad H_1 : \boldsymbol{\zeta} \neq \boldsymbol{\eta} \quad (6.11)$$

with  $\boldsymbol{\eta} = (\eta_1, \dots, \eta_k)^t \in \Xi^*$  defined by  $\eta_j = \ln(q_j)$ ,  $1 \leq j \leq k$ . Since the MLE of  $\boldsymbol{\zeta}$  in  $\Xi^*$  based on  $\mathbf{X}$  does not exist a.s., we focus on the Rao score test, which, in contrast to the LR test and the Wald test, does not depend on the MLE in  $\Xi^*$ .

**Theorem 6.7** *Under the assumptions of Theorem 6.4, the test statistic of the Rao score test for test problems (6.10) and (6.11) is given by*

$$R(\mathbf{x}) = \frac{1 - \sum_{i=1}^k e^{\eta_i}}{r_0} \left\{ \sum_{j=1}^k e^{-\eta_j} \left( x_j - \frac{r_0 e^{\eta_j}}{1 - \sum_{i=1}^k e^{\eta_i}} \right)^2 - \left[ \sum_{j=1}^k \left( x_j - \frac{r_0 e^{\eta_j}}{1 - \sum_{i=1}^k e^{\eta_i}} \right) \right]^2 \right\}$$

for  $\mathbf{x} = (x_1, \dots, x_k)^t \in \mathbb{N}_0^k$ .

**Proof** By using the same notation as in the proof of Theorem 6.5, we have

$$U_\zeta^t \mathbf{I}(\zeta)^{-1} U_\zeta = \frac{a}{r_0} \left\{ U_\zeta^t \mathbf{P}^{-1} U_\zeta - [(1, \dots, 1) U_\zeta]^2 \right\}$$

Moreover, we conclude from Theorem 3.29 together with formula (6.4) that the score statistic at  $\zeta$  is given by  $U_\zeta(\mathbf{x}) = \mathbf{x} - r_0 \mathbf{p}/a$ ,  $\mathbf{x} \in \mathbb{N}_0^k$ . Inserting for  $a$ ,  $p_1, \dots, p_k$ , and replacing  $\zeta$  by  $\eta$  then gives the representation for  $R$  defined in Remark 5.17.  $\square$

The Rao score test then rejects the null hypothesis if  $R$  in Theorem 6.7 exceeds a certain critical value. Since  $\mathbf{X}$  is discrete so is  $R$ , and to guarantee that the test has exact level  $\alpha \in (0, 1)$  therefore requires a randomized test based on  $R$ , in general. A non-randomized level- $\alpha$  test, which rejects  $H_0$  if and only if  $R$  exceeds the critical value, can be obtained if the critical value is determined as the  $(1 - \alpha)$ -quantile  $r_{1-\alpha}$ , say, of  $R$  under  $H_0$ , i.e., if  $\zeta = \eta$  is true. By doing so, the exact level of the test may be smaller than  $\alpha$ , in the case of which the test is also called *conservative*. To compute a numerical value for  $r_{1-\alpha}$ , Monte-Carlo simulation can be used, where a sufficiently large number of realizations of  $R$  are generated under the assumption that  $\zeta = \eta$ , and then the empirical  $(1 - \alpha)$ -quantile of the resulting data vector is chosen for  $r_{1-\alpha}$ . Here, note that the convergence of the empirical  $(1 - \alpha)$ -quantile to the true  $(1 - \alpha)$ -quantile for increasing simulation size is only guaranteed if the quantile function of  $R$  is continuous at  $1 - \alpha$  (see, e.g., [58, pp. 304/305]).

Finally, inverting the acceptance region of the test also yields a confidence region for  $\zeta$ , i.e.,

$$C(\mathbf{x}) = \{ \zeta \in \Xi^* : R(\mathbf{x}) \leq r_{1-\alpha} \}, \quad \mathbf{x} \in \mathcal{X},$$

with (at least) level  $1 - \alpha$ . A confidence region for the original parameter  $\mathbf{p}$  with (at least) level  $1 - \alpha$  is then given by

$$\tilde{C}(\mathbf{x}) = \left\{ (e^{\zeta_1}, \dots, e^{\zeta_k})^t : (\zeta_1, \dots, \zeta_k)^t \in C(\mathbf{x}) \right\}, \quad \mathbf{x} \in \mathcal{X}.$$

## 6.2 Dirichlet Distribution

The Dirichlet distribution is a multivariate extension of the beta distribution, and it is widely used in the literature (see, e.g., [37, Chapter 49] and [47]).

**Definition 6.8** The *Dirichlet distribution*  $\mathcal{D}(\boldsymbol{\beta})$  with parameter vector  $\boldsymbol{\beta} = (\beta_1, \dots, \beta_{k+1})^t \in (0, \infty)^{k+1}$ ,  $k \in \mathbb{N}$ , is defined as the distribution on the Borel sets of  $\mathcal{X} = \{(y_1, \dots, y_k)^t \in (0, 1)^k : \sum_{i=1}^k y_i < 1\}$  with Lebesgue density function

$$f_{\boldsymbol{\beta}}(\mathbf{x}) = \frac{1}{B(\boldsymbol{\beta})} \left( \prod_{j=1}^k x_j^{\beta_j-1} \right) \left( 1 - \sum_{j=1}^k x_j \right)^{\beta_{k+1}-1} \quad (6.12)$$

for  $\mathbf{x} = (x_1, \dots, x_k)^t \in \mathcal{X}$ , where  $B$  denotes the beta function, i.e.,

$$B(\boldsymbol{\beta}) = \frac{\prod_{i=1}^{k+1} \Gamma(\beta_i)}{\Gamma(\sum_{i=1}^{k+1} \beta_i)}, \quad \boldsymbol{\beta} \in (0, \infty)^{k+1}. \quad (6.13)$$

By setting  $k = 1$  in Definition 6.8,  $\mathcal{D}(\beta_1, \beta_2)$  coincides with the beta distribution  $\mathcal{B}(\beta_1, \beta_2)$ ,  $\beta_1, \beta_2 \in (0, \infty)$ .

According to Example 2.13, we may rewrite density function (6.12) as

$$f_{\boldsymbol{\beta}}(\mathbf{x}) = \frac{1}{B(\boldsymbol{\beta})} \exp \left\{ \sum_{j=1}^{k+1} \beta_j T_j(\mathbf{x}) \right\} h(\mathbf{x}) \quad (6.14)$$

with statistics

$$T_j(\mathbf{x}) = \ln(x_j), \quad 1 \leq j \leq k, \quad T_{k+1}(\mathbf{x}) = \ln \left( 1 - \sum_{i=1}^k x_i \right), \quad (6.15)$$

and function

$$h(\mathbf{x}) = \left[ \left( 1 - \sum_{j=1}^k x_j \right) \prod_{j=1}^k x_j \right]^{-1} \quad (6.16)$$

for  $\mathbf{x} \in \mathcal{X}$ . Hence,  $\mathcal{P} = \mathcal{D}(\bullet) = \{\mathcal{D}(\boldsymbol{\beta}) : \boldsymbol{\beta} \in (0, \infty)^{k+1}\}$  forms an EF. Note that formula (6.14) is already a canonical representation of  $\mathcal{P}$ .

