Gentle, James E.; Härdle, Wolfgang Karl; Mori, Yuichi

Working Paper

How computational statistics became the backbone of modern data science

SFB 649 Discussion Paper, No. 2011-020

Provided in Cooperation with:
Collaborative Research Center 649: Economic Risk, Humboldt University Berlin

Suggested Citation: Gentle, James E.; Härdle, Wolfgang Karl; Mori, Yuichi (2011) : How computational statistics became the backbone of modern data science, SFB 649 Discussion Paper, No. 2011-020, Humboldt University of Berlin, Collaborative Research Center 649 - Economic Risk, Berlin

This Version is available at:
https://hdl.handle.net/10419/56645

SFB 649 Discussion Paper 2011-020

How Computational Statistics Became the Backbone of Modern Data Science

James E. Gentle*
Wolfgang Karl Härdle**
Yuichi Mori***

* George Mason University Fairfax, USA
** Humboldt - Universität zu Berlin, Germany
*** Okayama University of Science, Japan

This research was supported by the Deutsche Forschungsgemeinschaft through the SFB 649 "Economic Risk".

http://sfb649.wiwi.hu-berlin.de
ISSN 1860-5664

SFB 649, Humboldt-Universität zu Berlin
Spandauer Straße 1, D-10178 Berlin

How Computational Statistics Became the Backbone of Modern Data Science

James E. Gentle[1], Wolfgang Härdle[2], and Yuichi Mori[3]

[1] George Mason University jgentle@gmu.edu
[2] Humboldt - Universität zu Berlin, CASE - Center for Applied Statistics and Economics haerdle@wiwi.hu-berlin.de
[3] Okayama University of Science mori@soci.ous.ac.jp

This first chapter serves as an introduction and overview for a collection of articles surveying the current state of the science of computational statistics. Earlier versions of most of these articles appeared in the first edition of *Handbook of Computational Statistics: Concepts and Methods*, published in 2004.

There have been advances in all of the areas of computational statistics, so we feel that it is time to revise and update this *Handbook*. This introduction is a revision of the introductory chapter of the first edition.

1 Computational Statistics and Data Analysis

To do data analysis is to do computing. Statisticians have always been heavy users of whatever computing facilities are available to them. As the computing facilities have become more powerful over the years, those facilities have obviously decreased the amount of effort the statistician must expend to do routine analyses. As the computing facilities have become more powerful, an opposite result has occurred, however; the computational aspect of the statistician's work has increased. This is because of paradigm shifts in statistical analysis that are enabled by the computer.

Statistical analysis involves use of observational data together with domain knowledge to develop a model to study and understand a data-generating process. The data analysis is used to refine the model or possibly to select a different model, to determine appropriate values for terms in the model, and to use the model to make inferences concerning the process. This has been the paradigm followed by statisticians for centuries. The advances in statistical theory over the past two centuries have not changed the paradigm, but they have improved the specific methods. The advances in computational power have enabled newer and more complicated statistical methods. Not only has the exponentially-increasing computational power allowed use of more

Keywords: Discrete time series models, continuous time diffusion models, models with jumps, stochastic volatility, GARCH
JEL classification: C15

The financial support from the Deutsche Forschungsgemeinschaft via SFB 649 "Ökonomisches Risiko", Humboldt-Universität zu Berlin is gratefully acknowledged.

detailed and better models, however, it has shifted the paradigm slightly. Many alternative views of the data can be examined. Many different models can be explored. Massive amounts of simulated data can be used to study the model/data possibilities.

When exact models are mathematically intractable, approximate methods, which are often based on asymptotics, or methods based on estimated quantities must be employed. Advances in computational power and developments in theory have made *computational inference* a viable and useful alternative to the standard methods of asymptotic inference in traditional statistics. Computational inference is based on simulation of statistical models.

The ability to perform large numbers of computations almost instantaneously and to display graphical representations of results immediately has opened many new possibilities for statistical analysis. The hardware and software to perform these operations are readily available and are accessible to statisticians with no special expertise in computer science. This has resulted in a two-way feedback between statistical theory and statistical computing. The advances in statistical computing suggest new methods and development of supporting theory; conversely, the advances in theory and methods necessitate new computational methods.

