DIMACS

Series in Discrete Mathematics
and Theoretical Computer Science

Volume 72

Data Depth:
Robust Multivariate Analysis, Computational Geometry and Applications

Regina Y. Liu
Robert Serfling
Diane L. Souvaine
Editors

American Mathematical Society

Data Depth:
Robust Multivariate Analysis, Computational Geometry and Applications

This page intentionally left blank

https://doi.org/10.1090/dimacs/072

DIMACS
Series in Discrete Mathematics
and Theoretical Computer Science

Volume 72

Data Depth:
Robust Multivariate Analysis, Computational Geometry and Applications

Regina Y. Liu
Robert Serfling
Diane L. Souvaine
Editors

Center for Discrete Mathematics
and Theoretical Computer Science
A consortium of Rutgers University, Princeton University,
AT&T Labs–Research, Bell Labs (Lucent Technologies),
NEC Laboratories America, and Telcordia Technologies
(with partners at Avaya Labs, HP Labs, IBM Research,
Microsoft Research, and Stevens Institute of Technology)

American Mathematical Society

This DIMACS volume presents the proceedings from the DIMACS workshop on Data Depth: Robust Multivariate Statistical Analysis, Computational Geometry and Applications held at Rutgers University, May 14–16, 2003.

2000 *Mathematics Subject Classification.* Primary 60D05, 62G10, 62G30, 62G35, 62G99, 62H10, 62H15, 62H30, 65D18, 68U05.

ISBN-10: 0-8218-3596-3
ISBN-13: 978-0-8218-3596-8

♾ The paper used in this book is acid-free and falls within the guidelines
established to ensure permanence and durability.
Visit the AMS home page at http://www.ams.org/

10 9 8 7 6 5 4 3 2 1 11 10 09 08 07 06

Dedication

Yehuda Vardi was one of the co-organizers of this workshop. Regrettably, he passed away in January 2005.

Yehuda was a leading statistician who had made outstanding contributions to many subjects, including biased-sampling, image analysis, network tomography, and others. More recently, he had been working in the area of data depth, where he and his collaborators had pioneered the notion of L_1 depth.

Yehuda had been the chair of the Department of Statistics of Rutgers University from 1996 until the time of his death. He led the department with energy and vision. He was a true champion for interdisciplinary research. We will all miss him.

This page intentionally left blank

Contents

Foreword

A workshop on Data Depth: Robust Multivariate Analysis, Computational Geometry and Applications was held on May 14-16, 2003, at Rutgers University. We would like to express our appreciation to Regina Liu, Robert Serfling, Diane Souvaine, and the late Yehuda Vardi for their efforts to organize and plan this highly successful conference.

The workshop was part of the 2001-2006 Special Focus on Data Analysis and Mining and the 2002-2005 Special Focus on Computational Geometry and Applications. We extend our thanks to Adam Buchsbaum, Pierre Hansen, Diane Lambert, David Madigan, Mario Szegedy, and the late Yehuda Vardi, as well as Steven Fortune, Bernard Chazelle, and Bill Steiger for their work as organizers of these two special foci.

The workshop brought together a group of statisticians and a group of computer scientists to jointly work on geometric aspects of multivariate data analysis and data depth from their respective vantage points. They investigated computational efficiency and new algorithms with applications to such diverse areas as statistical quality control, aviation safety data analysis and gene clustering. Geometric interpretations of data depth were used to develop algorithms that used depth based statistical tests.

DIMACS gratefully acknowledges the generous support that makes these programs possible. Special thanks go to the National Science Foundation, the New Jersey Commission on Science and Technology, and to DIMACS partners at Rutgers, Princeton, AT&T Labs - Research, Bell Labs, NEC Laboratories America, and Telcordia Technologies, and affiliate partners Avaya Labs, HP Labs, IBM Research, Microsoft Research, and Stevens Institute of Technology.

