Teaching Statistical Visualization

J. L. Solka [*]

Abstract

This paper examines some of the pleasures and challenges associated with teaching statistical visualization.

Key Words: teaching, statistical visualization.

1 HISTORY OF THE DISCIPLINE

Data visualization is one of those areas that had its seminal beginnings within statistics. It has not been widely embraced as a viable academic endeavor by the statistical academic community. In fact it seems to fall into the realm of a number of disciplines that had their beginnings in statistics but were later championed by other disciplines (Friedman, 1997).

It is my treatise that data visualization or statistical graphics should receive the full support of the statistical community. Following Friedman this means that they should publish articles about it in their journals, teach its practice in their undergraduate programs, teach relevant research topics in their gradate programs and provide recognition (jobs, tenure, awards) for those who do it well. My own personal experiences seem to indicate that the disciple is in general not being supported at this level.

2 MY BACKGROUND AND RESEARCH INTERESTS

I came to the discipline of statistical graphics with some previous work in scientific (physics and chemistry) visualization. In fact visualization has always been one of my great pleasures. My research interests have included scientific and statistical visualization, probability density estimation, clustering, and statistical pattern recognition. Some of my work has focused on image and video analysis, chemical agent detection, computer security, and gene expression analysis. I was previously trained at the doctoral level in the area of computational statistics. I obtained my doctorate from George Mason University in 1995 under the direction of Professor Edward Wegman. My dissertation was focused in the areas of mixture/kernel based probability density estimation and visualization of the density estimation process.

Since obtaining my doctorate I have taught in numerous capacities for George Mason and Johns Hopkins University. During this time I have taught a 3 credit hour course in scientific and statistical visualization twice, numerous 1 credit hour courses in scientific and statistical visualization, and a 3 day short course on MATLAB about 10 times. The third day of the 3 day short course focuses on visualization.

3 WHAT SHOULD WE BE TEACHING?

There are numerous topics that should be covered in a course on statistical visualization. The list of topics includes good practices, software systems, data/viewing transformations, one-dimensional analysis, m-dimensional analysis, and case studies. Doctoral students also need to know what constitutes research in statistical graphics, where they can publish their research, and what "bird of a feather" type groups exist within the community. I also believe that it is important that statistical visualization students also become acquainted with the visualization literature from various other academic visualization discipline. For example computer science has a rich history of work in the visualization arena.

My course topics list has been heavily influenced by the 1993 Wegman and Carr paper. Fusing the topics list from my 1996 GMU Scientific and Statistical Visualization course with that of my 2000 Johns Hopkins University Course one obtains the following list: the elements of graphing data, introduction to MATLAB and S–PLUS, graphics transformations, visualizing one-dimensional data, surface renderings, visualizing multivariate data, univariate density estimation, multivariate density estimation, dimension-

[*] J. L. Solka is a Senior Scientist, Advanced Computation Technology Division (Code B10), 17320 Dahlgren Rd., Dahlgren, Virginia, 22448-5100 (Email: jsolka@nswc.navy.mil). The author thanks Naomi Robbins for organizing the session.

ality reduction, and software systems. Students were graded based on several "out of class" pencil/software engineering assignments along with presentation of papers from the literature and student projects. The student projects usually resulted in a class presentation and paper by the student. The graphics development environments touched upon in the class included R/S–PLUS, MATLAB, JAVA, VTK, OPENGL, PV-WAVE/IDL, MATHEMATICA, and XGOBI/XGVIS.

4 HELP FROM THE LITERATURE

4.1 Handbooks

In this section we will discuss the references that have been of the most benefit to me in teaching statistical visualization. There are many wonderful books in this area but unfortunately length constraints allow me to discuss only a small subset. First we will discuss what I refer to as "handbooks". These works serve as standard references to "best practices" in the statistical visualization art. I think that one of the most comprehensive guides to modern statistical visualization can be found in Bertin, The Semiology of Graphics. This work first published in French in 1967 really foretells many of the techniques that were subsequently "rediscovered" by the statistical visualization community. Another reference that serves as a very good comprehensive overview is the seminal paper by Wegman and Carr published in the *The Handbook of Statistics.* This paper served as the basis of the curriculum in statistical graphics that I received at George Mason University during my doctoral training. A good overview of computer graphics in general can be found in Foley, Van Dam, Feiner, and Hughes. This tome serves as a *de facto* standard for the computer graphics community. A more modern treatment of statistical graphics philosophy can be found in the recent (1999) text by Wilkinson. This book although somewhat obtuse, has been well received by the community as a whole. Finally I mention that I usually cover a section on "best practices" in statistical graphics using Bill Cleveland's 1994 book.

