

OPTIMAL ESTIMATION OF PARAMETERS

This book presents a comprehensive and consistent theory of estimation. The framework described leads naturally to the maximum capacity estimator as a generalization of the maximum likelihood estimator. This approach allows the optimal estimation of real-valued parameters, their number and intervals, as well as providing common ground for explaining the power of these estimators.

Beginning with a review of coding and the key properties of information, the author goes on to discuss the techniques of estimation, and develops the generalized maximum capacity estimator, based on a new form of Shannon's mutual information and channel capacity. Applications of this powerful technique in hypothesis testing and denoising are described in detail.

Offering an original and thought-provoking perspective on estimation theory, Jorma Rissanen's book is of interest to graduate students and researchers in the fields of information theory, probability and statistics, econometrics, and finance.

JORMA RISSANEN was a member of research staff in the IBM Almaden Research Center from 1965 to 2001, and is currently Professor Emeritus at Tampere University of Technology, Finland. Among his main results are the introduction of the MDL principle for statistics, the invention of arithmetic coding, and the introduction of variable-length Markov chains with the associated Algorithm Context. He has received many awards, including the 2007 Kolmogorov Medal from the University of London's CLRC, and the 2009 Shannon Award from the Information Theory Society. He received two Outstanding Innovation Awards from IBM, one in 1980 and the other in 1988, and a Corporate Award in 1991.

The minimum description length (MDL) principle is a very universal principle of statistical modeling in estimation, prediction, testing, and coding. Jorma Rissanen, the pioneer of the MDL principle, evolves a new theory to reach the most general and complete notion, which he calls the complete MDL principle. In this book the author derives it by introducing the key notion of maximum capacity. The most fundamental methods of estimation such as maximum likelihood estimation and the MDL estimation are naturally derived as the maximum capacity estimators, and their optimality is justified within a unifying theoretical framework. Through the book, readers can revisit the meaning of estimation from the author's very original viewpoint, and will enjoy the most advanced version of the MDL principle.

Kenji Yamanishi, *The University of Tokyo*

In this splendid new book, Jorma Rissanen, the originator of the minimum description length (MDL) Principle, puts forward a comprehensive theory of estimation which differs in several ways from the standard Bayesian and frequentist approaches. During the development of MDL over the last 30 years, it gradually emerged that MDL could be viewed, informally, as a maximum *probability* principle that directly extends Fisher's classical maximum *likelihood* method to allow for estimation of a model's structural properties. Yet providing a formal link between MDL and maximum probability remained elusive until the arrival of this book. By making the connection mathematically precise, Rissanen now ties up the loose ends of MDL theory and at the same time develops a beautiful, unified, entirely original and fully coherent theory of estimation, which includes hypothesis testing as a special case.

Peter Grünwald, *Centrum voor Wiskunde en Informatica, The Netherlands*

OPTIMAL ESTIMATION OF PARAMETERS

JORMA RISSANEN

*Tampere University of Technology
Helsinki Institute for Information Technology*



CAMBRIDGE
UNIVERSITY PRESS



CAMBRIDGE
UNIVERSITY PRESS

Shaftesbury Road, Cambridge CB2 8EA, United Kingdom

One Liberty Plaza, 20th Floor, New York, NY 10006, USA

477 Williamstown Road, Port Melbourne, VIC 3207, Australia

314–321, 3rd Floor, Plot 3, Splendor Forum, Jasola District Centre, New Delhi – 110025, India

103 Penang Road, #05–06/07, Visioncrest Commercial, Singapore 238467

Cambridge University Press is part of Cambridge University Press & Assessment,
 a department of the University of Cambridge.

We share the University's mission to contribute to society through the pursuit of
 education, learning and research at the highest international levels of excellence.

www.cambridge.org

Information on this title: www.cambridge.org/9781107004740

© Cambridge University Press & Assessment 2012

This publication is in copyright. Subject to statutory exception and to the provisions
 of relevant collective licensing agreements, no reproduction of any part may take
 place without the written permission of Cambridge University Press & Assessment.

First published 2012

A catalogue record for this publication is available from the British Library

Library of Congress Cataloging-in-Publication data

Rissanen, Jorma.