Fred S. Roberts
Director

Robert Tarjan
Co-Director for Princeton

This page intentionally left blank

Preface

This book is a collection of some of the research work presented in the workshop on "DATA DEPTH: ROBUST MULTIVARIATE STATISTICAL ANALYSIS, COMPUTATIONAL GEOMETRY & APPLICATIONS". The workshop was held from May 14 to 16, 2003, at Rutgers University in New Jersey, and it was sponsored by DIMACS with support from the National Science Foundation. The workshop was co-organized by Regina Liu, Robert Serfling, Diane Souvaine and Yehuda Vardi. There were more than 100 participants from various fields, including: statistics, computer sciences, mathematics and operations research. This workshop brought together not only two DIMACS special foci but two different communities: statisticians and computational geometers. The result was a very exciting interdisciplinary workshop that laid the foundations for further interfaces and collaborations.

1. Goal of the Workshop

Multivariate data analysis plays a role of ever-increasing importance in scientific studies. In current developments of multivariate analysis, a more geometric point of view is being emphasized. Descriptive measures and sample statistics that can capture properly the higher-dimensional features of multivariate data are needed. Several geometric approaches have been proposed recently. Especially promising is the one founded on the concept of *data depth.* This new concept provides center-outward orderings of points in Euclidean space of any dimension. It also provides many new perspectives to aspects of probability as well as of computer science. In particular, the development of implementable computing algorithms for depth-based statistics has brought about many new challenges in computational geometry. The extensive development of data depth in recent years has spawned attractive depth-based tools for nonparametric multivariate data analysis, with a wide range of applications. The diversity in approaches, emphases, and concepts, however, makes it necessary to seek unified views and perspectives that would guide the further development of the depth-based approach.

2. Data Depth and Statistics

There has been a flurry of activities in the development of *data depth.* Different notions of data depth include: the projection of half-space, the counting of random simplices, the peeling of convex hulls, the summing over volumes of random simplices, the summing over absolute distance among random pairs, and others. Each notion of data depth provides a distinctive center-outward ordering of sample points in a multidimensional space. The probabilistic geometry underlying the definition of each data depth produces a particular ordering of the points in the space and makes it especially suitable for certain types of applications. Various studies have been carried out to investigate and compare the structural properties of various data depths and their corresponding geometric layout such as contours and central quantiles. Data analysis methodologies based on data depth have also been developed for constructing confidence regions, computing p-values for testing hypotheses, constructing rank tests, and regression, etc. Applications to statistical quality control, aviation safety data analysis, and gene clustering have been demonstrated with success. Many more applications are currently on-going or continue to be discovered. Although research on the subject of data depth has been on-going for more than a decade, many problems are still open. Developing efficient and implementable computer algorithms is also a crucial element of the development of data depth analysis methodology, and it requires special expertise in computer science. The need for collaborations between statisticians and computer scientists in this regard is evident and so was the need for this special workshop to bring together experts from both communities.

3. Computational Efficiency and Algorithms

The computational geometry community has long recognized that there are many important and challenging problems that lie at the interface of geometry and statistics. There are many instances in which geometric techniques have been utilized to give efficient algorithms for important problems in statistics. However, some of these algorithms, while guaranteeing good theoretical asymptotic performance, are difficult to implement in practice and thus are of limited utility for statisticians or general consumers of "scientific computing". There is a clear demand for modifying or creating implementable algorithms with good theoretical performance for statistical analysis of large scientific datasets.

The concept of *data depth* has been developed by statisticians for nonparametric multivariate data analysis. Among the numerous different notions of data depth, the so-called *halfspace depth* lends itself nicely to the creation of convex *depth contours* within the multivariate data set. Computational geometers and statisticians have provided algorithms and code for computing the halfspace depth contours in two dimensions. The running time is $O(n^2 \log n)$ for a dataset of size n. The theoretical result that requires only $O(n \log^5 n)$ does not lend itself easily to implementation. Recent improvements have used either the duality and topological sweep of an arrangement of lines in the dual or some randomized versions. Although *deepest points* have been widely studied in the computational geometry literature, the best known time for algorithms for center points in high dimensions is still of $O(n^{d+1})$. So far most computational algorithms are primarily centered around the *halfspace depth* in lower dimensions. Efficient computational algorithms are needed for other existing notions of depth, especially for higher dimensional settings. The main impediment to the widespread use of depth-induced data analysis is the computational bottleneck. Exploiting geometry to create implementable algorithms, both exact and approximate, would be extremely valuable, as evidenced by several of the papers in this volume.

