Group beats Trend!?
Testing feature hierarchy in statistical graphics

Susan VanderPlas, Heike Hofmann[*]

January 19, 2015

Abstract

abstract goes here

1 Introduction and background

Intro to lineups (Buja et al., 2009; Majumder et al., 2013; Wickham et al., 2010; Hofmann et al., 2012)

The change to lineups we make is to introduce a second target to each lineup. We then keep track of how many observers choose any one of the two targets (to assess the difficulty of a lineup), and additionally we record how often observers choose one target over the other one. This is information that we can use to evaluate how strong the signal of one target is compared to the other one.

A further extension of this testing framework are the use of color (in a qualitative color scheme), the use of shapes, and additional density lines - we anticipate that all of these features are going to emphasize the clustering component. On the other hand, regression lines should emphasize any linear trends in the data.

2 Design Choices

Perceptual kernels (Çağatay Demiralp et al., 2014)

3 Generating Model

We are working with two models M_C and M_T to generate data for the target plots. The null plots are showing data generate from a mixture model M_0. We made sure that data from the clustering model M_C shares the same correlation with the null data, while data from model M_T exhibits a similar amount of clustering as the null data.

[*] Department of Statistics and Statistical Laboratory, Iowa State University

3.1 Cluster Model

Explored options:

1. Generate cluster centers along a line, then generate points around the cluster center.
   Algorithm:
   Parameters N points, K clusters, q cluster cohesion, s cluster std. dev.

   (a) Generate cluster centers $(c_i^x, c_i^y), i = 1, ..., K$:

   i. $c_i^x = (i-1) + e_x,\ e_x \sim Unif(-0.2, 0.2)$

   ii. $c^z \sim Unif(-q * K, q * K)$
   $c_i^y = (c_i^z - \overline{c^z})/\sigma_{c^z} * q * K$

   (b) Determine groups: $g \sim Multinomial(K, p)$ where $p = p_1/\sum_1^K p_{1i}$ where $p_{1i} \sim N(\frac{1}{K}, \frac{1}{2K^2})$

   (c) Generate points around cluster centers:

   i. $x_i^* = c_{g_i}^x + e,\ e_i \sim N(0, s)$

   ii. $y_i^* = c_{g_i}^y + e,\ e_i \sim N(0, s)$

   (d) Scale points

   i. $x_i = (x_i^* - \overline{x^*})/sd_{x^*}$

   ii. $y_i = (y_i^* - \overline{y^*})/sd_{y^*}$

   Advantages:

   (a) Easy to manipulate underlying trend

   (b) Cluster variance/cohesion can be easily manipulated

   Disadvantages:

   (a) Works poorly for more than 3 groups

   (b) Difficult to easily specify cluster distribution along regression line while guaranteeing same underlying regression line

   (c) Cluster variance (away from line) doesn't show up well - not enough clusters to ensure similar overall variance compared to null plots

2. Generate points in K dimensions using random noise, then use LDA to get K clusters
   Advantages:

   (a) Clusters are separated in space

   (b) Variance is fairly easy to manipulate

   Disadvantages:

   (a) Difficult to translate to linear regression because of cluster distribution

   (b) Clusters are fairly easily identifiable

References

Buja, A., Cook, D., Hofmann, H., Lawrence, M., Lee, E. K., Swayne, D. F., and Wickham, H. (2009), "Statistical inference for exploratory data analysis and model diagnostics," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 367, 4361–4383.

Çağatay Demiralp, Bernstein, M., and Heer, J. (2014), "Learning Perceptual Kernels for Visualization Design," *IEEE Trans. Visualization & Comp. Graphics (Proc. InfoVis)*.

Hofmann, H., Follett, L., Majumder, M., and Cook, D. (2012), "Graphical Tests for Power Comparison of Competing Designs," *IEEE Transactions on Visualization and Computer Graphics (Proc. InfoVis)*, 18, 2441–2448, 25% acceptance rate.

Majumder, M., Hofmann, H., and Cook, D. (2013), "Validation of Visual Statistical Inference, Applied to Linear Models," *Journal of the American Statistical Association*, 108, 942–956.

Wickham, H., Cook, D., Hofmann, H., and Buja, A. (2010), "Graphical inference for infovis," *IEEE Transactions on Visualization and Computer Graphics (Proc. InfoVis)*, 16, 973–979, 26% acceptance rate. Best paper award.