Computing facilitates the development of statistical theory in two ways. One way is the use of symbolic computational packages to help in mathematical derivations (particularly in reducing the occurrences of errors in going from one line to the next!). The other way is in the quick exploration of promising (or unpromising!) methods by simulations. In a more formal sense also, simulations allow evaluation and comparison of statistical methods under various alternatives. This is a widely-used research method. For example, 66 out of 79 articles published in the Theory and Methods section of the *Journal of the American Statistical Association* in 2010 reported on Monte Carlo studies of the performance of statistical methods. (In 2002, this number was 50 out of 61 articles.) A general outline of many research articles in statistics is

1. State the problem and summarize previous work on it,
2. Describe a new approach,
3. Work out some asymptotic properties of the new approach,
4. Conduct a Monte Carlo study showing the new approach in a favorable light.

Much of the effort in mathematical statistics has been directed toward the easy problems of exploration of asymptotic properties. The harder problems for finite samples require different methods. Carefully conducted and reported Monte Carlo studies often provide more useful information on the relative merits of statistical methods in finite samples from a range of model scenarios.

While to do data analysis is to compute, we do not identify all data analysis, which necessarily uses the computer, as "statistical computing" or as "computational statistics".By these phrases we mean something more than just using a statistical software package to do a standard analysis. We use

the term "statistical computing" to refer to the computational methods that enable statistical methods. Statistical computing includes numerical analysis, database methodology, computer graphics, software engineering, and the computer/human interface. We use the term "computational statistics" somewhat more broadly to include not only the methods of statistical computing, but also statistical methods that are computationally intensive. Thus, to some extent, "computational statistics" refers to a large class of modern statistical methods. Computational statistics is grounded in mathematical statistics, statistical computing, and applied statistics. While we distinguish "computational statistics" from "statistical computing", the emergence of the field of computational statistics was coincidental with that of statistical computing, and would not have been possible without the developments in statistical computing.

One of the most significant results of the developments in statistical computing during the past few decades has been the statistical software package. There are several of these, but a relatively small number that are in widespread use. While referees and editors of scholarly journals determine what statistical theory and methods are *published*, the developers of the major statistical software packages determine what statistical methods are *used*. Computer programs have become necessary for statistical analysis. The specific methods of a statistical analysis are often determined by the available software. This, of course, is not a desirable situation, but, ideally, the two-way feedback between statistical theory and statistical computing dimishes the effect over time.

The importance of computing in statistics is also indicated by the fact that there are at least ten major journals with titles that contain some variants of both "computing" and "statistics". The journals in the mainstream of statistics without "computing" in their titles also have a large proportion of articles in the fields of statistical computing and computational statistics. This is because, to a large extent, recent developments in statistics and in the computational sciences have gone hand in hand. There are also two well-known learned societies with a primary focus in statistical computing: the International Association for Statistical Computing (IASC), which is an affiliated society of the International Statistical Institute (ISI), and the Statistical Computing Section of the American Statistical Association (ASA). There are also a number of other associations focused on statistical computing and computational statistics, such as the Statistical Computing Section of the Royal Statistical Society (RSS), and the Japanese Society of Computational Statistics (JSCS).

Developments in computing and the changing role of computations in statistical work have had significant effects on the curricula of statistical education programs both at the graduate and undergraduate levels. Training in statistical computing is a major component in some academic programs in statistics (see Nolan and Temple Lang, 2010; Gentle, 2004; Lange, 2004; Monahan, 2004). In all academic programs, some amount of computing instruction is necessary if the student is expected to work as a statistician. The

extent and the manner of integration of computing into an academic statistics program, of course, change with the developments in computing hardware and software and advances in computational statistics.

We mentioned above the two-way feedback between statistical theory and statistical computing. There is also an important two-way feedback between applications and statistical computing, just as there has always been between applications and any aspect of statistics. Although data scientists seek commonalities among methods of data analysis, different areas of application often bring slightly different problems for the data analyst to address. In recent years, an area called "data mining" or "knowledge mining" has received much attention. The techniques used in data mining are generally the methods of exploratory data analysis, of clustering, and of statistical learning, applied to very large and, perhaps, diverse datasets. Scientists and corporate managers alike have adopted data mining as a central aspect of their work. Specific areas of application also present interesting problems to the computational statistician. Financial applications, particularly risk management and derivative pricing, have fostered advances in computational statistics. Biological applications, such as bioinformatics, microarray analysis, and computational biology, are fostering increasing levels of interaction with computational statistics.

