STATISTICAL WORKSTATIONS FOR DATA ANALYSIS AND BIOMETRY[1]

G. Sawitzki, Heidelberg

Summary

The availability of workstations over the last years has contributed to the development of new statistical methods, in particular in the area of data analysis and in non-parametrics. Projection pursuit and CART are the most prominent examples. Besides stimulating new methods and approaches, workstations can be an effective tool for applied statistics. Technical requirements for the use in biometry are discussed.

Introduction

Workstations have been in use in statistical research now for a decade, and are becoming tools for applied fields, like biometry. Focussing on the aspect of computing power, a workstation can be defined as a high end personal computer. Already as such, workstations have opened new possibilities. We will illustrate this with two examples, PRIM and "linked windows" . Then we move to a more specific definition of a workstation and examine some requirements and consequences for its use in biometry.

Workstations: a Reduced Definition

Workstations evolved from attempts to design a working environment augmenting human information processing facilities (in the words of the fifties: "augment the human intellect"). Focussing on electronic media, especially on computing, a widely circulated Carnegie Mellon paper in the early 80s defined technical minimum requirements

- 1 million instruction per second (1 MIPS) computing power
- 1 million byte (1 MegaByte) internal memory
- 1 million pixel (1024*1024) graphical display

(for at most $10 000). This has become known as the 3M-definition of a workstation. It covers only a minor part of what is meant by a workstation, but it illustrates where the technical range of workstations is starting. Early models of workstations did about meet the 3M requirements (though at a higher price).

---

[1] Reprinted from Zs. Harnos (ed.), Proceedings of the XVth International Biometric Conference, Budapest 1990, with kind permission of the editor.

A guaranteed computing power of 1 MIPS, together with a high resolution display, allows substantial calculations. Moreover, it may allow interactive graphics: read input devices (keyboard or analog input like an input mouse), do appropriate calculations and display results without noticeable delay, thus giving the illusion of direct interaction with graphics.

Workstations and Development in Statistics

The use of interactive graphics in statistics, in particular in data analysis, has been reviewed repeatedly, for example Becker et al. (1987), Sawitzki (1990). Interactive graphics can be used for easier access to well known diagnostic plots. In residual analysis for example interactive graphics goes down to elementary actions like identifying individual influential observations or outliers simply by pointing at them. The ease of use is an important help in practical work. Beyond this, interactive graphics has been stimulating new theoretical developments.

Example 1: PRIM and projection pursuit. The PRIM project (Fisherkeller et al., 1974) introduced tools for interactive visualization of higher dimensional data. In the special case of the original PRIM -9, it allowed control of two dimensional projections from a 9 dimensional space, with free control over three dimensions. PRIM also included techniques like masking by additional variables or isolation of subgroups.

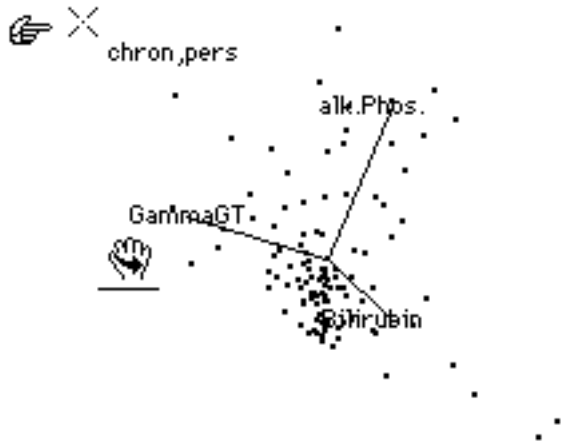

Figure 1: PRIM technique (projection, rotation, isolation, masking) allowing free movement in three dimensions from a higher dimensional data set. Insividual observations can be identified by pointing on them. The label of the identified case (here: a diagnosis) is shown as response. The example uses 3 variables from a clinical data set, observed on patients with various forms of hepatitis.

The idea to hunt for interesting projection was fixed in the idea of projection pursuit: given observations in $\mathbf{R}^{\mathbf{d}}$, you want to identify "interesting" projections A: $\mathbf{R}^{\mathbf{d}} \rightarrow \mathbf{R}^{\mathbf{k}}$. Since generically most projections are approximately normal as $d \rightarrow \infty$, the projections which are different from normal are the interesting ones. Projection pursuit uses various projection indices

Q:$\mathbf{R^d \rightarrow R}$ measuring deviation from normality in a certain direction, and hunts for projections *A* yielding a high projection index. A review of methods based on projection pursuit is given in (Huber, 1985).

