Group beats Trend!?
Testing feature hierarchy in statistical graphics

Susan VanderPlas, Heike Hofmann*

February 3, 2015

Abstract

abstract goes here

1 Introduction and background

Discussion of pre-attentive visual features (Healey and Enns, 2012) - with a focus on hierarchy of pre-attentive features: color trumps shape - do we also see this in our results, and if so, by how much?

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. Unfortunately, this limits the choice to the set used in the Tableau software. In order to produce experimental stimuli accessible to the approximately 4% of the population with red-green colorblindnessGegenfurtner and Sharpe (2001), we removed the grey hue from the palette. This modification produced maximally different color combinations which did not include red-green combinations, while also removing a color (grey) which is difficult to distinguish for those with color deficiency.

*Department of Statistics and Statistical Laboratory, Iowa State University

we didn't exclude red-green combinations; by excluding grey (which is a horrible color and similar to red, green, turquoise, etc. for colorblind people) we managed to produce palettes which were maximally different and didn't include red/green. We can formally exclude red/green if you'd like.

Shapes - there were some problems with reliable unicode representations in the lineups, which led to using slightly modified shapes. The left and right triangle shapes (available only in unicode using R) were excluded due to size differences between unicode and non-unicode shapes.

All shapes are non-filled shapes, which means that they are consistent with one of the simplest solutions to overplotting of points in the tradition of Tukey (1977) and Few (2009). For this reason we abstained from the additional use of alpha-blending of points to measure the amount of overplotting.

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.

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, .75].
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.
Yes - the issue is the variability in the range of x, because normalization would make certain plots with lower range more "squishy".

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
K	# Clusters	3, 5
N	# Points	$15 \cdot K$
σ_T	Scatter around trend line	.3, .4, .5
σ_C	Scatter around cluster centers	.20, .25, 0.30, .35, 0.40

Table 1: Parameter settings for Data Generation.

I'm going forth and back on the parameters for σ_C and σ_T, but I think I really like the ones in the table. For the trend line, we defintile get something along the lines 'easy', 'medium', and 'hard'; for the clustering we should be aiming for the same - it's not quite as clear cut, because there seems to be more variability in the results, so we need to double check the randomly generated results.

Emphasis	Aesthetics
Control	–
Group	Color, Shape
	Color + Shape, Color + Ellipse, Color + Shape + Ellipse
Trend	Line
	Line + Error band
Conflict	Color + Trend Line,
	Color + Ellipse + Trend Line + Error band

Table 2: Aesthetics and add-on design choices.

We can use the null model to get a distribution for each of the two quality measures for the targets, which gives us an objective measure to assess the difficulty of detecting each of the targets.

Figures 1 and 2 show densities ... The red lines show ten samples each from the trend model and the cluster model. The lines for the cluster are, relatively, further to the right of the overall distribution than the red lines for the trend model, indicating that $\sigma_C = 0.25$ is producing target plots that are a bit easier to spot thnt trend targets with a parameter value of $\sigma_T = 0.30$.