The hallmarks of computational statistics are the use of more complicated models, larger datasets with both more observations and more variables, unstructured and heterogeneous datasets, heavy use of visualization, and often extensive simulations.

2 The Emergence of a Field of Computational Statistics

Statistical computing is truly a multidisciplinary field and the diverse problems have created a yeasty atmosphere for research and development. This has been the case from the beginning. The roles of statistical laboratories and the applications that drove early developments in statistical computing are surveyed by Grier (1999). As digital computers began to be used, the field of statistical computing came to embrace not only numerical methods but also a variety of topics from computer science.

The development of the field of statistical computing was quite fragmented, with advances coming from many directions — some by persons with direct interest and expertise in computations, and others by persons whose research interests were in the applications, but who needed to solve a computational problem. Through the 1950's the major facts relevant to statistical computing were scattered through a variety of journal articles and technical reports. Many results were incorporated into computer programs by their authors and never appeared in the open literature. Some persons who contributed to the development of the field of statistical computing were not aware of the

work that was beginning to put numerical analysis on a sound footing. This hampered advances in the field.

2.1 Early Developments in Statistical Computing

An early book that assembled much of the extant information on digital computations in the important area of linear computations was by Dwyer (1951). In the same year, Von Neumann's NBS publication (Von Neumann, 1951) described techniques of random number generation and applications in Monte Carlo. At the time of these publications, however, access to digital computers was not widespread. Dwyer (1951) was also influential in regression computations performed on calculators. Some techniques, such as use of "machine formulas", persisted into the age of digital computers.

Developments in statistical computing intensified in the 1960's, as access to digital computers became more widespread. Grier (1991) describes some of the effects on statistical practice by the introduction of digital computers, and how statistical applications motivated software developments. The problems of rounding errors in digital computations were discussed very carefully in a pioneering book by Wilkinson (1963). A number of books on numerical analysis using digital computers were beginning to appear. The techniques of random number generation and Monte Carlo were described by Hammersley and Handscomb (1964). In 1967 the first book specifically on statistical computing appeared, Hemmerle (1967).

2.2 Early Conferences and Formation of Learned Societies

The 1960's also saw the beginnings of conferences on statistical computing and sections on statistical computing within the major statistical societies. The Royal Statistical Society sponsored a conference on statistical computing in December 1966. The papers from this conference were later published in the RSS's *Applied Statistics* journal. The conference led directly to the formation of a Working Party on Statistical Computing within the Royal Statistical Society. The first Symposium on the Interface of Computer Science and Statistics was held February 1, 1967. This conference has continued as an annual event with only a few exceptions since that time (see Goodman, 1993; Billard and Gentle, 1993; Wegman, 1993). The attendance at the Interface Symposia initially grew rapidly year by year and peaked at over 600 in 1979. In recent years the attendance has been slightly under 300. The proceedings of the Symposium on the Interface have been an important repository of developments in statistical computing. In April, 1969, an important conference on statistical computing was held at the University of Wisconsin. The papers presented at that conference were published in a book edited by Milton and Nelder (1969), which helped to make statisticians aware of the useful developments in computing and of their relevance to the work of applied statisticians.

In the 1970's two more important societies devoted to statistical computing were formed. The Statistical Computing Section of the ASA was formed in 1971 (see Chambers and Ryan, 1990). The Statistical Computing Section organizes sessions at the annual meetings of the ASA, and publishes proceedings of those sessions. The International Association for Statistical Computing (IASC) was founded in 1977 as a Section of ISI. In the meantime, the first of the biennial COMPSTAT Conferences on computational statistics was held in Vienna in 1974. Much later, regional sections of the IASC were formed, one in Europe and one in Asia. The European Regional Section of the IASC is now responsible for the organization of the COMPSTAT conferences.