Example 2: Brushing and CART: Classification and Regression Trees. Brushing in linked windows means continuously selecting subgroups in one plot with immediate feedback in other linked plots. If the plots are scatterplots showing different two-dimensional projections of the same data set, it is like applying colour with a brush from one face and looking at it appearing on the other sides of the scatterplot hypercube. The earliest reports on interactive graphics for linked windows seems to be Newton (1978, quoted from Becker et al. 1987). A very impressive implementation by John McDonald in the Orion project controlled a "brush" with a mouse. While the brush was moved across one plot, the points under the brush got coloured depending on their distance to the center of the brush. The corresponding representations of the data points in other windows (showing other variables) were coloured in the same colours.

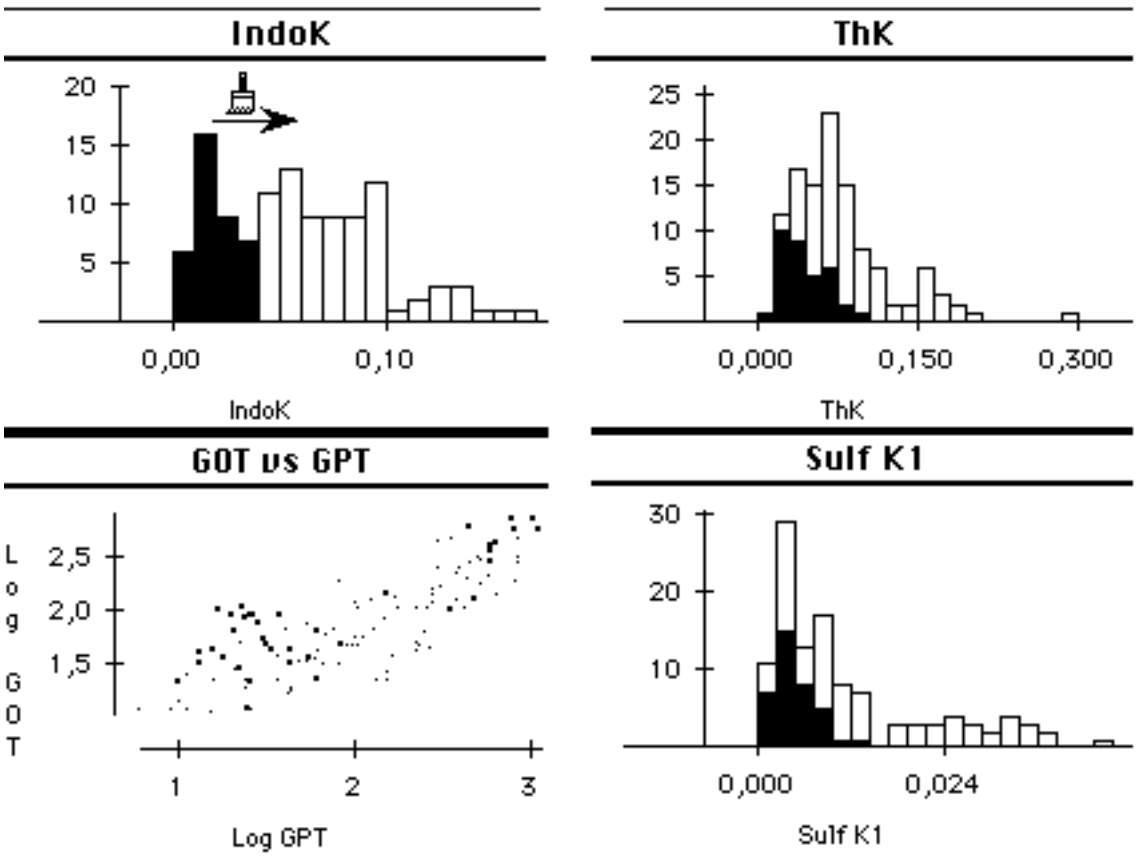

Figure 2: Brushing linked windows. A"brush" is moved across one of the plots (here: the top left histogram) technique. Corresponding observations are highlighted in the linked windows: for histograms, a corresponding proportion of the histogram bars is highlighted; for the scatterplot, the corresponding observations are marked as the brush is moved.
This example uses 5 variables from a clinical data set (kinetic constants for indocyanine, theophilline and BSP, and GOT/GPT levels).
One of the useages of brushing in linked windows is to search for ranges in some variables

which have homogeneous responses in others. This can be formalized to yield the idea of CART (classification and regression trees). In the regression context: the non-parametric regression problem is formulated as the problem to recover the regression function $\mathbf{E}(Y|X)$, where $X \in \mathbf{R}^{\mathbf{k}}$ are regressors and $Y \in \mathbf{R}^{\mathbf{m}}$ is the response of interest. The idea of CART is to tile the regressor space $\mathbf{R}^{\mathbf{k}}$ into a reasonably small number of tiles (blocks) $T_b=\{x: y_{b,i}<x_i \leq z_{b,i}\}$ to achieve maximum homogeneity within the tiles. The mean over a tile is used as an estimate for the local value of $\mathbf{E}(Y|X)$. The tiling is done by successively brushing over the marginals until an optimal split point is found. An extensive introduction is given in Breiman et al. (1984).

