COMMENTARY

Discussion of "Visualizing Statistical Models: Removing the Blindfold"

Catherine B. Hurley

Department of Mathematics & Statistics, Maynooth University, Maynooth, Co. Kildare, Ireland

DOI:10.1002/sam.11273
Published online 16 July 2015 in Wiley Online Library (wileyonlinelibrary.com).

1. COLLECTIONS, COMPARISONS, GRAPHS, AND SERIATION

I enjoyed reading the paper "Visualizing Statistical Models" and thank the authors Wickham, Cook, and Hofmann for an informative and stimulating contribution. The notion of using visualization techniques on model results is important as illustrated in the case studies. For those of us interested in taking on the challenge of developing "visualizations for more models in a wider array of statistical and graphical environments," the current paper (VSM) presents good strategies and a useful distinction between visualizing the data in model space and the model in data space. The original work presented here dates from 2007, yet the ideas are fresh.

A lot of careful thought goes into choices of visualization and then its design and implementation. When the presentation of ideas and visualizations are so clear as in the current paper, it all seems so easy, but it is in fact far from easy. The authors hint at this when they say "converting data-vis to model-vis is not always straightforward." Some new techniques are required, to find prediction boundaries as in Figure 5 of VSM, and to display dendrograms in feature space as in Figure 9. I looked forward to trying out the R packages classifly and clusterfly but they require rggobi which disappointingly no longer has a Mac version. Meifly installs but does not implement the techniques of Section 4.1. I think these packages have a lot to offer the statistical community and I would be delighted if they were updated to work on current hardware, with code for the visualizations of the current paper included as a vignette perhaps.

1.1. Collections and Comparisons

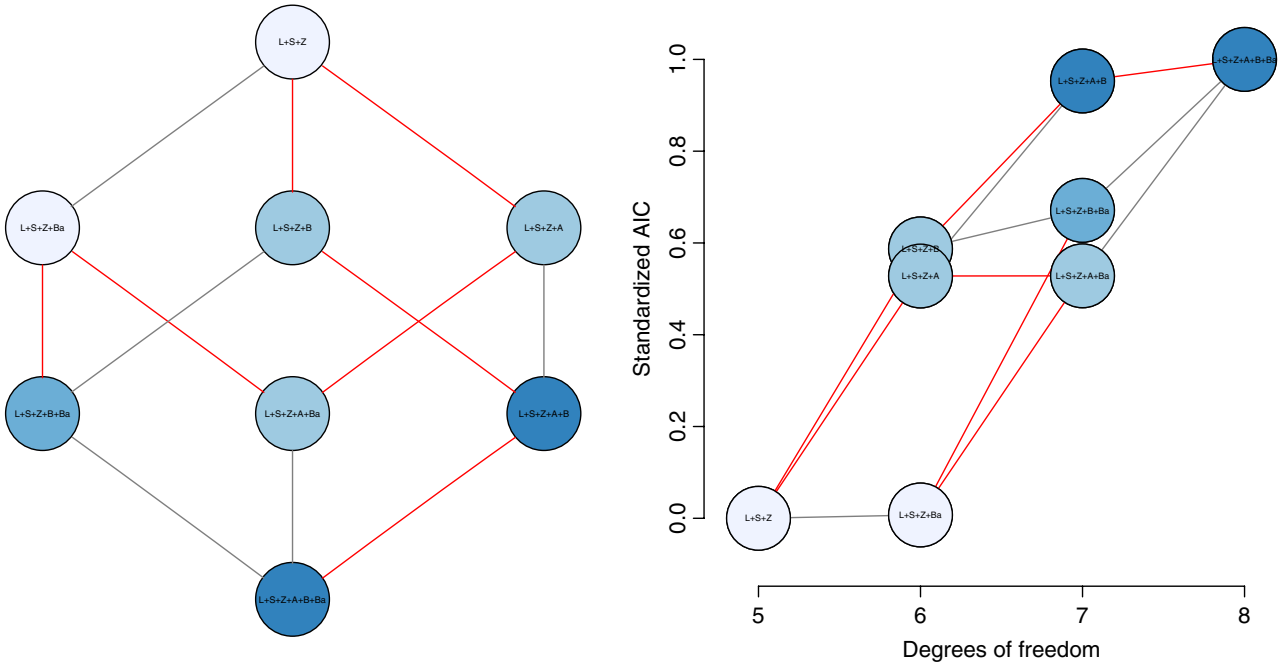
A theme of this paper is visualizing collections of models and iterative model fits. Perhaps there is a need for tools to help organize these collections that will facilitate comparisons. In their PairViz methodology Hurley and Oldford [1,2] proposed the use of a mathematical graph to organize collections of statistical objects, where edges in this graph represent comparisons of interest. Applying this principle to the collections of models proposed in Section 4 VSM, we have graphs whose nodes are models and edges represent model comparisons. For example, to explore the space of all possible regression models we could build a graph where each node represents a regression model, and edges connect models that differ by exactly one predictor.