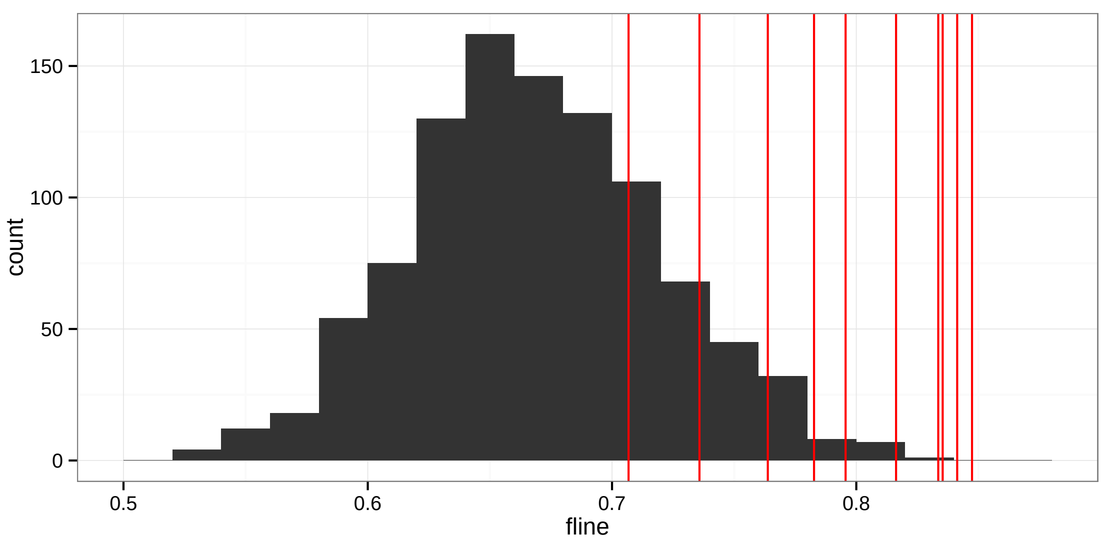


Figure 1: Histogram of R^2 values from 1000 data sets of the null model ($K = 3, N = 45, \sigma_C =$0.25, $\sigma_T =$0.30). The lines in red are R^2 values of ten sample data sets from the Trend model M_T.

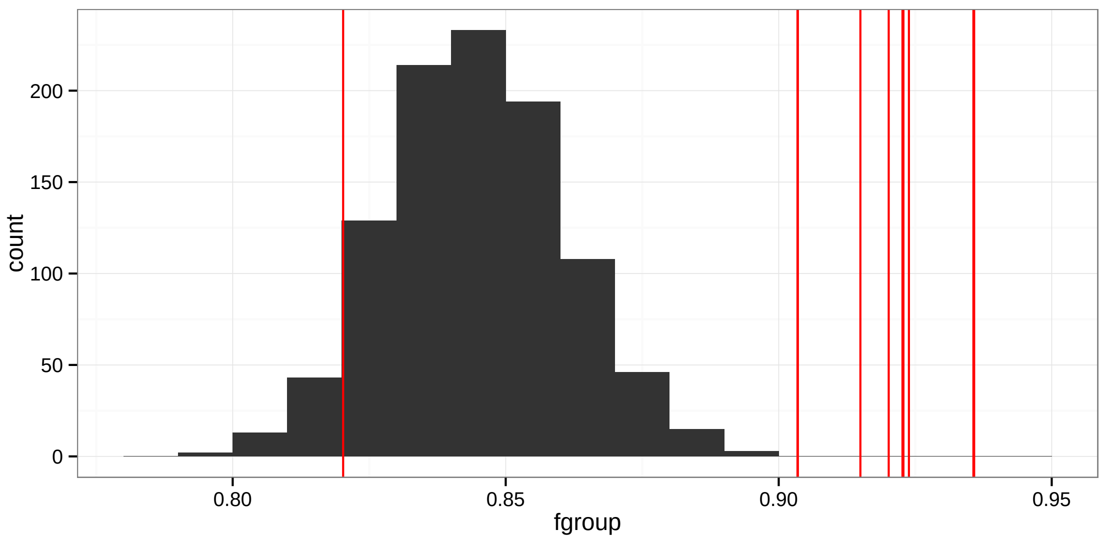


Figure 2: Histogram of F-statistics measuring the amount of clustering from 1000 data sets of the null model ($K = 3, N = 45, \sigma_C$ =0.25, σ_T =0.30). The lines in red are the F-statistics of ten sample data sets from the Clustering model M_C.

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

Initially, assume a fully factorial, balanced design, with r unique datasets per parameter set (replicates) and P evaluations per (aesthetic|dataset). The experiment is conducted at three levels: parameter sets (with replication, so EUs are data sets), plot types (i.e. a certain set of aesthetics), and participant evaluations. At the first level, there are three parameters: $K \in \{3, 5\}$, $\sigma_T \in \{.3, .4, .5, .6\}$, and $\sigma_C \in \{.2, .25, .3, .35, .4\}$. At the second level, there are blocks (by data set), and then 10 aesthetic combinations.

We'll have to use contrasts to measure the effect of color individually, etc., for now let's just consider the ANOVA evaluation

Finally, at the lowest level, there are participant effects.

At the participant level, we need to decide if we're going to fully randomize, try to block, etc. - are participants going to get 10 different data sets? 5? Not sure how to conceptualize that, and I would imagine it will affect how we organize model evaluation. Grr, I hate mixed models.

Modified from Table 10.6 (pg 181) of Design of Experiments by Dr. Morris. The table in the book has a four-factor split plot design with three levels (randomized, block, block).

