A joint newsletter of the Statistical Computing & Statistical Graphics Sections of the American Statistical Association.

April 97 Vol.8 No.1

Statistical COMPUTING & GRAPHICS

A WORD FROM OUR CHAIRS

Statistical Computing

Daryl Pregibon is the 1997 Chair of the Statistical Computing Section. While this is his first column for the Newsletter, he is familiar with the role of Section Chair.

This is my first column as Chair of the Statistical Computing Section, but alas, it is not my first tour of duty. Imagine if you will, a time when a well-groomed statistician could be sighted wearing bell-bottom trousers and a zipper-front shirt, perched atop 3 inch platform shoes. These fashion statements are so old that they are now back – so it goes with my tenure as Chair. And as before, I look forward to the responsibilities that the position brings – including the opportunity to write this column!

CONTINUED ON PAGE 2

Statistical Graphics

Sally Morton is the 1997 Chair of the Statistical Graphics Section. She welcomes your feedback and comments on any issue of interest to your section.

Let me begin my first column by thanking Bill DuMouchel for his leadership as our Statistical Graphics Chair during the past year. I'd also like to thank Deborah Donnell and Jane Gentleman for their service as Publications Officer and Council of Sections Representative respectively over the past three years.

In 1997, Michael Minnotte will be our new Publications Officer and Bob Newcomb continues as our Secretary/ Treasurer. Lorraine Denby and Colin Goodall continue

CONTINUED ON PAGE 3

TOPICS IN MEDICAL IMAGING

Functional Magnetic Resonance Imaging is a Team Sport

By William F. Eddy

It began like this. It was a cold, dreary day in November 1993. I was sitting in my office studying some boxplots when someone knocked. I looked up. It was a rather wild-looking man, long gray hair flying about. He was very animated, his intensity almost scary. "I'm a psychologist; I need some help," he said. I smiled; it was usually the other way around. He showed me a picture, a slice of a human brain, he said. I commented that it didn't seem like very good resolution; "I thought X-rays were a lot clearer than this fuzzy thing." He said it was done with magnetic resonance not x-rays and he was studying brain function not brain structure. I tried to absorb the idea. He said the picture was not actually a picture of a brain but rather a visual display of a collection of t-tests, one for each pixel in the slice. I smiled again thinking how silly it was to do thousands of t-tests. He noticed my smile and said "I don't know any statistics but someone told me about 'multiple comparisons.' I've read up on it and I did the Bonferroni corrections." My smile turned into outright laughter. "Nothing is significant," he said; "please help me." By now I was figuratively rolling on the floor. We talked a bit more; I decided this was not a problem I wanted to work on. I gave him the names of a few colleagues who, I thought, might straighten him out; I imagined a few Bayesian thoughts might thoroughly discombobulate him. I sent him away and promptly forgot all about the incident.

Aside: Just so you have some idea what this is about, I have included a couple of figures (page 17). The first is a functional image of a brain; it takes about 32 milliseconds to acquire the data for it (although the duty cycle of the machine and FDA regulations limit the acquisition

CONTINUED ON PAGE 17

EDITORIAL

On-Line PDF Version

This Newsletter marks our first full year as editors. As we commented in our maiden issue, we inherited a quality publication from Mike Meyer and Jim Rosenberger. During the past year, one of our goals has been to maintain the standards they set during their three year tenure as editors. So, you are probably asking yourself, why is the quality of the paper not as nice this time around? The answer is simple: because we wish to keep offering you a first-rate product. For many of the articles we publish, color figures constitute an essential ingredient, but, alas, they are expensive to print. With the money we save on the paper we can offset the increasing cost of color printing and still balance the budget.

There are, in fact, other avenues that could be explored in order to cut printing costs. One possibility would be to stop printing the newsletter altogether. This might be a drastic measure, but we are now offering a much improved on-line version at `http://cm.bell-labs.com/who/cocteau/newsletter/`. You can conveniently browse through the new PDF (Portable Document Format) version using the Acrobat reader available free from Adobe. The old PostScript version is still available, but we strongly encourage you to rely on the new PDF version, a much more interactive document due to the addition of article listings and various links. As we are contemplating polling our readers on the issue of phasing out the printed version, we would greatly appreciate receiving your suggestions on this matter. Please, let us know what you think and what questions we should ask.

In this issue, we feature an article by William Eddy on functional MRI. Bill tells us, in a lively and informal style, how he became interested in the study of brain function, what statistical issues are involved, and why this scientific work would be impossible without the cooperation of researchers from fields as varied as psychology, radiology and (of course!) statistics.

In another interesting article, John Salch and David Scott describe a novel exploratory tool that they developed to visualize high dimensional data. The method they propose, called the Density Grand Tour, is based on the combination of data exploration along continuous paths through lower dimensional projections and efficient density estimation techniques. John and David present examples of univariate and bivariate tours and discuss the difficulties associated with the implementation of trivariate tours.

Data visualization in the context of large-scale gene expression analysis is the topic of a paper by Daniel Carr, Roland Somogyi, and George Michaels. The tools they propose for looking at gene expression data include stereo plots, parallel coordinate (time series) plots and conditioned parallel coordinate plots.

Also, make sure you don't miss the article by Clive Loader describing Locfit, a software package designed to be used in S/Splus that performs local regression, likelihood and related smoothing procedures.

Among the several important notices and announcements that we hope you will find useful, let us point out the report on the results of the Computing Section's Student Paper Competition and the announcements for next year's competition and it's companion to be sponsored for the first time by the Graphics Section.

This issue should reach your mailboxes before the Joint Statistical Meetings in Anaheim and contains a timely article by our Program Chairs, James Rosenberger and Dianne Cook, outlining our sections' activities at the Meetings. The plans are very exciting and we hope to see all of you there.

Mark Hansen
Editor, Statistical Computing Section
Bell Laboratories
`cocteau@bell-labs.com`

Mario Peruggia
Editor, Statistical Graphics Section
The Ohio State University
`peruggia@stat.ohio-state.edu`

FROM OUR CHAIRS (Cont.). . .

Statistical Computing

CONTINUED FROM PAGE 1

Since my previous term, important changes have occurred on the statistical computing landscape and I'll try to highlight some of these. Perhaps the most important change to the Section is this Newsletter. While all of us now take it for granted as one of the benefits as Section members, it is a very recent innovation. In my opinion, it has evolved into the premier newsletter in the Association, combining interesting (and readable!) technical articles with timely announcements concerning activities relevant to the Section membership.

Many of us remember the good old days of the ARPANET, a world-wide web free of banners and junk email. But with all the bad comes a lot of good, namely the Section's web presence at `http://www-stat.montclair.edu/asascs`. This is the best place to look for information on Section activities, including online versions of the Newsletter.

The final bit of nostalgia concerns a subject near and dear to many of our hearts — namely data analysis. This subject gained respect in statistical circles in the early 70's. But this pales in comparison to the visibility it is now receiving in computer science circles under the name "data mining." Is it good or bad? In my view it is very good as it provides a tremendous opportunity for statistics to illustrate its central role in science and technology. While the end product is shared between statistical and CS researchers, the means to the end is often quite different. In order for both disciplines to better appreciate what each other has to offer, the ASA is cooperating with AAAI (American Association for Artificial Intelligence) by co-locating the premier data mining conference, KDD-97 (Knowledge Discovery and Data Mining) with the JSM. The two conferences will be back-to-back in southern California during the week of August 10-17. Depending on demand, there will be bus shuttle service on Thursday August 14 to transport JSM participants from Anaheim to nearby Newport Beach. I encourage anyone interested in attending this conference to register early since the already steep registration fees increase through time. (Please note that the registration includes admission to *all* tutorial sessions.)

I saw many of you at the very successful Interface '97 Conference held last month in Houston. In addition to the technical presentations and the timely conference theme, "Mining and Modeling Massive Data Sets", Interface '97 also provided an opportunity for the Executive Committee to meet. Three items that I would like to highlight are featured as articles elsewhere in this issue of the Newsletter: the successful student paper competition organized by the Section, the request for participation on the Executive Committee in the Continuing Education area, and the Section's Program at the JSM in Anaheim. The other relevant item of business is that the Executive Committee approved limited funds to assist Section members who want to attend KDD-97. If you are in this category, send me an email message outlining your financial constraints and the willingness of your institution to provide "matching funds".

Finally, I have just received the ASA election results. Please join me in congratulating our new Chair-elect, James L. Rosenberger from Pennsylvania State University; our Program Chair-Elect, Mark Hansen from Bell Labs; and our new Secretary/Treasurer, Merlise Clyde from Duke University. Complete election results will be available shortly from the ASA home page. I especially want to thank all of you who were candidates and all of you who took the time to send in your ballots.

On behalf of all the current and newly elected Section Officers, we look forward to seeing you in Anaheim in August – have a great summer!

Daryl Pregibon
Statistics Research AT&T Labs
`daryl@research.att.com`

Statistical Graphics

CONTINUED FROM PAGE 1

continue as Council of Sections representatives, and are joined by Roy Welsch this year. Mike Meyer is our incoming Chair-Elect. Both the Graphics and Computing Sections are ably served by our Newsletters Editors Mario Perruggia (Graphics) and Mark Hansen (Computing) who continue to maintain our high quality and informative section Newsletter. Section members are encouraged to discuss any issues or concerns regarding the section or ASA matters with any section officer. Contact information is provided on the inside back page of this issue.

Later in this issue, our ASA Program Chair Dianne Cook describes in detail the Graphics Program for the August Joint Statistical Meetings in Anaheim. Di and her session organizers are to be congratulated for putting together an innovative and exciting program for us. I hope to see you at the Graphic sessions, including the Data Exposition poster session which involves analyzing a hospital report card dataset. Though it's too late to submit a poster, you can access the data for this and past years' Expositions via the web at `http://lib.stat.cmu.edu/data-expo/`. Above all, be sure not to miss the Computing and Graphics mixer Monday night at the Meetings. We'll have door prizes and good food in addition to the latest section news. If you can stay on in Southern California, consider attending the Third International Conference on Knowledge Discovery and Data Mining (KDD-97) to be held in Newport Beach immediately following the ASA Meetings (`http://www.aig.jpl.nasa.gov/kdd97`).

Even though this August 1997 is still a few months away, it's not too early to start thinking about the August 1998 Joint Meetings in Dallas whose theme is "Statistics: A Guide to Policy." Ed Wegman is our Program Chair-Elect. Please contact Ed if you have ideas regarding the Graphics Program.

Your section dues have been put to good use in the past year. In particular, our section contributed to the Elizabeth Scott and George Snedecor Awards initiated by the Committee of Presidents of Statistical Societies. These awards are given in alternate years. The 1996 Elizabeth Scott award was given to Grace Wahba for fostering the role of women in statistics, and the 1995 Snedecor Award for an exceptional published paper in Biometry was given to Norman Breslow and David Clayton.

In terms of new initiatives, in the coming year our section will fund a student paper competition in parallel with the Computing Section's competition. Please see the article later in this issue describing the competition rules and deadlines, and encourage your students to enter. At the Interface '97 Meetings in May, we discussed our sponsorship of this competition and the Poster Competition for students in kindergarten through high school. In addition, we considered how our section might support the Journal of Computational and Graphical Statistics (JCGS), including the possibility of offering a prize for the best graphics paper. We especially need suggestions from you regarding Continuing Education Courses you'd like to take or teach, a dataset for the next Data Exposition, and how the section might serve the membership better.

We welcome ideas for other initiatives, and we can always use volunteers. If you'd like to get involved with the section or have comments, please contact me via email.

I look forward to seeing all of you at conferences and symposia throughout the year,

Sally Morton
RAND
Sally_Morton@rand.org

NEWS CLIPPINGS AND SECTION NOTICES

Results of the Student Paper Competition

By Daryl Pregibon and Trevor Hastie

The results are in! We have selected the four winners of the Statistical Computing Section's 1997 Student Paper Competition. In alphabetical order they are

- **Wenjiang J. Fu**, University of Toronto
  Penalized Regressions: the Bridge versus the Lasso
- **Alan Gous**, Stanford University
  Adaptive Estimation of Distributions using Exponential Sub-Families
- **Gareth James**, Stanford University
  The Error Coding Method and PaCT's
- **Ramani S. Pilla**, Pennsylvania State University
  New Cyclic Data Augmentation Approaches for Accelerating EM in Mixture Problems

Before we tell you more about each of the winners, a bit more about the competition itself. This is our third such competition, and this year we had 16 entries. There were many good submissions, but these four prevailed. If you plan to attend the ASA meetings this year, come and see for yourself!

As part of their prize, each of the four winners will present their papers in a special session at the ASA. The more tangible part of their prize is that their entire conference trip will be covered by the section – airfare, hotel and registration! The approximate value is $1000 each. They also get their names and faces in this newsletter as well as the Amstat news – publicity gives a great kick start to a career!

We would like to thank two colleagues who helped in the review process this year: Mark Hansen at Bell Labs and Bill DuMouchel at AT&T Labs. Mark, as you know, is the current Co-Editor of this Newsletter, and Bill is Past-Chair of the Section.

We intend to hold a student paper competition every year and in the upcoming year there is a strong possibility of joining forces with the Graphical Statistics Section. This would allow for even more winners and hopefully even more submissions. Registered students are eligible to submit papers. See the announcement on page 34 of this newsletter for more details of the competition and conditions for entry. Professors, please make a note in your calendar to get your students geared up in time for next year's submissions. The deadline for submissions for the 1998 competition is early January 1998.

Details on the Winners

Wenjiang J. Fu

Wenjiang is a fourth year Ph.D. student in biostatistics at the University of Toronto under the supervision of Professor Robert Tibshirani. He is graduating this year. His thesis is titled "A Statistical Shrinkage Model and Its Applications". Wenjiang is pursuing a career in applied statistical research. His research interest includes biostatistics and statistical computing, particularly in longitudinal models, survival models, cancer research, and AIDS studies. He received a B.Sc. in Mathematics from Peking University, an M.Sc. in Mathematics from Tsinghua University, and an M.Sc. in Statistics from the University of Toronto.

Penalized Regressions: the Bridge versus the Lasso

We consider Bridge regression, a special family of penalized regressions with penalty function $\sum |\beta_i|^\gamma$. A general approach to solving for the regression coefficients is developed. A simple and efficient algorithm for the Lasso ($\gamma = 1$) is obtained by studying the structure of Bridge solutions. The shrinkage parameters, λ and γ, are optimized via generalized cross-validation (GCV). Comparison among several models are made through a simulation study. The method is demonstrated through an application to prostate cancer data. Some computational advantages and applications are discussed.

Alan Gous

Alan is a third year Ph.D. student in the Statistics Department at Stanford University. His advisor is Brad Efron. Alan received a B.Sc.(hons) from the University of Natal, Pietermaritzburg, South Africa. His research interests include exponential families, splines, numerical optimization methods in statistics, and bioinformatics.

Adaptive Estimation of Distributions using Exponential Sub-Families

Suppose we are given a number of groups of data points on the real line. The observations are i.i.d. within each group, but the groups are drawn from different distributions, not well modeled as members of one of the standard distribution families. An algorithm is presented which finds an appropriate low-dimensional exponential family to model these data from among sub-families of an exponential family of much larger dimension. Model selection between sub-families of different dimension is discussed. The algorithm is implemented for the special case in which the large family is a logspline family of distributions. An example data set is analyzed using these methods.

Gareth James

Gareth is a third year Ph.D. student in the Statistics Department at Stanford University. His advisor is Associate Professor Trevor Hastie. Gareth plans to pursue an academic career because he enjoys both teaching and research. His research interests include classification problems with large numbers of classes, multivariate statistics and generalizations to bias and variance concepts.

The Error Coding Method and PaCT's

A new class of classification techniques have recently been developed in the statistics literature. A "plug in" classification technique (PaCT) is a method that takes a standard classifier (such as LDA or nearest neighbors) and plugs it into an algorithm to produce a new classifier. The standard classifier is known as the Plug in Classifier (PiC). These methods often produce large improvements over using a single classifier. In this paper we investigate one of these methods developed by Dietterich and Bakiri (1995). They use Error Correcting Coding Theory ideas to motivate their method but provide little hard theory. We give some further insight into the problem.

Ramani S. Pilla

Ramani is a final year Ph.D. student in the Department of Statistics at the Pennsylvania State University. Her thesis is entitled "EM-Based Methods in High-Dimensional Finite Mixtures: Theory & Applications," and her advisor is Professor Bruce G. Lindsay. Ramani's career plans include research, teaching, and consulting. Her statistical interests include likelihood methods in mixtures, missing data analysis, modern statistical computation, and statistical modeling with applications to mixture models, image analysis, genetics, and neural networks.

New cyclic data augmentation approaches for accelerating the EM algorithm in mixture problems.

Consider a general class of maximum likelihood (ML) problems in which one is interested in maximizing a mixture likelihood with finitely many known component densities over the set of unknown weight parameters. In this mixture setting, convergence of the EM algorithm will be extremely slow when the component densities are poorly separated and when the ML solution requires some of the weight parameters to be zero. To address this problem, two basic devices are used to construct new EM type algorithms. First, the "complete data" is constructed in two new ways, each designed to increase the convergence rate in the determination of the

relative weights of pairs of neighboring mixture components. These representations correspond to two different data augmentations; neither greatly increases the numerical complexity of the EM algorithm. Secondly, cyclic versions of these two basic methods are created by changing the missing data formulation between EM steps. Results on high dimensional mixtures of binomial, Poisson, and normal show that the new methods converge many (100 to 300) times faster than the conventional EM across these three cases.