Also, beginning in the late 1960's and early 1970's, most major academic programs in statistics offered one or more courses in statistical computing. More importantly, perhaps, instruction in computational techniques has permeated many of the standard courses in applied statistics.

As mentioned above, there are several journals whose titles include some variants of both "computing" and "statistics". The first of these, the *Journal of Statistical Computation and Simulation*, was begun in 1972. There are dozens of journals in numerical analysis and in areas such as "computational physics", "computational biology", and so on, that publish articles relevant to the fields of statistical computing and computational statistics.

By 1980 the field of statistical computing, or computational statistics, was well-established as a distinct scientific subdiscipline. Since then, there have been regular conferences in the field, there are scholarly societies devoted to the area, there are several technical journals in the field, and courses in the field are regularly offered in universities.

2.3 The PC

The 1980's was a period of great change in statistical computing. The personal computer brought computing capabilities to almost everyone. With the PC came a change not only in the number of participants in statistical computing, but, equally important, completely different attitudes toward computing emerged. Formerly, to do computing required an account on a mainframe computer. It required laboriously entering arcane computer commands onto punched cards, taking these cards to a card reader, and waiting several minutes or perhaps a few hours for some output — which, quite often, was only a page stating that there was an error somewhere in the program. With a personal computer for the exclusive use of the statistician, there was no incremental costs for running programs. The interaction was personal, and generally much faster than with a mainframe. The software for PCs was friendlier and easier to use. As might be expected with many non-experts writing software, however, the general quality of software probably went down.

The democratization of computing resulted in rapid growth in the field, and rapid growth in software for statistical computing. It also contributed to the changing paradigm of the data sciences.

2.4 The Cross Currents of Computational Statistics

Computational statistics of course is more closely related to statistics than to any other discipline, and computationally-intensive methods are becoming more commonly used in various areas of application of statistics. Developments in other areas, such as computer science and numerical analysis, are also often directly relevant to computational statistics, and the research worker in this field must scan a wide range of literature.

Numerical methods are often developed in an ad hoc way, and may be reported in the literature of any of a variety of disciplines. Other developments important for statistical computing may also be reported in a wide range of journals that statisticians are unlikely to read. Keeping abreast of relevant developments in statistical computing is difficult not only because of the diversity of the literature, but also because of the interrelationships between statistical computing and computer hardware and software.

An example of an area in computational statistics in which significant developments are often made by researchers in other fields is Monte Carlo simulation. This technique is widely used in all areas of science, and researchers in various areas often contribute to the development of the science and art of Monte Carlo simulation. Almost any of the methods of Monte Carlo, including random number generation, are important in computational statistics.

2.5 Reproducible Research

Reproducibility in the sense of replication within experimental error has always been a touchstone of science. In recent years, however, the term "reproducible research" (RR), or sometimes "reproducible analysis", has taken on a stronger meaning. The standards for RR include provision of computer codes (preferably in source) and/or data that would allow the reader to replicate the reported results (see Baggerly and Berry, 2011).

Many journals enforce these requirements, or at least facilitate the provisions. The *Journal of American Statistical Association*, for example, encourages authors to provide code and/or data, as well as other supporting material. This additional material is linked with an electronic version of the article at the journal's web site.

Many articles in computational statistics are written in LaTeX and the computations are done in R. The R code, together with any input data, allows the reader to perform the same computations for simulations and analyses that yielded the results reported in the accompanying text. The Sweave package facilitates the incorporation of code with text in the same file (see Leisch, 2002). Instructions for obtaining Sweave as well as the current user manual can be obtained at

`http://www.statistik.lmu.de/~leisch/Sweave/Sweave-manual.pdf`

2.6 Literature

Some of the major periodicals in statistical computing and computational statistics are listed below. Some of these journals and proceedings are refereed rather rigorously, some refereed less so, and some are not refereed. Although most of these serials are published in hardcopy form, most are also available electronically.

