Group beats Trend!?
Testing feature hierarchy in statistical graphics

Susan VanderPlas, Heike Hofmann*

January 25, 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. Both models generate data in the same range of values. We made also 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 M_C

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

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

   i. Generate vectors c^x and c^y as permutations of $\{1, ..., K\}$,

   ii. such that the correlation between cluster centers $\text{Cor}(c^x, c^y)$ falls into a range of [.25, .9].

   We might have to go up with the correlation a bit. I'm still worried that people will pick the cluster plot from the trend line lineup because of the lowest slope.

   (b) Center and standard-normalize cluster centers (c^x, c^y):

   $$\tilde{c}_i^x = \frac{c_i^x - \bar{c}}{s_c} \quad \text{and} \quad \tilde{c}_i^y = \frac{c_i^y - \bar{c}}{s_c},$$

   where $\bar{c} = K(K+1)/2$ and $s_c^2 = \frac{K(K+1)(2K+1)}{6} - \frac{K^2(K+1)^2}{4}$ for all $i = 1, ..., K$.

   (c) Determine group size g_i for groups $i = 1, ..., K$ as a random draw $g_i \sim \text{Multinomial}(K, p)$ where $p = p_1 / \sum_{i=1}^K p_{1i}$ for $p_{1i} \sim N(\frac{1}{K}, \frac{1}{2K^2})$.

   (d) Generate points around cluster centers:

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

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

   It may be reasonable to draw q from a distribution of some sort.

   Let's not worry about getting q from a random distribution, but let's rename it somehow, so it reflects the within cluster deviation a bit better ... σ_C?

3.2 Regression Model M_T

This model has the parameter σ_T to reflect the amount of scatter around the trend line.

3.3 Null Model M_0

The generative model for null data is created as a mixture model M_0 that draws $n_c \sim B_{N,\lambda}$ observations from the cluster model, and $n_T = N - n_c$ from the regression model M_T.

4 Experimental Setup

I would consider the values $\sigma_C = 0.3, .35, .4, .45$ for $K = 3$ clusters to be interesting. The actual values of σ_C don't make much sense - because they are only valid within the scaled data values. We might need to re-express the values of σ_C in terms of a percentage of the data or a percentage of the overall variability.

For $K = 5$ the parameters for q (now σ_C) and the standard deviation σ_T need to be smaller - we could start at 0.2 and 0.75, respectively.

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.