Some properties of  $\mathcal{D}(\bullet)$  are stated in the following theorem, where the proof of assertion (a) is technical and therefore omitted.

**Theorem 6.9** *For  $\mathcal{P} = \mathcal{D}(\bullet)$  as introduced above, we find:*

- (a)  $\text{ord}(\mathcal{P}) = k + 1$ .
- (b) The natural parameter space of  $\mathcal{P}$  is given by  $\Xi^* = (0, \infty)^{k+1}$ .
- (c)  $\mathcal{P}$  is regular with minimal canonical representation given by the Lebesgue density functions

$$f_{\boldsymbol{\beta}}(\mathbf{x}) = \frac{dP_{\boldsymbol{\beta}}}{d\lambda_{\mathcal{X}}}(\mathbf{x}) = e^{\boldsymbol{\beta}^t \mathbf{T}(\mathbf{x}) - \kappa(\boldsymbol{\beta})} h(\mathbf{x}), \quad \mathbf{x} \in \mathcal{X}, \quad (6.17)$$

for  $\boldsymbol{\beta} \in (0, \infty)^{k+1}$ , where  $\mathbf{T} = (T_1, \dots, T_{k+1})^t$  with  $T_j$ ,  $1 \leq j \leq k+1$ , as in formula (6.15),  $h$  as in formula (6.16), and cumulant function  $\kappa = \ln(B)$  on  $(0, \infty)^{k+1}$  with  $B$  as in formula (6.13).

**Proof**

- (b) Evidently, we have  $(0, \infty)^{k+1} \subset \Xi^*$ . Let  $\beta_2, \dots, \beta_{k+1} \in [0, \infty)$  and note that  $h(\mathbf{x}) \geq 1/x_1$  for  $\mathbf{x} \in \mathcal{X}$ . By introducing the  $k$ -dimensional rectangle  $\mathcal{Y} = I \times J^{k-1} \subset \mathcal{X}$  with  $I = (0, 1/(k+1))$  and  $J = (1/(k+2), 1/(k+1))$ , we have for  $\mathbf{x} \in \mathcal{Y}$  that

$$\begin{aligned} \exp \left\{ \sum_{j=2}^{k+1} \beta_j T_j(\mathbf{x}) \right\} h(\mathbf{x}) &\geq \left( \prod_{j=2}^k x_j^{\beta_j} \right) \left( 1 - \sum_{j=1}^k x_j \right)^{\beta_{k+1}} \frac{1}{x_1} \\ &\geq \left( \prod_{j=2}^k \left( \frac{1}{k+2} \right)^{\beta_j} \right) \left( 1 - \frac{k}{k+1} \right)^{\beta_{k+1}} \frac{1}{x_1} = \frac{d}{x_1}, \end{aligned}$$

say, where  $d > 0$ . Hence,

$$\begin{aligned} \int_{\mathcal{X}} \exp \left\{ \sum_{j=2}^{k+1} \beta_j T_j \right\} h d\lambda_{\mathcal{X}} &\geq \int_{\mathcal{Y}} \exp \left\{ \sum_{j=2}^{k+1} \beta_j T_j \right\} h d\lambda_{\mathcal{X}} \geq d \int_{\mathcal{Y}} \frac{1}{x_1} d\lambda_{\mathcal{X}}(\mathbf{x}) \\ &= d \left( \frac{1}{k+1} - \frac{1}{k+2} \right)^{k-1} \int_0^{1/(k+1)} \frac{1}{x_1} dx_1 = \infty. \end{aligned}$$

Hence,  $(0, \beta_2, \dots, \beta_{k+1})^t \notin \Xi^*$  for  $\beta_2, \dots, \beta_{k+1} \in [0, \infty)$ . For symmetry reasons, we even have that  $(\beta_1, \dots, \beta_{k+1})^t \notin \Xi^*$  for  $\beta_i \in [0, \infty)$ ,  $1 \leq i \leq k+1$ ,

if  $\beta_j = 0$  for some  $j \in \{1, \dots, k+1\}$ . Since  $\Xi^*$  is convex by Theorem 2.34, it follows that  $\Xi^* = (0, \infty)^{k+1}$ .

(c) The assertion follows directly from (a) and (b).  $\square$

The mean vector and the covariance matrix of  $\mathbf{T}$  are now readily derived.

**Theorem 6.10** *Let  $\mathcal{P} = \mathcal{D}(\bullet)$  with minimal canonical representation (6.17). Then, the mean vector and the covariance matrix of  $\mathbf{T}$  under  $P_\beta$  are given by*

$$E_\beta(\mathbf{T}) = (\psi(\beta_1) - \psi(\beta_\bullet), \dots, \psi(\beta_{k+1}) - \psi(\beta_\bullet))^t$$

$$\text{and} \quad \text{Cov}_\beta(T_i, T_j) = \psi'(\beta_i) \delta_{ij} - \psi'(\beta_\bullet), \quad 1 \leq i, j \leq k+1,$$

where  $\psi = [\ln(\Gamma)]' = \Gamma'/\Gamma$  denotes the digamma function,  $\delta_{ij}$  is the Kronecker delta, i.e.,  $\delta_{ij} = 1$  for  $i = j$ , and  $\delta_{ij} = 0$  for  $i \neq j$ , and the notation  $\beta_\bullet = \sum_{i=1}^{k+1} \beta_i$  is used for  $\beta \in (0, \infty)^{k+1}$ .

**Proof** Applying Corollary 2.50, we have for  $1 \leq j \leq k+1$  that  $E_\beta(T_j) = \partial/\partial\beta_j \kappa(\beta) = [\partial/\partial\beta_j B(\beta)]/B(\beta)$  and  $\text{Cov}_\beta(\mathbf{T}) = \mathbf{H}_\kappa(\beta)$ , from which the assertions follow by elementary calculus.  $\square$

As two by-products of Theorem 6.10, the mean value function of  $\mathcal{P}$  turns out to be

$$\pi(\beta) = (\psi(\beta_1) - \psi(\beta_\bullet), \dots, \psi(\beta_{k+1}) - \psi(\beta_\bullet))^t, \quad \beta \in (0, \infty)^{k+1}, \quad (6.18)$$

and the Fisher information matrix of  $\mathcal{P}$  at  $\beta \in (0, \infty)^{k+1}$  is given by

$$[\mathbf{I}(\beta)]_{i,j} = \psi'(\beta_i) \delta_{ij} - \psi'(\beta_\bullet), \quad 1 \leq i, j \leq k+1 \quad (6.19)$$

(see Theorems 2.56 and 3.29).

