

Monographs on Statistics and Applied Probability 120

Statistical Inference

The Minimum Distance Approach

Ayanendranath Basu

Indian Statistical Institute
Kolkata, India

Hiroyuki Shioya

Muroran Institute of Technology
Muroran, Japan

Chanseok Park

Clemson University
Clemson, South Carolina, USA



CRC Press

Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group an informa business

A CHAPMAN & HALL BOOK

Chapman & Hall/CRC
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

© 2011 by Taylor and Francis Group, LLC
Chapman & Hall/CRC is an imprint of Taylor & Francis Group, an Informa business

No claim to original U.S. Government works

Printed in the United States of America on acid-free paper
10 9 8 7 6 5 4 3 2 1

International Standard Book Number: 978-1-4200-9965-2 (Hardback)

This book contains information obtained from authentic and highly regarded sources. Reasonable efforts have been made to publish reliable data and information, but the author and publisher cannot assume responsibility for the validity of all materials or the consequences of their use. The authors and publishers have attempted to trace the copyright holders of all material reproduced in this publication and apologize to copyright holders if permission to publish in this form has not been obtained. If any copyright material has not been acknowledged please write and let us know so we may rectify in any future reprint.

Except as permitted under U.S. Copyright Law, no part of this book may be reprinted, reproduced, transmitted, or utilized in any form by any electronic, mechanical, or other means, now known or hereafter invented, including photocopying, microfilming, and recording, or in any information storage or retrieval system, without written permission from the publishers.

For permission to photocopy or use material electronically from this work, please access www.copyright.com (<http://www.copyright.com/>) or contact the Copyright Clearance Center, Inc. (CCC), 222 Rosewood Drive, Danvers, MA 01923, 978-750-8400. CCC is a not-for-profit organization that provides licenses and registration for a variety of users. For organizations that have been granted a photocopy license by the CCC, a separate system of payment has been arranged.

Trademark Notice: Product or corporate names may be trademarks or registered trademarks, and are used only for identification and explanation without intent to infringe.

Visit the Taylor & Francis Web site at
<http://www.taylorandfrancis.com>

and the CRC Press Web site at
<http://www.crcpress.com>

To Manjula, Srabashi, and Padmini – AB –

To my wife, my father, and mother – HS –

To Late Professor Byung Ho Lee – CP –

Contents

| | |
|--|------------|
| Preface | xv |
| Acknowledgments | xix |
| 1 Introduction | 1 |
| 1.1 General Notation | 3 |
| 1.2 Illustrative Examples | 4 |
| 1.3 Some Background and Relevant Definitions | 7 |
| 1.3.1 Fisher Information | 7 |
| 1.3.2 First-Order Efficiency | 9 |
| 1.3.3 Second-Order Efficiency | 9 |
| 1.4 Parametric Inference Based on the Maximum Likelihood Method | 10 |
| 1.4.1 Hypothesis Testing by Likelihood Methods | 11 |
| 1.5 Statistical Functionals and Influence Function | 14 |
| 1.6 Outline of the Book | 18 |
| 2 Statistical Distances | 21 |
| 2.1 Introduction | 21 |
| 2.2 Distances Based on Distribution Functions | 22 |
| 2.3 Density-Based Distances | 25 |
| 2.3.1 The Distances in Discrete Models | 26 |
| 2.3.2 More on the Hellinger Distance | 33 |
| 2.3.3 The Minimum Distance Estimator and the Estimating Equations | 34 |
| 2.3.4 The Estimation Curvature | 38 |
| 2.3.5 Controlling of Inliers | 39 |
| 2.3.6 The Robustified Likelihood Disparity | 40 |
| 2.3.7 The Influence Function of the Minimum Distance Estimators | 43 |
| 2.3.8 ϕ -Divergences | 45 |
| 2.4 Minimum Hellinger Distance Estimation: Discrete Models | 46 |
| 2.4.1 Consistency of the Minimum Hellinger Distance Estimator | 47 |
| 2.4.2 Asymptotic Normality of the Minimum Hellinger Distance Estimator | 52 |
| | ix |