Level	Factor	Source	DF	Sum of Squares
Dataset	K	α	1	$\sum_i (4)(5)(r)(10P)(\overline{y}_{i.....} - \overline{y}_{......})^2$
	σ^2_T	β	3	$\sum_j (2)(5)(r)(10P)(\overline{y}_{.j....} - \overline{y}_{......})^2$
	σ^2_C	γ	4	$\sum_i (2)(4)(r)(10P)(\overline{y}_{..k...} - \overline{y}_{......})^2$
		$(\alpha\beta)$	3	$\sum_{ij} (5)(r)(10P)(\overline{y}_{ij....} - \overline{y}_{i.....} - \overline{y}_{.j....} + \overline{y}_{......})^2$
		$(\alpha\gamma)$	4	$\sum_{ik} (4)(r)(10P)(\overline{y}_{i.k...} - \overline{y}_{i.....} - \overline{y}_{..k...} + \overline{y}_{......})^2$
		$(\beta\gamma)$	12	$\sum_{jk} (2)(r)(10P)(\overline{y}_{.jk...} - \overline{y}_{.j....} - \overline{y}_{..k...} + \overline{y}_{......})^2$
		$(\alpha\beta\gamma)$	3	$\sum_{ijk} (r)(10P)(\overline{y}_{ijk...} - \overline{y}_{i.....} - \overline{y}_{.j....} - \overline{y}_{..k...}$ $+\overline{y}_{ij....} + \overline{y}_{.jk...} + \overline{y}_{i.k...} - \overline{y}_{......})^2$
		Resid.	$(2)(4)(5)(r-1)$	$\sum_{ijkl} (10P)(\overline{y}_{ijkl..} - \overline{y}_{i.....} - \overline{y}_{.j....} - \overline{y}_{..k...} + \overline{y}_{ij....} + \overline{y}_{.jk...}$ $+\overline{y}_{i.k...} - \overline{y}_{ijk...} - \overline{y}_{ij.l..} - \overline{y}_{i.kl...} - \overline{y}_{.jkl...} + \overline{y}_{......})^2$
	Total		$(2)(4)(5)(r) - 1$	$\sum_{ijkl} (10P)(\overline{y}_{ijkl..} - \overline{y}_{......})^2$
Plot	Dataset	blocks	$(2)(4)(5)(r) - 1$	$\sum_{ijkl} (10P)(\overline{y}_{ijkl..} - \overline{y}_{......})^2$
	Aes.	δ	9	$\sum_m (2)(4)(5)(P)(\overline{y}_{....m.} - \overline{y}_{......})^2$
	Aes x K	$(\alpha\delta)$	9	$\sum_{im} (4)(5)(P)(\overline{y}_{i...m.} - \overline{y}_{i.....} - \overline{y}_{....m.} + \overline{y}_{......})^2$
	Aes x σ_T	$(\beta\delta)$	27	$\sum_{jm} (2)(5)(P)(\overline{y}_{.j..m.} - \overline{y}_{.j....} - \overline{y}_{....m.} + \overline{y}_{......})^2$
	Aes x σ_C	$(\gamma\delta)$	36	$\sum_{km} (2)(4)(P)(\overline{y}_{..k.m.} - \overline{y}_{..k...} - \overline{y}_{....m.} + \overline{y}_{......})^2$
	Others		9(31)	difference
	Resid		40(rP-1)-(40r-1)	
	Total		$400r - 1$	$\sum_{ijklm} (P)(\overline{y}_{ijklm.} - \overline{y}_{ijkl..})^2$
Trial	Picture	Sub-blocks	$400r - 1$	$\sum_{ijklm} (P)(\overline{y}_{ijklm.} - \overline{y}_{ijkl..})^2$
	Participants	τ	$P - 1$	$\sum_n (2)(4)(5)(r)(10)(\overline{y}_{.....n} - \overline{y}_{......})^2$
	Resid		$(400r-1)(P)$	difference
	Total		$400(r)(P) - 1$	$\sum_{ijklmn} (y_{ijklmn} - \overline{y}_{......})^2$

Table 3: Evaluation of sources of error in a full factorial version of the experiment, with r replicates of each parameter combination and P participant evaluations of each plot(data/aesthetic combination).

We have a couple of options:

- keep the full factorial experiment, use one (at most two) replicates, and use higher level factorial effects to beef up any error variance terms.
- Do a full factorial experiment for $K = 3$ and use a subset of the factorial experiment for $K = 5$ (either using a subset of cases for σ_T and σ_C, or a subset of combinations of the two cases/fractional factorial.)

