SpringerBriefs in Computer Science

Series editors

Stan Zdonik, Brown University, Providence, RI, USA
Shashi Shekhar, University of Minnesota, Minneapolis, MN, USA
Xindong Wu, University of Vermont, Burlington, VT, USA
Lakhmi C. Jain, University of South Australia, Adelaide, SA, Australia
David Padua, University of Illinois Urbana-Champaign, Urbana, IL, USA
Xuemin Sherman Shen, University of Waterloo, Waterloo, ON, Canada
Borko Furht, Florida Atlantic University, Boca Raton, FL, USA
V. S. Subrahmanian, Department of Computer Science, University of Maryland, College Park, MD, USA
Martial Hebert, Carnegie Mellon University, Pittsburgh, PA, USA
Katsushi Ikeuchi, Meguro-ku, University of Tokyo, Tokyo, Japan
Bruno Siciliano, Dipartimento di Ingegneria Elettrica e delle Tecnologie dell'Informazione, Università di Napoli Federico II, Napoli, Italy
Sushil Jajodia, George Mason University, Fairfax, VA, USA
Newton Lee, Institute for Education, Research and Scholarships, Los Angeles, CA, USA

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

Briefs allow authors to present their ideas and readers to absorb them with minimal time investment. Briefs will be published as part of Springer's eBook collection, with millions of users worldwide. In addition, Briefs will be available for individual print and electronic purchase. Briefs are characterized by fast, global electronic dissemination, standard publishing contracts, easy-to-use manuscript preparation and formatting guidelines, and expedited production schedules. We aim for publication 8–12 weeks after acceptance. Both solicited and unsolicited manuscripts are considered for publication in this series.

More information about this series at http://www.springer.com/series/10028

Daniil Ryabko

Asymptotic Nonparametric Statistical Analysis of Stationary Time Series

Springer

Daniil Ryabko
Fishlife Research
Belize City, Belize

ISSN 2191-5768 ISSN 2191-5776 (electronic)
SpringerBriefs in Computer Science
ISBN 978-3-030-12563-9 ISBN 978-3-030-12564-6 (eBook)
https://doi.org/10.1007/978-3-030-12564-6

Library of Congress Control Number: 2019933098

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2019
This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.
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, express 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

This book is about making statistical inference from stationary discrete-time processes. The assumption of stationarity alone is often considered too weak to make any meaningful inference. Here, this view is challenged by showing that, while some rather basic problems indeed can be proven not to admit any solution in this setting, surprisingly many are solvable without any further assumptions. These include such complex problems as clustering and change-point analysis. Some general results characterizing those problems that admit a solution are also presented.

The material in this volume is presented in a way that presumes familiarity with basic concepts of probability and statistics, up to and including probability distributions over spaces of infinite sequences. All the required background material can be found in the excellent monograph [16], which also contains a much deeper exposition of some of the key concepts used here, such as the distributional distance. Familiarity with ergodic theory is *not required* for understanding the material exposed in the present volume. Indeed, with two exceptions, the proofs do not rely on any facts deeper than the convergence of frequencies. One exception is Chap. 5, which deals with hypothesis testing and provides a characterization of hypotheses for which consistent tests exist; the required background material for this chapter can be found in Chap. 2. The other exception is Sect. 3.4, which establishes impossibility of discrimination between process distributions; this section is self-contained. The reader who is familiar with ergodic theory and feels the exposition in this volume is somewhat unorthodox can find all the necessary links to the more familiar framework in [58]; the latter book is also recommended to anyone seeking a deeper understanding of such results as the slow convergence of frequencies and entropy estimates, the classic ergodic theorem, and much more.

This book is organized as follows. Chapter 1 is introductory: besides providing some motivation for studying the problems addressed, it also introduces in an informal manner the main concepts used and the main results presented. Chapter 2 introduces the notation and definitions used in the subsequent chapters, as well as some necessary background material. Chapter 3 considers the most basic problems of statistical inference, on which the rest of the volume builds: estimating a distance

between processes (the distributional distance) and the problem of homogeneity testing or process discrimination, which, crucially for the subsequent problems addressed, is shown to be impossible to solve in the general setting of this book. Chapter 4 is devoted to clustering and change-point problems, which can be solved, or, in some cases, can be shown to admit no solution, based on the result of the preceding chapter. Chapter 5 addresses the problems of hypothesis testing in the general form: studying which pairs of hypotheses admit a consistent test. Finally, Chap. 6 discusses various generalizations of the presented results, as well as some directions for future research.

Acknowledgements Thanks to Léon Bottou for giving me the idea to write a book on this subject and for encouraging me to do it. Thanks to Boris Ryabko and Azadeh Khaleghi, in collaboration with whom some of the results presented here were obtained.

Santa Cruz de la Sierra, Bolivia Daniil Ryabko

Contents

1 Introduction 1
1.1 Stationarity, Ergodicity, AMS 2
1.2 What Is Possible and What Is Not Possible to Infer from Stationary Processes 4
1.3 Overview of the Inference Problems Covered 7

2 Preliminaries 9
2.1 Stationarity, Ergodicity 10
2.2 Distributional Distance 11

3 Basic Inference 13
3.1 Estimating the Distance Between Processes and Reconstructing a Process 13
3.2 Calculating $\hat{d}$ 15
3.3 The Three-Sample Problem 17
3.4 Impossibility of Discrimination 18
3.4.1 Setup and Definitions 19
3.4.2 The Main Result 19

4 Clustering and Change-Point Problems 29
4.1 Time-Series Clustering 30
4.1.1 Problem Formulation 31
4.1.2 A Clustering Algorithm and Its Consistency 32
4.1.3 Extensions: Unknown k, Online Clustering and Clustering with Respect to Independence 34
4.2 Change-Point Problems 37
4.2.1 Single Change Point 39
4.2.2 Multiple Change Points, Known Number of Change Points 41
4.2.3 Unknown Number of Change Points 42

5 Hypothesis Testing 45
5.1 Introduction 45
5.1.1 Motivation and Examples 46

5.2 Types of Consistency 47
5.2.1 Uniform Consistency 47
5.2.2 Asymmetric Consistency 48
5.2.3 Asymptotic Consistency 49
5.2.4 Other Notions of Consistency 49
5.3 One Example That Explains Hypotheses Testing 49
5.3.1 Bernoulli i.i.d. Processes 50
5.3.2 Markov Chains 51
5.3.3 Stationary Ergodic Processes 52
5.4 Topological Characterizations 53
5.4.1 Uniform Testing 53
5.4.2 Asymmetric Testing 54
5.5 Proofs 55
5.6 Examples 61
5.6.1 Simple Hypotheses, Identity or Goodness-of-Fit Testing 61
5.6.2 Markov and Hidden Markov Processes: Bounding the Order 62
5.6.3 Smooth Parametric Families 63
5.6.4 Homogeneity Testing or Process Discrimination 63
5.6.5 Independence 64
5.7 Open Problems 65
5.7.1 Relating the Notions of Consistency 65
5.7.2 Characterizing Hypotheses for Which Consistent Tests Exist 66
5.7.3 Independence Testing 67

6 Generalizations 69
6.1 Other Distances 69
6.1.1 sum Distances 69
6.1.2 Telescope Distance 70
6.1.3 sup Distances 71
6.1.4 Non-metric Distances 72
6.1.5 AMS Distributions 72
6.2 Piece-Wise Stationary Processes 73
6.3 Beyond Time Series 73
6.3.1 Processes Over Multiple Dimensions 73
6.3.2 Infinite Random Graphs 74

References 75