# Measuring Lineup Difficulty By Matching Distance Metrics with Subject Choices with Crowd-Sourced Data

**Abstract**

Revisit: Graphics play a crucial role in statistical analysis and data mining. Being able to quantify structure in data that is visible in plots, and how people read the structure from plots is an ongoing challenge. This paper describes metrics developed to assist with this, and utilizes them to evaluate how people select plots from lineups used for visual inference. Lineups embed the plot of the data among a set of null plots, and engage a human observer determine whether the data plot is different from the nulls. The data plot is treated like a test statistic, and protocol acts like a comparison with the sampling distribution of the nulls. Metrics are calculated in association with lineups to help understand what people see in the data plots, and assess the quality of a lineup, because the null plots represent a finite sample from a null distribution, and this finiteness may affect how observers read the lineups. The distance metrics are designed to describe how different the data plot is from the null plots, and the null plots from each other. Analysis of the distance metrics was conducted on data collected through a large number of Amazon Turk studies that have used the lineup protocol for studying an array of data analysis tasks.

*Keywords:* data visualization, statistical graphics, data mining, data science, information visualization

things to do:

- reduce paper to 30 pages + appendix (from 40 pages + 3 pages of appendix): strategy: explain first example in detail, include only conclusions for the other two examples, put supporting material into the appendix Di suggests the intro could be reduced, remove the Tukey quote, tighten up the distance descriptions but keep the plots, maybe fig 7, fig 14 could be moved to appendix Keep emphasis on sampling distribution, and re-state this in conclusions Explain the Turk studies in the intro DONE

- restructure introduction: DONE include findings from previous studies, such as permutation-invariance; include description of how data is collected; MAYBE NEEDS TO HAVE THIS, COULD BE IN APPENDIX

- be more precise in lineup description: null hypotheses

- add citations (Marron and Tsybakov, 1995) and (Hannig et al., 2013) in the text - pbly conclusions because these studies could benefit from the lineup protocol to vouch that the error metrics would be better than MSE for getting model fit to look more like we'd expect by eye DONE

- title change DONE

- conclusions don't match results, the specialist metric does not always outperform the binned DONE

# 1　Introduction

The lineup protocol was introduced by Buja et al. (2009) to bring data visualisation into the formal statistical inference framework. The fundamental idea is that a data plot is a test statistic, for a null hypothesis that is often implicitly specified by the choice of plot. Crowd-sourcing can be used to evaluate how far the data plot (test statistic) is from null plots. If the data plot is discoverable it is evidence against the null hypothesis. It would seem obvious that metrics might be also applied to comparing differences between plots that eventually may be used to supplant the crowd-sourcing, or to understand how the human visual system processes data plots. This paper explores the relationship between some common metrics for images with what subjects selected in several large Amazon Turk (MTurk) (Amazon, 2005-) crowd-sourcing studies that utilised the lineup protocol.

Figure 1 shows a lineup from a study described in Majumder et al. (2013). It is based on simulated data to examine the similarity of results derived from the lineup protocol and those derived from a classical $t$-test. Pseudo-data plots were generated with specific structure in a linear model, controlling $\beta_k, \sigma, n$ (slope, variation and sample size). And null plots were generated from a null hypothesis $\beta_k = 0$. In this field of plots one is a pseudo-data plot, $\beta_k \neq 0$, and the remaining 19 are null plots. Subjects were asked a very specific question in this study – "Which plot has the steepest slope?" – in order to make comparisons with the classical test. (In practice, when using the lineup protocol, subjects are generically asked to select the plot that is most different from the others, so that the value of graphics to discover the unexpected can be utilized.) The pseudo-data plot in this lineup is #2. In the MTurk study, 66 out of 70 subjects selected this plot. This would produce a (visual) $p$-value of 0, leading to the conclusion that the null hypothesis should be rejected.