| | | |
|----------|---|------------|
| 2.5 | Minimum Distance Estimation Based on Disparities: Discrete Models | 55 |
| 2.6 | Some Examples | 67 |
| 3 | Continuous Models | 73 |
| 3.1 | Introduction | 73 |
| 3.2 | Minimum Hellinger Distance Estimation | 75 |
| 3.2.1 | The Minimum Hellinger Distance Functional | 75 |
| 3.2.2 | The Asymptotic Distribution of the Minimum Hellinger Distance Estimator | 78 |
| 3.3 | Estimation of Multivariate Location and Covariance | 83 |
| 3.4 | A General Structure | 87 |
| 3.4.1 | Disparities in This Class | 93 |
| 3.5 | The Basu–Lindsay Approach for Continuous Data | 94 |
| 3.5.1 | Transparent Kernels | 98 |
| 3.5.2 | The Influence Function of the Minimum Distance Estimators for the Basu–Lindsay Approach | 100 |
| 3.5.3 | The Asymptotic Distribution of the Minimum Distance Estimators | 102 |
| 3.6 | Examples | 107 |
| 4 | Measures of Robustness and Computational Issues | 115 |
| 4.1 | The Residual Adjustment Function | 116 |
| 4.2 | The Graphical Interpretation of Robustness | 118 |
| 4.3 | The Generalized Hellinger Distance | 126 |
| 4.3.1 | Connection with Other Distances | 129 |
| 4.4 | Higher Order Influence Analysis | 129 |
| 4.5 | Higher Order Influence Analysis: Continuous Models | 136 |
| 4.6 | Asymptotic Breakdown Properties | 137 |
| 4.6.1 | Breakdown Point of the Minimum Hellinger Distance Estimator | 137 |
| 4.6.2 | The Breakdown Point for the Power Divergence Family | 139 |
| 4.6.3 | A General Form of the Breakdown Point | 141 |
| 4.6.4 | Breakdown Point for Multivariate Location and Covariance Estimation | 144 |
| 4.7 | The α -Influence Function | 147 |
| 4.8 | Outlier Stability of Minimum Distance Estimators | 149 |
| 4.8.1 | Outlier Stability of the Estimating Functions | 152 |
| 4.8.2 | Robustness of the Estimator | 153 |
| 4.9 | Contamination Envelopes | 156 |
| 4.10 | The Iteratively Reweighted Least Squares (IRLS) | 160 |
| 4.10.1 | Development of the Algorithm | 160 |
| 4.10.2 | The Standard IREE | 163 |
| 4.10.3 | Optimally Weighted IREE | 164 |

| | | |
|----------|--|------------|
| 4.10.4 | Step by Step Implementation | 166 |
| 5 | The Hypothesis Testing Problem | 167 |
| 5.1 | Disparity Difference Test: Hellinger Distance Case | 167 |
| 5.2 | Disparity Difference Tests in Discrete Models | 172 |
| 5.2.1 | Second-Order Effects in Testing | 175 |
| 5.3 | Disparity Difference Tests: The Continuous Case | 180 |
| 5.3.1 | The Smoothed Model Approach | 182 |
| 5.4 | Power Breakdown of Disparity Difference Tests | 184 |
| 5.5 | Outlier Stability of Disparity Difference Tests | 186 |
| 5.5.1 | The GHD and the Chi-Square Inflation Factor | 189 |
| 5.6 | The Two-Sample Problem | 191 |
| 6 | Techniques for Inlier Modification | 195 |
| 6.1 | Minimum Distance Estimation: Inlier Correction in Small Samples | 195 |
| 6.2 | Penalized Distances | 197 |
| 6.2.1 | The Penalized Hellinger Distance | 198 |
| 6.2.2 | Minimum Penalized Distance Estimators | 200 |
| 6.2.3 | Asymptotic Distribution of the Minimum Penalized Distance Estimator | 201 |
| 6.2.4 | Penalized Disparity Difference Tests: Asymptotic Results | 206 |
| 6.2.5 | The Power Divergence Family versus the Blended Weight Hellinger Distance Family | 207 |
| 6.3 | Combined Distances | 212 |
| 6.3.1 | Asymptotic Distribution of the Minimum Combined Distance Estimators | 216 |
| 6.4 | ϵ -Combined Distances | 222 |
| 6.5 | Coupled Distances | 225 |
| 6.6 | The Inlier-Shrunk Distances | 227 |
| 6.7 | Numerical Simulations and Examples | 230 |
| 7 | Weighted Likelihood Estimation | 235 |
| 7.1 | The Discrete Case | 236 |
| 7.1.1 | The Disparity Weights | 237 |
| 7.1.2 | Influence Function and Standard Error | 242 |
| 7.1.3 | The Mean Downweighting Parameter | 244 |
| 7.1.4 | Examples | 245 |
| 7.2 | The Continuous Case | 249 |
| 7.2.1 | Influence Function and Standard Error: Continuous Case | 251 |
| 7.2.2 | The Mean Downweighting Parameter | 252 |
| 7.2.3 | A Bootstrap Root Search | 253 |
| 7.2.4 | Asymptotic Results | 254 |