Now, we turn to inferential results for the Dirichlet distribution based on an iid sample from  $P_\beta$ . We start with maximum likelihood estimation of the natural parameters  $\beta_1, \dots, \beta_{k+1}$ . To apply the results of Sect. 4.1, note first that the boundary of the convex support of  $\mathbf{T}$  is nothing but the support  $\{\mathbf{T}(\mathbf{x}) : \mathbf{x} \in \mathcal{X}\}$  of  $\mathbf{T}$ , i.e.,

$$\partial M = \left\{ \left( \ln(x_1), \dots, \ln(x_k), \ln\left(1 - \sum_{j=1}^k x_j\right) \right)^t : \mathbf{x} \in (0, 1)^k, \sum_{j=1}^k x_j < 1 \right\} \quad (6.20)$$

Theorem 4.3 then yields that an ML estimate based on a single observation  $\mathbf{x}$  from  $P_\beta$  does not exist, since  $\mathbf{T}(\mathbf{x}) \in \partial M$   $P_\beta$ -a.s.. However, for a sample size larger than 1, we have the following theorem (see also [20, pp. 150–152] for the particular case of the beta distribution).

**Theorem 6.11** Let  $\mathcal{P} = \mathcal{D}(\bullet)$  with minimal canonical representation (6.17), and let  $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)} \in \mathcal{X}$  be realizations of iid random vectors  $\mathbf{X}^{(1)}, \dots, \mathbf{X}^{(n)}$  with distribution  $P_{\boldsymbol{\beta}}$  for some  $\boldsymbol{\beta} \in (0, \infty)^{k+1}$ . If  $n \geq 2$  and  $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}$  are not all identical, then an ML estimate of  $\boldsymbol{\beta}$  in  $(0, \infty)^{k+1}$  based on  $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}$  uniquely exists and is given by the only solution of the equations

$$\psi(\beta_j) - \psi(\beta_{\bullet}) = \frac{1}{n} \sum_{i=1}^n \ln(x_j^{(i)}), \quad 1 \leq j \leq k,$$

$$\text{and} \quad \psi(\beta_{k+1}) - \psi(\beta_{\bullet}) = \frac{1}{n} \sum_{i=1}^n \ln \left( 1 - \sum_{l=1}^k x_l^{(i)} \right)$$

with respect to  $\boldsymbol{\beta} \in (0, \infty)^{k+1}$ . Moreover, the MLE  $\hat{\boldsymbol{\beta}}^{(n)}$  of  $\boldsymbol{\beta}$  in  $(0, \infty)^{k+1}$  based on  $\mathbf{X}^{(1)}, \dots, \mathbf{X}^{(n)}$  uniquely exists (a.s.) for all  $n \geq 2$ .

**Proof** Let  $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)} \in \mathcal{X}$  with  $\mathbf{x}^{(i)} = (x_1^{(i)}, \dots, x_k^{(i)})^t$ ,  $1 \leq i \leq n$ , be not all identical. By Corollary 4.5 and Remark 4.4, we have to show that

$$\frac{1}{n} \sum_{i=1}^n \mathbf{T}(\mathbf{x}^{(i)}) \notin \partial M,$$

where  $\partial M$  is given by formula (6.20). For this, suppose that there exists some  $\mathbf{y} = (y_1, \dots, y_k)^t \in \mathcal{X}$  with

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n T_j(\mathbf{x}^{(i)}) &= \frac{1}{n} \sum_{i=1}^n \ln(x_j^{(i)}) \\ &= \ln(y_j), \quad 1 \leq j \leq k. \end{aligned}$$

Then, two-fold application of the inequality of arithmetic and geometric mean yields that

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n T_{k+1}(\mathbf{x}^{(i)}) &= \frac{1}{n} \sum_{i=1}^n \ln \left( 1 - \sum_{j=1}^k x_j^{(i)} \right) = \ln \left( \left[ \prod_{i=1}^n \left( 1 - \sum_{j=1}^k x_j^{(i)} \right) \right]^{1/n} \right) \\ &\leq \ln \left( \frac{1}{n} \sum_{i=1}^n \left[ 1 - \sum_{j=1}^k x_j^{(i)} \right] \right) = \ln \left( 1 - \sum_{j=1}^k \frac{1}{n} \sum_{i=1}^n x_j^{(i)} \right) \\ &< \ln \left( 1 - \sum_{j=1}^k \left[ \prod_{i=1}^n x_j^{(i)} \right]^{1/n} \right) = \ln \left( 1 - \sum_{j=1}^k y_j \right), \end{aligned}$$

i.e.,  $\sum_{i=1}^n \mathbf{T}(\mathbf{x}^{(i)})/n \notin \partial M$ , and an ML estimate of  $\boldsymbol{\beta}$  based on  $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}$  uniquely exists. Since  $\mathbf{x}^{(1)} = \dots = \mathbf{x}^{(n)}$  occurs with  $P_{\boldsymbol{\beta}}$ -probability 0, the proof is done.  $\square$

For  $n \geq 2$ , the MLE  $\hat{\boldsymbol{\beta}}^{(n)}$  of  $\boldsymbol{\beta}$  in Theorem 6.11 cannot be stated explicitly, and an ML estimate based on (non-identical) realizations  $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}$  has to be obtained via numerical procedures (see [37, pp. 505/506] and also [27, pp. 223–225] in case of the beta distribution). In contrast, we have the explicit MLE  $\sum_{i=1}^n \mathbf{T}(\mathbf{X}^{(i)})/n$  of  $\pi(\boldsymbol{\beta})$  based on  $\mathbf{X}^{(1)}, \dots, \mathbf{X}^{(n)}$ , which is moreover efficient by Corollary 4.27.

Since the MLE of the natural parameter  $\boldsymbol{\beta}$  is not available in explicit form, we cannot derive its exact distribution analytically, which however is required, e.g., for the construction of statistical tests and confidence regions for  $\boldsymbol{\beta}$ . We therefore now assume to have a sequence of iid random vectors following a Dirichlet distribution and focus on asymptotic inferential results.