Congratulations!

Congratulations to our four winners, and to everyone who participated in this year's competition.

Trevor Hastie
Stanford University
trevor@playfair.stanford.edu

Daryl Pregibon
Statistics Research AT&T Labs
daryl@research.att.com

ASA Election Results

Recently you were asked to vote for a number of ASA Officers, including ASA President and Vice-President. In a memorandum from Ray A. Waller, Executive Director, the ASA has just released the election results:

President-Elect, 1998 (President 1999)
Jonas H. Ellenberg, Westat, Inc.

Vice President, 1998–2000
Mary Ellen Bock, Purdue University

In addition, several officers of the Statistical Computing and Graphics Sections were elected during this round of balloting.

Statistical Computing

Chair-Elect
James L. Rosenberger,
The Pennsylvania State University

Program Chair-Elect:
Mark H. Hansen, Bell Laboratories

Secretary/Treasurer (1998–99):
Merlise Clyde, Duke University

Statistical Graphics

Chair-Elect:
Dianne Cook, Iowa State University

Program Chair-Elect:
Deborah Swayne, Statistics Research AT&T Labs

Council of Sections Representative (1998–2000):
David W. Scott, Rice University

Newly elected Section Officers should attend our Business meeting/mixer in Anaheim.

We especially want to thank those of you who were candidates, as well as those of you who took the time to vote. Your continuing interest in and support for our Sections' activities are greatly appreciated.

Buja Named JCGS Editor

The Journal of Computational and Graphical Statistics management committee is pleased to announce the appointment of Andreas Buja, as the new editor of the journal 1998-2000. He will succeed William J. Kennedy, who has held the editorship since 1995. For 1997 Buja will serve as Editor-Elect, and then he will serve as Editor for a three year term, from 1998 through 2000. The appointment was recommended by the management committee and approved by the boards of the three sponsoring organizations: the American Statistical Association, the Institute of Mathematical Statistics, and the Interface Foundation of North America.

Manuscripts and correspondence should continue to be sent to the current editor:

William J. Kennedy, Editor JCGS
Department of Statistics
117 Snedecor Building
Iowa State University
Ames, IA 50011
Email: JCGS@iastate.edu

until August 31, 1997. From September 1 on, new submissions should be sent to the new JCGS editor:

Andreas Buja, Editor JCGS
AT&T Laboratories
Email: andreas@research.att.com
http://www.research.att.com/~andreas

The mailing address and phone numbers are available from the URL above.

Andreas Buja's current mailing address is:

AT&T Laboratories
180 Park Ave
P.O. Box 971
Florham Park, NJ 07932-0971

The management committtee extends our thanks to Stephen Fienberg, outgoing chair of the JCGS management committee, for conducting a vigorous search resulting in this important appointment.

James L. Rosenberger
Chair, JCGS Management Committee
The Pennsylvania State University
jlr@stat.psu.edu

Invitation to Membership on SCS Continuing Education Committee

By Tom Devlin

The Statistical Computing Section has had a very successful continuing education program over the past five years. We have offered exciting courses at the Joint Statistical Meetings (JSM) and have co-sponsored courses with other organizations such as NIBS and the Interface Foundation. These offerings have included:

- "Extending the Cox Model" by Terry Therneau
- "Introduction to Complex Bayesian Modeling using BUGS" by David Spiegelhalter and Nicole Best
- "Multivariate Density Estimation and Visual Clustering" by David Scott
- "Modern Nonparametric Regression and Classification" by Trevor Hastie and Rob Tibshirani
- "Resampling-Based Multiple Testing" by Peter Westfall and Stanley Young

I have been appointed the Section's Electronic Communication Liaison and am stepping down as chair of the Continuing Education (CE) Committee. The Section invites people with strong interests in education and continuing education to membership on the CE Committee and to help coordinate the CE program. Responsibilities of the CE Committee include:

- Solicitation of proposals for presentation at the JSM and at Section co-sponsored events, e.g. the Interface;
- Review and selection of proposals;
- Management of the CE student scholarship program.

Continuing education is a very important service that the Section provides to its members and to the profession. I have truly enjoyed participating in the development of the program and think that it offers a great opportunity to help promote statistical education and thinking. If you have an interest in serving on the Statistical Computing Section CE committee, please contact me via email.

Tom Devlin
Montclair State University
devlin@mozart.montclair.edu

TOOLS FOR DATA ANALYSIS

Data Exploration with the Density Grand Tour

By John D. Salch and David W. Scott

Exploratory analysis can often lead to effective modeling and powerful inference. Interactive tools, such as XGobi (Swayne, Cook, and Buja, 1990), provide a unified environment to work with techniques available for this purpose. However, these methods become infeasible when applied to moderately large ($n \approx 10,000$) data sets. It is pondering this deficiency that has driven the work discussed herein.

The reliance on scatterplots, as a means of visualizing data, represents one computational stumbling block for standard exploratory methods. With the resolution of modern CRT's it seems very unlikely that visualizing data points themselves will remain feasible without subsetting. Other graphical methods, such as binning (Carr 1991), glyphs (Chambers et. al. 1983), and high resolution density plots (Scott 1996), provide practical alternatives.

One commonly used method for exploring high-dimensional data is the Grand Tour (Asimov 1985). The basic premise is to explore data of high-dimension by continuous paths through projections in 1, 2, or 3-dimensions. This method involves viewing a sequence of scatterplots in time, allowing the user to "tour" the data in search of interesting structure. This method works well if the data set is reasonably small and if its dimension is not too large.

We propose combining efficient density estimation, via the averaged-shifted histogram (ASH) (Scott 1985), with the Grand Tour to create the Density Grand Tour, or DGT. Here we describe software that we are making available for the SGI and demonstrate it with real data.

Univariate Tour: The Iris Data Revisited

The 1-D DGT can be described as an attempt to view all possible one-dimensional projections of the data in a semi-orderly fashion. In the case of Fisher's Iris Data, there are 4 dimensions and 150 data points, which represent measurements on 3 different species of irises (setosa, virginica, versicolor), 50 points per species.

To start, the user first invokes Geomview (Phillips 1993), an interactive environment for displaying geometric entities, produced and distributed by researchers at the NSF Geometry Center (`http://www.geom.umn.edu`). This environment includes many modules providing such features as zoom, rotation, and rearranging of light sources as well as PostScript snapshots. The ability to easily add new modules is very powerful and facilitated the coding of the DGT.

Once Geomview is started, the user can then select the "DGT 1D" from the "modules" menu. The Density Grand Tour operates as a module of Geomview, written entirely in C, thus eliminating the need to write our own display code.

After invoking the DGT (1-D in this case) a Graphical User Interface (GUI) appears allowing interactive control of the tour. Features include control over the ASH through the number of bins and the smoothing parameter. Besides flowing forward in time, the user can stop the tour, rewind it, and step through at at leisure. Figure 1 demonstrates the GUIs for both the 1-D and 2-D tours.

Three frames from the tour for the Iris data are displayed in Figure 2. These frames represent a small but interesting part of a complete tour and hint at the dynamic nature of the tour. As one can notice, in the first frame there appears to be only one mode present, but by the third frame the multimodality of the data becomes clear. In this example, the frames are 5 time units, or "steps," apart, which represents a reasonably smooth transition. However, the smoothness, or "step size," of the animation can be adjusted as the user desires.

Other features of the DGT include "brushing" and the visual display of the data points and projection (weight) vector, extensions of XGobi capabilities. Brushing is implemented by computing a separate density estimate for each unique group. Then all of these densities are drawn on the same plot with different colors as shown in Figure 3. In the case of the 1-D DGT, the different group-specific densities are plotted with an offset so as to prevent overlap. The user can also select "Draw Points" to have the projected data points drawn adjacent to the 1-D ASH, while the tour is running. "Add Weights" can be selected to display the projection vector currently being used.

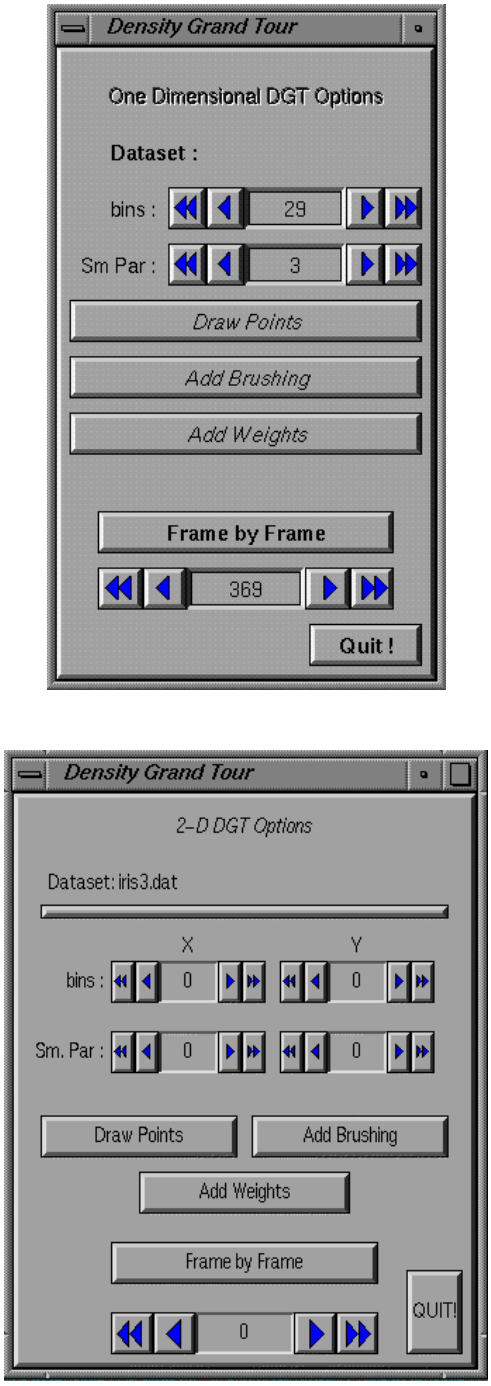


Figure 1. Interactive GUIs for the DGT 1-D (top panel) and the DGT 2-D (lower panel).

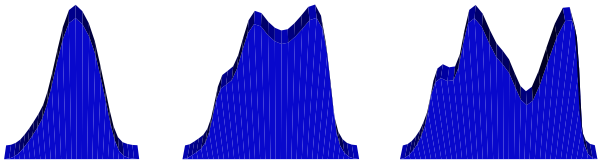


Figure 2. Three frames from the univariate tour of the Iris Data.

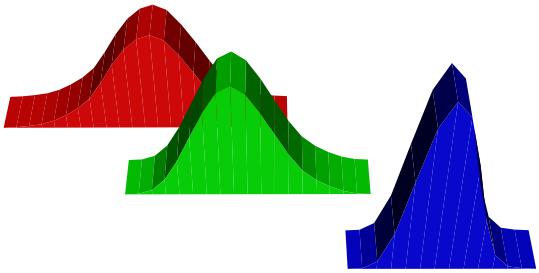


Figure 3. The effect of brushing on the univariate tour.

Bivariate Tour: Landsat Image Data

A demonstration of the 2-D DGT is given using a subset of a Landsat satellite image. The data represent $8,034$ pixels from three groups of farm crops (sunflower, spring wheat, and spring barley). Subsetting was performed only to reduce the data down to three distinct subgroups (from 32), so as to reduce confusion in the figures below. Within each of the three subgroups the data are otherwise complete.

The 2-D DGT is invoked similarly to the 1-D version, but the GUI is slightly different (see Figure 1). Since the ASH is based on two dimensional data, there are controls for the number of bins and smoothing parameter in both the "X" and "Y" directions.

Figure 4 illustrates 5 frames (sequential in time) from a tour of the Landsat data. After looking at these frames one might suspect that there is at least one crop contributing to the multimodality. Using ground truth, which was available for these data, brushing was activated giving a clear picture of how the individual groups are contributing to the whole.

As before, brushing is implemented by plotting a separate density estimate for each brushed group on the same graph (see Figure 5). However, for the 2-D DGT, the plots are allowed to overlap thus obscuring some of the density for each. Despite the overlap, we can still observe which group has the highest density at each set of coordinates. We have explored ideas such as utilizing hardware transparency to avoid the visual confusion, but these have proven unsatisfactory.

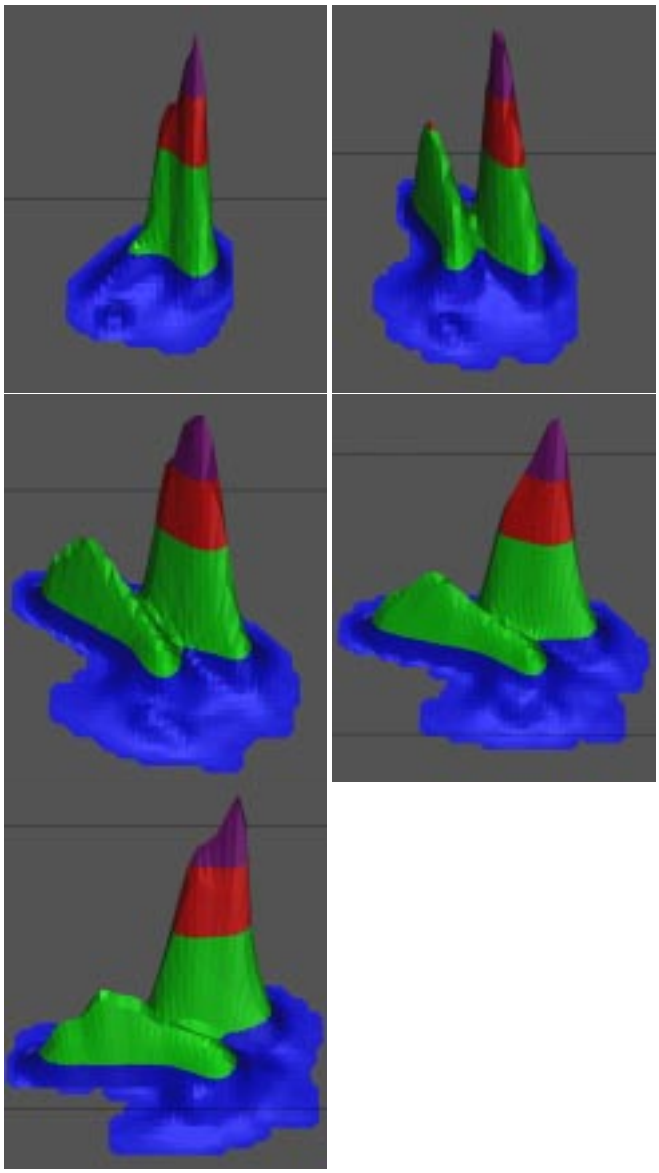


Figure 4. Five frames from the bivariate tour of the Landsat Image Data. Time progresses from left to right and from top to bottom.

The user can also select "Draw Points" to superimpose a scatterplot of the data. The points are plotted with a fixed Z-coordinate so as to appear above the density plot in the graphics window. "Add Weights" displays both projection (weight) vectors used to obtain the projected data at the current time.

Trivariate Tour

A working module of a 3-D DGT was written using Splus and the Animate module of Geomview. We discovered that the visualization overhead was too high for all but a few of the fastest machines. While conceptually more powerful than the bivariate tour, the trivariate tour is much more difficult to follow from frame to frame.

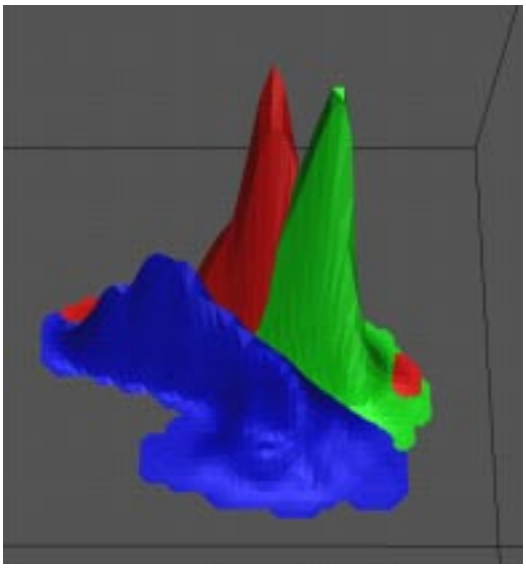


Figure 5. The effect of brushing on the bivariate tour.

This problem is magnified if the data have any non-trivial structure. Thus, we think that this option will not be available for a few more years (see Figure 6).

Future Work

There are several areas we are currently exploring as to improve this software. Extending the idea of the Projection Pursuit Guided Tour (Cook et. al. 1991) could be very valuable as the dimensionality of the data become more massive. The time cost of exploring all possible 2-dimensional projections of, say, 7-dimensional data could be greatly reduced by locally optimizing a PP-index. However, the indices we have reviewed tend to be too computationally expensive for use in a dynamic setting such as this one. Thus, research into an efficient index is underway.

We are also currently working on a new method to obtain better ASH estimates while reducing the amount of data storage required. This could be very helpful in increasing the feasibility of the DGT for larger data sets. We also think that high-dimensional binning could be done efficiently and current results look promising.

We are also hoping that the DGT could possibly be included in a future version of XGobi. This could combine both the power of the XGobi interface and its computational efficiency with our method of exploration.