| | | |
|----------|--|------------|
| 7.2.5 | Robustness of Estimating Equations | 255 |
| 7.3 | Examples | 256 |
| 7.4 | Hypothesis Testing | 261 |
| 7.5 | Further Reading | 263 |
| 8 | Multinomial Goodness-of-Fit Testing | 265 |
| 8.1 | Introduction | 265 |
| 8.1.1 | Chi-Square Goodness-of-Fit Tests | 266 |
| 8.2 | Asymptotic Distribution of the Goodness-of-Fit Statistics . . | 267 |
| 8.2.1 | The Disparity Statistics | 268 |
| 8.2.2 | The Simple Null Hypothesis | 268 |
| 8.2.3 | The Composite Null Hypothesis | 270 |
| 8.2.4 | Minimum Distance Inference versus Multinomial Goodness-of-Fit | 272 |
| 8.3 | Exact Power Comparisons in Small Samples | 273 |
| 8.4 | Choosing a Disparity to Minimize the Correction Terms . . . | 277 |
| 8.5 | Small Sample Comparisons of the Test Statistics | 280 |
| 8.5.1 | The Power Divergence Family | 280 |
| 8.5.2 | The BWHF Family | 282 |
| 8.5.3 | The BWCS Family | 283 |
| 8.5.4 | Derivation of $F_S(y)$ for a General Disparity Statistic . | 284 |
| 8.6 | Inlier Modified Statistics | 286 |
| 8.6.1 | The Penalized Disparity Statistics | 287 |
| 8.6.2 | The Combined Disparity Statistics | 288 |
| 8.6.3 | Numerical Studies | 290 |
| 8.7 | An Application: Kappa Statistics | 294 |
| 9 | The Density Power Divergence | 297 |
| 9.1 | The Minimum L_2 Distance Estimator | 298 |
| 9.2 | The Minimum Density Power Divergence Estimator | 300 |
| 9.2.1 | Asymptotic Properties | 303 |
| 9.2.2 | Influence Function and Standard Error | 308 |
| 9.2.3 | Special Parametric Families | 309 |
| 9.3 | A Related Divergence Measure | 311 |
| 9.3.1 | The JHHB Divergence | 311 |
| 9.3.2 | Formulae for Variances | 314 |
| 9.3.3 | Numerical Comparisons of the Two Methods | 316 |
| 9.3.4 | Robustness | 316 |
| 9.4 | The Censored Survival Data Problem | 317 |
| 9.4.1 | A Real Data Example | 318 |
| 9.5 | The Normal Mixture Model Problem | 322 |
| 9.6 | Selection of Tuning Parameters | 323 |
| 9.7 | Other Applications of the Density Power Divergence | 324 |