Figure 2 illustrates the difference between the classical test, and the lineup protocol, for the example shown in Figure 1. In the lineup protocol, a finite set of draws from the sampling distribution is used for comparison, as opposed to the full distribution for the classical test. To be clear, in practice we do not know the sampling distribution – this is a special case, where the lineup was produced under a classical scenario, and we know where the pseudo-data plot and the null plots fall in regard to the known sampling distribution.
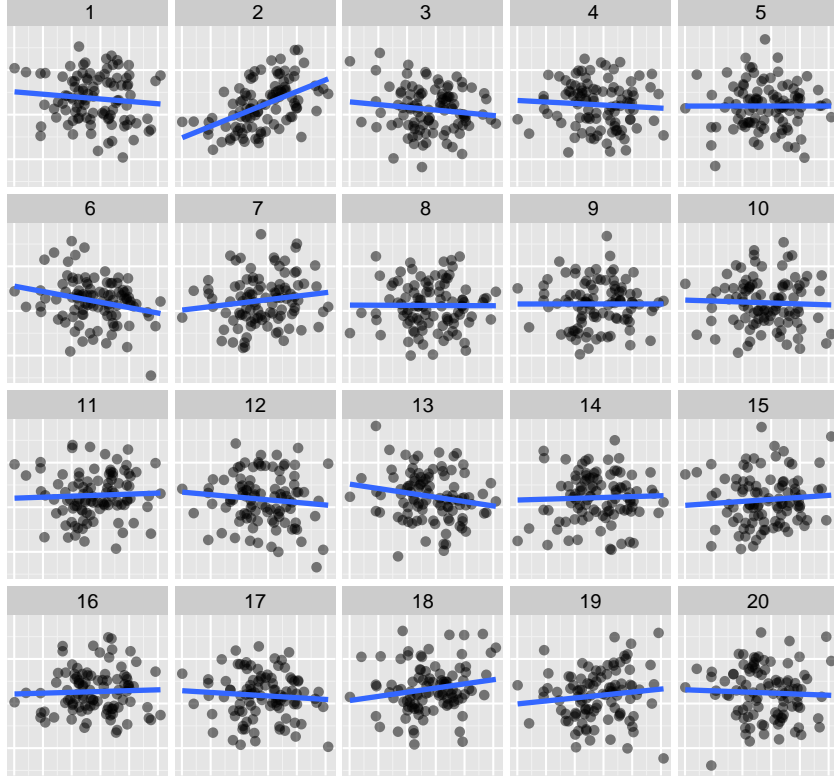
3

Figure 1: Example lineup plot of size $m = 20$ of scatterplots with overlaid regression line, fro a simulation study to compare the protocol with the classical $t-$test of $H_o : \beta_k = 0$, where covariate $X_k$ is continuous. One of the plots in the lineup is the plot of the pseudo-data, generated from a model where $\beta_k \neq 0$, and the others are null plots generated by simulating data from the null distribution $\beta_k = 0$. Which plot has the steepest slope?

Also, in practice with crowd-sourcing many lineups can be generated, so the data plot can be compared to more than 19 nulls. The left plot shows the classical testing setup, where the null hypothesis is rejected is the calculated test statistic falls in the shaded region. In contrast, with the lineup protocol, the null hypothesis is rejected in the human observers unqeuivocally identify the data plot. In this example, we would expect the observers would identify the data plot, marked by # 16, because it corresponds to as extreme value on the sampling distribution.

To date, more than 20 studies have been conducted utilizing the lineup protocol, primarily using MTurk. The first three experiments (described in Majumder et al. (2013))

4
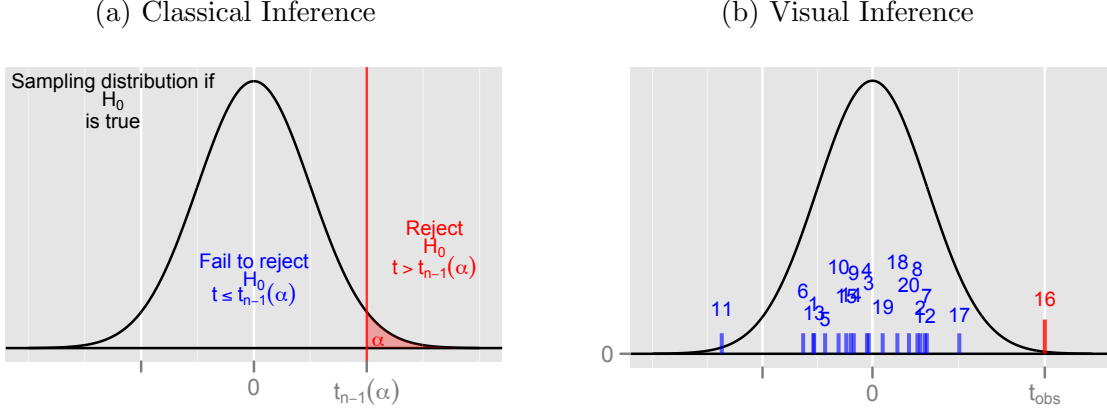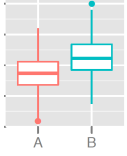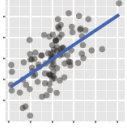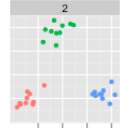
(a) Classical Inference    (b) Visual Inference