**Theorem 6.12** *Let  $\mathcal{P} = \mathcal{D}(\bullet)$  with minimal canonical representation (6.17), and let  $\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \dots$  be a sequence of iid random vectors with distribution  $P_{\boldsymbol{\beta}}$  for some  $\boldsymbol{\beta} \in (0, \infty)^{k+1}$ . Then, the sequence  $\hat{\boldsymbol{\beta}}^{(n)}$ ,  $n \in \mathbb{N}$ ,  $n \geq 2$ , of MLEs for  $\boldsymbol{\beta}$  is strongly consistent and asymptotically efficient for  $\boldsymbol{\beta}$ , i.e., we have  $\hat{\boldsymbol{\beta}}^{(n)} \rightarrow \boldsymbol{\beta}$   $\mathbb{P}_{\boldsymbol{\beta}}$ -a.s. and*

$$\sqrt{n}(\hat{\boldsymbol{\beta}}^{(n)} - \boldsymbol{\beta}) \xrightarrow{d} \mathcal{N}_{k+1}(\mathbf{0}, \mathbf{I}(\boldsymbol{\beta})^{-1})$$

for  $n \rightarrow \infty$ , where the inverse of the Fisher information matrix at  $\boldsymbol{\beta}$  is given by

$$[\mathbf{I}(\boldsymbol{\beta})^{-1}]_{i,j} = \frac{\delta_{ij}}{\psi'(\beta_i)} - \left( \sum_{l=1}^{k+1} \frac{1}{\psi'(\beta_l)} - \frac{1}{\psi'(\boldsymbol{\beta}_{\bullet})} \right)^{-1} \frac{1}{\psi'(\beta_i) \psi'(\beta_j)}$$

for  $1 \leq i, j \leq k+1$ .

**Proof** Strong consistency and asymptotic efficiency of the sequence  $\hat{\boldsymbol{\beta}}^{(n)}$ ,  $n \geq 2$ , are obtained from Theorems 4.31, 4.34, and Remark 4.35. To obtain the inverse of  $\mathbf{I}(\boldsymbol{\beta})$  given by formula (6.19), note that  $\mathbf{I}(\boldsymbol{\beta}) = \mathbf{D} + b\mathbf{v}\mathbf{v}^t$  for the diagonal matrix  $\mathbf{D}$  with diagonal entries  $d_j = \psi'(\beta_j) > 0$ ,  $1 \leq j \leq k+1$ , real number  $b = -\psi'(\boldsymbol{\beta}_{\bullet}) < 0$ , and vector  $\mathbf{v} = (1, \dots, 1)^t \in \mathbb{R}^{k+1}$ . Applying a commonly known matrix identity (see, e.g., [45, p. 458]) and setting  $\mathbf{w} = (1/d_1, \dots, 1/d_{k+1})^t$  then yields

$$\begin{aligned} (\mathbf{D} + b\mathbf{v}\mathbf{v}^t)^{-1} &= \mathbf{D}^{-1} - \mathbf{D}^{-1}\mathbf{v} \left( \frac{1}{b} + \mathbf{v}^t \mathbf{D}^{-1} \mathbf{v} \right)^{-1} \mathbf{v}^t \mathbf{D}^{-1} \\ &= \mathbf{D}^{-1} - \left( \frac{1}{b} + \sum_{l=1}^{k+1} \frac{1}{d_l} \right)^{-1} \mathbf{w} \mathbf{w}^t. \end{aligned}$$

Inserting for  $d_1, \dots, d_{k+1}$  and  $b$  then gives the representation.  $\square$



Now, let us consider the multivariate test problem

$$H_0 : \boldsymbol{\beta} = \boldsymbol{\eta} \quad \leftrightarrow \quad H_1 : \boldsymbol{\beta} \neq \boldsymbol{\eta} \quad (6.21)$$

for some fixed  $\boldsymbol{\eta} = (\eta_1, \dots, \eta_{k+1})^t \in (0, \infty)^{k+1}$ .

**Theorem 6.13** *Under the assumptions of Theorem 6.12, the test statistics  $\Lambda_n = \Lambda_n(X^{(1)}, \dots, X^{(n)})$  and  $W_n = W_n(X^{(1)}, \dots, X^{(n)})$  of the LR test and the Wald test (with  $g(\boldsymbol{\beta}) = \boldsymbol{\beta} - \boldsymbol{\eta}$ ,  $\boldsymbol{\beta} \in (0, \infty)^{k+1}$ ) for test problem (6.21) based on  $X^{(1)}, \dots, X^{(n)}$ ,  $n \geq 2$ , are given by*

$$\Lambda_n = 2n \left[ \ln \left( \frac{B(\boldsymbol{\eta})}{B(\hat{\boldsymbol{\beta}}^{(n)})} \right) + \sum_{j=1}^{k+1} (\hat{\beta}_j^{(n)} - \eta_j) \left( \psi(\hat{\beta}_j^{(n)}) - \psi(\hat{\beta}_{\bullet}^{(n)}) \right) \right]$$

$$\text{and} \quad W_n = n \left[ \sum_{j=1}^{k+1} \psi'(\hat{\beta}_j^{(n)}) (\hat{\beta}_j^{(n)} - \eta_j)^2 - \psi'(\hat{\beta}_{\bullet}^{(n)}) (\hat{\beta}_{\bullet}^{(n)} - \eta_{\bullet})^2 \right].$$

Under  $H_0$ , both test statistics are asymptotically  $\chi^2(k+1)$ -distributed for  $n \rightarrow \infty$ .

**Proof** From Theorem 5.19 and Remark 5.21,  $\Lambda_n$  and  $W_n$  both have an asymptotic  $\chi^2(k+1)$ -distribution if  $H_0$  is true, such that we only have to verify the representations of the test statistics. From Theorem 5.16 and Lemma 3.37, we obtain that

$$\begin{aligned} \Lambda_n &= 2n D_{KL}(\hat{\boldsymbol{\beta}}^{(n)}, \boldsymbol{\eta}) \\ &= 2n \left[ \kappa(\boldsymbol{\eta}) - \kappa(\hat{\boldsymbol{\beta}}^{(n)}) + (\hat{\boldsymbol{\beta}}^{(n)} - \boldsymbol{\eta})^t \pi(\hat{\boldsymbol{\beta}}^{(n)}) \right], \end{aligned}$$

and the representation follows from  $\kappa = \ln(B)$  and formula (6.18). Here, note that  $nD_{KL}$  is the Kullback-Leibler divergence corresponding to the product family  $\mathcal{P}^{(n)}$ .

To obtain the formula for  $W_n$  as introduced in Remark 5.17 note first that by Theorem 3.29 the Fisher information matrix at  $\boldsymbol{\beta}$  for the product family  $\mathcal{P}^{(n)}$  is given by  $n\mathbf{I}(\boldsymbol{\beta})$ . Hence, by using the same notation as in the proof of Theorem 6.12, we find

$$(\boldsymbol{\beta} - \boldsymbol{\eta})^t [n\mathbf{I}(\boldsymbol{\beta})] (\boldsymbol{\beta} - \boldsymbol{\eta}) = n \left\{ (\boldsymbol{\beta} - \boldsymbol{\eta})^t \mathbf{D} (\boldsymbol{\beta} - \boldsymbol{\eta}) + b [(\boldsymbol{\beta} - \boldsymbol{\eta})^t \mathbf{v}]^2 \right\}. \quad (6.22)$$

Inserting for  $d_1, \dots, d_{k+1}$  and  $b$  and finally setting  $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$  in formula (6.22) leads to the representation for  $W_n$ .  $\square$

For  $\alpha \in (0, 1)$ , the LR test and the Wald test with asymptotic level  $\alpha$  then reject null hypothesis (6.21) if and only if the respective test statistic exceeds the critical value  $\chi_{1-\alpha}^2(k+1)$ , i.e., the  $(1-\alpha)$ -quantile of  $\chi^2(k+1)$ . For fixed  $n \in \mathbb{N}$ , the actual level of every test can be obtained by Monte-Carlo simulation. Here, a sufficiently large number of realizations of the test statistic under  $H_0$ , i.e., under the assumption that  $\boldsymbol{\beta} = \boldsymbol{\eta}$  is true, are generated. The actual level of the test is then estimated by the relative frequency of exceeding  $\chi_{1-\alpha}^2(k+1)$ , i.e., of (falsely) rejecting  $H_0$ .