4.2 Software Systems

Next we will discuss those references that detail the inner workings of particular software systems. Much of the analysis performed in my class can be accomplished using S–PLUS or MATLAB. Recently I have become a proponent of the public domain version of S–PLUS that is known as R. A good introduction to S-PLUS is contained in *A Handbook of Statistical Analysis Using S–PLUS.* This short text provides the student with the necessary S–PLUS skills needed for an undergraduate course. An in-depth treatment of MATLAB can be found in *Mastering MATLAB 6* by Hanselman and Littlefield. I have successfully used this book for several of the MATLAB short courses that I taught. If one is interested in analyzing a moderately sized high-dimensional data set then I would recommend the use of XGOBI. XGOBI is an X-WINDOWS based software system for the analysis of said data sets. The reader is referred to Swayne, Cooke, and Buja (1998) for a thorough description of the inner workings of XGOBI. The same team developed a similar package that provides for multi-dimensional scaling. This package XGVIS is discussed in Buja, Swayne, Littman, and Dean, (1998).

4.3 Probability Density Estimation and Clustering

Density estimation and clustering are methods of exploratory data analysis that allow the user to examine the structure of a data set and possibly to formulate hypotheses about the data set. The references discussed in this section reflect primarily my own exposure to the domain area more than any thing else. My own interests include kernel and mixture based probability density estimation and clustering and this is clearly reflected in my literature choices. A good overall treatment of kernel and other nonparametric methods of density estimation can be found in Scott's *Multivariate Density Estimation.* This text does a good job of covering histogram, frequency polygon, average shifted histogram, and kernel based density estimation methods. Unfortunately there is not much treatment of multivariate density estimation despite the title of the text. Titterington's *Statistical Analysis of Finite Mixture Distributions* provides a wonderful introduction to the application of finite mixture models in density estimation and clustering. I also try to cover during the class some of the more recent work in these areas. Priebe, (1994) discusses a semiparametric extension of finite mixture models which he deems adaptive mixtures. Solka, Poston, and Wegman, (1995) discuss visualization of the time evolution of finite and adaptive mixture models. Solka et al., (1998) discuss cluster structure assessment in high-dimensional spaces.

The last two references to be discussed particularly demonstrate the benefits of a close coupling between visualization methods and clustering. Minnotte and West, (1999) formulated a visual method of cluster assessment based on an idea first proposed by Wegman, (1990). This method can be used on moderately sized, $n \leq 1000$, data sets in fairly high dimensional spaces, $p \leq 1000$. This type of approach, termed the data image, has recently been championed by the biological community. The data image has proven particularly effective in the analysis of gene expression patterns, see Eisen et al, (1998).

4.4 Dimensionality Reduction, Tours, and Alternate Coordinate Systems

In this section we will discuss methods to reduce the dimensionality of a data set prior to its analysis along with non-Cartesian coordinate systems for the representation of the data. When the inherent dimensionality of a data set is larger than 3 one looses the ability to place the planar coordinate axes perpendicular to one another. Wegman, (1990) proposed placing the coordinate axes parallel to one another as a way to circumvent this problem. This method is beneficial for moderately sized data sets $n < 1000$ with dimensionality $p < 20$. Wegman has found a method using color saturation that allows one to push this method to data sets with $n \leq 100000$.

Given a set of observations in $p > 2$ dimensional space one is often interested in forming fortuitous linear combinations of the observations. Asimov, (1985) formulated several methods to effectively explore the space of all possible projections from p dimensional space to two dimensional space. These methods have collectively been designated "grand tour" methods. Solka et al., (1998) developed a new grand tour method based on the use of space filling curves.

The real power of this approach comes from the synergy of the tour-based methods and the alternative coordinate systems. Wegman, (1990) cleverly proposed the marriage of the grand tour methodology with the parallel coordinates framework. This allows the user to interactively investigate the inner structure of a high-dimensional data set.