Fig. 1 shows such graphs for the New Haven Housing Data, as in Section 4 of VSM. Only the best eight models according to AIC are shown. The left-hand side figure shows a conventional graph layout, the right-hand side shows node position given by (standardized) AIC and degrees of freedom, similar to Figure 11. Paths through the graph such as the one shown in red could be used as a basis for comparing summary measures for collections of models. We could then easily see whether adding or removing a predictor changes other coefficients, far more effectively than pouring over pages of tables. For such displays to be truly useful they could be supplemented by interactive tools for filtering nodes and edges. Interactive tools for selecting paths through the graph could then be used to drive model comparison visualizations. In their RnavGraph package Waddell and Oldford [3] (see also ref. [4]) used similar ideas to drive exploration of high-dimensional data.

1.2. Role of Seriation

Paths through graphs could be found algorithmically, using Eulerian paths that visit all edges or Hamiltonian paths that visit all nodes [1]. Seriation techniques (see e.g., refs. [5-7]) produce paths visiting all nodes. Such paths may be chosen specifically to improve comparison

* *Correspondence to:* Catherine B. Hurley (catherine.hurley@nuim.ie)



(a) Graph layout

(b) Scatterplot layout

Fig. 1 Top eight models according to AIC. Edges connect models that differ by exactly one predictor. (a) Graph layout (b) Scatterplot layout [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

of statistical objects (cases, variables, and models) through better visualizations.

It is interesting to consider how seriation techniques might be used to improve some of the visualizations presented by the authors. The dendrogram of Figure 8 of VSM is an obvious candidate. The three clusters shown correspond roughly to the three wine classes with just a small number of wines misclassified.

This visualization shows the data (leaves) in model space, but potentially in a misleading way. A few of the variety A wines shown in purple are placed at the far end of the predominantly B variety green cluster, giving the impression that these variety A wines are at the edge of the green cluster and are extreme outliers relative to their class. A number of the variety B (green) wines appear within the predominantly C variety orange cluster. Again their position gives the impression that these variety B wines are extreme outliers relative to their class. But, in fact the default arrangement of dendrogram leaves is somewhat arbitrary and these wines may not be extreme outliers. Seriation improves the dendrogram and reduces its potential to mislead. In Fig. 2, we use the DendSer seriation algorithm [6,5] to re-arrange the leaves of the dendrogram so that leaves close in the dendrogram are close in feature space (as measured by the distance matrix). The misclassified variety A wines now appear near the cluster with the other variety A wines. Similarly, two of the misclassified variety B wines are moved adjacent to the other B variety wines. This dendrogram gives the impression that perhaps two variety B wines might be outliers relative to their class.

We can check this in feature space. The displays in Figures 9 and 17 of VSM show some overlap between the clusters and that misclassified points are not extreme outliers. Incidentally, as the hierarchy of joins is not evident in Figure 9 even with close scrutiny, I am not convinced that plotting dendrograms in feature space is a useful m-in-ds visualization.

1.3. Enhance Teaching

I strongly agree with the authors on the value of model visualization in teaching. As a profession we could make much better use of (preferably interactive) graphics in teaching, even to explain and explore concepts such as tables of numeric regression output and ANOVA tables. Simple tools such as nomograms allow students to play around with models which should deepen their understanding. The DynNom R package [8] is a nice nomogram implementation based on Shiny [9] for general linear models. I would welcome more and better tools for interactive visualization of models. The diverse case studies and methodologies used by Wickham, Cook, and Hofmann present a convincing argument for the value of such tools,