- *ACM Transactions on Mathematical Software*, published quarterly by the ACM (Association for Computing Machinery), includes algorithms in Fortran and C. Most of the algorithms are available through `netlib`. The ACM collection of algorithms is sometimes called *CALGO*.
  `www.acm.org/toms/`
- *ACM Transactions on Modeling and Computer Simulation*, published quarterly by the ACM.
  `www.acm.org/tomacs/`
- *Applied Statistics*, published quarterly by the Royal Statistical Society. (Until 1998, it included algorithms in Fortran. Some of these algorithms, with corrections, were collected by Griffiths and Hill, 1985. Most of the algorithms are available through `statlib` at Carnegie Mellon University.)
  `www.rss.org.uk/publications/`
- *Communications in Statistics — Simulation and Computation*, published quarterly by Marcel Dekker. (Until 1996, it included algorithms in Fortran. Until 1982, this journal was designated as *Series B*.)
  `www.dekker.com/servlet/product/productid/SAC/`
- *Computational Statistics* published quarterly by Physica-Verlag (formerly called *Computational Statistics Quarterly*).
  `comst.wiwi.hu-berlin.de/`
- *Computational Statistics. Proceedings of the xxth Symposium on Computational Statistics* (COMPSTAT), published biennially by Physica-Verlag/Springer.
- *Computational Statistics & Data Analysis*, published by Elsevier Science. There are twelve issues per year. (This is also the official journal of the International Association for Statistical Computing and as such incorporates the *Statistical Software Newsletter*.)
  `www.cbs.nl/isi/csda.htm`
- *Computing Science and Statistics*. This is an annual publication containing papers presented at the Interface Symposium. Until 1992, these proceedings were named *Computer Science and Statistics: Proceedings of the xxth Symposium on the Interface*. (The 24$^{\text{th}}$ symposium was held in 1992.) In 1997, Volume 29 was published in two issues: Number 1, which contains the papers of the regular Interface Symposium; and Number 2, which contains papers from another conference. The two numbers are not sequentially paginated. Since 1999, the proceedings have been published only in CD-ROM form, by the Interface Foundation of North America.
  `www.galaxy.gmu.edu/stats/IFNA.html`

- *Journal of Computational and Graphical Statistics*, published quarterly as a joint publication of ASA, the Institute of Mathematical Statistics, and the Interface Foundation of North America.
  `www.amstat.org/publications/jcgs/`
- *Journal of the Japanese Society of Computational Statistics*, published once a year by JSCS.
  `www.jscs.or.jp/oubun/indexE.html`
- *Journal of Statistical Computation and Simulation*, published in twelve issues per year by Taylor & Francis.
  `www.tandf.co.uk/journals/titles/00949655.asp`
- *Journal of Statistical Software*, a free on-line journal that publishes articles, book reviews, code snippets, and software reviews.
  `www.jstatsoft.org/`
- *Proceedings of the Statistical Computing Section*, published annually by ASA.
  `www.amstat.org/publications/`
- *SIAM Journal on Scientific Computing*, published bimonthly by SIAM. This journal was formerly *SIAM Journal on Scientific and Statistical Computing*.
  `www.siam.org/journals/sisc/sisc.htm`
- *Statistical Computing & Graphics Newsletter*, published quarterly by the Statistical Computing and the Statistical Graphics Sections of ASA.
  `www.statcomputing.org/`
- *Statistics and Computing*, published quarterly by Chapman & Hall.

In addition to literature and learned societies in the traditional forms, an important source of communication and a repository of information are computer databases and forums. In some cases, the databases duplicate what is available in some other form, but often the material and the communications facilities provided by the computer are not available elsewhere.

3 This Handbook

The purpose of this handbook is the same as that of the first edition of *Concepts and Fundamentals*. It is to provide a survey of the basic concepts of computational statistics. A glance at the table of contents reveals a wide range of articles written by experts in various subfields of computational statistics. The articles are generally expository, taking the reader from the basic concepts to the current research trends. The emphasis throughout, however, is on the concepts and fundamentals. Most chapters have been revised to provide up-to-date references to the relevant literature.

We have retained the organization of the in three parts. Part II on "statistical computing" addresses the computational methodology; Part III on "statistical methodology" covers techniques of applied statistics that are computer-intensive, or otherwise that make use of the computer as a tool of discovery,

rather than as just a large and fast calculator; and, finally, Part IV describes a number of application areas in which computational statistics plays a major role are surveyed.

3.1 Summary and Overview; Part II: Statistical Computing

Statistical computing is in the interface of numerical analysis, computer science, and statistics. This interface includes computer arithmetic, algorithms, database methodology, languages and other aspects of the user interface, and computer graphics.

