Group beats Trend!?
Testing feature hierarchy in statistical graphics

Susan VanderPlas, Heike Hofmann*

January 31, 2015

Abstract

abstract goes here

1 Introduction and background

Discussion of preattentive visual features (Healey and Enns, 2012)

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

We choose colors and shapes for the lineups in our study to be the most different from a set of ten choices as evaluated by participants in the study by Çağatay Demiralp et al. (2014) on the so called perceptual kernels.

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

We compute the correlation coefficient for all of the plots to assess the amount of linearity in each panel. As a measure of clustering, we can use the F statistic of between versus within group variation.

3.1 Cluster Model M_C

We begin by generating cluster centers along a line, then we generate points around the cluster center.
Algorithm:
Parameters N points, K clusters, σ_C cluster standard deviation

1. Generate cluster centers (c_i^x, c_i^y) for each of the K clusters, $i = 1, ..., K$:
   (a) Generate vectors c^x and c^y as permutations of $\{1, ..., K\}$,
   (b) 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.

2. 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 $i = 1, ..., K$.

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

4. Generate points around cluster centers:
   (a) $x_i^* = c_{g_i}^x + e$, $e_i \sim N(0, \sigma_C^2)$
   (b) $y_i^* = c_{g_i}^y + e$, $e_i \sim N(0, \sigma_C^2)$

3.2 Regression Model M_T

This model has the parameter σ_T to reflect the amount of scatter around the trend line.
Algorithm:
Parameters N points, σ_T standard deviation around the line, slope a (1 by default)

1. Generate x_i, $i = 1, ..., N$, a sequence of evenly spaced points from $[-1, 1]$ (σ_T added and subtracted to match the range of cluster points in x)

2. Jitter x_i: $x_i = x_i + \eta_i$, $\eta_i \sim Unif(-z, z)$, $z = 1/5 * (2/(N-1)$

3. Generate y_i: $y_i = a * x_i + e_i$, $e_i \sim N(0, \sigma_T^2)$

Would the pictures change dramatically, if you used $x \sim U[-1, 1]$ to start out with? that would be easier to explain.

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

4.1 Design

Factors:

Parameter	Description	Choices
N	# Points	45, 75
K	# Clusters	3, 5
σ_T	Scatter around trend line	.35, .45, .55
σ_C	Scatter around cluster centers	($K = 3$) .25, .35, .45 ($K = 5$) .2, .3, .4

Table 1: Data Generation Options

<table>
<tr><th>Emphasis</th><th>Aesthetics</th></tr>
<tr><td>Control</td><td>–</td></tr>
<tr><td>Group</td><td>Color, Shape, Ellipse
Color + Shape, Color + Ellipse</td></tr>
<tr><td>Trend</td><td>Line, Error band
Line + Error band</td></tr>
<tr><td>Conflict</td><td>Color + Trend Line, Color + Trend Line + Error band</td></tr>
</table>

Table 2: Plot Generation Options

HH: I would think of ellipses as the analogue to the error bands in trends - so rather not show them alone. I'm not so sure that we need both conflict situations - the trend lines against color are already bad enough - if we additionally have an error band we get into problems with the band on top of the color.
SVP: I fixed this by changing to lines instead of ribbons. I think it makes it better.

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.

ugh, that sounds like a pain.

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

Why would σ_T change? It shouldn't matter for changes in K.

Design choices

1. Plain: two targets with data from one of each of the two generative models are included in a set of eighteen panels of null data.
2. Color/Shape: points in each of the panels are colored/marked based on the results of a hierarchical clustering .
3. Trend line: a line of the least square fit is drawn through the points.
4. Color & Shape
5. Color & trend line: this emphasises both the clustering and the regression - it is not clear, which signal will be stronger.
6. Color & Ellipsoids: around the groups of the same color, ellipsoids are drawn to reflect the 95% density estimate.

4.2 Hypothesis

The plot most identified as the ”target” will change based on plot aesthetics which emphasize linear features or cluster features. This effect will be mediated by the signal strength of the line and cluster features.

- Increasing N will increase signal strength for both line and clusters
- Increasing K will decrease signal strength for clusters (fewer points per cluster, thus lower visual cohesiveness)
- Increasing σ_T will decrease signal strength for lines
- Increasing σ_C will decrease signal strength for clusters

Plot features will emphasize either lines or clusters as follows:

- None (control)
- Color (cluster emphasis)
- Shape (cluster emphasis)
- Color + shape (double cluster emphasis)
- Ellipse + color (doubly cluster emphasis)
- Line
- Line + Prediction Interval (double line emphasis)
- Color + line (conflict)
- Color + line + Prediction Interval (conflict)

In a more organized representation:

		Line Emphasis		
		0	1	2
Cluster Emphasis	0	None	Line	Line + Prediction
	1	Color, Shape	Color + Line	
	2	Color + Shape Color + Ellipse		Color + Ellipse + Line + Prediction
	3	Color + Shape + Ellipse		

4.3 Experimental Design

Starting with the assumption of a fully factorial, balanced design, with M datasets per parameter set (replicates) and P evaluations per (aesthetic|dataset)

If we then do one replicate of $K = 5$ and one of $N = 75$ (reducing the full factorial experiment) we should be able to make broad generalizations about the effect of K and N without a fully factorial design. As we don't care about all of the interactions, we can add many of those terms into the error term as well.

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)*.

Healey, C. G. and Enns, J. (2012), "Attention and Visual Memory in Visualization and Computer Graphics," *IEEE Transactions on Visualization and Computer Graphics*, 18, 1170–1188.

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.

Level	Factor	Source	DF	SS
Dataset	N	α	1	
	K	β	1	
	σ_T^2	γ	3	
	σ_C^2	δ	4	
		$(\alpha\beta)$	1	
		$(\alpha\gamma)$	3	
		$(\alpha\delta)$	4	
		$(\beta\gamma)$	3	
		$(\beta\delta)$	4	
		$(\gamma\delta)$	12	
		$(\alpha\beta\gamma)$	3	
		$(\alpha\beta\delta)$	4	
		$(\beta\gamma\delta)$	12	
		$(\alpha\gamma\delta)$	12	
		$(\alpha\beta\gamma\delta)$	12	
		Dataset Error	(M-1)(2)(2)(4)(5)	
Plot	Aesthetic	τ	10	
		color	1	
		shape	1	
		line	1	
		color + shape	1	
		color + ellipse	1	
		color + shape + ellipse	1	
		line + pred. interval	1	
		color + line	1	
		color + ellipse + line + pred. interval	1	
	Data x Aesthetics		(10)((2)(2)(4)(5)-1)	
	error		$(P-1)(11)(M)(2)(2)(4)(5)$	
	Total		I think $P(11)(M)(2)(2)(4)(5)-1$	