Workstations: a Proper Definition

Of course it is interesting to have sufficient computing power. But only a small part of the biometrical work consist amounts to statistical analysis or mere computing. Biometrical work includes other parts as well, like experimental planning, administration, control of experiments and documentation.

Workstation originated in investigations on "augmentation technologies", technologies supporting the general process of scientific work. Computing power and visual feedback is not enough to help to this aim. The process which has led to workstations has strongly been influenced by work done at the XEROX' Palo Alto Research Center (XEROX PARC) in the 70s. The PARC studies recognized the basic need for effective information access, both in the sense of information retrieval from common or project oriented data bases, and in the sense of information exchange like in research teams. Ever since, effective means for information access have become fundamental features of workstations.

Going beyond the rudimentary 3M-definition, a workstation is a dedicated computing facility which satisfies the immediate needs of its user for

- access to electronic information
- electronic communication
- data processing
- data presentation

It has to be adaptable to the requirements and preferences of its user, and should support integration into the ambient team infrastructure.

Hardware and software requirements cannot be separated. A workstation is defined by its functionality, and both hardware and software have to contribute to guarantee this functionality. Moreover, critical requirements cannot be stated in a general way. The intended use defines the requirements, and requirements can only be discussed with respect to a specific field. Typically in biometry, there are different requirement profiles.

The consulting situation is most demanding. An important task in consulting is to communicate the pre-conditions and the consequences of modelling choices and decisions. If the client is not

very well educated (as seen from the biometrician), we would like means to illustrate the importance of the assumptions, and the possible directions of analysis. But this would demand extensive flexibility and computing power. The aim would be to allow for free simulation of experiments and analysis in real time, i.e. adapted to the time scale of human communication.

Exploratory data analysis comes next in the level of requirements. Here interactive work is of main importance. Short delay times and flexible control become dominating, including flexible access to external sources of information. Journalling and editable play back facilities are necessary components for a workstation for exploratory work.

Classical experimental design or statistical analysis are relatively modest. For a routine evaluation, the proper calculation in general can be planned in advance and does not define an immediate need - even batch processing might be satisfactory for a routine evaluation. But reliability may be a feature of high priority for routine evaluation. What enters is the need for documentation including a flexible editing of mixed texts, tables and graphical output.

A workstation has to satisfy the *immediate* needs. Immediate means at once, which requires a guaranteed computing power and rules out general time sharing solution. But immediate does not include tasks which may be finished later. That is, a workstations is not intended to be a universal computing environment. If there is an immediate need to run a model calculation online, for example in a consulting situation to discuss the outcome with the client, then the workstation should have sufficient computing power. But if there is an immediate need to define a query for a data base, or to ask for an excerpt from an epidemiological base, or to launch a calculation on a mass data base, then the workstation needs only support the definition of the task. After that, for example it should communicate on its own with a remote mainframe to do the bulk work. In this case, we would not require computing power from the workstation, but machine-to-machine communication. In general, it is not adequate to isolate one component from the definition of a workstation. We need it all.

Estimating Required Power

Critical requirements cannot be stated in a general way. So the required power must be estimated for specific use. We did extensive measurements to try to relate *formal* power parameters to the *effective* power for the user. However the technology still seems to be so inhomogeneous that it is too early to give fixed formulae. In this situation, a possible strategy is to analyze the requirements for typical applications. We will illustrate this strategy with two examples and recommend a similar proceeding adapted to the choices in question. The requirement profile we have in mind is that of a consulting session, an immediate dialog between consultant and client.

Example: Computing Power. By mere guesswork, we selected regression as a typical application. We took a sample size of 100 as a mediate sample size, and 5000 observations as a high sample size for immediate evaluation. Remember that we only included immediate needs in

the definition of a workstation. We assumed that samples of a larger sample size would be worth a strategic planning. So it would not be critical to the requirements of a workstation. Assuming 5 regressors to be a typical medium count, we have fixed a situation we can check practically. Experiments with simulations on a 2 MIPS machine, gave a time of 2 seconds for 100 observations, ≈3 s for 5000 observations. This is a sufficient speed to do the calculation even in a consulting session, without interrupting the dialog. Since the regression algorithm must at least have a complexity depending on sample size, the small difference in time shows that the computing power is not yet the limiting factor, and 2 MIPS computing power are sufficient for the problems in question.

