Scagnostics

Katrin Grimm[1,⋆], **Antony Unwin**[1]

1. University of Augsburg
⋆Contact author: katrin-grimm@web.de

Keywords: Visualization, Scagnostics, Scatterplots

The neologism scagnostics is derived from the words scatterplots and diagnostics. The term was first mentioned by John and Paul Tukey in the middle of the nineteen-eighties. Scagnostics are measures for characterising scatterplots and identifying different features. The aim is to give an initial overview of unknown datasets with a large number of variables. Concrete measures were proposed by Wilkinson et al. and implemented in the *R* package **scagnostics**. The package assumes that simplifications are necessary in order to make the measures applicable for large datasets and it uses hexagonal binning to speed up calculations. For eight of the nine measures from Wilkinson et al. the graphs' minimum spanning trees, complex hulls and alpha hulls are used as basis for the calculations. This can lead to some imprecision, as the measures are limited by the utilized graphs. Some alternative concepts and measures will be presented with the focus on finding interesting scatterplot structures quickly. Although computers continually become faster, the calculation of two-dimensional measures is still computationally intensive and methods for reducing calculation time are important.

An obvious idea to reduce the computing time is to exclude discrete variables. Additionally, variables with significant anomalies in 1-D — for example multimodal or skew variables — can also be left out of the calculation of two-dimensional measures, as a scatterplot in two dimensions will almost always be dominated by an anomaly in one dimension.

Another important question in the context of computer analyzed graphics is, how a good selection of graphics can be found. In this case Wilkinsons idea was to present scatterplots which are significant different from the others (*Outliers*) on the one hand and to do a clustering of the measures on the other hand. Each cluster stands for a group of similar scatterplots and it is sufficient to look at one representative from each cluster (*Exemplars*). Although this approach is optimal in theory, there are difficulties in practice which can be explained by the measures themselves. We propose a different approach for automatically selecting scatterplots using the measures.

The talk includes alternative criteria for functional dependencies — based on splines and distance correlation — and a presentation of an implementation of some of these ideas in *R*, using a German election and demographic dataset with 70 different variables.

References

[1] Szkély, G. J., M. L. Rizzo, and N. K. Bakirov (2007). Measuring and testing dependence by correlation of distances. *The Annals of Statistics 35*(6), 2769–2794.

[2] Wilkinson, L., A. Anand, and R. Grossman (2005). Graph-theoretic scagnostics. *Proceedings of the 2005 IEEE Symposium on Information Visualization*, 157–164.