4. Acknowledgements

This workshop was held under the auspices of the DIMACS Special Focus on Data Analysis and Mining and the DIMACS Special Focus on Computational Geometry and Applications, with support from NSF DMS 02-05136 and NSA MDA

04-02-0101. Many people in DIMACS and the Department of Statistics of Rutgers University had worked hard to help make this workshop a reality. In particular, we would like to thank the Department of Statistics of Rutgers University and Ms. Pat Wolf from the Department of Statistics, Rutgers University, for her enduring patience and efficiency in handling all the details from abstracts to those papers in this volume.

This page intentionally left blank

Titles in This Series

72 **Regina Y. Liu, Robert Serfling, and Diane Souvaine, Editors,** Data Depth: Robust Multivariate Analysis, Computational Geometry and Applications

71 **Zhilan Feng, Ulf Dieckmann, and Simon Levin, Editors,** Disease Evolution: Models, Concepts, and Data Analyses

70 **James Abello and Graham Cormode, Editors,** Discrete Methods in Epidemiology

69 **Siemion Fajtlowicz, Patrick W. Fowler, Pierre Hansen, Melvin F. Janowitz, and Fred S. Roberts, Editors,** Graphs and Discovery

68 **A. Ashikhmin and A. Barg, Editors,** Algebraic Coding Theory and Information Theory

67 **Ravi Janardan, Michiel Smid, and Debasish Dutta, Editors,** Geometric and Algorithmic Aspects of Computer-Aided Design and Manufacturing

66 **Piyush Gupta, Gerhard Kramer, and Adriaan J. van Wijngaarden, Editors,** Advances in Network Information Theory

65 **Santosh S. Vempala,** The Random Projection Method

64 **Melvyn B. Nathanson, Editor,** Unusual Applications of Number Theory

63 **J. Nešetřil and P. Winkler, Editors,** Graphs, Morphisms and Statistical Physics

62 **Gerard J. Foschini and Sergio Verdú, Editors,** Multiantenna Channels: Capacity, Coding and Signal Processing

61 **M. F. Janowitz, F.-J. Lapointe, F. R. McMorris, B. Mirkin, and F. S. Roberts, Editors,** Bioconsensus

60 **Saugata Basu and Laureano Gonzalez-Vega, Editors,** Algorithmic and Quantitative Real Algebraic Geometry

59 **Michael H. Goldwasser, David S. Johnson, and Catherine C. McGeoch, Editors,** Data Structures, Near Neighbor Searches, and Methodology: Fifth and Sixth DIMACS Implementation Challenges

58 **Simon Thomas, Editor,** Set Theory: The Hajnal Conference

57 **Eugene C. Freuder and Richard J. Wallace, Editors,** Constraint Programming and Large Scale Discrete Optimization

56 **Alexander Barg and Simon Litsyn, Editors,** Codes and Association Schemes

55 **Ding-Zhu Du, Panos M. Pardalos, and Jie Wang, Editors,** Discrete Mathematical Problems with Medical Applications

54 **Erik Winfree and David K. Gifford, Editors,** DNA Based Computers V

53 **Nathaniel Dean, D. Frank Hsu, and R. Ravi, Editors,** Robust Communication Networks: Interconnection and Survivability

52 **Sanguthevar Rajasekaran, Panos Pardalos, and D. Frank Hsu, Editors,** Mobile Networks and Computing

51 **Pierre Hansen, Patrick Fowler, and Maolin Zheng, Editors,** Discrete Mathematical Chemistry

50 **James M. Abello and Jeffrey Scott Vitter, Editors,** External Memory Algorithms

49 **Ronald L. Graham, Jan Kratochvíl, Jaroslav Nešetřil, and Fred S. Roberts, Editors,** Contemporary Trends in Discrete Mathematics

48 **Harvey Rubin and David Harlan Wood, Editors,** DNA Based Computers III

47 **Martin Farach-Colton, Fred S. Roberts, Martin Vingron, and Michael Waterman, Editors,** Mathematical Support for Molecular Biology

46 **Peng-Jun Wan, Ding-Zhu Du, and Panos M. Pardalos, Editors,** Multichannel Optical Networks: Theory and Practice

45 **Marios Mavronicolas, Michael Merritt, and Nir Shavit, Editors,** Networks in Distributed Computing