4.5 Application Case Studies

The application case studies presented in the class are heavily focused on my own personal interests. Much of my work in recent years has focused on discriminant analysis, probability density estimation, and clustering and this slant is made clear during my in-class case studies discussions. The paper by Solka and Green, et al., (1998) is discussed. This paper details the application of statistical graphics to the discriminant analysis process. I also cover some of my recent work in chemical agent detection, (1998). Some of the most recent work that I have discussed involves the application of statistical visualization to intrusion detection, Solka, Marchette, and Wallet, (2000).

5 RECOMMENDATIONS

5.1 Teaching Recommendations

My limited experience has led me to make the following teaching recommendations. First consider teaching the class with a colleague from an other discipline. This type of "cross fertilization" can prove very beneficial. Second consider the use of "nonstandard" texts. I find the recent work by R. Spence, *Information Visualization*, (2001) to be very interesting. If possible try to enforce prerequisites for the students that take the class. I know that this can be a tenuous issue at times but it can help improve the overall quality of the class. Try to make sure that the students enter the class with advance knowledge of at least one computer language. Knowledge of a high level language such as S–PLUS/R or MATLAB is particularly beneficial. Try to provide the students with solutions to class problems in more than one language. Finally, encourage the students to work together on class projects.

5.2 Advising Recommendations

This section covers my recommendations with regard to advising our students. First I think that we need to sensitize our students with regard to statistical research issues. It is good to discuss with them what exactly constitutes fundamental research in statistical graphics. I believe that new visualization methods, new visualization frameworks, case studies, visualization methods tagging along with new algorithms, and computational improvements to existing algorithms are all publishable work. Sometimes I wonder if there are really any new visualization methods after the ground breaking work of Bertin. I think of XGOBI as good example of a visualization framework. A visualization framework is a system for visualizing data/models that has been assembled from a set of existing components. None of the individual components may be new but their assemblage is novel. I define case studies as the application of existing visualization methods to new data sets. The novelty of this type of analysis may be determined in part by the insights obtained on the data set. I define the last two types of research areas as follows.

A researcher may often develop a new visualization method to study/debug a new algorithm that he has developed. This type of visualization research can prove particularly beneficial to the researcher. The last research type example relates to improvements to existing visualization algorithms. This type of work is often published in the computer science visualization literature.

The second advising recommendation is to educate our students as to publication arenas. This job may not be as straight forward as one might imagine. My own experiences seem to indicate that it may not be as easy to publish statistical graphics research as say research in probability density estimation. With this caveat in mind I recommend the following journals in the order of the likelihood that one can successfully publish in the journal. First I think that the Proceedings of the Interface is a marvelous place to publish. This conference is one of my favorites and it is fairly well attended at least by those researchers in computational statistics and statistical visualization. My next recommendation would be the Statistical Computing and Graphics Newsletter. This publication often features many good articles and it clearly seems focused at least in part on the topic of statistical visualization. Next I would recommend the Journal of Computational and Graphic Statistics (JCGS). This journal would have received a more glowing recommendation if it were not for the fact that the focus of the journal in recent years has been more in the computational arena. Next I would recommend special issues of statistical journals. Sometimes a given journal such as Computational Statistics and Data Analysis will publish a special issue of their journal that is specifically focused on statistical visualization. Next I would recommend other publication venues such as the Journal of Data Mining and Knowledge Discovery, IEEE Pattern Analysis and Machine Intelligence, and even the Proceedings of the National Academy of Science. The last publication outlet that I recommend is the Journal of the American Statistical Association. This of course is one of the premiere journals in statistics, however I believe that one could count the number of statistical visualization papers that have appeared in this journal on one or perhaps two hands.

5.3 Community Recommendations

I close with my recommendations to the statistical visualization community. First I believe that a new statistical graphics journal should be established. This journal should be solely focused on the rapid dissemination of research in statistical graphics. I don't think, based on my comments above, that this journal will be a total duplication of JCGS. My last recommendation is for the statistical visualization community to forgo infighting. This research community is too small for there to be continued ego driven disagreements among the major players in the community.