Software Availability

Software can be obtained via anonymous ftp from `ftp.stat.rice.edu` in the subdirectory `/pub/scottdw/DGT.code`. The software is written entirely in C with the interface to Geomview. Note that versions of Geomview do exist for other computing platforms that run X-windows including PC's and Sun workstations. The DGT code (written in standard C) has been compiled and tested on several platforms as well. Thus, it is possible to try the code on a variety of machines. However, due to the dynamic graphical nature of the product, we recommend using high-end SGI machines.

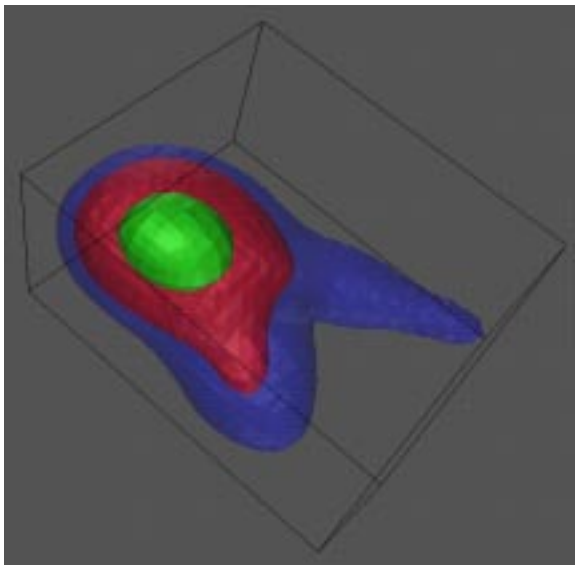


Figure 6. Contours from a trivariate ASH. This would represent a single frame from a trivariate tour.

References

Asimov, D., (1985), "The grand tour: A tool for viewing multidimensional data." *SIAM Journal on Scientific and Statistical Computing*, 6, 128-143.

Carr, D.B., (1991), "Looking at large data sets using binned data plots." *Computing and Graphics in Statistics*, 7-39.

Chambers, J.M., Cleveland, W.S., Kleiner, B., Tukey, P.A., (1983), *Graphical Methods for Data Analysis*. Wadsworth, Belmont, CA.

Cook, D., Buja, A., Cabrera, J., (1991), "Direction and motion control in the grand tour." In *Computing Science and Statistics. Proceedings of the 23rd Symposium on the Interface*, 180-183.

Phillips, M., (1993), *The Geomview Manual* `ftp://geom.umn.edu`.

Scott, D.W., (1985), "Averaged shifted histograms: Effective nonparametric density estimators in several dimensions." *The Annals of Statisics*, 13, 1024-1040.

Scott, D.W., (1995), "Incorporating density estimation into other exploratory tools." *ASA Proceedings of Statistical Graphics Section*.

Support

This research was supported in part by the National Science Foundation (DMS-9626187) and the National Security Agency (MDA 904-95-C-2203).

John D. Salch
Rice University
jsalch@stat.rice.edu

David W. Scott
Rice University
scottdw@stat.rice.edu

Locfit: An Introduction

By Clive R. Loader

What is Locfit?

Locfit is a software package performing local regression, likelihood and related smoothing procedures. It is designed to be used in S/Splus, making extensive use of the data management and graphical facilities in that language. The interface uses the modeling language and classes and methods of S, so users familiar with the existing modeling software in S (Chambers and Hastie 1992) should find Locfit easy to use. Most of the numerical routines of Locfit are written in C, and it is also possible to use Locfit as a stand-alone C program.

Locfit can be obtained via the WWW at the URL `http://cm.bell-labs.com/stat/project/locfit`; substantial online documentation can also be found at this address.

Local Regression

Local regression was applied in a variety of fields in late 19th and early 20th centuries; see for example Henderson (1916). The current popularity of local regression as a statistical procedure is largely due to the Lowess procedure (Cleveland 1979) and Loess (Cleveland and Devlin 1988).

The underlying model for local regression is

$$Y_i = \mu(x_i) + \epsilon_i;$$

the function $\mu(x)$ is assumed to be smooth and is estimated by fitting a polynomial model (most commonly, linear or quadratic) within a sliding window. That is, for each fitting point x, we consider a locally weighted least squares criterion:

$$\sum_{i=1}^{n} W\left(\frac{x_i - x}{h}\right)(Y_i - (a_0 + a_1(x_i - x)))^2 \qquad (1)$$

By default, Locfit uses the weight function

$$W(v) = \begin{cases} (1-|v|^3)^3 & |v| < 1 \\ 0 & \text{otherwise} \end{cases}.$$

The bandwidth h controls the smoothness of the fit. A large h may result in oversmoothing, or miss important features in the data, while a small h may result in a fit that is too noisy. The simplest choice is to take h constant; often it may be desirable to vary h with the fitting point x.

The local least squares criterion (1) is easily minimized to produce estimates $\hat{a}_0$ and $\hat{a}_1$. The local linear estimate of $\mu(x)$ is

$$\hat{\mu}(x) = \hat{a}_0.$$

Note that each least squares problem produces $\hat{\mu}(x)$ for a single point x; to estimate at additional points, the local weights change and a new least squares problem must be solved.

Our example uses the ethanol dataset, studied extensively in Cleveland (1993), and fits a local quadratic model:

```
> fit.et <- locfit(NOx~E, data=ethanol,
+   alpha=0.5)
> plot(fit.et, get.data = T)
```

Three arguments are given to the `locfit()` function. The model formula, `NOx~E`, specifies a response variable `NOx` and predictor `E`. The `data=ethanol` argument provides a data frame, where the variables may be found. The smoothing parameter is given by `alpha=0.5`; this gives a nearest-neighbor based bandwidth covering 50% of the data. Figure 1 shows the plot of the fit.

Bivariate local regression arises when there are two predictor variables: $x_i = (x_{i,1}, x_{i,2})$. The localization weights then become

$$w_i(x) = W\left(\frac{\|x_i - x\|}{h}\right),$$

where $\|\cdot\|$ denotes the Euclidean norm. The local quadratic model around a point $x = (u, v)$ is

$$\begin{aligned}\mu(x_i) \approx\ & a_0 + a_1(x_{i,1} - u) + a_2(x_{i,2} - v) \\ & + a_3(x_{i,1} - u)^2 + a_4(x_{i,1} - u)(x_{i,2} - v) \\ & + a_5(x_{i,2} - v)^2.\end{aligned}$$

These are substituted into the local least squares criterion (1). Again, we minimize over the coefficients, and take $\hat{\mu}(x) = \hat{a}_0$.

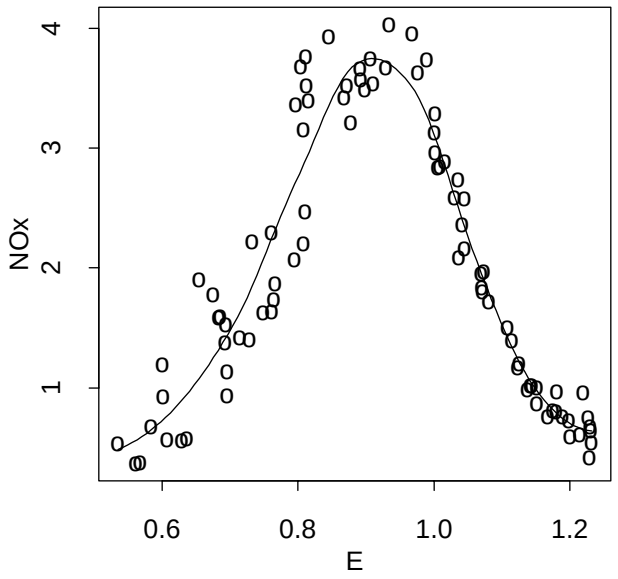


Figure 1. Local quadratic fit for the ethanol dataset.

In Locfit, multivariate local regression is requested by adding additional terms on the right hand side of the formula. For example, the ethanol dataset contains a second predictor variable `C`:

```
> fit.et2<-locfit(NOx~E+C, data=ethanol,
+   alpha=0.5, scale=0)
> plot(fit.et2, type="persp")
```

The `scale` argument allows the user to specify different scales for each variable, used in computing neighborhood weights. When `scale=0` is given, each variable is scaled by its standard deviation. The two dimensional fit can be displayed in several formats: contour plots (the default); perspective plots (Figure 2) and cross sections, using trellis displays (Cleveland 1993).

Local Likelihood

Local likelihood fitting was developed by Tibshirani (1984) and Tibshirani and Hastie (1987). The procedure is applicable in situations such as binary data, when an additive Gaussian model is inappropriate as an error structure. In local likelihood, we simply replace the local least squares criterion by an appropriate local log-likelihood criterion. For binary data, the local log-likelihood is

$$\sum_{i=1}^{n} w_i(x)\left(Y_i \log(p_i) + (1-Y_i)\log(1-p_i)\right)$$

where $p_i = p(x_i)$. We could model $p(x)$ directly using local polynomials; however, it is usually preferable (and the Locfit default) to model via the logistic link function, $\theta(x) = \log(p(x)/(1-p(x)))$. As in the local regression case, we approximate $\theta(x)$ locally by a polynomial, then choose the polynomial coefficients to maximize the likelihood.

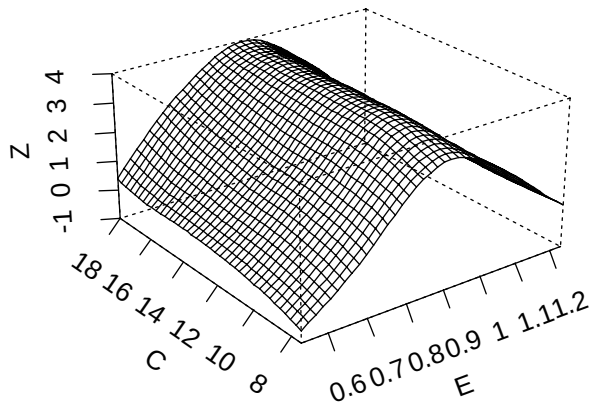


Figure 2. Bivariate local quadratic fit for ethanol dataset.

By changing the likelihood and weighting scheme, local likelihood estimates can be obtained in numerous different settings. Those supported in Locfit include likelihood regression models; density estimation; conditional hazard rate estimation and censored likelihood models.

As an example, we consider a mortality dataset from Henderson and Sheppard (1919). This consists of three variables: age; number of patients of each age, and number of deaths for each age. Local logistic regression is requested by the `family="binomial"` argument, and the number of patients is passed as the `weights` argument:

```
> fit.mo <- locfit(deaths~age,weights=n,
+   family="binomial",data=morths,
+   alpha=0.5)
> plot(fit.mo, get.data=T)
```

Figure 3 displays the result. Note that while estimation is performed on the logistic scale, the result is automatically back-transformed to the probability scale.

For density estimation, the appropriate local likelihood criterion is

$$\sum_{i=1}^{n} w_i(x)\log(f(x_i)) - n\int W\left(\frac{u-x}{h}\right) f(u)du;$$

(see Loader 1996). By default, we use the log-link; that is, $\log(f(x))$ is modeled by local polynomials.

In Locfit, density estimation is requested with `family="density"`; this becomes the default if no response is given in the formula. When `link="ident"` is given, a local polynomial model for the density is used. Local quadratic fitting with the identity link is one construction of the fourth order kernel estimate discussed in section 6.2.3.1 of Scott (1992).

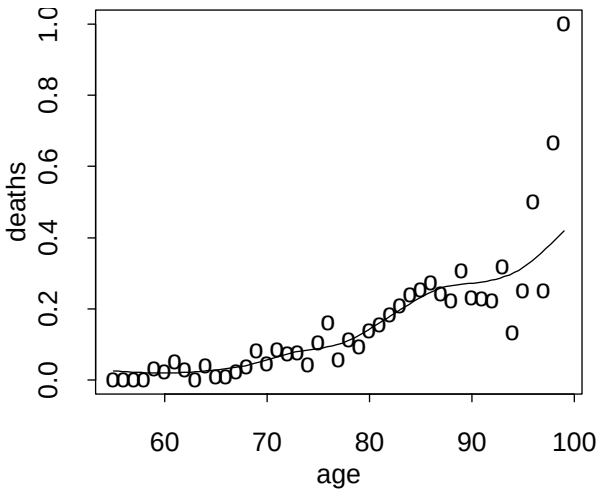


Figure 3. Mortality data of Henderson and Shepherd: Local logistic fit.

Our example estimates the density of the durations of 107 eruptions of the Old Faithful geyser:

```
> fit.of <- locfit(~geyser,flim=c(1,6),
+   alpha=c(0.15,0.9))
> plot(fit.of,get.data=T,mpv=200)
> fit.og <- locfit(~geyser,flim=c(1,6),
+   alpha=c(0.15,0.7),link="ident")
> plot(fit.og,get.data=T,mpv=200)
```

Note the two components to the smoothing parameter `alpha`: the first is a nearest neighbor component, and the second a fixed component. At each fitting point, both components are evaluated, and the larger bandwidth is used in the local likelihood.

From Figure 4 the log-link provides a visually more appealing estimate; Loader (1995) provides substantial evidence that it is also a better estimate. While asymptotic theory suggests the choice of link should have little effect on the estimate, it is often advantageous in practice to choose a link mapping the parameter space to $(-\infty, \infty)$. The Locfit default satisfies this for all models.

Model Assessment

It is well known that smoothing parameters have a critical influence on the smooth curve: A large bandwidth leads to an oversmoothed curve that may inadequately model or completely miss important features, while a small bandwidth may undersmooth the curve, resulting in a fit that is visually too noisy.

A number of tools are available to help assess the performance of smooths. Global criteria such as cross validation and generalized cross validation Craven and Wahba (1979) estimate the average squared prediction error, while the M statistic of Cleveland and Devlin (1988)

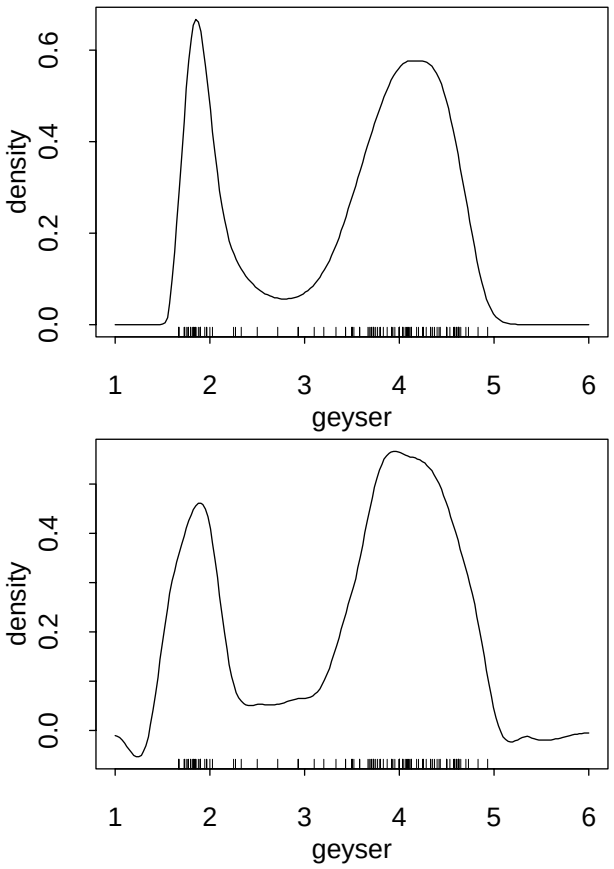


Figure 4. Local quadratic density estimates for Old Faithful data: log link (top) and identity link (4th order kernel) (bottom).

estimates the average squared estimation error. Other tools, such as residual plots and confidence bands, attempt to assess the importance of individual features and check other modeling assumptions.

For example, in Figure 3, the smooth seems to fit the data nicely up to age 90, while the data becomes wild for larger ages. But the number of patients is very small for these larger ages, making it impossible to tell from Figure 3 whether there is any problem. Instead, we examine residuals:

```
> plot(morths$age, residuals(fit.mo),
+   xlab="age", ylab="residual")
```

Several possible definitions of residuals for generalized linear models are given by McCullagh and Nelder (1989), section 2.4. By default, Locfit uses deviance residuals; when the smooth is adequate, the distribution is very approximately $N(0, 1)$. Figure 5 shows there is no problem at the right end; if anything, a small amount of overdispersion in the range $60 <$ `age` < 80 is possible.