44 **Laura F. Landweber and Eric B. Baum, Editors,** DNA Based Computers II

TITLES IN THIS SERIES

43 **Panos Pardalos, Sanguthevar Rajasekaran, and José Rolim, Editors,** Randomization Methods in Algorithm Design

42 **Ding-Zhu Du and Frank K. Hwang, Editors,** Advances in Switching Networks

41 **David Aldous and James Propp, Editors,** Microsurveys in Discrete Probability

40 **Panos M. Pardalos and Dingzhu Du, Editors,** Network Design: Connectivity and Facilities Location

39 **Paul W. Beame and Samuel R Buss, Editors,** Proof Complexity and Feasible Arithmetics

38 **Rebecca N. Wright and Peter G. Neumann, Editors,** Network Threats

37 **Boris Mirkin, F. R. McMorris, Fred S. Roberts, and Andrey Rzhetsky, Editors,** Mathematical Hierarchies and Biology

36 **Joseph G. Rosenstein, Deborah S. Franzblau, and Fred S. Roberts, Editors,** Discrete Mathematics in the Schools

35 **Dingzhu Du, Jun Gu, and Panos M. Pardalos, Editors,** Satisfiability Problem: Theory and Applications

34 **Nathaniel Dean, Editor,** African Americans in Mathematics

33 **Ravi B. Boppana and James F. Lynch, Editors,** Logic and Random Structures

32 **Jean-Charles Grégoire, Gerard J. Holzmann, and Doron A. Peled, Editors,** The SPIN Verification System

31 **Neil Immerman and Phokion G. Kolaitis, Editors,** Descriptive Complexity and Finite Models

30 **Sandeep N. Bhatt, Editor,** Parallel Algorithms: Third DIMACS Implementation Challenge

29 **Doron A. Peled, Vaughan R. Pratt, and Gerard J. Holzmann, Editors,** Partial Order Methods in Verification

28 **Larry Finkelstein and William M. Kantor, Editors,** Groups and Computation II

27 **Richard J. Lipton and Eric B. Baum, Editors,** DNA Based Computers

26 **David S. Johnson and Michael A. Trick, Editors,** Cliques, Coloring, and Satisfiability: Second DIMACS Implementation Challenge

25 **Gilbert Baumslag, David Epstein, Robert Gilman, Hamish Short, and Charles Sims, Editors,** Geometric and Computational Perspectives on Infinite Groups

24 **Louis J. Billera, Curtis Greene, Rodica Simion, and Richard P. Stanley, Editors,** Formal Power Series and Algebraic Combinatorics/Séries Formelles et Combinatoire Algébrique, 1994

23 **Panos M. Pardalos, David I. Shalloway, and Guoliang Xue, Editors,** Global Minimization of Nonconvex Energy Functions: Molecular Conformation and Protein Folding

22 **Panos M. Pardalos, Mauricio G. C. Resende, and K. G. Ramakrishnan, Editors,** Parallel Processing of Discrete Optimization Problems

21 **D. Frank Hsu, Arnold L. Rosenberg, and Dominique Sotteau, Editors,** Interconnection Networks and Mapping and Scheduling Parallel Computations

20 **William Cook, László Lovász, and Paul Seymour, Editors,** Combinatorial Optimization

19 **Ingemar J. Cox, Pierre Hansen, and Bela Julesz, Editors,** Partitioning Data Sets

18 **Guy E. Blelloch, K. Mani Chandy, and Suresh Jagannathan, Editors,** Specification of Parallel Algorithms

17 **Eric Sven Ristad, Editor,** Language Computations

For a complete list of titles in this series, visit the
AMS Bookstore at **www.ams.org/bookstore/**.

The book is a collection of some of the research presented at the workshop of the same name held in May 2003 at Rutgers University. The workshop brought together researchers from two different communities: statisticians and specialists in computational geometry. The main idea unifying these two research areas turned out to be the notion of *data depth*, which is an important notion both in statistics and in the study of efficiency of algorithms used in computational geometry. Many of the articles in the book lay down the foundations for further collaboration and interdisciplinary research.

ISBN 0-8218-3596-3

9 780821 835968

DIMACS/72

AMS *on the* Web
www.ams.org