References

[1] Asimov, D. (1985), "The grand tour: a tool for viewing multidimensional data", *SIAM J. Sci. Stat. Comput.* , pp. 128- 143.

[2] Bertin, J. (1967), *Semiologie Graphique* , Editions Gauthier- Villars, (English translation by W. J. Berg as Semiology of Graphics , University of Wisconsin Press, Madison, 1983).

[3] Bujas, A., Swayne, D., Littman, M. and Dean, N. (1998), "Xgvis: Interactive Data Visualization with Multidimensional Scaling," *Journal of Computational and Graphical Statistics.*

[4] Cleveland, W. S. (1994), *The Elements of Graphing Data* , CRC Press.

[5] Eisen, M., Spellman, P., Brown, P., and Botstein, D. (1998) "Cluster analysis and display of genome- wide expression patterns", *PNAS*, Vol. 95, pp. 14863–14868.

[6] Everitt, B. (1994), *A Handbook of Statistical Analysis using S- Plus* , Chapman and Hall.

[7] Foley, J., Van Dam, A., Feiner, S. K., Hughes, J. (1990), *Computer Graphics: Principle and Practice , Second Edition, Addison- Wesley.*

[8] Friedman, J. R. (1997) "Data Mining and Statistics: What's the Connection?," Proceedings of the 29th symposium of the Interface (David Scott Editor.

[9] Hanselman, D. and Littlefield, B. (2000), Mastering MATLAB 6 , Prentice Hall.

[10] Minnotte, M., and West, R. (1999), "The data image: a tool for exploring high dimensional data sets", 1998 Proceedings of the ASA Section on Statistical Graphics.

[11] Priebe, C. E. (1994), "Adaptive Mixtures", JASA, Vol 89, No. 427, pp. 786–806.

[12] Scott, D. (1992), Multivariate Density Estimation , John Wiley and Sons.

[13] Solka, J. L., Marchette, D. J., and Wallet, B. W. (2000) "Statistical Visualization Methods in Intrusion Detection", presented at and appearing in the Proceedings of Interface 2000.

[14] Solka,J. L., Marchette, D. J., Poston, W. L. and Wallet, B. C. (1998) "Visualization Methods for Cluster Assessment in High Dimensional Spaces", Proceedings of the Joint Conference on Information Sciences.

[15] Solka, J. L., Green, J. E., Norton, J. J., Guidry, R. J., and Marchette, D. J. (1998), "Applications of Statistical Graphics to Discriminant Analysis", Proceedings of the Section on Statistical Graphics of the American Statistical Association , pp. 1–5.

[16] Solka, J. L. and Marchette, D. J. (1998) "Discriminant Analysis for the Detection of Chemical Agents", unpublished manuscript .

[17] Solka, J., Wegman, E., Reid, L., and Poston, W. (1998) "Exploration of the Space of Orthogonal Transformations from R^p to R^p Using Space–Filling Curves", Computing Science and Statistics, Vol. 30, pp. 494–498.

[18] Solka, J. L., Poston, W. L., and Wegman, E. J. (1995), "A Visualization Technique for Studying the Iterative Estimation of Mixture Densities", Journal of Computational and Graphical Statistics, 4(3), pp. 180–197.

[19] Spence, R. (2001) Information Visualization, Addison-Wesley.

[20] Swayne, D., Cook, D., and Buja, A. (1998) "Xgobi: Interactive Dynamic Data Visualization in the X Window System", Journal of Computational and Graphical Statistical , Vol 7, Number 1, pp. 113- 130.

[21] Titterington, D. (1994) Statistical Analysis of Finite Mixture Distributions , John Wiley and Sons.

[22] Wegman, E. J. and Carr, D. B. (1993), "Statistical Graphics and Visualization.," in Handbook of Statistics, Computational Statistics, Vol. 9. ed. C.R. Rao, North Holland, New York. pp. 857–958.

[23] Wegman, E. (1990) "The grand tour in k dimensions", Computing Science and Statistics Proceedings of the 22nd Symposium on the Interface , pp. 127- 136.

[24] Wegman, E. (1990) "Hyperdimensional Data Analysis Using Parallel Coordinates", JASA, pp. 664- 675, 1990.

[25] Wilkinson, L. (1999) The Grammar of Graphics , Springer.