Global goodness of fit criteria are readily computed. For example, the generalized cross validation statistic is

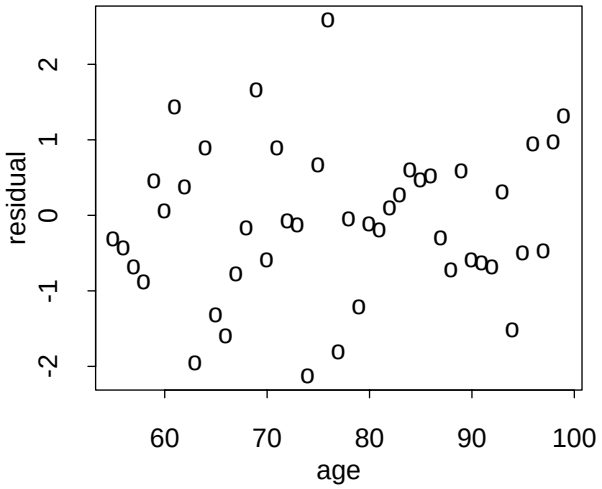


Figure 5. Deviance Residuals for the mortality dataset.

$$\text{GCV} = n\frac{\sum_{i=1}^{n}(Y_i - \hat{\mu}(x_i))^2}{(n - \text{tr}(\mathbf{H}))^2}$$

where $\mathbf{H}$ is the hat matrix. The `"locfit"` object contains all the necessary components to compute this; the `gcv()` function provided with Locfit makes a call to `locfit()` and extracts the necessary components:

```
> gcv(NOx~E,data=ethanol,alpha=0.5)
      lik     infl     vari      gcv
 -4.53376 7.013307 6.487449 0.1216589
```

The `lik` component is -0.5 times the residual sum of squares; `infl` is $\text{tr}(\mathbf{H})$; `vari` is $\text{tr}(\mathbf{H}^\mathbf{T}\mathbf{H})$ and `gcv` is the GCV score. We can easily loop through `gcv()` for several smoothing parameters:

```
> alpha <- seq(0.2,0.8,by=0.05)
> g <- matrix(nrow=length(alpha),
+   ncol=4)
> for(i in 1:length(alpha))
+   g[i, ] <- gcv(NOx~E, data=ethanol,
+   alpha=alpha[i])
> plot(g[,3], g[,4], ylim=c(0,0.2),
+   xlab="Fitted DF", ylab="GCV")
```

The plot is displayed in Figure 6. Note the plot is fairly flat from about 6 to 16 degrees of freedom. This situation is not uncommon, and reflects the difficulty of purely data-based bandwidth selection. Looking at Figure 1, one might argue that the peak should be much flatter than the smooth displays, or possibly even bimodal. The flatness of GCV simply reflects this uncertainty.

Recent literature on bandwidth selection (e.g. Ruppert, Sheather, and Wand 1995) has strongly criticized cross validation and related procedures as being too variable and unreliable. In fact, a careful analysis shows the variability of cross validation is not the problem, but rather

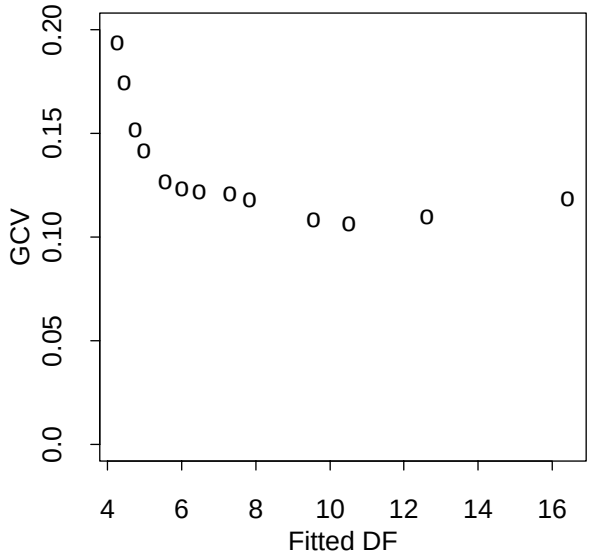


Figure 6. Generalized cross validation plot for the ethanol dataset.

a symptom of how difficult data-based model selection is; for example, reflecting the uncertainty as to the correct amount of smoothing of the peak in Figure 1. Plug-in selectors, often claimed to be less variable, have not magically answered this uncertainty, but effectively make strong prior assumptions as to the correct amount of smoothing. This point is discussed further in Loader (1995), where it is shown overreliance on bandwidth selectors has led to questionable conclusions on some standard examples.

Locally Adaptive Fitting

Sometimes, we may be blessed with large datasets with low noise. In such cases, we can try to choose a separate bandwidth for each smoothing point. Locfit provides one such method for doing so, based on a localized version of AIC.

Figure 7 shows an example, using one of the four examples popularized by Donoho and Johnstone (1994). The S commands producing this example are

```
> x <- seq(0, 1, length.out=2048)
> y <- 20*sqrt(x*(1-x))*
+   sin((2*pi*1.05)/(x+0.05))+
+   rnorm(2048)
> plot(y~x)
> fit.ad <- locfit(y~x, maxk=500,
+   alpha=c(0,0,log(2048)))
> plot(fit.ad, mpv = 2048)
> plot(predict(fit.ad, what="band"),
+   type="p")
```

The locally adaptive fit is requested by providing a third component to the smoothing parameter `alpha`; this specifies a penalty for the 'number of parameters' in the

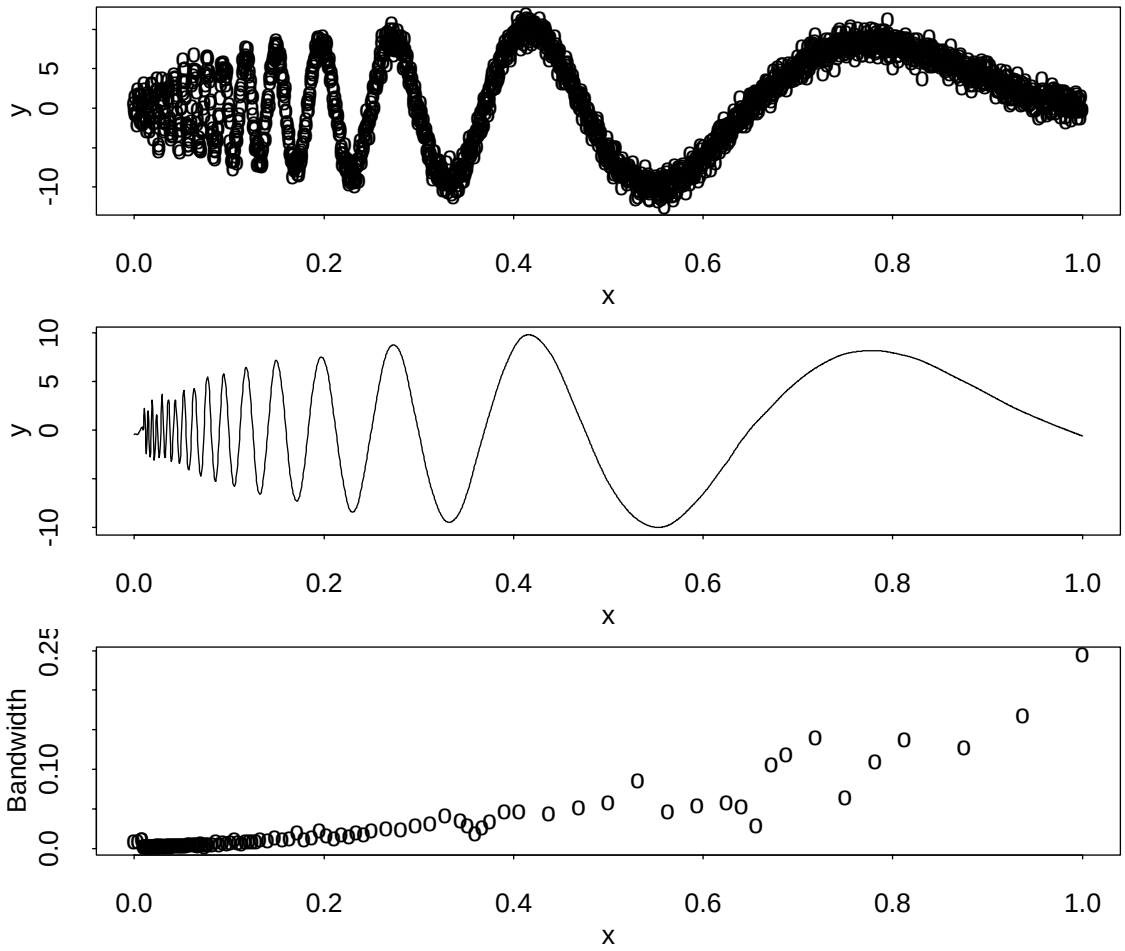


Figure 7. Locally adaptive fit. Dataset (top); Smooth fit (middle) and bandwidths used (bottom).

local AIC criterion. Usually, a value slightly larger than $2\sigma^2$ produces the most satisfactory results.

While examples like Figure 7 are challenging from an algorithmic viewpoint (it is hard to make a computer draw a smooth curve through the data), they are quite simple from a statistical viewpoint (the structure is quite obvious, and a perfectly adequate smooth would be obtained with a pen). For the types of data statisticians often face — more noise, and less obvious structure — locally adaptive smoothing is likely to be less satisfactory. The model selection uncertainty identified in the previous section applies equally to locally adaptive selection, and there can be no guarantee that the smooth produced by Locfit (or any other locally adaptive smoother) matches what the user desires.

Classification

Classification problems, where one attempts to classify observations into two or more classes, can be addressed using either logistic regression or density estimation. As an example, we consider classifying the versicolor and virginica species from Fisher's Iris data set, based on the petal measurements. Local logistic regression (the default for a `T/F` response) is used:

```
> fit.ir<-locfit(
+  I(species=="virginica")~petal.wid+
+  petal.len,scale=0,data=iris)
> plot(fit.ir, v = 0.5)
> plotbyfactor(petal.wid,petal.len,
+  species,data=iris,pch=c("O","+"),
+  col=c(1,1),add=T,lg=c(1,7))
```

Figure 8 shows the resulting fit, with the single contour plotting the classification boundary.

We can also estimate the error rate, using cross validation:

```
> table(fitted(fit.ir,cv=T)>0.5,
+   iris$species)
      versicolor virginica
FALSE         47         2
 TRUE          3        48
```

Here, we estimate the cross validated fit, and hence the misclassification rate is estimated as $5/100$.

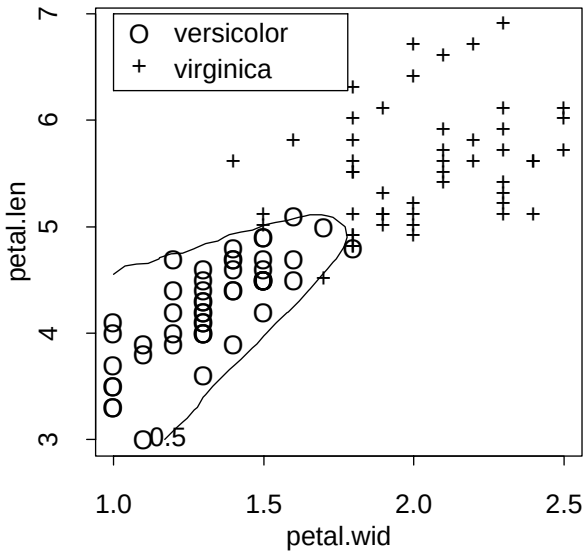


Figure 8. Classification boundary for Fisher's Iris data.

Computational Algorithms

Modern work stations and PC's are sufficiently fast that direct implementation of local regression for datasets with a few hundred points is not a problem. For larger datasets, or iterative procedures such as local likelihood, computational approximations become useful.

The computational methods used in Locfit develop the ideas used in Loess (Cleveland and Grosse 1991). Roughly, the local regression/likelihood is performed directly at a small number of points, and the fit at these points is smoothly interpolated to obtain the fit at remaining points. Locfit differs from Loess in the way points are selected for direct fitting. While Loess bases its choice on the density of data points, Locfit uses the bandwidths at fitting points. This allows Locfit to adapt to different bandwidth schemes: fixed, nearest-neighbor, and locally adaptive. The power of this approach becomes apparent in the third panel of Figure 7: the direct fit is performed at just 108 points; far less than the 2048 data points, and most of the fitting points are in the interval $0 \leq x \leq 0.2$, where the smallest bandwidths are used and the locally adaptive procedure is relatively cheap.

Conclusions

This article has outlined the main ideas underlying Locfit, and presented examples showing some of the main capabilities. The web pages referred to in Section 1 contain a number of other applications, including several models with censored data, and details of many more options. Of course, the best way for readers to decide whether Locfit is useful is to download and try it!

Acknowldegments

I thank Mark Hansen for inviting this article, and for helpful comments on presentation.

References

Chambers, J. M. and Hastie, T. J., (1992), *Statistical Models in S*, Wadsworth & Brooks-Cole, Pacific Grove, California.

Cleveland, W. S. (1979), "Robust locally weighted regression and smoothing scatterplots," *Journal of the American Statistical Association*, 74, 829–836.

Cleveland, W. S. (1993), *Visualizing Data*, Hobart Press, Summit, New Jersey.

Cleveland, W. S. and Devlin, S. J. (1988), "Locally weighted regression: An approach to regression analysis by local fitting," *Journal of the American Statistical Association*, 83, 596–610.

Cleveland, W. S. and Grosse, E. H. (1991), "Computational methods for local regression," *Statistics and Computing*, 1, 47–62.

Craven, P. and Wahba, G. (1979), "Smoothing noisy data with spline functions," *Numerische Mathematik*, 31, 377–403.

Donoho, D. L. and Johnstone, I. M. (1994), "Ideal spatial adaptation by wavelet shrinkage," *Biometrika*, 81, 425–455.

Henderson, R. (1916), "Note on graduation by adjusted average," *Transactions of the Actuarial Society of America*, 17, 43–48.

Henderson, R. and Sheppard, H. N. (1919), *Graduation of Mortality and other Tables*, Actuarial Society of America, New York.

Loader, C. R. (1995), "Old Faithful erupts: Bandwidth selection reviewed," Technical Report, Bell Laboratories, Murray Hill, New Jersey.
`http://cm.bell-labs.com/stat/doc/95.9.ps`

Loader, C. R. (1996), "Local likelihood density estimation," *The Annals of Statistics*, 24, 1602–1618.

McCullagh, P. and Nelder, J. A. (1989), *Generalized Linear Models*, Chapman and Hall, London.

Ruppert, D., Sheather, S. J., and Wand, M. P. (1995), "An effective bandwidth selector for local least squares regression," *Journal of the American Statistical Association*, 90, 1257–1270.

Scott, D. W. (1992), *Multivariate Density Estimation: Theory, Practice and Visualization*, John Wiley & Sons, New York.

Tibshirani, R. J. (1984), *Local Likelihood Estimation*, Ph.D. Thesis, Department of Statistics, Stanford University.

Tibshirani, R. J. and Hastie, T. J. (1987), "Local likelihood estimation," *Journal of the American Statistical Association*, 82, 559–567.

Clive R. Loader
Lucent Technologies
clive@bell-labs.com

TOPICS IN MEDICAL IMAGING (Cont.)

CONTINUED FROM PAGE 1

rate to about 4 such images per second). The second image is a "t-map" in which those voxels in the functional map which have either increased (black) or decreased (white) by more than some fixed amount between two experimental conditions are highlighted. These are the "significant" voxels the psychologist is seeking. You have probably seen such maps in various scientific and popular publications. The main difference here is that I have not "cleaned it up" by removing the stray voxels which exceed the threshold by chance. A week went by and there he was again, even wilder looking and more animated. He said he had talked to the colleagues I had suggested and they said that I (Me; Bill Eddy) was the person who could help him. I laughed yet again and said "No; multiple comparisons is not my kind of problem. Go away." I gave him the names of more colleagues …on the other side of town.

Another week and there he was again. Persistent son-of-a-gun, I thought. Maybe I should listen more carefully. "Tell me more about this magnetic resonance machine." He did. I was pretty impressed with his explanation; it ranged over physics, electrical engineering, biophysics, neurobiology, neurology, anatomy, cognitive psychology, computer programming, and, of course, statistics.

Then he took me to see the MR machine over in the hospital. It was pretty impressive. A huge hollow cylinder in a quite large room. (I later found out the reason the machine was so large was it contained an internal magnetic shield made of steel about one foot thick. I also found out that the room was large so the uncontained field would be small outside the room.) Next door was the control room. Actually much less impressive, it only had a fancy touch screen console and a few computer monitors. And next door to that was the radio equipment that made it all work. (It was extremely cold in there. I later found out that was to reduce the noise from the electronics.) And then the whole setup was repeated. They had two of these multi-million dollar machines. Just for research! Into the functioning of the brain???

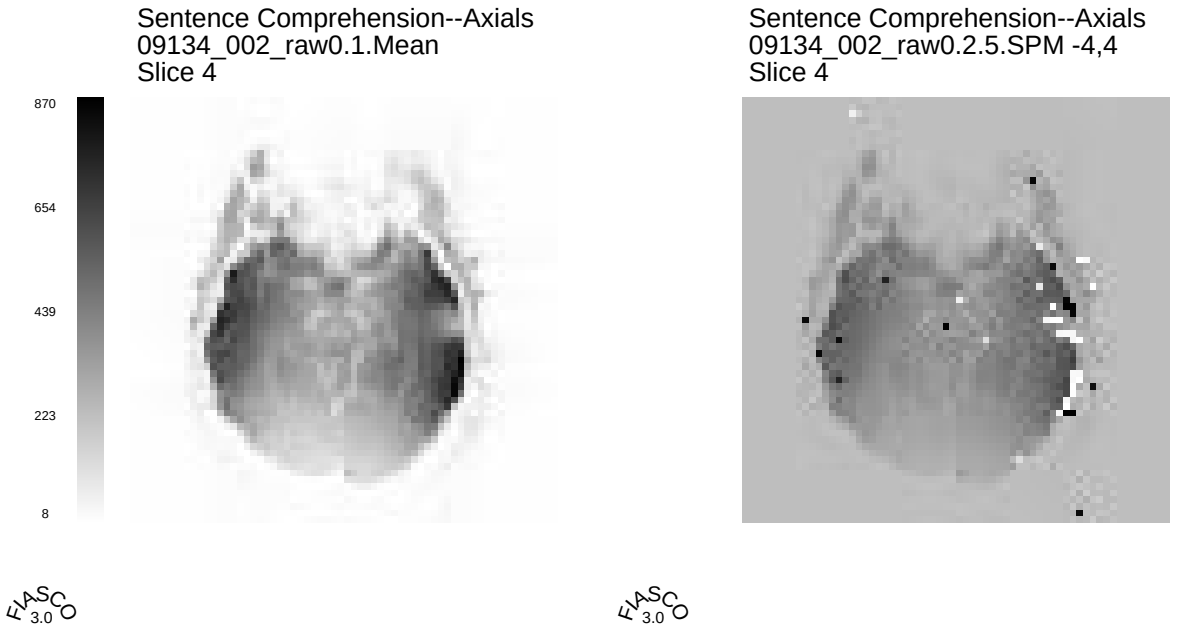


Figure 1. A functional image of a brain, and the associated t-map in which voxels in the functional map have either increased (black) or decreased (white) by more than some fixed amount between two experimental conditions.