Since the above tests are non-randomized, they may also serve to construct confidence regions for  $\beta$ , which asymptotically meet a desired confidence level. For instance, in case of the LR test,

$$C_{\Lambda_n}(\tilde{\mathbf{x}}) = \left\{ \beta \in (0, \infty)^{k+1} : \Lambda_n(\tilde{\mathbf{x}}) \leq \chi_{1-\alpha}^2(k+1) \right\}, \quad \tilde{\mathbf{x}} \in \mathcal{X}^n,$$

forms a confidence region for  $\beta$  with asymptotic level  $1 - \alpha$  (for  $n \rightarrow \infty$ ).

### 6.3 Generalized Order Statistics

Generalized order statistics have been introduced as a general semi-parametric model for ordered random variables, which contains, for instance, common order statistics, record values and  $k$ -record values, sequential order statistics, Pfeifer record values, and progressively type-II censored order statistics (see [31, 32]). What is remarkable here is the fact that joint density functions of generalized order statistics with a fixed but arbitrarily chosen absolutely continuous baseline distribution function  $F$  (see formula (6.23)) lead to a multiparameter EF, the support of which is a cone of  $\mathbb{R}^n$ .

**Definition 6.14** Let  $F$  denote a distribution function with Lebesgue density function  $f$  and quantile function  $F^{-1}$ , and let  $\gamma_1, \dots, \gamma_n$  be positive parameters. Then, the random variables  $X_1, \dots, X_n$  are called *generalized order statistics* based on  $F$  and  $\gamma_1, \dots, \gamma_n$  if their joint density function is given by

$$g_{F, \gamma}(\mathbf{x}) = \left( \prod_{j=1}^n \gamma_j \right) \left( \prod_{j=1}^{n-1} (1 - F(x_j))^{\gamma_j - \gamma_{j+1} - 1} f(x_j) \right) (1 - F(x_n))^{\gamma_n - 1} f(x_n) \quad (6.23)$$

for  $\mathbf{x} \in X = \{(x_1, \dots, x_n)^t \in \mathbb{R}^n : F^{-1}(0+) < x_1 < \dots < x_n < F^{-1}(1)\}$ .

For  $\gamma_j = n - j + 1$ ,  $1 \leq j \leq n$ , formula (6.23) is the joint density function of  $n$  common order statistics based on  $F$ , and setting  $\gamma_j = 1$ ,  $1 \leq j \leq n$ , yields the joint density function of  $n$  record values based on  $F$ .

In the context of sequential order statistics and Pfeifer record values, the parameters  $\gamma_1, \dots, \gamma_n$  and/or the distribution function  $F$  are typically unknown and have to be estimated based on data (see, e.g., [21, 32, 10, 11, 46]). In [8], it has been shown that formula (6.23) can be rewritten as

$$g_{F, \gamma}(\mathbf{x}) = C(\gamma) e^{\gamma' T(\mathbf{x})} h(\mathbf{x}), \quad \mathbf{x} \in X,$$

where

$$C(\boldsymbol{\gamma}) = \prod_{j=1}^n \gamma_j, \quad \boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_n)^t \in \Xi^* = (0, \infty)^n,$$

vector  $\mathbf{T} = (T_1, \dots, T_n)^t$  of statistics is defined by

$$T_j(\mathbf{x}) = \ln \left( \frac{1 - F(x_j)}{1 - F(x_{j-1})} \right), \quad \mathbf{x} = (x_1, \dots, x_n)^t \in \mathcal{X}, \quad 1 \leq j \leq n,$$

with  $F(x_0) = 0$ , and the function  $h$  is given by  $h(\mathbf{x}) = \prod_{j=1}^n f(x_j)/(1 - F(x_j))$  on  $\mathcal{X}$ . Hence, for fixed distribution function  $F$ ,

$$\mathcal{P} = \{g_{F,\boldsymbol{\gamma}} \lambda_{\mathcal{X}} : \boldsymbol{\gamma} \in \Xi^*\}$$

forms an EF, which can be proven to be regular and of order  $n$  (see [8]). These findings are helpful to establish inferential properties for generalized order statistics.

First, by Lemma 3.1(a), the moment generating function of  $-\mathbf{T}$  turns out to be

$$m_{-\mathbf{T}}(\mathbf{a}) = m_{\mathbf{T}}(-\mathbf{a}) = \frac{C(\boldsymbol{\gamma})}{C(\boldsymbol{\gamma} - \mathbf{a})} = \prod_{j=1}^n \frac{\gamma_j}{\gamma_j - a_j}$$

for  $\mathbf{a} = (a_1, \dots, a_n)^t \in \mathbb{R}^n$  with  $a_j < \gamma_j$ ,  $1 \leq j \leq n$ . Since the mapping  $t \mapsto \lambda/(\lambda - t)$ ,  $t < \lambda$ , is the moment generating function of the exponential distribution  $\mathcal{E}(1/\lambda)$  with scale parameter  $1/\lambda$ , we obtain from the one-to-one correspondence of distribution and moment generating function that  $T_1, \dots, T_n$  are independent and  $-T_j \sim \mathcal{E}(1/\gamma_j)$  for  $1 \leq j \leq n$ . Moments of  $\mathbf{T}$  are therefore at hand; in particular, the mean value function is given by

$$\pi(\boldsymbol{\gamma}) = E_{\boldsymbol{\gamma}}[\mathbf{T}] = -(1/\gamma_1, \dots, 1/\gamma_n)^t, \quad \boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_n)^t \in \Xi^*,$$

and the Fisher information matrix  $\mathbf{I}(\boldsymbol{\gamma})$  at  $\boldsymbol{\gamma} \in \Xi^*$  is a diagonal matrix with diagonal elements  $1/\gamma_j^2$ ,  $1 \leq j \leq n$  (see Theorems 2.56 and 3.29).