| | |
|--|------------|
| 10 Other Applications | 327 |
| 10.1 Censored Data | 327 |
| 10.1.1 Minimum Hellinger Distance Estimation in the Random Censorship Model | 327 |
| 10.1.2 Minimum Hellinger Distance Estimation Based on Hazard Functions | 329 |
| 10.1.3 Power Divergence Statistics for Grouped Survival Data | 330 |
| 10.2 Minimum Hellinger Distance Methods in Mixture Models . . | 331 |
| 10.3 Minimum Distance Estimation Based on Grouped Data . . . | 332 |
| 10.4 Semiparametric Problems | 335 |
| 10.4.1 Two-Component Mixture Model | 335 |
| 10.4.2 Two-Sample Semiparametric Model | 336 |
| 10.5 Other Miscellaneous Topics | 337 |
| 11 Distance Measures in Information and Engineering | 339 |
| 11.1 Introduction | 339 |
| 11.2 Entropies and Divergences | 340 |
| 11.3 Csiszár's f -Divergence | 341 |
| 11.3.1 Definition | 341 |
| 11.3.2 Range of the f -Divergence | 343 |
| 11.3.3 Inequalities Involving f -Divergences | 345 |
| 11.3.4 Other Related Results | 346 |
| 11.4 The Bregman Divergence | 346 |
| 11.5 Extended f -Divergences | 347 |
| 11.5.1 f -Divergences for Nonnegative Functions | 347 |
| 11.5.2 Another Extension of the f -Divergence | 351 |
| 11.6 Additional Remarks | 352 |
| 12 Applications to Other Models | 353 |
| 12.1 Introduction | 353 |
| 12.2 Preliminaries for Other Models | 354 |
| 12.3 Neural Networks | 356 |
| 12.3.1 Models and Previous Works | 356 |
| 12.3.2 Feed-Forward Neural Networks | 356 |
| 12.3.3 Training Feed-Forward Neural Networks | 357 |
| 12.3.4 Numerical Examples | 360 |
| 12.3.5 Related Works | 360 |
| 12.4 Fuzzy Theory | 361 |
| 12.4.1 Fundamental Elements of Fuzzy Sets | 361 |
| 12.4.2 Measures of Fuzzy Sets | 362 |
| 12.4.3 Generalized Fuzzy Divergence | 364 |
| 12.5 Phase Retrieval | 365 |
| 12.5.1 Diffractive Imaging | 365 |
| 12.5.2 Algorithms for Phase Retrieval | 367 |
| 12.5.3 Statistical-Distance-Based Phase Retrieval Algorithm | 368 |

| | |
|------------------------------------|------------|
| 12.5.4 Numerical Example | 369 |
| 12.6 Summary | 371 |
| Bibliography | 373 |
| Index | 403 |

Preface

In many ways, estimation by an appropriate minimum distance method is one of the most natural ideas in statistics. A parametric model imposes a certain structure on the class of probability distributions that may be used to describe real life data generated from a process under study. There hardly appears to be a better way to deal with such a problem than to choose the parametric model that minimizes an appropriately defined distance between the data and the model.

The issue is an important and complex one. There are many different ways of constructing an appropriate “distance” between the “data” and the “model.” One could, for example, construct a distance between the empirical distribution function and the model distribution function by a suitable measure of distance. Alternatively, one could minimize the distance between the estimated data density (obtained, if necessary, by using a nonparametric smoothing technique such as kernel density estimation) and the parametric model density. And when the particular nature of the distances has been settled (based on distribution functions, based on densities, etc.), there may be innumerable options for the distance to be used within the particular type of distances. So the scope of study referred to by “Minimum Distance Estimation” is literally huge.

Statistics is a modern science. In the early part of its history, minimum distance estimation was not a research topic of significant interest compared to some other topics. There may be several reasons for this. The neat theoretical development of the idea of maximum likelihood and its superior performance under model conditions meant that any other competing procedure would have had to make a real case for itself before being proposed as a viable alternative to maximum likelihood. Until other considerations such as robustness over appropriate neighborhoods of the parametric model came along, there was hardly any reason to venture outside the fold of maximum likelihood, particularly given the computational simplicity of the maximum likelihood method in most common parametric models, which minimum distance methods in general do not share.

The growth of the area of research covered by the present book can be attributed to several factors. Two of them require special mention. The first one is the growth of computing power. As in all other areas of science, research in statistical science got a major boost with the advent of computers. Previously intractable problems became numerically accessible. Approximate methods could be applied with enhanced degrees of precision. Computational complex-

ity of the procedure became a matter of minor concern, rather than the major deciding factor. This made the construction of distances and the computation of the estimators computationally feasible. The second major reason is the emergence of the area of robust statistical inference. It was no longer sufficient to have a technique which was optimal under model conditions but had weak robustness properties. Several minimum distance techniques have natural robustness properties under model misspecifications. Thus, the computational advances and the practical requirements converged to facilitate the growth of research in minimum distance methods.

Among the class of minimum distance methods we have focused, in this book, on density-based minimum distance methods. Carrying this specialization further, our emphasis, within the class of density-based distances has been on the chi-square type distances. Counting from Beran's path breaking 1977 paper, this area has seen a major spurt of research activity during the last three decades. In fact, the general development of the chi-square type distances began in the 1960s with Csiszár (1963) and Ali and Silvey (1966), but the robustness angle in this area probably surfaced with Beran. The procedures within the class of " ϕ -divergences" or "disparities" are popular because many of them combine strong robustness features with full asymptotic model efficiency.