I decided this deserved some more careful study. I got a couple of my students and we started reading about magnetic resonance, magnetic resonance imaging, and functional magnetic resonance imaging. The latter was pretty easy; at the time there were only four papers on the subject. I was very quickly hooked. I am, by nature, a sort of data junkie, the larger the data set the more interesting it is a priori. It turns out that a magnetic resonance scanner is a multi-million dollar machine which can crank out data at the rate of 500MBph (that's megabytes per hour, folks).

I guess I should say that the animated psychologist was Jon Cohen from CMU. He and I have now, together with others, had a couple of grants proposals funded and have written several papers. He is one of the very smart people I have met in this field, which makes the research a lot of fun.

My students and I spent the next year bothering every one in sight asking question after question about this mysterious process. (I want to especially thank Doug Noll of the Department of Radiology at the University of Pittsburgh for his clear and patient explanations; without his efforts we would be doing something else now.) My personal favorite question was to the Chief of MR Research (Keith Thulborn); these are his machines. One day (wondering if the noise was truly Gaussian, constant variance, and isotropic, as I had been told repeatedly) I asked him if we could get some images while there was nothing in the scanner. He laughed at me: "You don't even understand that it's resonance; there has to be something in the scanner to resonate or you'll get no image." The following week I asked him again with the same result. And again the following week …Then one week he said "The real reason I won't let you do that is that it would damage the electronics." Now I had him; he couldn't keep his story straight. I pressed and eventually he allowed me to take almost 15,000 images with the scanner empty; we did have to turn the gain down on the receiver. A few simple statistical summaries later and they called the repairman. No, we didn't damage the scanner; we found a minor flaw in the hardware. And forget Gaussian or constant variance or isotropic; none seem to be true.

So how does it all work? At the lowest level are the atomic nuclei. Each nucleus which is affected by magnetism (it has to have an odd number of protons or neutrons or both) spins like a gyroscope when in a magnetic field. Remember how the axis of rotation rotates itself (precession); same thing happens to the nuclei. The frequency of precession is proportional to the strength of the magnetic field, the constant of proportionality being determined by the atomic species. For hydrogen, the one that is currently the basis of fMRI, the constant is around 42MHz per Tesla. A Tesla is a measure of the strength of the magnetic field; one Tesla is 10,000 times the strength of the earth's magnetic field. A standard "high-field" MR scanner used for fMRI has a main field of 1.5 Tesla.

Alright, so the nuclei precess. So what? Well, the distribution of the directions the axes of all the nuclei are pointing is determined by the strength of the field. The stronger the field the more concentrated the distribution pointing in the direction of the field (or in the opposite direction to the field; there is a local minimum in the energy function in exactly the opposite direction). This, of course, depends on the energy state of the system; the more energy the nuclei have the less they have to pay attention to the magnetic field and the less concentrated is the distribution. And this is the secret to magnetic resonance; inject a little energy into the system. At the frequency of precession, it is absorbed by the atoms. (They resonate.) However, they prefer to return to their ground state, which they do by emitting the absorbed energy.

If you remember "the right hand rule" from elementary physics, you probably remember that associated with every electric field there is a magnetic field and vice versa. That means that by transmitting pulses of electric current through wires carefully wrapped around the hollow tube of the magnet one can cause spatially varying fluctuations in the strength of the magnetic field and hence spatially varying fluctuations in the frequency of precession. Thus when the atoms emit the absorbed energy the frequency of the emitted energy tells where they are and the strength of the emitted energy tells how many there are. Of course, it is a bit trickier than this description would indicate. The one important point is that because of the physics, the data that are acquired from the emitted signal are actually sampled at locations in k-space (Fourier space). Thus the images must be reconstructed by Fourier transform (the Faster the better) from the data.

That really is it, although a lot of important detail has been left out. So let's turn now to the "functional" part of fMRI. It had been known since the 1930's that "blue" blood (deoxygenated hemoglobin) was more paramagnetic than "red" blood. In the early 1980's it was shown that the MR signal from blue blood was stronger. In the late 1980's (using positron emission tomography) it was shown that the active brain required an increased blood supply but did not utilize the increase in available oxygen. This set the stage for the discovery in the early 1990's that the MR signal in the vicinity of active brain

tissue showed an apparent increase (because the blood leaving the region was more oxygenated and hence less paramagnetic and hence interfered less with the local magnetic field). Whew!

Consider an experiment, carefully designed by a cognitive psychologist to cause activation of a certain brain function in one condition and deactivation in another condition. Images gathered in the first condition will show a stronger signal at locations associated with the function that images gather in the second condition. Thus we are naturally led to compare the average of the images in the first condition with the average of the images in the second condition.

So where is the statistics in all this? It's simple really: the data are very, very noisy. There are many sources of variation we think we understand and, of course, many we don't understand. At the lowest level there is the "thermal" vibration of the atoms which can only be controlled by cooling the material toward absolute zero. There are two main sources of variation over which we have some control: the hardware and the subject. In the hardware one source of variation is the lack of uniformity of the main magnetic field, B_0; another source is lack of linearity in the gradient field, B_1. There is miscalibration of the analog-to-digital converter; there is drift in the receiver electronics; mistimings of resonant gradients cause "N/2 Nyquist ghosts;" the list goes on and on.

One of the main sources of variation cause by the subject is movement of the brain. Movement of the brain is itself the result of many sources. There is the rigid movement of the whole head; there is the almost periodic compression and vertical movement of the brain caused by the diastole of the cardiac cycle; and there is the complex distortion of the brain caused by respiration.

And what can we do about the noise? There are two general approaches: engineering (remove the noise at its source) and statistical (model the data and remove the variation it accounts for). We believe very strongly in both approaches. Keeping with statistical tradition, we'd really like to make a model which relates the data to the parameters of interest. That seems very difficult and the computations required by any plausible estimation technique are daunting. As a stop gap measure we have chosen to estimate (and remove from the data) each effect successively. So, in short summary, we currently correct for (see Eddy et al. 1996 for details): a) analog-to-digital converter miscalibration; b) gradient mis-timings; c) receiver drift; d) subject head motion; e) shot noise; and then we reconstruct the images (by a Fast Fourier Transform) and remove the (unexplained) voxel-wise trend over time. Finally we are at the point that statistics is traditionally invoked: it is time to decide on the effect of the experimental paradigm. A t-test is often powerful enough. As MR techniques and the cognitive questions evolve there is greater demand for subtler methods.

There is much work to do. Every question we ask leads to something new. The question that got me into this: "How can I retain some power after allowing for multiple comparisons?" led to Forman et al. (1995). It is not the "solution" but it is a step in the right direction. The question "What happens if you run the same experiment tomorrow?" led to Genovese et al. (1997) and some related papers. The question "Why not correct the raw data (instead of the reconstructed images) for head motion?" led to Eddy et al. (1996). And, of course, there is the "science." For example, "Does reading more complex sentences cause more brain activity?" led to Just et al. (1996).

If you are at all interested in fMRI, I suggest you walk over to your local MR center and begin to get to know the people and what they are doing. If you don't have the energy for that you might want to read Lange (1996); it is a good introduction to fMRI for statisticians although it is already a bit dated. Change in this field is extremely rapid; e.g., we are already using 8-fold higher spatial resolution than Lange describes.

Addendum

On rereading this I see that I have failed to capture the collaborative nature of fMRI work. I regularly interact with a large team of physicists, electrical engineers, psychologists, neurologists, computer scientists, and, of course, fellow statisticians. And, this doesn't count the workhorses: the MR technologists, computer programmers, registered nurses, etc. fMRI is truly a team sport.

Annotated References

Eddy, W.F., Fitzgerald, M., Genovese, C.R., Mockus, A., and Noll, D.C. (1996), "Functional Imaging Analysis Software - Computational Olio," *Proceedings in Computational Statistics* (A. Prat, Ed.), Physica-Verlag, Heidelberg, 39-49. *(A slightly more technical description of the data processing and the software we have developed.)*

Eddy, W.F., Fitzgerald, M., and Noll, D.C. (1996), "Improved Image Registration By Using Fourier Interpolation," *Magnetic Resonance in Medicine*, 36, 6, 923-931. *(Our method for two dimensional motion correction.)*

Forman, S.D., Cohen, J.D., Fitzgerald, M., Eddy, W.F., Mintun, M.A., and Noll, D.C. (1995), "Improved Assessment of Significant Change in Functional Magnetic Resonance Imaging (fMRI): Use of a Cluster Size Threshold," *Magnetic Resonance in Medicine*, 33, 636-647. *(Our first attempt to increase the power of the t-testing procedures.)*

Genovese, C.R., Noll, D.C., and Eddy, W.F. (1997), "Estimating Test-Retest Reliability in Functional MR Imaging I: Statistical Methodology," *Magnetic Resonance in Medicine* (in press). *(What happens when you repeat an fMRI experiment?)*

Just, M.A., Carpenter, P.A., Keller, T.A., Eddy, W.F., and Thulborn, K.R. (1996), "Brain Activation Modulated by Sentence Comprehension," *Science*, 274, October 4, 1996, 114-116. *(An example application of fMRI.)*

Lange, N. (1996), "Statistical Approaches to Human Brain Mapping by Functional Magnetic Resonance Imaging," *Statistics in Medicine*, 15, 389-324. *(A very long introduction to fMRI.)*

Willian F. Eddy
Carnegie Mellon University
bill@cmu.edu

TOPICS IN INFORMATION VISUALIZATION

Templates for Looking at Gene Expression Clustering

By Daniel B. Carr, Roland Somogyi and George Michaels

1. Introduction

In this paper, we describe the design of graphical displays for investigating clustering evident in gene expression data. The displays include stereo plots, parallel coordinate (time series) plots and conditioned parallel coordinate plots. These basic templates are subject to numerous variations and are potentially useful in many other cluster analysis settings.

As an application of our approach, we will consider gene expression mapping of the developing spinal cord in rats, focusing on only 112 genes. Because tens of thousands of interacting genes control spinal cord development, this study is really addressing just the tip of the iceberg. Some background information is appropriate to place this study in a larger context.

We are in a new era in which biologists can selectively disrupt genes and design genes to perform specific tasks. However, genes do not function in isolation. In gene knock-out experiments, the deletion of a gene can have a disastrous effect in some cases, while in others the working constellation of genes compensates quite well for the gene's absence (see Galli-Taliadoros et al. 1995). This suggests redundancy of gene function and combinatorial regulation of genes: a complex *genetic network*.

The study of genetic networks is one of the topics in recent books and journals addressing complexity (see Kauffman 1993, Somogyi and Sniegoski 1996, and `http://rsb.info.nih.gov/mol-physiol/homepage.html`). The introduction of networks into this area has effectively split the ranks of biologists. The old guard continues to focus on the study of individual genes using an evolving but relatively mature methodology. Those taking on the challenge of genetic network studies are pioneers who must modify and create conceptual models and methodologies to view this new landscape of molecular interactions. This work is much different than studying electronic communications networks where one can at least consult with the engineers who designed the network!

Initial work in studying genetic networks conceptualizes two types of communication paths (Somogyi and Sniegoski 1996). The first consists of proximal paths operating through "cis regions" and "trans elements." Cis are control regions of DNA proximal to gene coding sequences and trans elements are gene products that regulate by interacting with cis regions. The second type of communication pathway is composed of extended paths involving protein-protein and protein-signaling factor interactions governing intra- and extra-cellular communication. Genes encoding the participating proteins control this communication.

Our starting point is very simple; we will look for clusters in gene expression time series data. When the output patterns of different genes are very similar, there is hope that they are a part of a constellation of communicating genes, receiving similar control signals. Again, this is only a starting point. Since some genes turn other genes off, things get complicated quickly. Cluster measures such as mutual information provide clues about additional members of the constellation. (Mutual information is also referred to as the rate of transmission, and is related to conditional entropy.) As this is work in process, we welcome insights into how to proceed in the face of our limited understanding and the apparent com-

plexity of genetic networks. In the next section we provide more details about the nature of our gene expression data.

2. The Observations and Clustering Methods

As mentioned above, we identified 112 genes involved in the development of a rat's spinal cord. This collection includes genes deemed important for the development of the central nervous system: neurotransmitter receptors and metabolizing enzymes, intracellular signaling proteins, peptide factors and their receptors and growth factors. Genes for marker proteins were also selected so that we can associate gene patterns with cell differentiation.

The observations on each gene constitute a time series of length nine. The nine developmental times studied were gestation days 11, 13, 15, 18, 21; and after birth days 0, 7, 14, and 90 (adult). Each point in the series is a value between zero and one. When a gene is not functioning the value is zero, and when it is maximally functioning the value is one. The observations themselves are obtained through a process known as RT-PCR, reverse transcription-polymerase chain reaction (see Somogyi et al1995). The values are determined from digital records of gel images and are actually means of triplicate observations. The variability of the triplicates is very small and not shown in the graphs below. Scaling forces the means to range from zero to one.

Figure 1 (page 27) shows the gene expression patterns by functional groups and gene sequence families. The group names appear to the left of the clusters of panels. The groups are members of four general functional categories described later in Figure 4. Since our focus here is on graphics, we will not describe the gene families.

The design of Figure 1 makes heavy use of perceptual grouping and warrants some comment. The scale for the axes appears in the top left panel. Genes in the same functional group appear in a consecutive grouping of panels. Each panel within a gene function group shows the times series for four or fewer genes. The representation is a parallel coordinates plot (see Inselberg 1985 and Wegman 1990) with omitted axes.

Each times series in a panel has its own color. While there is overplotting, the reader can quickly infer values for overplotted points. The color key and corresponding gene label appears at the right of the panel. The color has no meaning other than to serve as a link (see also Carr and Pierson 1996). One can additionally sort the rows of the key by values for the last time period. This positional linking makes more lines run into the rectangles of their own color and linking becomes trivial. However, the current example emphasizes reading labels in order and sorting would scramble the order.

In Figure 1 the plotting order for color is consistent in all panels: cyan, green, orange, and red. We use graphical ordering and plot from the bottom up within each functional group. This might be argued since there is a clash of conventions. Graph reading is bottom up while table reading is top down. Figure 1 is much like a table so there is ambiguity about which convention to apply. Note that with four genes in a panel, the red line, which should appear closest based on wavelength considerations, plots on top. The color selection also makes red the darkest color on a lightness scale and hence it contrasts the most against the light background. The graph convention is slightly advantageous because red appears on top in the panel and at the top of the color key.

The left to right sequence of gene functional group name, panels, and then gene names can also be argued. The task can motivate a different order. If communicating membership in functional groups were much more important than looking at the time series within function groups, then putting the functional group names and gene names together would be the logical design. Putting the text in one place has merit in its own right. However, putting the gene names on the right panels allows the names to be left aligned and to be uniformly close to the key and panel. Since finding the names within the function groups is still easy, we show this variation.

The selection of four genes per panel follows Kosslyn's (1994) advice for creating small perceptual groups. The apparent simplicity of the panels deteriorates quickly as number of time series in each panel increases. Figure 1 appears simple while showing the thousand means in this data set. The four time series per panel design has many applications. In landscape orientation and without the two column format, the design readily accommodates the much longer time series that occur in manufacturing and other applications.

Sorting the time series can make the plot appear simpler (see Carr and Olsen 1996). However, data ordering can serve other purposes. In Figure 1 we ordered the functional groups based on page layout considerations. We kept the provided label ordering within functional groups, because that simplified finding a specific gene within a function group. A time-series sorted version would be interesting. In an interactive setting (see Carr, Valliant and Rope 1996) one might try visual clustering by dragging and dropping time series into different panels. An automated approach can use clustering as a nominal basis for sorting.

In this case my co-authors came to me (Dan) with FITCH (Felsenstein 1993) clustering results. FITCH is an n4 algorithm and produces graphics like that in Figure 2. They had decided that their Euclidean distance clustering was better if they included the differences between consecutive observations in the time series. In other words the input vectors were of length 17, 9 time series values plus 8 differences. This is equivalent to using a weighted distance that emphasizes the seven internal points of the time series. The co- authors note that another option is to add slopes based on the actual spacing in days. They also used mutual information clustering.

Like many statisticians, I am aware of hierarchical clustering algorithms, maximum likelihood clustering, and refinement of clusters using the K-means algorithm. However, I am far from being an expert. I had never heard of FITCH nor mutual information clustering. My co-authors gave me every opportunity to recommend a clustering algorithm that would provide the truth, or one that all scientists would recognize the best of the available choices, but I declined. My early participation was simply to help them look at the data and the results of clustering.