Since computing time (and sample size) proved to be uncritical for the problem size in question, we need not vary these parameters. However computing speed is determined by both computing power and data access. So we varied parameters affecting speed or frequency of data access. For the first run, we had allocated 1.5 megabyte of memory for the process, allowing for abundant internal tables. For a second run, we reduced the allocated memory to 0.5 megabyte. Obviously this would lead to more memory management/swapping operations. The observed time to completion for 5000 observations increased to 5s-6s. So the available memory is critical, and there is a trade off between time and memory. A free/accessible memory of 1-2 megabyte, and 2 Mips computing power, seems to be necessary for the regression problems of this size.

The situation changes drastically if we go from regression to bootstrapping. Trying to bootstrap upto a level of 5% with a variation coefficient of 10%, that is doing a rough bootstrap, would require about 2000 bootstrap runs. If a simple bootstrap replication is used, the above measurements would need about 1 hour of computing time at 2 MIPS, 100 observations. To get bootstrapping down to a time scale which can be used in a consulting situation (seconds), 100 MIPS-120 MIPS would be necessary.

Example: Display Power. To determine the necessary display power, we use a similar strategy. We first define a critical event, and then we perform our estimate based on this event. Regression requires very little output. So it is not useful as a general critical element. As a critical element we choose the performance of classical BMDP output. This is in general a well structured output, and we do not want to fall behind what is provided by e.g. BMDP.

A standard BMDP output page consists of 72 limes by 128 chars (blanks included). Browsing means scanning about 4 pages/second. So the net throughput equivalent to browsing BMDP output would be 36 K byte/second. By usual overheads, this means 1 MBaud nominal throughput would be necessary. Note that this easily can become a critical requirement. If you are not working in a decentralized way, 10 stations may eat up all the bandwidth of a standard ethernet installation !

For high resolution graphics, you have about 1 megabit/colourbit screen information. Browsing 4 picture/second with 4 bit colour (16 colours) means a net throughput of 2 megabyte/second.

To keep a memory of the pictures requires 0.5 megabyte/picture, so for a minimum storage for 4 pictures you need 2 megabyte more of memory in access to the memory you need for your computing tasks. Adding to the figures from the example above this gives a requirement of 3-4 megabyte of free memory.

Situation changes drastically if you want interactive high resolution graphics in full colour. A single 32 bit picture occupies about 4 megabyte. So for a refresh rate of 20 pictures per second, just about enough to give the impression of running pictures, you need a net throughput of 80 megabyte/s.

Model application:
linear regression,
5 regressors,
100…5000 observations,
immediate (≈3s) response.

Minimal requirements
4 megabyte memory
2 MIPS computing power
36 K byte/second effective transfer over network
2 megabyte/second data transfer to display

Table 1: Minimal requirements

Model application: bootstrap for regression, as above

Requirements
80 megabyte memory
120 MIPS computing power
80 megabyte/second data transfer to display

Table 2: Bootstrap & graphics requirements

Conclusion

Hardware and software requirements cannot be separated. Some estimates of hardware requirements for typical applications are given here. An attempt towards listing minimal software requirements is made in Rasch & Sawitzki, these proceedings. The hardware components to give a useful workstation are available. The software still shows considerable problems. There are software systems which allow to use the flexibility and computing power available on workstations, and there are steps towards enhancing classical statistical packages to make use of the workstaion technology.

But is still up to the user to define the software features needed to make an effective integrated working environment, and then it will be up the user again to urge software producers to make this environment available.

Literature:

Becker, R.A., Cleveland, W.S., Wilks, A.R. (1987) Dynamic Graphics for Data Analysis (with discussion). Statistical Science 2 355-395

Breiman, L., Friedman, J.H., Olshen, R.A. and Stone, C.J (1984) Classification and Regression Trees. Wadsworth, Belmont

Fisherkeller, M.A., Friedman, J.H. and Tukey, J.W. (1974) Prim-9, an Interactive Multidimensional Data Display and Analysis System. In: Proceedings of the Pacific ACM Regional Confrerence 1974

Huber, P. J. (1985) Projection Pursuit. Annals of Statistics 13 435-525

Sawitzki, G. (1990) Data Analysis: Tools and Concepts. In: F. Faulbaum et al. (eds.), Softstat '89, Fischer 1990

Statistical Workstations for Data Analysis and Biometry

Dr. Günther Sawitzki
Institut für Angewandte Mathematik
Im Neuenheimer Feld 294
D 6900 Heidelberg