For statistical numerical analysis, it is important to understand how the computer does arithmetic, and more importantly what the implications are for statistical (or other) computations. In addition to understanding of the underlying arithmetic operations, the basic principles of numerical algorithms, such as divide and conquer, must be in the working knowledge of persons writing numerical software for statistical applications. Although many statisticians do not need to know the details, it is important that all statisticians understand the implications of computations within a system of numbers and operators that is not the same system that we are accustomed to in mathematics. Anyone developing computer algorithms, no matter how trivial the algorithm may appear, must understand the details of the computer system of numbers and operators.

One of the important uses of computers in statistics, and one that is central to computational statistics, is the simulation of random processes. This is a theme is central to several chapters of this handbook, but in Part II, the basic numerical methods relevant to simulation are discussed. These include the basics of random number generation, including assessing the quality of random number generators, and simulation of random samples from various distributions, as well as the class of methods called Markov chain Monte Carlo. Statistical methods using simulated samples are discussed further in Part III.

Some chapters of Part II address specific numerical methods, such as methods for linear algebraic computations, for optimization, and for transforms. Separate chapters in Part II discuss two specific areas of optimization, the EM algorithm and its variations, and stochastic optimization. Another chapter describes transforms, such as the well-known Fourier and wavelet transforms, that effectively restructure a problem by changing the domain are important statistical functionals.

Other chapters of Part II focus on efficient usage of computing resources. Specific topics include parallel computing, database management methodology, issues relating to the user interface, and even paradigms, such as an object orientation, for software development.

Statistical graphics, especially interactive and dynamic graphics, play an increasingly prominent role in data analysis. Two chapters of Part II are devoted to this important area.

3.2 Summary and Overview; Part III: Statistical Methodology

Part III covers several aspects of computational statistics. In this part the emphasis is on the statistical methodology that is enabled by computing. Computers are useful in all aspects of statistical data analysis, of course, but in Part III, and generally in computational statistics, we focus on statistical methods that are computationally intensive. Although a theoretical justification of these methods often depends on asymptotic theory, in particular, on the asymptotics of the empirical cumulative distribution function, asymptotic inference is generally replaced by computational inference.

The first few chapters of this part deal directly with techniques of computational inference; that is, the use of cross validation, resampling, and simulation of data-generating processes to make decisions and to assign a level of confidence to the decisions. Selection of a model implies consideration of more than one model. As we suggested above, this is one of the hallmarks of computational statistics: looking at data through a variety of models. Cross validation and its generalizations and resampling are important techniques for addressing the problems. Resampling methods also have much wider applicability in statistics, from estimating variances and setting confidence regions to larger problems in statistical data analysis. Computational inference depends on simulation of data-generating processes. Any such simulation is an *experiment*, and in Part III, principles for design and analysis of experiments using computer models are discussed.

Estimation of a multivariate probability density function is also addressed in Part III. This area is fundamental in statistics, and it utilizes several of the standard techniques of computational statistics, such as cross validation and visualization methods.

The next few chapters of Part III address important issues for discovery and analysis of relationships among variables. One class of models are asymmetric, that is, models for the effects of a given set of variables ("independent variables") on another variable or set of variables. Smoothing methods for these models, which include use of kernels, splines, and orthogonal series, are generally nonparametric or semiparametric. Two important types of parametric asymmetric models discussed in Part III are generalized linear models and nonlinear regression models. In any models that explore the relationships among variables, it is often desirable to reduce the effective dimensionality of a problem. All of these chapters on using models of variate relationships to analyze data emphasize the computational aspects.

One area in which computational inference has come to play a major role is in Bayesian analysis. Computational methods have enabled a Bayesian approach in practical applications, because no longer is this approach limited to simple problems or conjugate priors.

Survival analysis, with applications in both medicine and product reliability, has become more important in recent years. Computational methods for analyzing models used in survival analysis are discussed in Part III.