Figure 2: If the lineup protocol was to be used instead of classical inference this is what it would look like. (a) Rejection region (shaded in red) for classical inference for $H_0 : \mu = \mu_0$ vs $H_a : \mu > \mu_0$ and (b) values corresponding to the true value (red) and the null plots (blue) in a single lineup of size $m = 20$ that would be used to test the same null hypothesis. The actual data plot is extreme relative to the null plots, and observers would likely be able to pick it out, resulting in a decision to reject the null hypothesis. In practice, the lineup protocol would not be used if a classical test can be used.

compared the protocol to classical tests associated with fitting linear models. A later experiment examined the use of the protocol for detecting clusters in high-dimensional data (Roy Chowdhury et al., 2015). Each of these experiments involved a carefully controlled study where structure in plots was generated using simulation, and multiple replicates for each treatment were used. Subjects participating in the studies were experienced MTurk workers, and evaluated blocks of ten lineups. The blocks were chosen using stratified sampling. For each lineup subjects selected a plot from the lineup that they felt best matched the question asked, rated how sure they had found the best match, and why they decided on their choice. (Full details of the experimental setups and data collection are provided in the Appendix.) It is the data from these experiments (Table 1) that we now examine with a different purpose, to assess how distance metrics compare with human evaluation. All combined, this amounts to information from 1047 subjects examining 206 lineups. The questions arising from these studies are:

- Using distance metrics applied to data plots, can we get the computer choose like a human does?

- Do some distance metrics match how subjects choose plots better than others?

- Could distance metrics be used as a first pass in designing simulation experiments for the lineup protocol to group lineups into levels of "easy", "moderate", "difficult"?

Table 1: Overview of the three MTurk experiments, from which data was extracted to investigate distance metrics in relation to subject choices.

| ID | Experiment | Test Statistic | Lineup question |
|----|-----------|----------------|-----------------|
| I | Box plot |  | Which set of box plots shows biggest vertical difference between group A and B? (Majumder et al., 2013) |
| II | Scatter plot |  | Of the scatter plots below which one shows data that has steepest slope? (Majumder et al., 2013) |
| III | Group separation |  | Which of these plots has the most separation between the coloured groups? (Roy Chowdhury et al., 2015) |

The paper is organized as follows. Section 2 starts by defining distance measures and discussing different choices of measures. The distribution of the distance measures are studied in Section 2.2. Section 4.1 describes the effect of the plot type and the question of interest on the distance measure while Section 2.3 talks about the distance evaluations. Section 3 presents a comparison of the distance measures to the performance of human subjects in several experiments conducted with MTurk.

## 2  Distance Measures

The goal of the metrics is to measure the "distance" between two plots. In particular, we would like to know how similar or different is the data plot from the surrounding null plots.

A naive approach would be to use existing goodness-of-fit statistics that compare data with reference probability distributions, for example, the Kolmogorov-Smirnov (Stephens,

1974), Anderson-Darling (Stephens, 1974), Shapiro-Wilk (Shapiro and Wilk, 1965) tests and Bhattacharyya distance (Bhattacharyya, 1946). However, these measure differences between univariate distributions which limits their applicability to distances between plots, generally. Bhattacharyya distance is used for image processing but not in manner useful for plot comparisons. Hausdorff distance (Huttenlocher et al., 1993) has been successfully used for comparing images. It effectively matches points between sets and computes the distances between the matched points, but it is prohibitive computationally.

Along these lines, Hannig et al. (2013) developed a metric for comparing two images, based on one developed by Baddeley (1992). Comparing two images is different from comparing two data plots, where only essential plot elements are compared, so this metric is not useful comparing data plots generally. Marron and Tsybakov (1995) describes metrics for comparing two curves developed because the results from smoothing, based on classical mathematical norms do not match how the eye perceives the difference. This is too specialized. Actually, the lineup protocol could be utilized to rigorously substantiate the work in these papers.