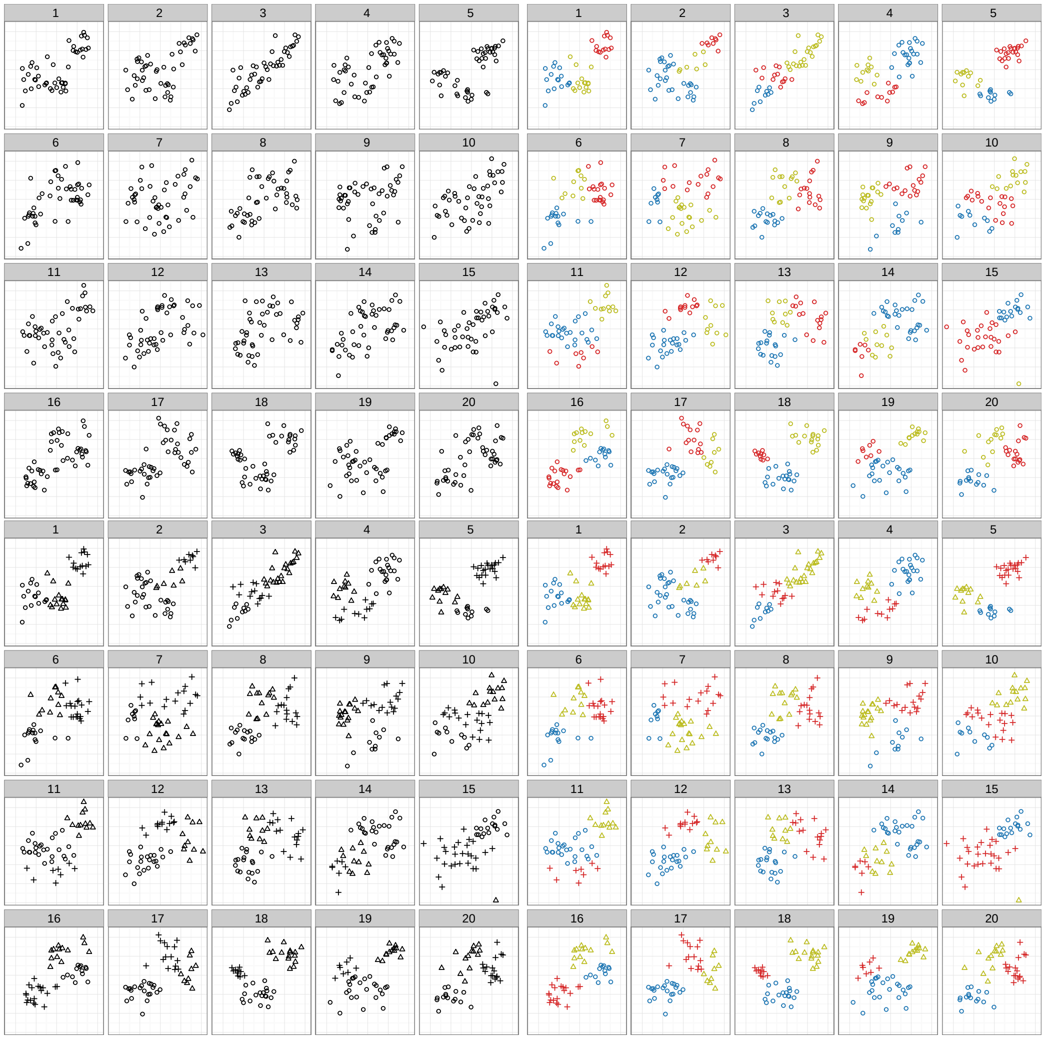
The fractional factorial option will be a pain to explain when we write things up; it will be simpler to explain using a subset of cases. Given that we don't particularly care about the third-order effects (and possibly not even the second-order effects) for the parameters, I'm inclined to say that the single-replicate option is the easiest way to go (and lets us keep the simple SSQ in the table, which is a huge bonus in my opinion). Even if we just use the third-order interaction effect as error, we still have 12 degrees of freedom; that should be plenty - we'd only need F=2.69 to get a significant result for even the $(\sigma_T\sigma_C)$ test.