3. Cluster Plots and Stereo Plot Construction

Figure 2 shows a cluster tree produced by FITCH. The follow-the-line distance between points approximates the multivariate interpoint distance. FITCH minimizes a measure of stress that differs somewhat from the measure minimized in traditional 2-D nonmetric multidimensional scaling (MDS). Figure 2 shows the average time series profile for each of the six resulting clusters. The labels for the clusters derive from the profiles and do not necessarily have any deep meaning. Figure 2 is good in that it gets all the labels into the plot and provides a feel for clusters and subclusters. The extra freedom provided by using connecting-line length rather than direct interpoint distance should allow significant reduction in any measure of stress. In this sense Fitch cluster tree views should be better than MDS or first-two principal component views. However, tracking lines for each pair in the cluster to assess interpoint distance is a complicated visual operation. How can we judge the clustering if we can not easily judge interpoint distances? A first reaction is to stick with views that represent interpoint point distances directly even though the distance approximations are not as good.

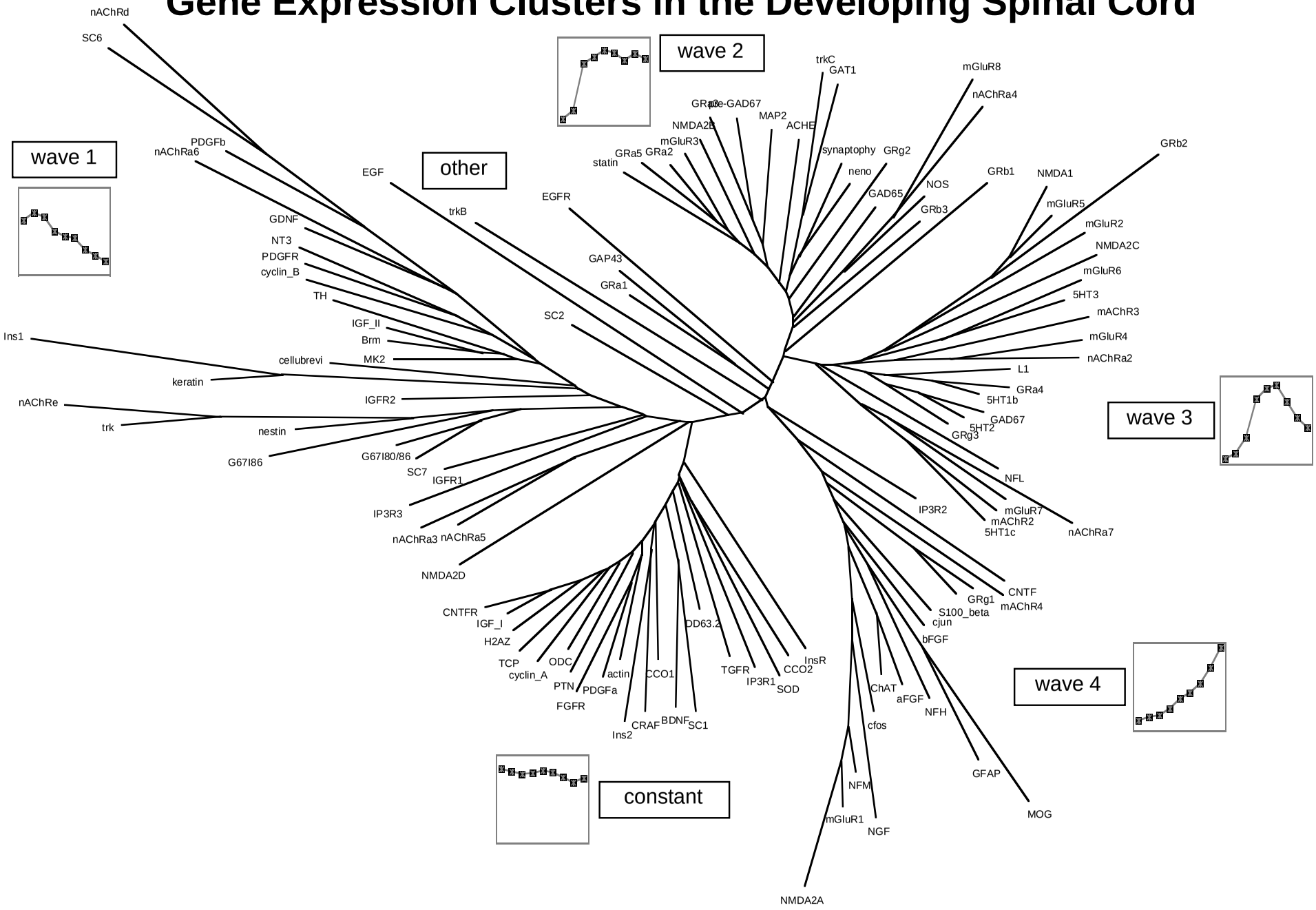
Conceptually, higher-dimensional plots reduce the measure of stress and hence provide a better representation of interpoint distances than low dimensional plots. In practice the analyst must translate the differences between encoded multivariate points into interpoint distances. The merits of a higher-dimensional representations can be more than counterbalanced by the difficulty and inaccuracy of the decoding process. In fact a definitive test for multivariate representations should be how well the user can assess the distance between two points and the ratio of two such distances. The position of Carr et al 1986 is that 3-D stereo plots (and possibly 4-D stereo ray glyph plots) allow quick distance judgements that are good enough to be worthwhile. In the principal components context, if a third or fourth component adds little to the percent of variability explained, then one might get by focusing on a 2-D plot. However, a 3-D or 4-D plot is often a better starting point. Those busy interpreting a 2-D plot can seem naive when important structure is obvious in a 3-D view. Of course naiveté is relative. Those that can incorporate and understand even more information in the graphics have an advantage.

Figure 3 (page 28) is a side-by side stereo plot that distinguishes the six groups using color and symbol. Carr (1990) discusses stereo projections. The slightly rotated view in Figure 3 seems to help image fusion over a directly facing view. Many people can learn to fuse side-by-side images without the aid of a view device. This learned skill involves the decoupling of eye-convergence and lens focusing that normally work together in a process called accommodation. Proper fusion results in the square dot appearing in the back left corner of the plot frame.

The plot axes in Figure 3 are the first three principal components of the 17 variables. The three coordinates capture 65 percent of the variability. The figure uses global scaling for the three coordinates and the plot frame reflects the range of the principal components. That is, the x-axis represents the first principal component and the frame is largest in the x direction. The y-axis represents the second principal component. Representing the third principal component with stereo depth reduces overplotting and the complications of looking through many layers of data. The analyst should view the stereo plot from the correct distance to perceive interpoint distances properly. The assignment of variables to the axes gives an important clue. If the frame appears deeper than the frame is tall, then the analyst is too far away.

The cluster and color pairing are: Constant = red, Wave–1 = orange, Wave–2 = green, Wave–3 = cyan, Wave–4 = magenta, and Other = black. A minimal spacing tree based on the three axes connects the points in each cluster. This helps to constrain visual traversal paths in

Gene Expression Clusters in the Developing Spinal Cord



repeated viewing, and the perceptual grouping makes the plot look simpler (Carr et al 1986). Given fusion, we see plausible clusters (red triangles and green octagons) at a glance. The scale for the red and green clusters raises serious doubts about some of the other clusters such as the orange squares and magenta x's.

Another way to look at cluster results is to use local averages of the times series rather than principal components. The averaging of adjacent times series values is a standard dimension reduction technique. The advantage is that when patterns appear, interpretation often less complicated than for a principal components view. (A disadvantage is that researchers don't like loosing temporal resolution especially when the experiments were laborious.) After grouping the nine values into sets of three and averaging, we used the shuttering glass stereo in ExplorN (Carr, Wegman, and Luo 1997) to look at the resulting three coordinates. (ExplorN also supports touring in parallel coordinate and scatterplot matrix views of up to 20 dimensions.) Color and ray angle represented the cluster membership. The clusters were plausibly coherent just as they are in Figure 3. However there was one bad exception in our initial look at the clustering. The data for that gene had a transcription error. The principal component view in Figure 3 shows the corrected data. Before assessing clusters further, we pause for more comments on stereo views and plot construction.

4. Stereo Viewing and Plot Production

Small side-by-side stereo plots are less than optimal. A good stereo viewer with appropriate mirrors and lenses allows use of much larger left-eye and right-eye plots. Stereo workstations using shuttering eye-glasses work quite well, although they sacrifice a bit in terms of spatial and brightness resolution. Rotation of points provides a good depth cue, motion parallax. However, rapidly rotating plots are hard to study. In our use of ExplorN we found very slowly rotating stereo views to be a desirable compromise.

The move from the workstation stereo graphics to printed side-by-side views raises the issue of color overplotting inconsistencies. In non-translucent stereo mode, our SGI graphics workstation uses a z-buffer methods to make sure that whatever is closest to the viewer plots on top. One could utilize the workstation graphics by accessing the separate eye views and copying the low resolution bit maps to a high resolution printer. For small side-by-side views the size reduction ameliorates the problem of limited resolution in workstation views. For the graphics here we sought to use more conventional software. Production of Figure 3 is straightforward using high resolution black lines. However, color inconsistencies arise using conventional vector graphics. The wrong color overplots when drawing a distant line of one color after drawing a close line of a different color.

The partial solution used to construct Figure 3, broke the line segments into a sequence of short line segments based on a large number of depth planes. The algorithm used the closest of the short segment endpoints as the measure of the segment's depth. The algorithm then sorted both points and lines back to front before plotting. This procedure, while computationally tedious, corrects the problem except for overplotting of segments and points at almost identical depth.

5. Cluster Interpretation and Assessment

In an unsupervised clustering problem, one hopes to use corroborating scientific information as well as cluster tightness to assess the clustering. Here the genes tyrosine hydroxylase (Th), insulin 1 (Ins1) and insulin-like growth factor II (IFGII), appear as a tight subgroup in the Wave–1 cluster. They turn out to be located on the same human cytogenetic band (11p15.5) and are close together on mouse chromosome 7 (see Mouse Genome Database). This suggests the genes are regulated in parallel due to their close proximity on the chromosomes.

Figure 4 shows the residuals from the cluster means by gene functional group and cluster. (Using a different scale, one can also show the panel means in row and column margins.) Both Waves–2 and 3 are notably confined to neurotransmitter signaling. Wave–4 cluster genes primarily belong to several functional families. Genes showing largely constant expression (the Constant group) originate from diverse families but strictly exclude the neurotransmitter signaling and neuroglial markers. The variability in Figure 4 raises concern about the adequacy of the clustering and the stability of the clusters variations when using more data or other algorithms. It seems doubtful that all of Wave–1 is just one constellation of genes. With more data, Wave–1 may break into several defensible clusters. Currently the subcluster indicated above could begin to define a constellation. The co-location of genes on a chromosome is reasonable confirmation.

6. Cluster Comparison

As indicated above my co-authors also brought results from a mutual information clustering algorithm. Again they selected six classes. A natural step is to compare the clustering.

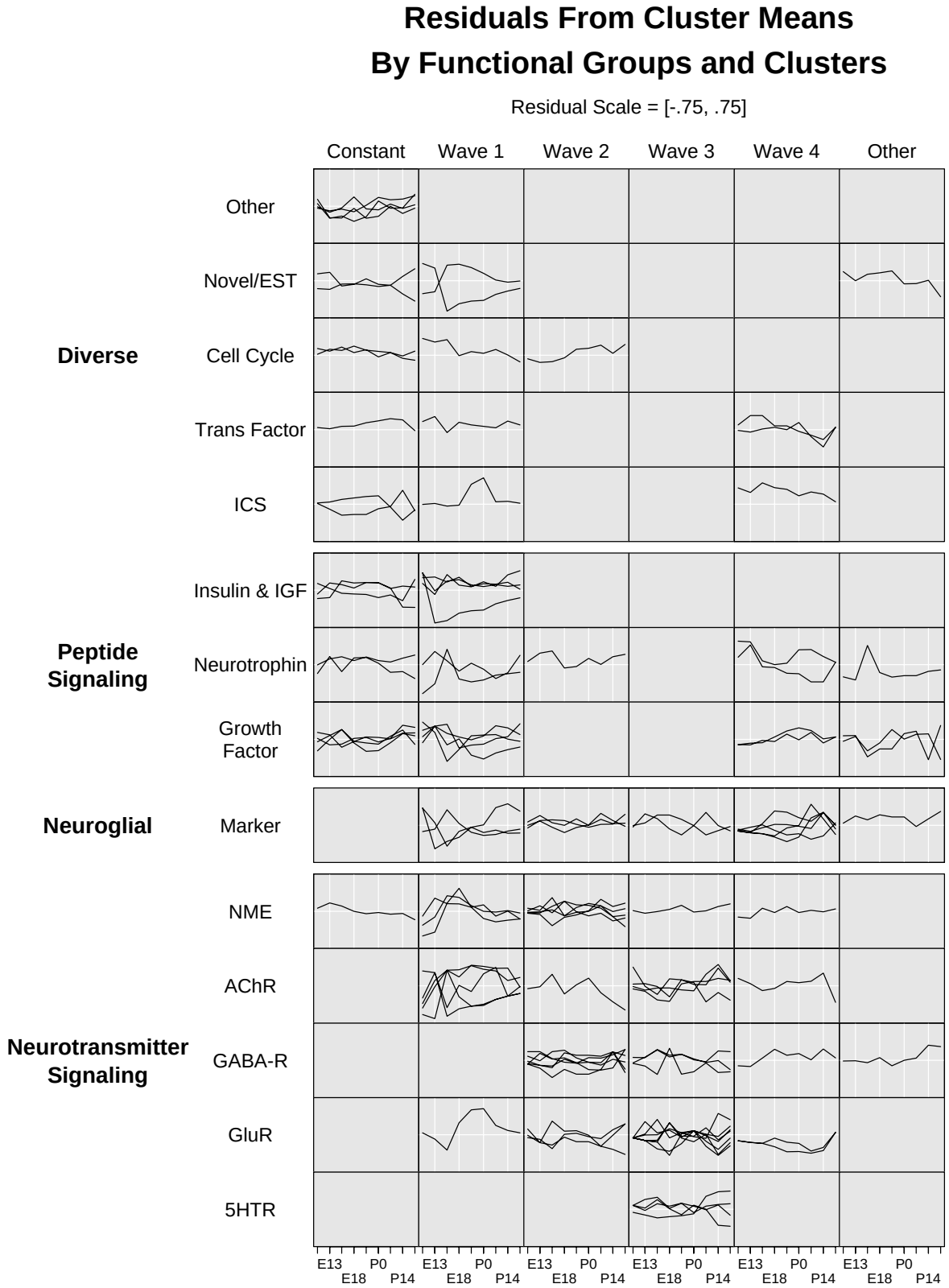


Figure 4. Multivariate residuals in a two-way layout.

Comparison graphics can take many forms. One template is a two-way panel layout. The rows are class membership for one algorithm and the columns are the class membership for other. Each panel contains the corresponding times series (if any). After suitable rearrangement of rows and columns, a strong diagonal pattern would indicate effective similarity of the clustering algorithms. The shape of the series in off diagonal panels provides insight into algorithm differences. One can augment such graphics. In the current cases, there would be a 6 x 6 panel layout and color could indicate membership in one of the four major functional groups diverse, peptide signaling, neuroglial and neurotransmitter signaling. Another variation incorporates the information using a (4 x 6) x 6 layout. The 24 rows of panels result from crossing the major functional groups with one of the classifications. The potential variations using conditioning and color to incorporate additional information are numerous.

Figure 5 (page 28) shows a cluster comparison approach based on parallel coordinates. Here the time series are omitted and the plot emphasizes three classifications: gene function, Euclidean distance clusters and mutual information clusters. The gene function axis appears at both the top and bottom of the plot. The regular spacing between a small number of crossing points distinguishes the classification axes.

The Figure 5 design introduces two unique-case axes between each classification axis. Every case (gene) has its own unique plotting position on these axes. With only a hundred or so lines, all lines are visibly distinct. George's idea behind using two such axes was to confine messy line crossing to the region between the unique-id axes. This creates regular patterns of lines reaching the classification axes.

The choice of color in the figure serves two purposes. First it collapses the 13 gene function groups in the four functional families. Second the color selection purposely calls attention to the peptide signaling (high contrast yellow) and down plays the distinction between neuroglial and neurotransmitter signaling.

The precursors to Figure 5 raised an interesting sorting issue. The crossing lines made the plots look complicated. The challenge then is to order classifications, order classes within each classification, subclasses within nested classifications, and genes within (sub)classes to minimize line crossings. An all permutations approach works for small problems. Unfortunately the combinatorics become overwhelming in a general table setting.

For two classifications there is a convenient approximate approach. Wegman (1990) observes that few crossings correspond to high correlations. Kendall and Stuart (1979) describe an eigenvector scoring approach for categorical data that will maximize the correlation for two classification. This provides a basis for ordering the classes within each classification. Unfortunately, we have not been able to generalize this approach to three variables and fear that the general case may be np-complete.

Some find the string art in Figure 5 appealing but for others the plot is still too complicated. The advantage of the small panel approach described previous is that it shows the times series in addition to the classification. We present Figure 5 because it illustrates one way of converting classification tables into graphs and because it raises a sorting challenge.

7. Closing Remarks

Alternatives and extensions to the above templates for viewing clustering results are numerous. Hierarchical cluster tree views are common. Visually connecting the cluster trees to the multivariate data can help provide insights about the clustering. Buja, Cook, and Swayne (1996) used color brushing in XGobi to link cluster tree branches to other scatterplot views of data. Such interactivity facilitates following the visual clues provided by graphics. The idea of joining multiple window brushing capability with adaptable thoughtfully-designed multiple panel plots has occurred to many but implementations have been slow to appear.