The final chapters of Part III address an exciting area of computational statistics. The general area may be called "data mining", although this term has a rather anachronistic flavor because of the hype of the mid-1990's. Other terms such as "knowledge mining" or "knowledge discovery in databases" ("KDD") are also used. To emphasize the roots in artificial intelligence, which is a somewhat discredited area, the term "computational intelligence" is also used. This is an area in which machine learning from computer science and statistical learning have merged.

3.3 Summary and Overview; Part IV: Statistical Methodology

Many areas of applications can only be addressed effectively using computationally-intensive statistical methods. This is often because the input datasets are so large, but it may also be because the problem requires consideration of a large number of possible alternatives. In Part IV, there are separate chapters on some areas of applications of computational statistics. One area is finance and economics, in which heavy-tailed distributions or models with nonconstant variance are important.

Human biology has become one of the most important areas of application, and many computationally-intensive statistical methods have been developed, refined, and brought to bear on problems in this area. Two important question involve the geometrical structure of protein molecules and the functions of the various areas in the brain. While much is known about the order of the components of the molecules, the three-dimensional structure for most important protein molecules is not known, and the tools for discovery of this structure need extensive development. Understanding the functions of different areas in the brain will allow more effective treatment of diseased or injured areas and the resumption of more normal activities by patients with neurological disorders.

Another important area of application of computational statistics is computer network intrusion detection. Because of the importance of computer networks around the world, and because of their vulnerability to unauthorized or malicious intrusion, detection has become one of the most important — and interesting — areas for data mining.

The articles in this handbook cover the important subareas of computational statistics and give some flavor of the wide range of applications. While the articles emphasize the basic concepts and fundamentals of computational statistics, they provide the reader with tools and suggestions for current research topics. The reader may turn to a specific chapter for background reading and references on a particular topic of interest, but we also suggest that the reader browse and ultimately peruse articles on unfamiliar topics. Many surprising and interesting tidbits will be discovered!

3.4 The Ehandbook

A unique feature of this handbook is the supplemental ebook format. Our ebook design offers a HTML file with links to world wide computing servers. This HTML version can be downloaded onto a local computer via a licence card included in this handbook.

3.5 Other Handbooks in Computational Statistics

This handbook on concepts and fundamentals sets the stage for future handbooks that go more deeply into the various subfields of computational statistics. These handbooks will each be organized around either a specific class of theory and methods, or else around a specific area of application.

The development of the field of computational statistics has been rather fragmented. We hope that the articles in this handbook series can provide a more unified framework for the field.

In the years since the publication of the first volume in the series of Handbooks in Computational Statistics, which covered general concepts and methods, three other volumes have appeared. These are on somewhat more narrow topics within the field of computational statistics: data visualization, partial least squares, and computational finance.

References

Baggerly, K. A., and Berry, D. A. (2011). Reproducible research, *AmStatNews*, January, `http://magazine.amstat.org/blog/2011/01/01/scipolicyjan11/`.

Billard, L. and Gentle, J.E. (1993). The middle years of the Interface, *Computing Science and Statistics*, 25:19–26.

Chambers, J.M. and Ryan, B.F. (1990). The ASA Statistical Computing Section, *The American Statistician*, 44(2):87–89.

Dwyer, P.S. (1951). *Linear Computations*, John Wiley and Sons, New York.

Gentle, J.E. (2004). Courses in statistical computing and computational statistics, *The American Statistician*, 58:2–5.

Goodman, A. (1993). Interface insights: From birth into the next century, *Computing Science and Statistics*, 25:14–18.

Grier, D.A. (1991). Statistics and the introduction of digital computers, *Chance*, 4(3):30–36.

Grier, D.A. (1999). Statistical laboratories and the origins of statistical computing, *Chance*, 4(2):14–20.

Hammersley, J.M. and Handscomb, D.C. (1964). *Monte Carlo Methods*, Methuen & Co., London.

Hemmerle, W.J. (1967). *Statistical Computations on a Digital Computer*. Blaisdell, Waltham, Massachusetts.

Lange, K. (2004). Computational Statistics and Optimization Theory at UCLA, *The American Statistician*, 58:9–11.

Leisch, F. (2002) Sweave: Dynamic Generation of Statistical Reports Using Literate Data Analysis. In *Compstat 2002 - Proceedings in Computational Statistics*, edited by W. Härdle and B. Rönz, 575–580.