There is no single book which tries to provide a comprehensive documentation of the development of this theory over the last 30 years or so. Our primary intention here has been to fill in this gap. Our development has mainly focused on the problem for independently and identically distributed data. But we have tried to be as comprehensive as possible in this regard in establishing the basic structure of this inference procedure so that the reader is sufficiently prepared to grasp the applications of this technique to more specialized scenarios. We have discussed the estimation and hypothesis testing problems for both discrete and continuous models, extensively described the robustness properties of the minimum distance methods, discussed the inlier problem and its possible solutions, described weighted likelihood estimators and considered several other related topics. We trust that this book will be a useful resource for any researcher who takes up density-based minimum distance estimation in the future.

Apart from minimum distance estimation based on chi-square type distances, on which we have spent the major part of this book, we have briefly looked at three other topics. These may be described as (i) minimum distance estimation based on the density power divergence; (ii) some recent developments on goodness-of-fit tests based on disparities and their modifications, and (iii) a discussion of the applications of these minimum distance methods in information theory and engineering. We believe that the last item will make the book useful to scientists outside the mainstream statistics area.

In this connection it is appropriate to mention some closely related books that are available in the literature. The book by Pardo (2006) gives an excellent description of minimum ϕ -divergence procedures and is a natural resource for

this area. However, Pardo deals almost exclusively, although thoroughly, with discrete models. In our book we have also provided an extensive description of continuous models. Besides, the robustness angle is a driving theme of our book, unlike Pardo's case.

Our discussion of the multinomial goodness-of-fit testing problem has been highly influenced by the classic by Read and Cressie (1988). However, we have made every effort not to be repetitive, and only described such topics not covered extensively by Cressie and Read (or extended their findings beyond the power divergence family). Unlike the minimum distance inference case where we have tried to be comprehensive, in the goodness-of-fit testing problem we have been deliberately selective.

We have also kept the description to a level where it will be easily accessible to students who have been exposed to first-year graduate courses in statistics. Our presentation, although sufficiently technical, does not assume a measure theoretic background for the reader and, except in Chapter 11, the rare references to measures do not arrest the flow of the book. The book can very well serve as the text for a one-semester graduate course in minimum distance methods.

We take this opportunity to acknowledge the help we have received from many colleagues, teachers, and students while completing the book. We should begin by acknowledging our intellectual debt to Professor Bruce G. Lindsay, two of the three authors of the current book being his Ph.D. advisees. Discussions with Professors Leandro Pardo, Marianthi Markatou and Claudio Agostinelli have been very helpful. Discussion with Professor Subir Bhandari has helped to make many of our mathematical derivations more rigorous. Many other colleagues, too innumerable to mention here, have helped us by drawing our attention to related works. We also thank Professor Wen-Tao Huang, who was instrumental in bringing this group of authors together.

Special thanks must be given to Dr. Abhijit Mandal; his assistance in working out many of the examples in the book and constructing the figures has been invaluable. Dr. Rohit Patra and Professor Biman Chakraborty also deserve thanks in this connection.

Finally, we wish to thank all our friends and family members who stood by us during the sometimes difficult phase of manuscript writing.

Ayanendranath Basu
Indian Statistical Institute
India

Hiroyuki Shioya
Muroran Institute of Technology
Japan

Chanseok Park
Clemson University
USA

Acknowledgments

Some of the numerical examples and figures represented here are from articles copyrighted to different journals or organizations. They have been reproduced with the permission of the appropriate authorities. A list is presented below, and their assistance in permitting these reproductions is gratefully acknowledged.

The simulated results in Example 5.5 have been reproduced from *Sankhya*, Series B, Volume 64, Basu, A. (author), Outlier resistant minimum divergence methods in discrete parametric models, pp. 128–140 (2002), with kind permission from *Sankhya*.

The simulated results in Example 6.1, together with Figures 6.1, 6.2 and Table 6.1, are reproduced from *Statistica Sinica*, Volume 8, Basu, A. and Basu, S. (authors), Penalized minimum disparity methods for multinomial models, pp. 841–860 (1998), with kind permission from *Statistica Sinica*.

The real data example in Section 9.4.1, together with Figures 9.2 and 9.3 have been reproduced from *The Annals of the Institute of Statistical Mathematics*, Volume 58, Basu, S., Basu, A. and Jones, M. C. (authors), Robust and efficient estimation for censored survival data, pp. 341–355 (2006), with kind permission from the *Institute of Statistical Mathematics*.