Many multiple panel-designs scale to much large sample sizes. Density methods apply to parallel coordinate plots (Wegman and Luo 1997). The data need not necessarily be time series. Plots like Figure 4 have many extensions.

As usual, S-PlusT functions and scripts for producing the graphics are available via anonymous ftp to `galaxy.gmu.edu`. Change directory to `pub/dcarr/newsletter/gene`. In contrast to the past, the data provided is artificial. The real data will be substituted when my co-authors have published in a refereed journal.

Splus users may find the matrix layout functions of particular interest. Currently there is a Bureau of Labor Statistics technical report describing the functions and selected connections to TrellisT graphics. Contact Dan for a copy.

Those with Silicon Graphics workstations may be interested in ExplorN. A tar file containing an executable and sample data sets is available. Conversion to OpenGL and available on other OpenGl compatible platforms may happen later in the year.

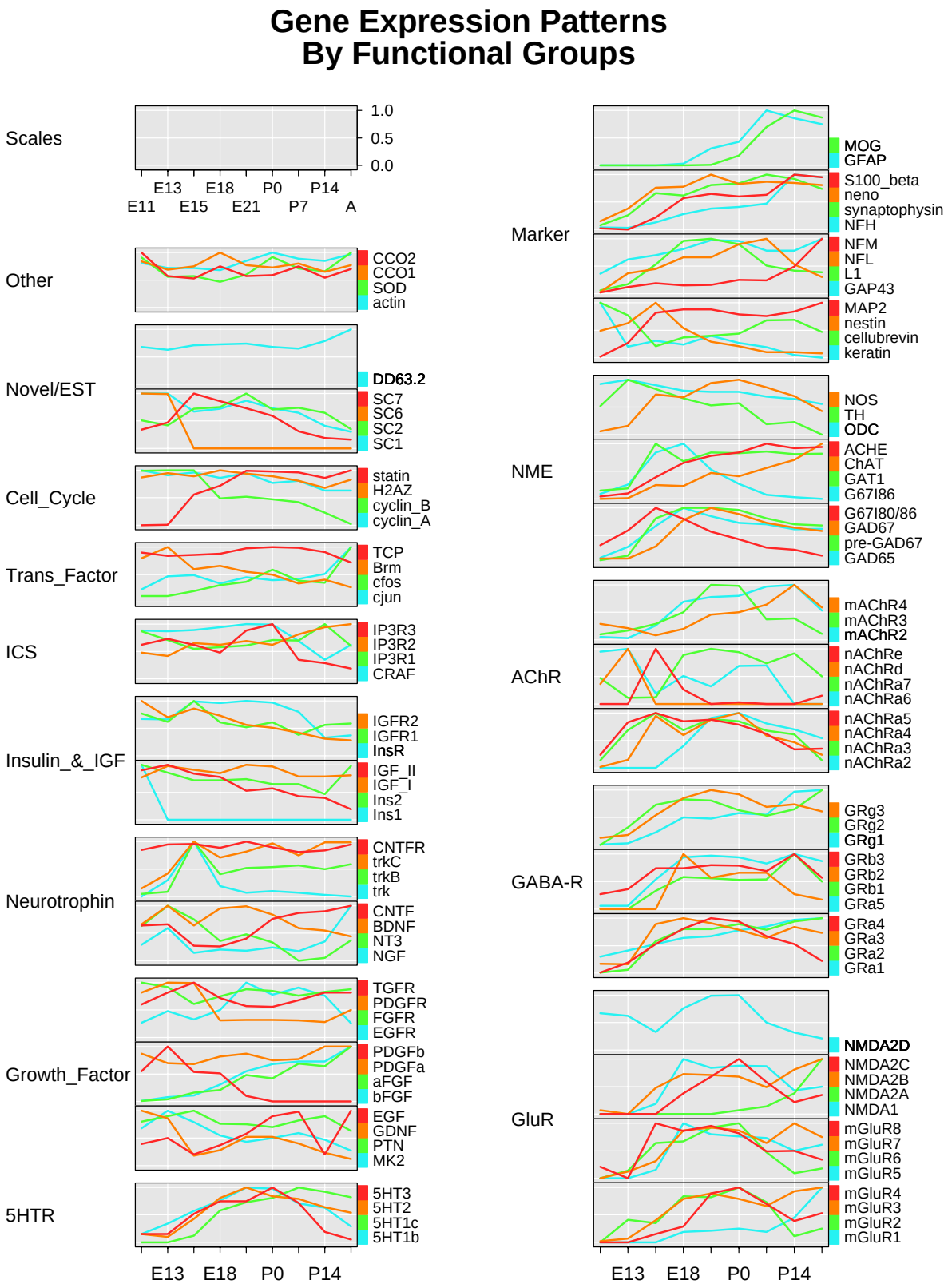


Figure 1. Labeled time series with controlled overplotting. Here, the colors have no meaning, but serve only as a linking device within each panel.

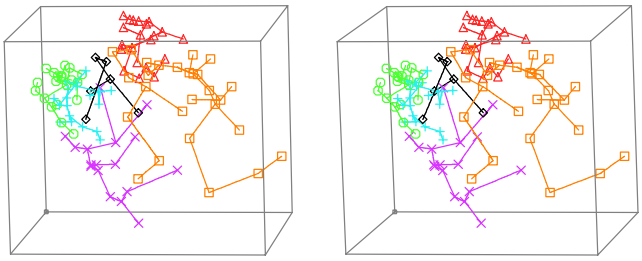


Figure 3. Stereo pairs with careful color overplotting. In this case, colors correspond to clusters (see the text for a detailed description).

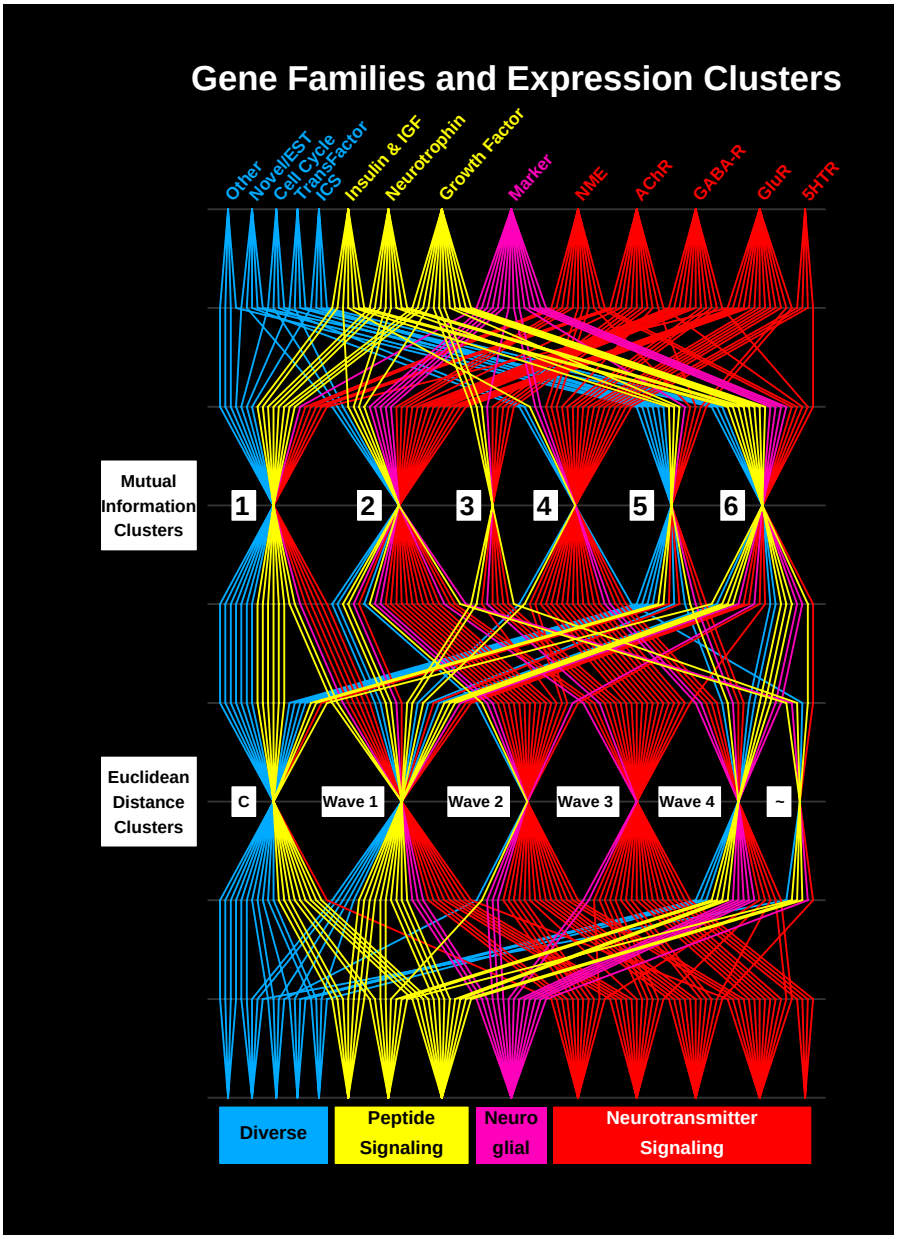


Figure 5. A sorted parallel coordinates view of multiway classified cases.

As always, the authors are open to gentle constructive suggestions. Comments on graphics are best addressed to Dan and comments on the study or on genetic networks are best addressed to Roland and George.

Acknowledgements

The authors thank Drs. Xiling Wen and Stefanie Fuhrman, Laboratory of Neurophysiology, NINDS, for their outstanding work in the experimental data acquisition and analysis. Thanks also go to Andreas Buja for a discussion on ordering classes.

References

Buja, A., Cook, D. and Swayne, D. F. (1996), "Interactive High-Dimensional Data Visualization," *Journal of Computational and Graphical Statistics*, 5(1), 78–99.

Carr, D. B. (1993), "Production of Stereoscopic Displays for Data Analysis," *Statistical Computing & Graphics Newsletter*, 4(1), 2–7.

Carr, D. B. and Olsen, A. R. (1996), "Simplifying Visual Appearance By Sorting: An Example Using 159 AVHRR Classes," *Statistical Computing & Graphics Newsletter*, 7(1), 10–16.

Carr, D. B., Nicholson, W. L., Littlefield, R. J. and D. L. Hall (1986), "Interactive Color Display Methods for Multivariate Data," *Statistical Image Processing and Graphics*, eds. E. J. Wegman and D. J. DePriest, Marcel Dekker, New York, pp. 215–250.

Carr, D. B. and Pierson, S. M. (1996), "Emphasizing Statistical Summaries and Showing Spatial Context with Micromaps," *Statistical Computing & Graphics Newsletter*, 7(3), 16–23.

Carr, D. B., Valliant, R. and Rope, D. (1996), "Plot Interpretation and Information Webs: A Time-Series Example From the Bureau of Labor Statistics," *Statistical Computing & Graphics Newsletter*, 7(2), 19–26.

Carr, D. B., Wegman, E. J., and Luo, Q. (1997), "ExplorN: Design Considerations Past and Present," Center for Computation Statistics Technical Report No. 137, George Mason University, Fairfax, Va. 22030.

Felsenstein, J. (1993), *PHYLIP (Phylogeny Inference Package), version 3.5c*, distributed by the author, Department of Genetics, University of Washington, Seattle.

Galli-Taliadoros, L.A., Sedgwick, J. D., Wood, S. A., and Korner, H. (1995), *J. Immunol. Methods*, 181, 1–15.

Inselberg, A. (1985), "The Plane with Parallel Coordinates," *The Visual Computer*, 1, 69–96.

Kauffman S. A. (1993), *The Origins of Order, Self-Organization and Selection in Evolution*, Oxford University Press, London.

Kendall,M. and Stuart, A. (1979), *The Advanced Theory of Statistics, Vol. 2, Inference and Relationship*, Charles Griffin & Company, London.

Kosslyn, S. M. (1994), *Elements of Graph Design*, W. H. Freeman and Company, New York.

Mouse Genome Database, Mouse Genome Informatics, The Jackson Laboratory, Bar Harbor, Maine.
`http://www.informatics.jax.org/`

Somogyi R. and C. A. Sniegoski (1996), "Modeling the complexity of genetic networks: understanding multigenic and pleiotropic regulation," *Complexity* 1(6), 45–63.

Somogyi R, Wen, X., Ma, W., and Barker, J. L. (1995), "Developmental kinetics of GAD family mRNAs parallel neurogenesis in the rat spinal cord," *J. Neurosci*, 15, 2575–2591.

Wegman, E. J. (1990), "Hyperdimensional Analysis Using Parallel Coordinates," *Journal of the American Statistical Association*, 85, 664–675.

Wegman, E. J. and Luo, Q. (1997), "High-Dimensional Clustering Using Parallel Coordinates and the Grand Tour," *Computing Science and Statistics*, 28, 361–368.

Dan Carr
Institute for Computational Sciences and Informatics
George Mason University
`dcarr@voxel.galaxy.gmu.edu`

Roland Somogyi
National Institute of Neurological Disorders and Stroke
National Institutes of Health
`rolands@helix.nih.gov`

George Michaels
Institute for Computational Sciences and Informatics
George Mason University
`gmichael@gmu.edu`

The Joint Statistical Meetings, Anaheim 1997

By Dianne Cook and James L. Rosenberger

Statistical Computing Activities at the JSM

The Statistical Computing Section has organized a busy schedule of invited and contributed sessions, posters and tutorials, for the members attending the Joint Statistical Meetings, August 10-14, 1997, in Anaheim, California.

The theme of "data mining and massive data sets" provides the focus for a number of sessions, and fits nicely with the overall conference theme of **Shaping Statistics for Success in the 21st Century**. The invited program begins with a "Data Mining Tutorial" on Sunday at 2:00PM by Usama Fayyad of Microsoft Research. Monday at 2:00PM a co-sponsored session with Statistics in Sports discusses Data Mining and its Application to the NBA, with Daryl Pregibon of AT&T Labs giving the statisticians perspective and Ed Colet and I.S. Bhandari of IBM Research Center demonstrating the advanced scout software. On Tuesday at 10:30AM Peter J. Huber presents a Program Chair's special lecture on **Statistics and Massive Data Sets** with discussion by Ed Wegman and others.

Other futuristic looking invited sessions include: "Computational Science and the Internet: A Seamless Web" on Monday at 10:30AM organized by Mark Hansen and showcasing S and JAVA for data analysis, use of the graphics production library, and dynamic electronic research publications; and invited sessions on Tuesday at 8:30AM on "Quantile Regression", on Wednesday at 2:00PM on "The Bootstrap and Empirical Likelihood", and on Thursday at 8:30AM on "Statistical Software Engineering", which presents a statistical approach to software development and maintenance. A special contributed session of interest to program developers is titled: "Statistical Reference Datasets for Assessing the Numerical Accuracy of Statistical Software."

The winners of our section's student paper competiton present their talks on Tuesday at 2:00PM. Please see Wenjiang Fu, Toronto; Ramani Pilla, Penn State; Alan Gous, Stanford; and Gareth James, Stanford, in this special contributed paper session.

In addition to the seven invited sessions and special sessions mentioned above, the section has organized nine contributed paper sessions on topics including Modeling and ANOVA, Robustness and Nonparametrics, Categorical and Latent Variable Methods, Density Estimation, Markov chain Monte Carlo, Algorithms, Computational Innovations, and Outlier Detection.

Everyone should browse the poster presentations on Tuesday at noon. For those of you who would like to interact with others in greater depth, we will also sponsor four roundtable lunches on Tuesday on the topics of "Multiple Comparisons", with Peter Westfall, "Data Mining" with Daryl Pregibon, "Computational Science and the Internet" with Mark Hansen, and "Markov chain Monti Carlo' with Peter Mueller. Sign up and meet other section members with common interests.

We're hoping to see a good turnout in Anaheim. Be sure to attend the mixer on Monday evening following a short business meeting at 7:30PM. Good food, door prizes, and a chance to meet newcomers and oldtimers.

Statistical Graphics at the JSM

The Statistical Graphics program at this year's Joint Statistical Meetings has a wide array of talks and posters. Late on Sunday afternoon there are several invited posters with graphical content: immersive virtual reality environments, the production graphics library, and spatial/environmental visualization. On Tuesday early afternoon the contributed poster session has many posters with graphical content, and in particular several on the topic of graphics for biometrics. This is a year that we also have a statistical graphics data exposition: there are over 15 posters that will display graphical methods applied to examining costs and outcomes of hospital visits (Wednesday 3:00-5:00pm).

There are 4 invited, 2 special contributed, and 2 contributed paper sessions. Bright and early at 8:30am Monday is a panel discussion on publishing in the electronic age, and then there are 2 sessions on spatial data analysis tools at 10:30am and 2:00pm. Tuesday at 8:30am there is a session on multivariate data and at 10:30am there is a session, with a discussant, on interactive graphics. In the early afternoon at 2:00pm there is the invited session "Natural Selection: Advances in Brushing". Wednesday morning features two invited sessions: "Recent Developments in Projection Pursuit" at 8:30am and "Lost in Space: Assessing Multivariate Missing Data" at 10:30am.

There are also many sessions outside of the main sponsorship of the Stat Graphics section that have graphics content. For example, on Monday at 10:30am there is a session on graphics and publishing interactive material: "Computational Science and the Internet". There is a session on graphics and teaching at 10:30am Tuesday.