Now, let an iid sample  $\mathbf{X}^{(1)}, \dots, \mathbf{X}^{(s)}$  from the distribution with density function  $g_{F,\boldsymbol{\gamma}}$  be given, and let  $\tilde{\mathbf{X}} = (\mathbf{X}^{(1)}, \dots, \mathbf{X}^{(s)})$ . Then, applying Theorem 3.13(b), Corollaries 3.17 and 3.21 to the product EF  $\mathcal{P}^{(s)}$  yields that  $\mathbf{T}^{(s)} = (T_1^{(s)}, \dots, T_n^{(s)})^t$  with

$$T_j^{(s)}(\tilde{\mathbf{x}}) = \sum_{i=1}^s T_j(\mathbf{x}^{(i)}), \quad \tilde{\mathbf{x}} = (\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(s)}) \in \mathcal{X}^s,$$

is a minimal sufficient and complete statistic for  $\boldsymbol{\gamma}$ . Evidently,  $T_1^{(s)}, \dots, T_n^{(s)}$  are independent with  $-T_j \sim \mathcal{G}(1/\gamma_j, s)$ ,  $1 \leq j \leq n$ . By Corollary 4.5 with  $\pi(\Xi^*) = (-\infty, 0)^n$ , the MLE  $\hat{\boldsymbol{\gamma}}^{(s)}$  of  $\boldsymbol{\gamma}$  in  $\Xi^*$  based on  $\tilde{X}$  uniquely exists and is obtained as

$$\hat{\boldsymbol{\gamma}}^{(s)} = \pi^{-1} \left( \frac{1}{s} \mathbf{T}^{(s)}(\tilde{X}) \right) = - \left( s/T_1^{(s)}(\tilde{X}), \dots, s/T_n^{(s)}(\tilde{X}) \right)^t$$

with independent components, which are inverse gamma-distributed (since the statistics  $T_1^{(s)}, \dots, T_n^{(s)}$  are independent and gamma-distributed). According to Corollary 4.27, the vector of the reciprocals of  $\gamma_1, \dots, \gamma_n$  can efficiently be estimated by  $-\mathbf{T}^{(s)}(\tilde{X})/s$ . Moreover, Theorems 4.31 and 4.34 guarantee strong consistency and asymptotic efficiency of the sequence  $\hat{\boldsymbol{\gamma}}^{(s)}$ ,  $s \in \mathbb{N}$ , of MLEs for  $\boldsymbol{\gamma}$ .

Optimal tests for various test problems concerning a single component  $\gamma_1$ , say, of  $\boldsymbol{\gamma}$  are quickly derived by using the findings of Sect. 5.3, since  $T_1^{(s)}, \dots, T_n^{(s)}$  are independent. The conditional distribution of  $T_1^{(s)}$  given  $T_2^{(s)}, \dots, T_n^{(s)}$  is thus nothing but the marginal distribution of  $T_1^{(s)}$ , i.e., a gamma distribution. For instance, a UMPU level- $\alpha$  test for the one-sided test problem

$$H_0 : \gamma_1 \leq \eta \quad \leftrightarrow \quad H_1 : \gamma_1 > \eta$$

for some fixed  $\eta \in (0, \infty)$  is given by

$$\varphi(\tilde{\mathbf{x}}) = \mathbb{1}_{(0, \chi_{\alpha}^2(2s))} \left( -2\eta T_1^{(s)}(\tilde{\mathbf{x}}) \right), \quad \tilde{\mathbf{x}} \in \mathcal{X}^s$$

(see [7]). A UMPU test for the two-sided test problem with null hypothesis  $H_0 : \gamma_1 = \eta$  can be obtained by applying Theorem 5.13 leading to a similar situation as in Example 5.9. The corresponding confidence intervals for  $\gamma_1$  with minimum coverage probabilities of false parameters are shown in [12].

Multivariate test problems with a simple and composite null hypothesis have been studied in the literature. Tests for homogeneity of the parameters including the LR test are shown in [21]. In [9], the Rao score statistic and the Wald statistic have been addressed, the test statistics of which are easily computed, since the Fisher information matrix has diagonal shape (see also [13]). Moreover, confidence regions for  $\boldsymbol{\gamma}$  are obtained in [12] by inverting the acceptance regions of tests for a simple null hypothesis. Other confidence regions are constructed in [60] via divergence measures for generalized order statistics. In particular, for  $P_1, P_2 \in \mathcal{P}$  with parameter vectors  $\boldsymbol{\gamma}^{(1)} = (\gamma_1^{(1)}, \dots, \gamma_n^{(1)})^t$  and  $\boldsymbol{\gamma}^{(2)} = (\gamma_1^{(2)}, \dots, \gamma_n^{(2)})^t$ , say, the Kullback-Leibler divergence  $D_{KL}(\boldsymbol{\gamma}^{(1)}, \boldsymbol{\gamma}^{(2)})$  is determined by applying

Lemma 3.37:

$$\begin{aligned}
 D_{KL}(\boldsymbol{\gamma}^{(1)}, \boldsymbol{\gamma}^{(2)}) &= \kappa(\boldsymbol{\gamma}^{(2)}) - \kappa(\boldsymbol{\gamma}^{(1)}) + (\boldsymbol{\gamma}^{(1)} - \boldsymbol{\gamma}^{(2)})^t \boldsymbol{\pi}(\boldsymbol{\gamma}^{(1)}) \\
 &= -\sum_{j=1}^n \ln(\gamma_j^{(2)}) + \sum_{j=1}^n \ln(\gamma_j^{(1)}) - \sum_{j=1}^n \frac{\gamma_j^{(1)} - \gamma_j^{(2)}}{\gamma_j^{(1)}} \\
 &= \sum_{j=1}^n \left[ \frac{\gamma_j^{(2)}}{\gamma_j^{(1)}} - \ln\left(\frac{\gamma_j^{(2)}}{\gamma_j^{(1)}}\right) - 1 \right].
 \end{aligned}$$

Obviously, the latter expression depends on component-wise ratios of parameters, only. This finding can be used, e.g., to obtain an exact LR test for the test problem

$$H_0 : \boldsymbol{\gamma} = \boldsymbol{\gamma}^{(0)} \quad \leftrightarrow \quad H_1 : \boldsymbol{\gamma} \neq \boldsymbol{\gamma}^{(0)}$$

for fixed  $\boldsymbol{\gamma}^{(0)} \in \Xi^*$ , which has test statistic  $\Lambda_s = 2D_{KL}(\hat{\boldsymbol{\gamma}}^{(s)}, \boldsymbol{\gamma}^{(0)})$  by Theorem 5.16. Based on Cressie-Read divergences, multivariate tests with simple null hypothesis are examined in [60].