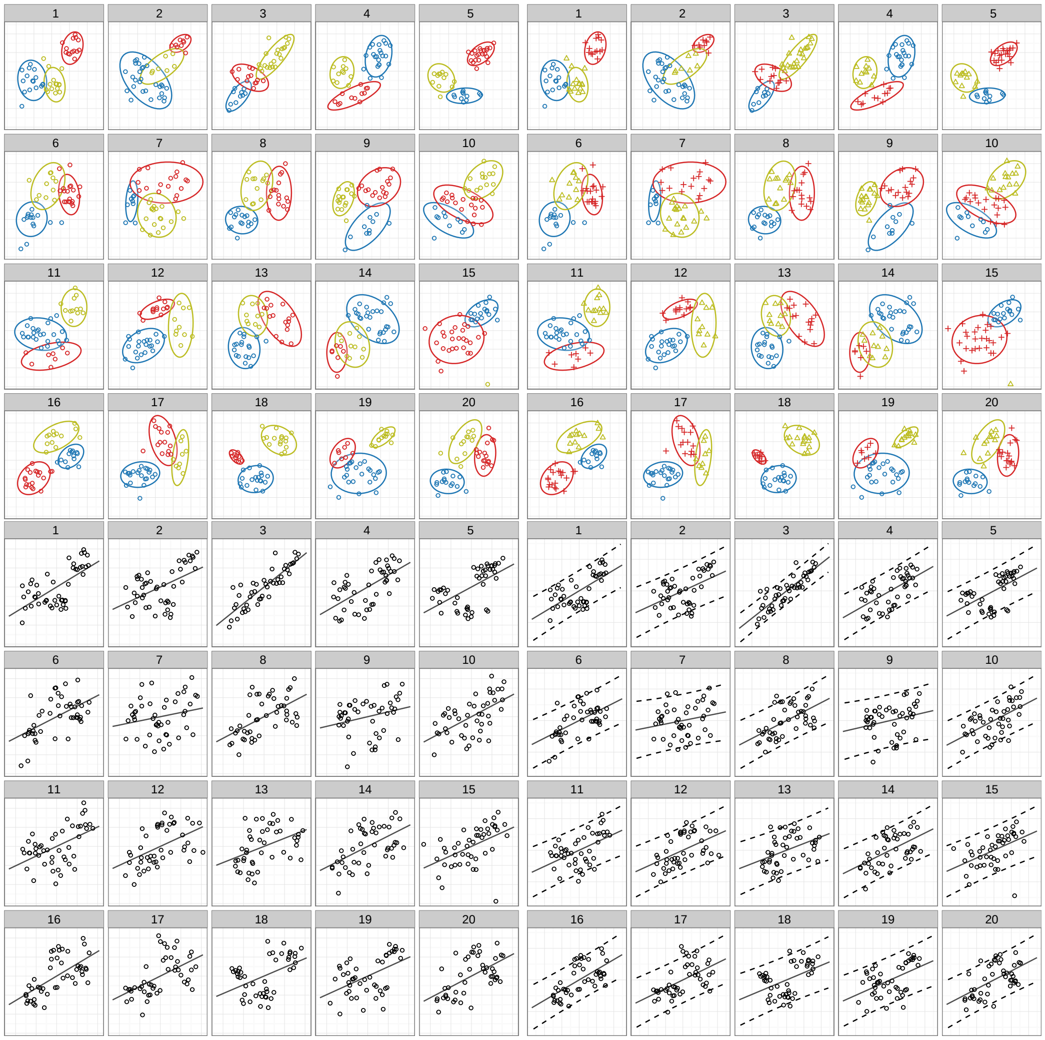
Table 4: ANOVA table - only one replicate. Evaluation of sources of error in a full factorial version of the experiment, with one replicate of each parameter combination and P participant evaluations of each plot(data/aesthetic combination).