On Wednesday there is a session with several papers on graphics for outlier detection at 10:30am and another at the same time on graphics in teaching. On Thursday at 10:30am look for the session on using the WWW for classroom instruction; it has a decidedly graphical flavor.

Finally don't forget the roundtable luncheons tentatively scheduled for Wednesday. This year there are discussions on visualizing spatial data, graphics with JAVA, virtual reality, projection pursuit/neural networks and visualizing massive data. Oh, and one last major event is that the popular Stat Computing and Graphics mixer is planned for Monday night.

For more details point your web browser to: `http://www.public.iastate.edu/~dicook/graphicsprog97.html`

Dianne Cook
Statistical Graphics 1997 Program Chair
Iowa State University
dicook@iastate.edu

James L. Rosenberger
Statistical Computing 1997 Program Chair
The Pennsylvania State University
jlr@stat.psu.edu

CONFERENCE ANNOUNCEMENTS

Case Studies in Bayesian Statistics Workshop 4 – 1997

September 26-27, 1997
Carnegie Mellon University

The Fourth Workshop on Case Studies in Bayesian Statistics will take place September 26-27, 1997, on the campus of Carnegie Mellon University in Pittsburgh, Pennsylvania.

Contributed paper abstracts will be accepted through August 1, 1997 for presentation in a poster session on the first evening of the Workshop. The Morris H. DeGroot Lecture and Reception will precede the poster session and a sit down dinner will follow. The organizers plan to accept 25 to 30 papers. The posters will be displayed and discussed with the individual authors in an informal presentation.

Limited travel support is available to workshop participants and the organizers especially want to encourage young researchers, women, underrepresented minorities, and the handicapped to attend and apply for support.

The Fourth Morris H. DeGroot Memorial Lecture

The Fourth Morris H. DeGroot Memorial Lecture and Reception will be held on the first day of the Workshop, Friday, September 26. The guest lecturer is Brad Efron, Professor of Statistics at Stanford University. Past lecturers have been Adrian F.M. Smith, A. Philip Dawid, and James O. Berger.

Invited Case Studies

Four invited papers will be presented at the workshop, organized into two sessions per day. Each presentation is scheduled to last three hours, which includes time for several invited discussants. Below is a list of authors and titles. Full abstracts are available from the Bayes '97 home page given at the end of this announcement.

- "Modeling Risk of Breast Cancer and Decisions about Genetic Testing," by Giovanni Parmigiani, *ISDS, Duke University*; Joellen Schildkraut, *Cancer Control Unit, Duke University Medical Center*; Eric Winer, *Department of Medicine, Duke University Medical Center*; and Don Berry, Ed Iversen, and Peter Mueller, *ISDS, Duke University*.

- "Functional Connectivity in the Cortical Circuits Subserving Eye Movements," by Christopher R. Genovese of the *Department of Statistics, Carnegie Mellon*; and John A. Sweeney of the *Western Psychiatric Institute and Clinic, University of Pittsburgh*.

- "Population Pharmacokinetic Modeling in Drug Development," by Jon Wakefield, *Department of Epidemiology and Public Health, Imperial College School of Medicine at St. Mary's, London, UK*; Leon Aarons, *Department of Pharmacy, University of Manchester, UK*; and Amy Racine-Poon, *Biometrics Group, Ciba-Geigy, Basel, Switzerland*.

- "Modeling Customer Survey Data," by Linda Clark, Bill Cleveland, Lorraine Denby, and Chuanhai Liu of *Bell Labs*

Invited papers from the Workshops I and II, together with a selected subset of contributed papers, were published by Springer-Verlag in refereed volumes of the mini-series "Case Studies In Bayesian Statistics." A third volume will appear soon, covering Workshop III, and a similar volume will be published for Workshop IV.

Organizers

Bradley P. Carlin	*University of Minnesota*
Alicia L. Carriquiry	*Iowa State University*
Constantine Gatsonis	*Brown University*
Andrew Gelman	*Columbia University*
Robert E. Kass	*Carnegie Mellon University*
Isabella Verdinelli	*Carnegie Mellon University*
Mike West	*Duke University*

Contact Bayes 97 via email at `bayes@stat.cmu.edu` or through the URL `http://www.stat.cmu.edu/meetings/Bayes97.html`

Statistics Week at Duke

October 9-13, 1997

Under sponsorship from various organisations, the Institute for Decision Sciences (ISDS) is hosting three linked research workshops during the first couple of weeks of October this year.

Stochastic Model Building and Variable Selection

October 9-10, 1997

Contact Organiser: Giovanni Parmigiani, ISDS

The workshop will bring together in an informal and productive way researchers active in the area of stochastic model building and variable selection. Advances in statistical methodology and computing are opening opportunities for statistical analysis and modeling of larger and more complex data sets. In this endeavor, it is becoming increasingly important to use computer-based tools for guiding the initial specification of the main features of statistical models. Recently, a new generation of stochastic algorithms has been emerging as an important augmentation to traditional deterministic strategies. Examples include stochastic search methods for variable selection, graphical models, selection of variable transformations and interactions, wavelet thresholding, ARMA modeling, CART, MARS, neural networks, data mining and more.

Among the motivations for the increasing usage of stochastic methods are (1) searching large spaces of model specifications more thoroughly than "greedy" deterministic algorithms, (2) implementing practical generalizations of standard model selection strategies, such as adaptive thresholding of wavelets, (3) incorporating subject matter knowledge, expert judgment, and existing evidence in a non-binding way, via prior distributions and utilities, and (4) providing a quantitative framework for accounting for the uncertainty in the model selection process. Imaginative solutions are being developed in specific application areas and there is wide potential for application of many of these solutions in other areas or disciplines. Similarly, a high level of activity and innovation is registered in methodological developments.

The goal of the proposed workshop is to promote interaction between researchers involved in diverse aspects of this field. We believe that interaction will: a) help elicit and provide focus on current issues and methods from a host of different application areas and disciplines; b) promote wider utilization of worthy practical approaches and solutions developed in specific fields; c) advance understanding of relative merits of existing tools and approaches, and d) point to directions for future methodological developments. Dissemination of stochastic model search techniques to practitioners and development of practical communication strategies for addressing model uncertainty in statistical consulting and teaching are also among the objectives of the meeting.

The 1997 NBER/NSF Time Series Seminar

October 10-11, 1997

Contact Organiser: Mike West, ISDS

Wavelets and Statistics

October 12-13, 1997

Contact Organiser: Brani Vidakovic, ISDS

The workshop is planned to focus on the developing interface between wavelets and statistical science. It will bring together a small group of leading researchers, applied workers, and new investigators, from various disciplinary backgrounds, to explore and summarize the current status of research on wavelets in statistics, to explore the applied impact that statistical wavelet modeling and wavelet methods in statistics are having, and to suggest and stimulate novel theoretical, methodological and computational research directions. It is an opportune time for the wavelet/statistics communities to come together to assess and address the dual questions about the status and future prospects: for the nature and potential contributions of wavelet ideas to statistical science and application, and for furthering wavelet based technology through the use of statistical concepts and models.

The workshop follows two recent meetings on wavelet methods in statistics: (i) Wavelets and Statistics, Institute IMAG-LMC, Grenoble, France, November 16-18, 1994, and (ii) the ANU Wavelet Workshop, Canberra, Australia, June 26-30, 1995.

The Center for Imaging Science

The Army Center for Imaging Science was established in 1995 with funding from the Army Research Office. It is a consortium which includes Washington University, Brown, MIT, the Universities of Texas at Austin and El Paso, Yale and industrial partners Lincoln Laboratory and Smith Kettlewell Institute. The activity is centered at Washington University under the Direction of Michael Miller, Professor of Electrical Engineering. The research of the Center is devoted to the development of representation and understanding of complex remotely sensed scenes, and reflects the broad multidisciplinary nature of imaging science, encompassing physics, mathematics, electrical engineering, computer vision, computer science and cognitive science.

Mission for the Center for Imaging Science

The overall goal of the Center is to develop the fundamental foundations of image understanding for recognizing and describing complex objects contained in natural and cluttered scenes. The thrust of the Center is towards the development of fundamental limits of performance of automated target recognition systems. This includes the development of detection and recognition bounds, as well as Cramer-Rao bounds on the pose of objects, and on complexity and information capacity bounds describing model databases and sensor/channel fusion characterizations.

Theoretical Foundation of the Center for Imaging Science

The scientific and intellectual foundations of the Center build on the Pattern Theories which have emerged over the past several decades from the mathematics, computational vision, engineering and statistics communities. The mathematical methodology being pursued by Center members is organized into the three principal components of image understanding: the representation of complex scenes, image formation and sensor modeling, and characterization of recognition and decoding strategies for image understanding.

A central theoretical theme involves the use of geometry for the representation of natural and cluttered scenes, focusing on both low-dimensional Lie groups for rigid bodies, as well as high-dimensional groups for deformable objects in clutter.

Call for Participation in the Center for Imaging Science

The Center for Imaging Science is amassing data-bases of research articles, simulators and remote sensor data associated with image understanding, computer-vision and automated target recognition. We invite the statistics and graphics community to participate in the Center for Imaging Science. Registering through the mailing list as a participant provides updates on activities and current research and availability of new information in the Center's data base.

Come register at `http://cis.wustl.edu/`

Director:

Michael I. Miller
Department of Electrical Engineering
Electronic Signals and Systems Research Laboratory
Washington University, Campus Box 1127
St. Louis, Missouri 63130
(314) 935-6195 (tel), (314) 935-7500 (fax)
`mim@cis.wustl.edu`

Contact Point:

Dr. David Skatrud
U.S. Army Research Office
Attn: AMXRO-PR/SFIA
P.O. Box 12211
Research Triangle Park, NC 27709-2211
(919) 549-4315 (tel), (919) 549-4310 (fax)

Call for JSM 1998 Proposals

The Statistical Graphics Section is seeking proposals for short courses for presentation at the 1998 Joint Statistical Meetings next August in Dallas, Texas. We are interested in courses that involve innovative uses of statistics in new application areas. We are especially interested in courses describing innovative uses of graphics in the design of user interfaces, statistics on the world wide web (WWW), visualization in conjunction with database mining, bioinformatics, and drug discovery. Proposals in other areas and proposals that would be cosponsored with other sections are encouraged. For more information or to informally present ideas, please contact Russ Lenth (`rlenth@stat.uiawa.edu`) or Perry Haaland (`pdh@bdrc.bd.com`). All proposals received before October, 1997 will be considered.

The Third Annual Student Paper Competition

The Computing Section announces its Student Paper Competition for 1998

The Statistical Computing Section of the ASA will again sponsor a Student Paper Session at the Joint Statistical Meetings in 1998. The winners present their papers in a special session at the annual ASA meetings. See the article on page 4 concerning the winners of this year's competition.

The topic of the session is *Statistical Computing*. Four students will be selected to participate in this session. All fees associated with registration, accommodation, and travel to the conference will be awarded to the participants in this Session.

Students at all levels (undergraduate, Masters, and Ph.D.) are encouraged to participate. To be eligible, an applicant must be a registered student in the fall of 1997. The applicant must be the first author of the paper.

To be considered for selection in the session, students must submit an abstract, a six page manuscript, a resume, and a letter of recommendation from a mentor familiar with their work. The manuscript should be single-spaced in a 10 point font with one inch margins (this is consistent with ASA's Proceedings guidelines.) All figures, tables and references should be included in the six-page limit. In the case of joint authorship, the mentor should indicate what fraction of the contribution is attributable to the applicant.

All application materials MUST BE RECEIVED by January 9, 1998. They will be reviewed by the Statistical Computing Section Student Paper Competition Award committee, which is made up of the Section's representatives to the ASA's Council of Sections. The topic of the paper should be in the area of statistical computing, and might be original methodological research, some novel application, or any other suitable contribution (for example, a software related project). Selection will be based on a variety of criteria at the discretion of the selection committee, and will include novelty and significance of contribution, amongst others. Award announcements will be made in late January, 1998. The selection committee's decision will be final.

Students not selected for inclusion in the Session may submit their abstract and a registration fee to ASA by February 1, 1998 if they plan to attend the Joint Meetings. Those abstracts must be submitted according to the ASA abstract submission instructions described in AMSTAT News. Students selected for inclusion in the session will receive further information about abstract submission and fee waivers from the award committee.

Inquiries and materials should be emailed or mailed to:

Terry M. Therneau
Student Paper Selection Committee
Statistical Computing Section
Section of Biostatistics
Mayo Clinic
Rochester, Minn 55905
`therneau.terry@mayo.edu`

All electronic submissions of papers should be in postscript.

ATTENTION, STUDENTS!

Statistical Graphics Section Student Paper Competition

New for 1998!

Are you interested in attending the 1998 annual ASA meeting? Are you looking for sponsorship to attend this meeting? Well, submit an entry to the Statistical Graphics Section Student Paper Competition!

The emphasis of the competition is statistical graphics. This can mean original methodological research, some novel application, or any other suitable contribution (for example, a data analysis project which employs an extensive use of graphics). The judging committee will select up to 3 winning papers to be presented by their authors at a discussion session at the 1998 annual ASA meeting. The winners will receive up to a $1000 award for expenses involved in attending this meeting.

The deadline for submission is January 9, 1998, but it is not too early to start thinking of an appropriate project. For more information, check the Graphics Section web page that can be reached through `http://www.amstat.org/` or contact Lorraine Denby (`ld@bell-labs.com`) or Mike Minnotte (`minnotte@math.usu.edu`).

SECTION OFFICERS

Statistical Graphics Section - 1997

Sally C. Morton, Chair
310-393-0411 ext 7360
The Rand Corporation
Sally_Morton@rand.org

Michael M. Meyer, Chair–Elect
412-268-3108
Carnegie Mellon University
MikeM@stat.cmu.edu

William DuMouchel, Past–Chair
908-582-7180
AT&T Laboratories
dumouchel@research.att.com

Dianne H. Cook, Program Chair
515-294-8865
Iowa State University
dicook@iastate.edu

Edward J. Wegman, Program Chair–Elect
703-993-1680
George Mason University ewegman@gmu.edu

Mario Peruggia, Newsletter Editor (96-97)
614-292-0963
Ohio State University
peruggia@stat.ohio-state.edu

Robert L. Newcomb, Secretary/Treasurer (97-98)
714-824-5366
University of California, Irvine
rnewcomb@uci.edu

Michael C. Minnotte, Publications Liaison Officer
801-797-1844
Utah State University
minnotte@math.usu.edu

Lorraine Denby, Rep.(96-98) to Council of Sections
908-582-3292
Bell Laboratories
ld@bell-labs.com

Colin R. Goodall, Rep.(95-97) to Council of Sections
Health Process Management
colin@hdsys.com

Roy E. Welsch, Rep.(97-99) to Council of Sections
617-253-6601
MIT, Sloan School of Management
rwelsch@mit.edu

Statistical Computing Section - 1997

Daryl Pregibon, Chair
908-582-3193
AT&T Laboratories
daryl@research.att.com

Karen Kafadar, Chair–Elect
303-556-2547
University of Colorado-Denver
kk@tiger.cudenver.edu

Sallie Keller-McNulty, Past–Chair
913-532-6883
Kansas State University
sallie@cecil.stat.ksu.edu

James L. Rosenberger, Program Chair
814-865-1348
The Pennsylvania State University
JLR@stat.psu.edu

Russel D. Wolfinger, Program Chair–Elect
919-677-8000 SAS
sasrdw@sas.com

Mark Hansen, Newsletter Editor (96-98)
908-582-3869
Bell Laboratories
cocteau@bell-labs.com

Evelyn M. Crowley, Secretary/Treasurer (97-98)
317-494-6030
Purdue University
crowley@purdue.edu

James S. Marron, Publications Liaison Officer
919-962-5604
University of North Carolina, Chapel Hill
marron@stat.unc.edu

MaryAnn H. Hill, Rep.(95-97) to Council of Sections
312-329-2400 SPSS
hill@spss.com

Janis P. Hardwick, Rep.(96-98) Council of Sections
313-769-3211
University of Michigan
jphard@umich.edu

Terry M. Therneau, Rep.(97-99) Council of Sections
507-284-1817
Mayo Clinic
therneau@mayo.edu

Naomi S. Altman, Rep.(97-99) to Council of Sections
607-255-1638
Cornell University
naomi_altman@cornell.edu

INSIDE

Statistical COMPUTING & GRAPHICS

The *Statistical Computing & Statistical Graphics Newsletter* is a publication of the Statistical Computing and Statistical Graphics Sections of the ASA. All communications regarding this publication should be addressed to:

Mark Hansen
Editor, Statistical Computing Section
Statistics Research
Bell Laboratories
Murray Hill, NJ 07974
(908) 582-3869 • FAX: 582-3340
cocteau@bell-labs.com
http://cm.bell-labs.com/who/cocteau

Mario Peruggia
Editor, Statistical Graphics Section
Department of Statistics
The Ohio State University
Columbus, OH 43210-1247
(614) 292-0963 • FAX: 292-2096
peruggia@stat.ohio-state.edu
http://stat.ohio-state.edu/~peruggia

All communications regarding membership in the ASA and the Statistical Computing or Statistical Graphics Sections, including change of address, should be sent to:

American Statistical Association
1429 Duke Street
Alexandria, VA 22314-3402 USA
(703) 684-1221 • FAX (703) 684-2036
asainfo@amstat.org

Nonprofit Organization
U. S. POSTAGE
PAID
Permit No. 50
Summit, NJ 07901

American Statistical Association
1429 Duke Street
Alexandria, VA 22314-3402
USA

This publication is available in alternative media on request.