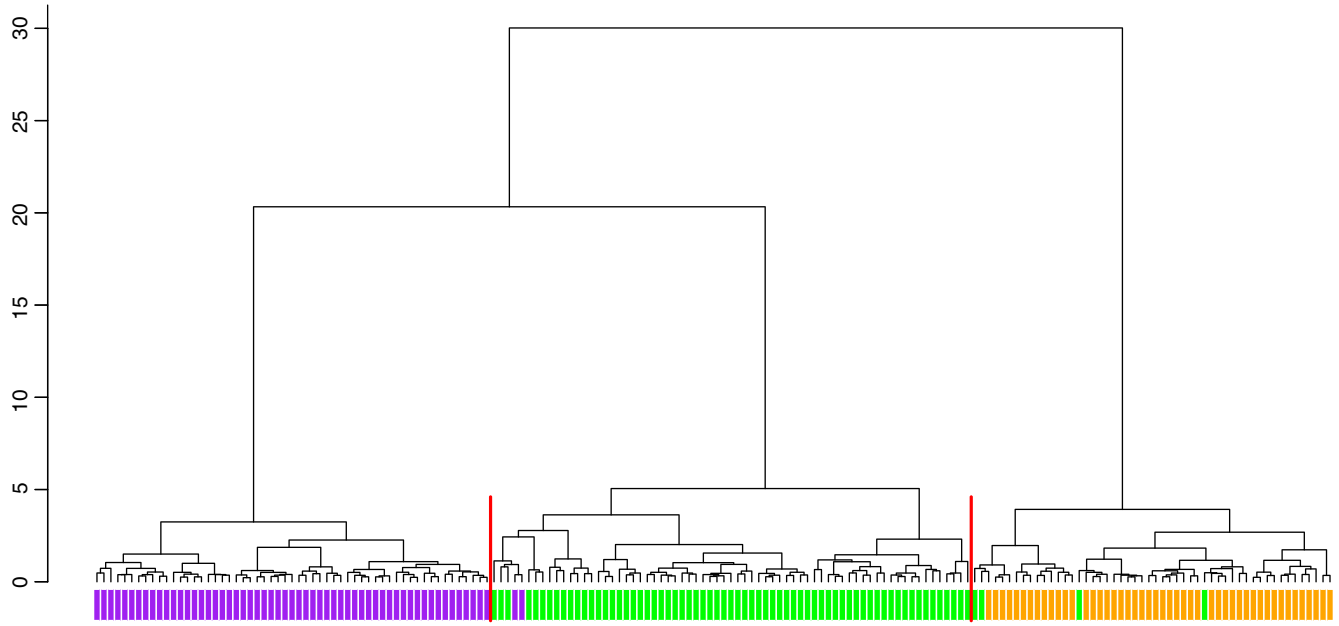


Fig. 2 Top eight models according to AIC. Edges connect models that differ by exactly one predictor. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

and should serve as motivation for those of us hoping to advance research in this area.

REFERENCES

[1] C.Hurley and R. Oldford, Eulerian tour algorithms for data visualization and the PairViz package, Comput Stat 26 (2011a), 613–633.
[2] C. Hurley and R. Oldford, PairViz: Visualization using Eulerian tours and Hamiltonian decompositions, R package version 1.2.1, 2011b.
[3] A. R. Waddell and R. W. Oldford, RnavGraph: Using Graphs as a Navigational Infrastructure, r package version 0.1.8, 2014.
[4] R. Oldford and A. Wadell, Visual Clustering of high-dimensional data by navigating low-dimensional spaces, In Proceedings of ISI, Dublin, Ireland, 2011.
[5] D.Earle and C. B. Hurley, Advances in dendrogram seriation for application to visualization, J Comput Graph Stat 24 (2015), 1–25.
[6] C. B. Hurley and D. Earle, DendSer: Dendrogram seriation, R package version 1.0.1, 2013.
[7] M. Hahsler, K. Hornik, and C. Buchta, Getting things in order: an introduction to the R package seriation, J Stat Softw 25 (2008), 1–34.
[8] A. Jalali, A. Alvarez-Iglesias, and J. Newell, DynNom: a dynamic nomogram for linear and generalized linear models as shiny applications, r package version 1.0.1, 2015.
[9] W. Chang, J. Cheng, J. Allaire, Y. Xie, and J. McPherson, shiny: Web Application Framework for R, r package version 0.11.1, 2015.