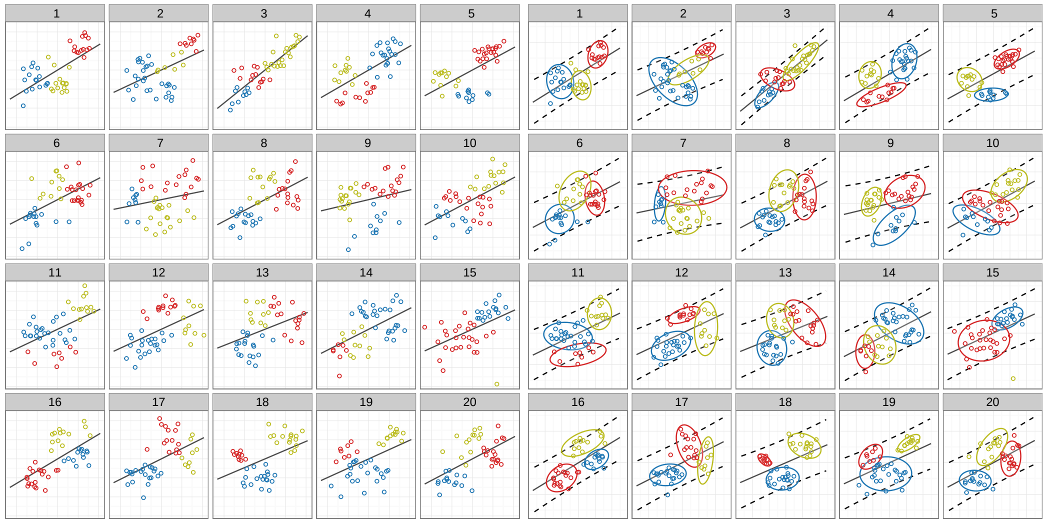
Level	Factor	Source	DF	Sum of Squares
Dataset	K	α	1	$\sum_i (4)(5)(10P)(\overline{y}_{i....} - \overline{y}_{.....})^2$
	σ_T^2	β	3	$\sum_j (2)(5)(10P)(\overline{y}_{.j...} - \overline{y}_{.....})^2$
	σ_C^2	γ	4	$\sum_i (2)(4)(10P)(\overline{y}_{..k..} - \overline{y}_{.....})^2$
		Resid.	22	difference
	Total		39	$\sum_{ijk} (10P)(\overline{y}_{ijk..} - \overline{y}_{.....})^2$
Plot	Dataset	blocks	39	$\sum_{ijk} (10P)(\overline{y}_{ijk..} - \overline{y}_{.....})^2$
	Aes.	δ	9	$\sum_m (2)(4)(5)(P)(\overline{y}_{...m.} - \overline{y}_{.....})^2$
	Aes x K	$(\alpha\delta)$	9	$\sum_{im} (4)(5)(P)(\overline{y}_{i..m.} - \overline{y}_{i....} - \overline{y}_{...m.} + \overline{y}_{.....})^2$
	Aes x σ_T	$(\beta\delta)$	27	$\sum_{jm} (2)(5)(P)(\overline{y}_{.j.m.} - \overline{y}_{.j...} - \overline{y}_{...m.} + \overline{y}_{.....})^2$
	Aes x σ_C	$(\gamma\delta)$	36	$\sum_{km} (2)(4)(P)(\overline{y}_{..km.} - \overline{y}_{..k..} - \overline{y}_{...m.} + \overline{y}_{.....})^2$
	Resid		9(31)	difference
	Total		399	$\sum_{ijkm} (P)(\overline{y}_{ijkm.} - \overline{y}_{ijk..})^2$
Trial	Picture	Sub-blocks	399	$\sum_{ijkm} (P)(\overline{y}_{ijkm.} - \overline{y}_{ijk..})^2$
	Participants	τ	$P-1$	$\sum_n (2)(4)(5)(10)(\overline{y}_{....n} - \overline{y}_{.....})^2$
	Resid		$399(P-1)$	difference
	Total		$400P-1$	$\sum_{ijkmn} (y_{ijkmn} - \overline{y}_{....})^2$

4.4 Sample Pictures

The following plots use $\sigma_T = .4$ and $\sigma_C = .3$.




1 2 3 4 5 1 2 3 4 5
6 7 8 9 10 6 7 8 9 10
11 12 13 14 15 11 12 13 14 15
16 17 18 19 20 16 17 18 19 20
1 2 3 4 5 1 2 3 4 5
6 7 8 9 10 6 7 8 9 10
11 12 13 14 15 11 12 13 14 15
16 17 18 19 20 16 17 18 19 20



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)*.

Few, S. (2009), *Now You See It: Simple Visualization Techniques for Quantitative Analysis*, Burlingame, CA: Analytics Press, 1st ed.

Gegenfurtner, K. R. and Sharpe, L. T. (2001), *Color vision: From genes to perception*, Cambridge University Press.

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.

Tukey, J. W. (1977), *Exploratory Data Analysis*, Lebanon, IN: Addison Wesley.

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.