Optimal estimation of parameters / Jorma Rissanen, Tampere University of Technology,
 Helsinki Institute for Information Technology.

pages cm

ISBN 978-1-107-00474-0 (hardback)

1. Estimation theory. I. Title.

QA276.8.R57 2012

519.5'44 – dc23 2012007503

ISBN 978-1-107-00474-0 Hardback

Cambridge University Press & Assessment has no responsibility for the persistence
 or accuracy of URLs for external or third-party internet websites referred to in this
 publication and does not guarantee that any content on such websites is, or will
 remain, accurate or appropriate.

Contents

	<i>Preface</i>	<i>page vii</i>
1	Introduction	1
	Part I Coding and information	9
2	Basics of coding	11
	2.1 Coding and probability	12
	2.2 Coding and entropy	15
	2.3 Huffman’s code	16
	2.4 Codes as number representations	18
	2.4.1 Arithmetic codes	21
3	Basics of information	24
	3.1 Main definitions	24
	3.2 Mutual information	25
	3.2.1 Expected mutual information	25
	3.2.2 Maximum estimation information	27
	Part II Estimation	33
4	Modeling problems	35
	4.1 Models	36
	4.2 General comments on estimation	39
	4.3 Maximum capacity	41
	4.4 Necessary conditions for optimality	47
	4.5 General and complete MDL principles	51
5	Other optimality properties	57
	5.1 Minmax problems	57
	5.2 Consistency	63
6	Interval estimation	70
	6.1 Optimum intervals	71

vi	Contents	
	6.2 Maximum capacity partition	75
	6.3 Error probability	77
	6.3.1 Asymptotic distinguishability	78
	6.3.2 Partition algorithm for $k = 2$	81
7	Hypothesis testing	83
	7.1 General plan	84
	7.2 Test statistics	87
	7.3 Characteristic histograms	88
	7.4 Main tests	92
	7.4.1 Simple hypothesis	96
	7.4.2 Composite hypothesis	98
	7.4.3 Likelihood ratio	102
8	Denoising	104
	8.1 Hard thresholding	105
	8.2 Soft thresholding	109
9	Sequential models	112
	9.1 Bernoulli class	114
	9.2 Variable order Markov chains	116
	9.2.1 Algorithm Context	119
	9.2.2 Tree pruning	120
	9.2.3 Universal code	123
	9.2.4 Extension to time series	124
	9.3 Linear quadratic regression models	127
	9.3.1 Fixed variance	129
	9.3.2 Free variance	129
	9.4 Autoregressive moving average (ARMA) models	133
	9.4.1 Prediction	134
	9.4.2 Prediction bound with estimated parameters	141
	Appendices	
A	Elements of algorithmic information	144
	A.1 Recursive and partial recursive functions	145
	A.1.1 Universal computers	147
	A.1.2 Relative randomness and typicality	149
	A.2 Kolmogorov structure function	150
B	Universal prior for integers	153
	<i>References</i>	156
	<i>Index</i>	161

Preface

I have a long lasting interest in estimation, which started with attempts to control industrial processes. It did not take long to realize that the control part is easy if you knew the behavior of the process you want to control, which meant that the real problem is estimation. When I was asked by the Information Theory Society to give the 2009 Shannon Lecture I thought of giving a coherent survey of estimation theory. However, during the year given to prepare the talk I found that it was not possible, because there was no coherent theory of estimation. There was a collection of facts and results but they were isolated with little to connect them. To my surprise this applied even to the works of some of the greatest names in statistics, such as Fisher, Cramér, and Rao, which I had been familiar with for decades, but which I had never questioned until now that I was more or less forced to do so. As an example, the famous maximum likelihood estimator due to Fisher [12] had virtually no formal justification. The celebrated Cramér–Rao inequality gives it a non-asymptotic justification only for special models and for more general parametric models only an asymptotic justification. Clearly, no workable theory should be founded on asymptotic behavior. About the value of asymptotics, we quote Keynes’ famous quip that “asymptotically we all shall be dead.”

In trying to prepare the talk I was faced with a large jigsaw puzzle in which the pieces did not quite fit, and some crucial pieces were completely missing. After a considerable struggle I was able to make the pieces fit but to do so I had to alter virtually all of them and ignore the means and concepts introduced by the masters, who of course did not have access to information and coding theory. Their followers, however, did, but since this theory did not immediately solve their problems it was regarded as irrelevant and ignored. About the only concept that survived was the maximum likelihood estimator, and even that was as a special case of a more powerful *maximum capacity* estimator, which permits estimation of the number of parameters as well as intervals. The result is a theory of estimation which covers all the aspects of estimation that I can think of including, even hypothesis testing, which in this treatment is founded on estimation, and is quite different from the usual.