In many practical uses of the lineup protocol, permutation can be used to generate data for the null plots in a lineup. This was employed in the study done in experiment III. Hamming distance (Hamming, 1950) calculates how different the permutations are, by measuring the minimum number of substitutions it takes to get from one permutation to another. It was one of our initial choices of distances, but it did not do well at distinguishing data plots from null plots, probably because it does not compare differences between the actual numbers, so we abandoned it.

Interpoint distance metrics, developed for cluster analysis, are fast to compute. These can be adapted to measuring structure in plots. For example, when the purpose is to compare two scatterplots containing clusters, we might calculate the smallest intercluster distance for both plots, and then use this to determine which plot exhibits the strongest clustering. This distance, together with various other measures to calculate distances between groups is implemented in the R package `fpc` (Hennig, 2015). Figure 3 illustrates choices of metrics of for measuring clustering. These are especially applicable to data from experiment III. The method can be adapted more broadly to different types of plots.
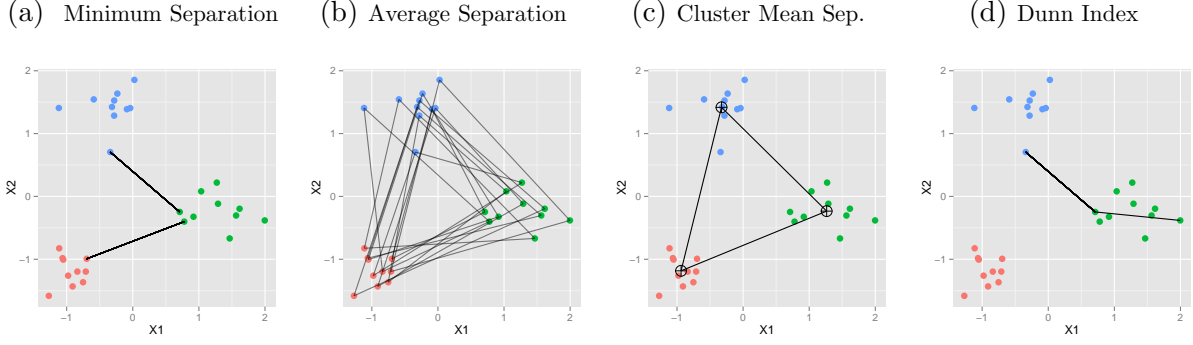
Figure 3: Illustration of four different distance metrics for cluster separation. Minimum Separation (a) calculates the minimum distance between points of each cluster from the other clusters. Average separation (b) calculates the average distance of each point in a cluster to the other clusters. Cluster mean distance (c) sums the distances between the means of each cluster. The Dunn index (d) is based on a comparison of minimal separation between clusters (as shown in (a)) and maximal cluster diameter.

For example, to measure the structure in a graph of side-by-side boxplots of two groups. The graphical features of the plot are the two five number summaries represented by the boxplots. An interpoint distance measuring the difference between the two sets of five points, can determine which plot exhibits the biggest difference between the two groups.

After trialing many of these distance, we report on the results from just a few metrics, that performed best. One approach was the interpoint distances, adapted to the displays used in experiments I-III, and the other a simple relatively simple distance based on binned frequencies (illustrated by Figure **??**), which is generalizable to most data plots. A summary of distance metrics follows, along with a procedure for computing the empirical distribution of the distance metrics, that can is used to measure the distance between data plot and null plots.

## 2.1 Distance metrics

A sketch of the distance metrics is provided here, with full details available in the Appendix.

1. **Binned Distance (BN):** compares the cell counts from binned data. This distance can be calculated for univariate continuous data, bivariate data with two categorical variables, or data with one continuous and one categorical variable. For a categorical