An important submodel of generalized order statistics results from the general setting by assuming that successive differences of the parameters are identical, i.e.,

$$\gamma_1 - \gamma_2 = \dots = \gamma_{n-1} - \gamma_n = m + 1, \text{ say.} \quad (6.24)$$

The model  $\mathcal{P}_0 \subset \mathcal{P}$  thus obtained is called  $m$ -generalized order statistics, and any distribution in  $\mathcal{P}_0$  is determined by specifying  $m$  and  $\gamma_n$ . The particular settings  $m = \gamma_n - 1$  and  $m = -1$  correspond to order statistics and record values, respectively. By utilizing formula (6.24) to rewrite density function  $g_{F,\boldsymbol{\gamma}}$  in formula (6.23),  $\mathcal{P}_0$  turns out to form a regular EF of order 2 with minimal canonical representation

$$g_{F,m,\gamma_n}(\mathbf{x}) = \left( \prod_{j=1}^n [(n-j)(m+1) + \gamma_n] \right) e^{mU(\mathbf{x}) + \gamma_n V(\mathbf{x})} \frac{\prod_{j=1}^n f(x_j)}{1 - F(x_n)}$$

and minimal sufficient and complete statistic  $(U, V)$  defined by  $U(\mathbf{x}) = \sum_{j=1}^{n-1} \ln(1 - F(x_j))$  and  $V(\mathbf{x}) = \ln(1 - F(x_n))$  for  $\mathbf{x} \in \mathcal{X}$ . Here, the natural parameter space is given by

$$\Xi_0^* = \{(m, \gamma_n)^t \in \mathbb{R} \times (0, \infty) : (n-1)(m+1) + \gamma_n > 0\}.$$

More results on this two-parameter EF concerning, in particular, existence, uniqueness, and asymptotic properties of the MLE of  $(m, \gamma_n)^t$  can be found in [10].

While the restriction to  $m$ -generalized order statistics again results in a regular EF, other assumed relations between the parameters may lead to curved EFs. This

is the case, for instance, by connecting  $\gamma_1, \dots, \gamma_n$  via the log-linear link function

$$\ln(\gamma_j) = \tau_1 + \tau_2 z_j, \quad 1 \leq j \leq n,$$

where  $\tau_1, \tau_2 \in \mathbb{R}$  are link function parameters and  $z_1, \dots, z_n$  are known real numbers, which are not all identical. Properties of curved EFs may then be applied to establish inferential results for  $(\tau_1, \tau_2)^t$ . For this special link function as well as for other types of link functions, maximum likelihood estimation of the link function parameters is discussed in [59].

# References

1. Akahira, M. (2017). *Statistical estimation for truncated exponential families*. Singapore: Springer.
2. Amari, S. I., Barndorff-Nielsen, O. E., Kass, R. E., Lauritzen, S. L., & Rao, C. R. (1987). *Differential geometry in statistical inference*. Hayward: Institute of Mathematical Statistics.
3. Bar-Lev, S. K., & Kokonendji, C. C. (2017). On the mean value parametrization of natural exponential families-A revisited review. *Mathematical Methods of Statistics*, 26(3), 159–175.
4. Barndorff-Nielsen, O. (2014). *Information and exponential families in statistical theory*. Chichester: Wiley.
5. Barndorff-Nielsen, O. E., & Cox, D. R. (1994). *Inference and asymptotics*. Boca Raton: Chapman & Hall.
6. Bauer, H. (2001). *Measure and integration theory*. Berlin: de Gruyter.
7. Bedbur, S. (2010). UMPU tests based on sequential order statistics. *Journal of Statistical Planning and Inference*, 140(9), 2520–2530.
8. Bedbur, S., Beutner, E., & Kamps, U. (2012). Generalized order statistics: An exponential family in model parameters. *Statistics*, 46(2), 159–166.
9. Bedbur, S., Beutner, E., & Kamps, U. (2014). Multivariate testing and model-checking for generalized order statistics with applications. *Statistics*, 48(6), 1297–1310.
10. Bedbur, S., & Kamps, U. (2017). Inference in a two-parameter generalized order statistics model. *Statistics*, 51(5), 1132–1142.
11. Bedbur, S., & Kamps, U. (2019). Testing for equality of parameters from different load-sharing systems. *Stats*, 2(1), 70–88.
12. Bedbur, S., Lennartz, J. M., & Kamps, U. (2013). Confidence regions in models of ordered data. *Journal of Statistical Theory and Practice*, 7(1), 59–72.
13. Bedbur, S., Müller, N., & Kamps, U. (2016). Hypotheses testing for generalized order statistics with simple order restrictions on model parameters under the alternative. *Statistics*, 50(4), 775–790.
14. Berk, R. H. (1972). Consistency and asymptotic normality of MLE's for exponential models. *The Annals of Mathematical Statistics*, 43(1), 193–204.
15. Bhatia, R. (1997). *Matrix analysis*. New York: Springer.
16. Bickel, P. J., & Doksum, K. A. (2015). *Mathematical statistics: Basic ideas and selected topics* (2nd ed., Vol. I). Boca Raton: Taylor & Francis.
17. Billingsley, P. (1995). *Probability and measure* (3rd ed.). New York: Wiley.
18. Billingsley, P. (1999). *Convergence of probability measures* (2nd ed.). New York: Wiley.

19. Boyd, S., & Vandenberghe, L. (2004). *Convex optimization*. Cambridge: Cambridge University Press.
20. Brown, L. D. (1986). *Fundamentals of statistical exponential families*. Hayward: Institute of Mathematical Statistics.
21. Cramer, E., & Kamps, U. (2001). Sequential  $k$ -out-of- $n$  systems. In N. Balakrishnan & C. R. Rao (Eds.), *Advances in reliability, handbook of statistics* (Vol. 20, pp. 301–372). Amsterdam: Elsevier.
22. Davis, P. J. (1965). Gamma functions and related functions. In M. Abramowitz & I. A. Stegun (Eds.), *Handbook of mathematical functions: With formulas, graphs, and mathematical tables* (pp. 253–294). New York: US Government Printing Office, Dover.
23. Dunn, P. K., & Smyth, G. K. (2018). *Generalized linear models with examples in R*. New York: Springer.
24. Ferguson, T. S. (1967). *Mathematical statistics: A decision theoretic approach*. New York: Academic.
25. Garren, S. T. (2000). Asymptotic distribution of estimated affinity between multiparameter exponential families. *Annals of the Institute of Statistical Mathematics*, 52(3), 426–437.
26. Johnson, N. L., Kemp, A. W., & Kotz, S. (2005). *Univariate discrete distributions* (3rd ed.). New York: Wiley.
27. Johnson, N. L., Kotz, S., & Balakrishnan, N. (1995). *Continuous univariate distributions* (2nd ed., Vol. 2). New York: Wiley.
28. Johnson, N. L., Kotz, S., & Balakrishnan, N. (1997). *Discrete multivariate distributions*. New York: Wiley.
29. Jørgensen, B. (1997). *The theory of dispersion models*. London: Chapman & Hall.
30. Kaas, R., Goovaerts, M., Dhaene, J., & Denuit, M. (2008). *Modern actuarial risk theory* (2nd ed.). Berlin: Springer.
31. Kamps, U. (1995). A concept of generalized order statistics. *Journal of Statistical Planning and Inference*, 48(1), 1–23.
32. Kamps, U. (2016). Generalized order statistics. In N. Balakrishnan, P. Brandimarte, B. Everitt, G. Molenberghs, W. Piegorsch, & F. Ruggeri (Eds.), *Wiley StatsRef: Statistics reference online* (pp. 1–12). Chichester: Wiley.
33. Katzur, A., & Kamps, U. (2016). Classification into Kullback-Leibler balls in exponential families. *Journal of Multivariate Analysis*, 150, 75–90.
34. Katzur, A., & Kamps, U. (2016). Homogeneity testing via weighted affinity in multiparameter exponential families. *Statistical Methodology*, 32, 77–90.
35. Keener, R. W. (2010). *Theoretical statistics: Topics for a core course*. New York: Springer.
36. Kollo, T., & von Rosen, D. (2005). *Advanced multivariate statistics with matrices*. Dordrecht: Springer.
37. Kotz, S., Balakrishnan, N., & Johnson, N. L. (2000). *Continuous multivariate distributions, models and applications* (2nd ed.). New York: Wiley.
38. Krätschmer, V. (2007). The uniqueness of extremum estimation. *Statistics & Probability Letters*, 77(10), 942–951.
39. Lehmann, E. L., & Casella, G. (1998). *Theory of point estimation* (2nd ed.). New York: Springer.
40. Lehmann, E. L., & Romano, J. P. (2005). *Testing statistical hypotheses* (3rd ed.). New York: Springer.
41. Letac, G. (1992). *Lectures on natural exponential families and their variance functions*. Rio de Janeiro: Instituto de Matemática Pura e Aplicada (IMPA).
42. Liese, F., & Miescke, K. J. (2008). *Statistical decision theory: Estimation, testing, and selection*. New York: Springer.
43. Liese, F., & Vajda, I. (1987). *Convex statistical distances*. Leipzig: Teubner.
44. Lindsey, J. K. (2006). *Parametric statistical inference*. Oxford: Clarendon Press.
45. Mardia, K. V., Kent, J. T., & Bibby, J. M. (1979). *Multivariate analysis*. London: Academic.
46. Mies, F., & Bedbur, S. (2020). Exact semiparametric inference and model selection for load-sharing systems. *IEEE - Transactions on Reliability*, 69(3), 863–872.