Milton, R. and Nelder, J. (eds) (1969). *Statistical Computation*, Academic Press, New York.

Monahan, J. (2004). Teaching Statistical Computing at NC State, *The American Statistician*, 58:6–8.

Nolan, D. and Temple Lang, D. (2010). Computing in the Statistics Curriculum, *The American Statistician*, 64:97–107.

Von Neumann, J. (1951). *Various Techniques Used in Connection with Random Digits*, National Bureau of Standards Symposium, NBS Applied Mathematics Series 12, National Bureau of Standards (now National Institute of Standards and Technology), Washington, DC.

Wegman, E.J. (1993). History of the Interface since 1987: The corporate era, *Computing Science and Statistics*, 25:27–32.

Wilkinson, J. H. (1963). *Rounding Errors in Algebraic Processes*, Prentice-Hall, Inc., Englewood Cliffs, New Jersey.

SFB 649 Discussion Paper Series 2011

For a complete list of Discussion Papers published by the SFB 649, please visit http://sfb649.wiwi.hu-berlin.de.

001 "Localising temperature risk" by Wolfgang Karl Härdle, Brenda López Cabrera, Ostap Okhrin and Weining Wang, January 2011.
002 "A Confidence Corridor for Sparse Longitudinal Data Curves" by Shuzhuan Zheng, Lijian Yang and Wolfgang Karl Härdle, January 2011.
003 "Mean Volatility Regressions" by Lu Lin, Feng Li, Lixing Zhu and Wolfgang Karl Härdle, January 2011.
004 "A Confidence Corridor for Expectile Functions" by Esra Akdeniz Duran, Mengmeng Guo and Wolfgang Karl Härdle, January 2011.
005 "Local Quantile Regression" by Wolfgang Karl Härdle, Vladimir Spokoiny and Weining Wang, January 2011.
006 "Sticky Information and Determinacy" by Alexander Meyer-Gohde, January 2011.
007 "Mean-Variance Cointegration and the Expectations Hypothesis" by Till Strohsal and Enzo Weber, February 2011.
008 "Monetary Policy, Trend Inflation and Inflation Persistence" by Fang Yao, February 2011.
009 "Exclusion in the All-Pay Auction: An Experimental Investigation" by Dietmar Fehr and Julia Schmid, February 2011.
010 "Unwillingness to Pay for Privacy: A Field Experiment" by Alastair R. Beresford, Dorothea Kübler and Sören Preibusch, February 2011.
011 "Human Capital Formation on Skill-Specific Labor Markets" by Runli Xie, February 2011.
012 "A strategic mediator who is biased into the same direction as the expert can improve information transmission" by Lydia Mechtenberg and Johannes Münster, March 2011.
013 "Spatial Risk Premium on Weather Derivatives and Hedging Weather Exposure in Electricity" by Wolfgang Karl Härdle and Maria Osipenko, March 2011.
014 "Difference based Ridge and Liu type Estimators in Semiparametric Regression Models" by Esra Akdeniz Duran, Wolfgang Karl Härdle and Maria Osipenko, March 2011.
015 "Short-Term Herding of Institutional Traders: New Evidence from the German Stock Market" by Stephanie Kremer and Dieter Nautz, March 2011.
016 "Oracally Efficient Two-Step Estimation of Generalized Additive Model" by Rong Liu, Lijian Yang and Wolfgang Karl Härdle, March 2011.
017 "The Law of Attraction: Bilateral Search and Horizontal Heterogeneity" by Dirk Hofmann and Salmai Qari, March 2011.
018 "Can crop yield risk be globally diversified?" by Xiaoliang Liu, Wei Xu and Martin Odening, March 2011.
019 "What Drives the Relationship Between Inflation and Price Dispersion? Market Power vs. Price Rigidity" by Sascha Becker, March 2011.
020 "How Computational Statistics Became the Backbone of Modern Data Science" by James E. Gentle, Wolfgang Härdle and Yuichi Mori, May 2011.

SFB 649, Ziegelstraße 13a, D-10117 Berlin
http://sfb649.wiwi.hu-berlin.de

This research was supported by the Deutsche Forschungsgemeinschaft through the SFB 649 "Economic Risk".