(a) Dataset $X$ with two variables $X_1$ and $X_2$
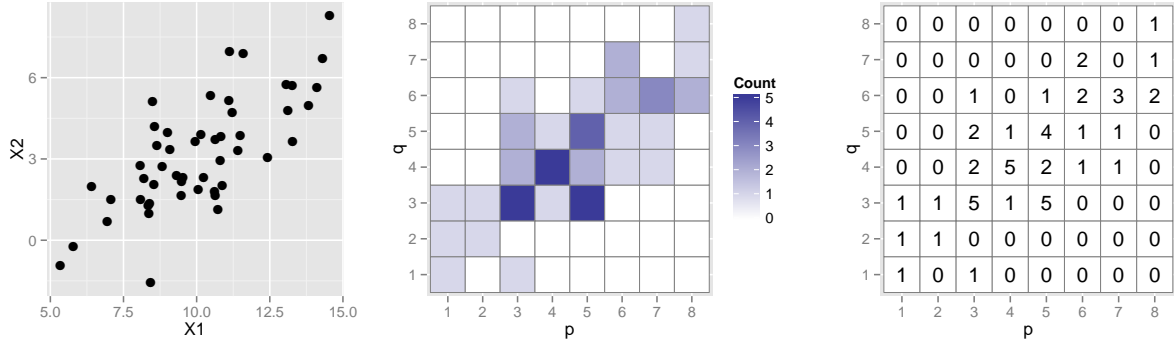
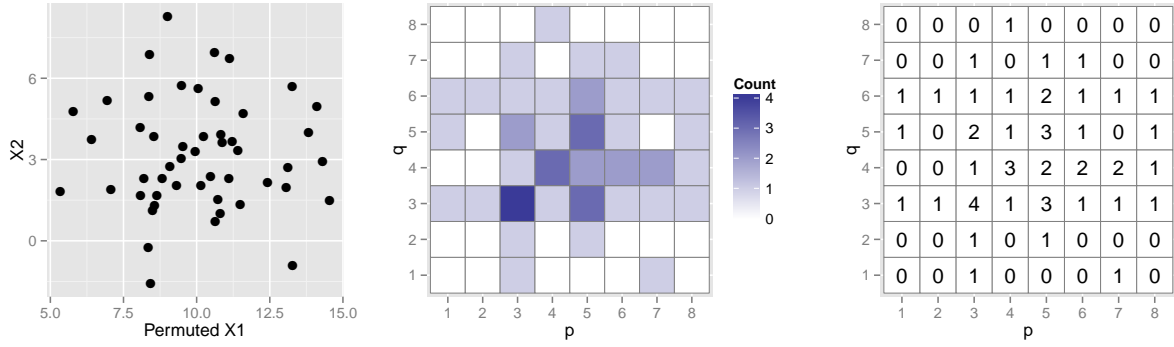(b) Dataset $Y$ with permuted $X_1$ and original $X_2$

Figure 4: Illustration of binned distance for data with a strong positive association (a), and the same data where variable $X_1$ has been permuted (b). The scatterplot of the data is shown (left) along with a binned view of the data (center) and the cell count matrix $C$ (right). Binned distance is the Euclidean distance of these counts. The binned distance between these plots is 6.4807.

variable, we choose the number of bins to be equal to the number of categories.

Binned distance is highly susceptible to small differences in values and depends on the number of bins as well as the anchor point (bottom left corner of first cell). It is necessary to find the optimal number of bins in each direction. For our purposes 'optimal' was defined as the number of bins that produced the largest detectable difference between data plot and null plots, compared to the biggest difference between any pair of null plots. Details of these choices on various different data sets can be found in the Appendix.

Several variations to this distance are possible, such as a change to using kernel density estimates, a change from $L_2$ to $L_p$ based distance, or using transformations on

the counts. All of these changes will affect distances, and might lead to qualitatively different conclusions. Hausdorff distance (Huttenlocher et al., 1993) was also examined, but the binned distance is computationally efficient and performed as well as the Hausdorff as a rough, generic measure of similarity of plots.

2. **Distance based on boxplots (BX):** is specifically designed for side-by-side boxplots based on only their graphical elements corresponding to the three points defining the boxes of a boxplot. It makes the assumption that subjects focus on the difference in the boxes. Variations on this might include adding whiskers' values, the number or values of outliers, or including higher-order letter values (Tukey, 1977; Hofmann et al., 2015) for more exact tail specifications.

3. **Distance based on the regression line (RG):** Many times, to examine association between two variables a regression line is overplotted on the points of a scatterplot. This distance was developed to help assess if the observer is paying attention to the line or the spread of points. The metric bins the data, and compares the intercept and slope of the regression line computed in the bins. Variations might include using slope alone, or absolute value of slope.