47. Ng, K. W., Tian, G. L., & Tang, M. L. (2011). *Dirichlet and related distributions: Theory, methods and applications*. Chichester: Wiley.
48. Pardo, L. (2006). *Statistical inference based on divergence measures*. Boca Raton: Chapman & Hall/CRC.
49. Pázman, A. (1986). On the uniqueness of the m.l. estimate in curved exponential families. *Kybernetika*, 22(2), 124–132.
50. Pfanzagl, J. (1994). *Parametric statistical theory*. Berlin: de Gruyter.
51. Rockafellar, R. T. (1970). *Convex analysis*. Princeton: Princeton University Press.
52. Rudin, W. (1976). *Principles of mathematical analysis* (3rd ed.). Auckland: McGraw-Hill.
53. Sen, P. K., & Singer, J. M. (1993). *Large sample methods in statistics: An introduction with applications*. Boca Raton: Chapman & Hall/CRC.
54. Serfling, R. J. (1980). *Approximation theorems of mathematical statistics*. New York: Wiley.
55. Shao, J. (2003). *Mathematical statistics* (2nd ed.). New York: Springer.
56. Sundberg, R. (2019). *Statistical modelling by exponential families*. Cambridge: Cambridge University Press.
57. Vajda, I. (1989). *Theory of statistical inference and information*. Dordrecht: Kluwer.
58. van der Vaart, A. W. (1998). *Asymptotic statistics*. Cambridge: Cambridge University Press.
59. Volovskiy, G., Bedbur, S., & Kamps, U. (2021). Link functions for parameters of sequential order statistics and curved exponential families. *Probability and Mathematical Statistics*, 41(1), 115–127.
60. Vuong, Q. N., Bedbur, S., & Kamps, U. (2013). Distances between models of generalized order statistics. *Journal of Multivariate Analysis*, 118, 24–36.
61. Wei, B. C. (1998). *Exponential family nonlinear models*. Singapore: Springer.
62. Wellek, S. (2010). *Testing statistical hypotheses of equivalence and noninferiority* (2nd ed.). Boca Raton: Chapman & Hall/CRC.
63. Witting, H. (1985). *Mathematische statistik I*. Stuttgart: Teubner.

# Index

## A

Affinely independent, 17

## C

Characteristic function, 43

Convex

function, 29

hull, 40

set, 23

Cramér-Rao inequality, 79

Cumulant, 47

Cumulant generating function, 47

## D

Distribution

beta, 12

binomial, 7

bivariate normal, 107

chi-square, 12

conditional, 47

Dirichlet, 12

exponential, 11

gamma, 11

geometric, 7

Good, 8

inverse gamma, 12

inverse normal, 11

joint, 50

Kemp discrete normal, 8

Liouville, 12

log-normal, 12

marginal, 47

multinomial, 9

multivariate normal, 14

multivariate power series, 9

negative binomial, 7

negative multinomial, 117

normal, 10

Poisson, 6

support of, 13

Wishart, 12

Divergence measure, 60

Cressie-Read, 61

Hellinger, 61

Jeffrey, 61

Kullback-Leibler, 60

Rényi, 61

## E

Entropy measure, 60

Shannon, 60

Estimator, 65

asymptotically efficient, 86

consistent, 85

efficient, 81

maximum likelihood, 69

strongly consistent, 85

unbiased, 79

uniformly minimum variance unbiased, 79

Exponential dispersion model, 53

Exponential family

canonical representation of, 23

cumulant function of, 29

curved, 22

cut of, 49

- full, 23
- generating measure of, 24
- mean value function of, 37
- mean value parametrization of, 40
- minimal representation of, 17
- natural, 24
- natural parameter space of, 23
- natural parametrization of, 23
- order of, 17
- regular, 23
- steep, 39
- truncated, 14
- variance function of, 42

**F**

Fisher information matrix, 58

**H**

Hypothesis

- alternative, 94
- of equivalence, 104
- null, 94
- one-sided, 96
- simple, 109
- two-sided, 101

**L**

Likelihood function, 66

Log-likelihood function, 66

**M**

Maximum likelihood estimate, 66

Moment generating function, 43

**N**

Neyman-Pearson lemma, 95

**P**

Parameter

- dispersion, 53
- identifiable, 21

Product measure, 49

**S**

Score statistic, 58

Statistic

- complete, 55
- minimal sufficient, 54
- sufficient, 54

**T**

Test, 93

- error probabilities of the first/second kind of, 94
- homogeneity, 109
- likelihood-ratio, 110
- non-randomized, 94
- power function of, 94
- randomized, 94
- Rao score, 111
- significance level of, 94
- unbiased, 94
- uniformly most powerful, 95
- uniformly most powerful level- $\alpha$ , 95
- uniformly most powerful unbiased level- $\alpha$ , 95

Wald, 111