4. **Distance based on separation between multiple groups (MS, AS, DS, CM):** using minimum separation (MS), average separation (AS), Dunn separation (DS) and distance between cluster means (CM) are considered.

Generally to compare different distance measures their empirical distribution are used.

## 2.2   Empirical Distribution of Distance Metrics

For a given lineup of size $m$, the empirical distribution of distance metrics is obtained by calculating the distances between the null plots among themselves. One null data is generated using the null generating mechanism, and labelled to be the "true" data set, then a number of null data sets are generated and the distances between these datasets are calculated. Averaging all these distances yields one single distance value. This process is repeated a large number of times, say, $N$ between 1,000 to 10,000. Finally $N$ mean or

average distances are obtained which gives the empirical distribution of the distance. For comparing data plot with nulls using the empirical distribution of the distance metric, we use the following algorithm:

1. Calculate the distance between the true data and all the null datasets and take the average of these distances.

2. For each of the $(m - 1)$ null datasets, calculate the distance between the null data and all the other $(m - 2)$ null datasets and obtain the average distance. Hence, we obtain $(m - 1)$ distances, one corresponding to each null plot.

3. Generate a lineup of size $m$ using the null generating mechanism. Single out one of these nulls as the 'data' plot, and calculate the distances as described in steps (1) and (2). Repeat this procedure $N$ times.

4. The $N$ distance values then represent the empirical distribution of the distance metric and are used for making comparisons

The observed test statistic is compared to the empirical distribution, as shown in Figure 2. The distance measures for the true dataset and the null datasets are plotted on the empirical distribution. If the distance measure of the true plot is larger than any of the null plots, the lineup might be regarded as "easy". Otherwise, we consider it to be a "difficult" lineup. For easy lineups, we would expect that most observers could detect the true data plot amongst the decoys, but that far fewer observers to be able to do so with a difficult lineup. This gives us a way to compare the actual results from the MTurk human subject studies with what we might expect given the distance metric assessment.

The empirical distribution of the distance based on regression is shown in Figure 5 using $N = 1000$ simulation runs. Figure 5a shows the lineup plot for $m = 20$ for testing whether there exists a significant linear relationship between $X_1$ and $X_2$. The 19 null plots are generated by fitting the null model and generating from the null model. Figure 5b shows the general empirical distribution of distance measures based on the null model. For the particular lineup on the left, mean distances are shown by overlaid line segments for the true plot (in orange) and the null plots (in black). The true plot is easily identifiable from the lineup (Figure 5a, in the experiment 40 out of 45 observers identified the data). This

is backed by the regression based distance measure seen in Figure 5b, as the orange line is on the extreme right tail compared to the black lines.

Figure 6a shows a lineup of size $m = 20$ for testing whether there exists a significant difference in the group medians between A and B. The 19 null plots are generated from a null model, consisting of draws from a normal distribution. Figure 6b shows the empirical distribution of the distance based on the boxplots with the mean distance for the true plot (in orange) and the null plots (in black). It is hard to identify the true plot from the lineup. During the study, only 2 out of 26 observers picked the data plot, indicating little to no evidence of a deviation from the null hypothesis. This is also evident from the boxplot based distance measure: the orange line corresponding to the true data is mixed in with the mass of the black lines, with one null plot (16) exhibiting a lot more signal than the true plot.

## 2.3   Metric Evaluation

For a lineup of size $m = 20$, the average distance of the true plot from all null plots is compared to 18 average distances between the null plots. This high dimensionality of the comparison can sometimes complicate things. A logical solution is to derive a single statistic for each lineup. Such a statistic should take both the mean distance of the true plot into account as well as the maximum of the mean distances for the null plots. Hence we define,

1. $\delta$-**Difference:** let $\bar{d}_.$ be the average difference of plot . to all of the (other) null plots. We define the difference between the mean distance for the true plot and the maximum of the mean distances for the null plots as a measure of lineup difficulty, more specifically,

$$\delta(\ell) = \bar{d}_{\text{true}} - \max_j \left( \bar{d}_{\text{null}_j} \right) \tag{1}$$

for $j = 1, \ldots, (m-1)$ defines the lineup difference for lineup $\ell$. A positive difference indicates that the mean distance of the true plot is larger than the maximum of the mean distances of the null plots. Hence the true plot is more extreme compared to the set of null plots in the lineup. A larger difference should make data plot identification

12

(a) Lineup: which slope is the steepest?
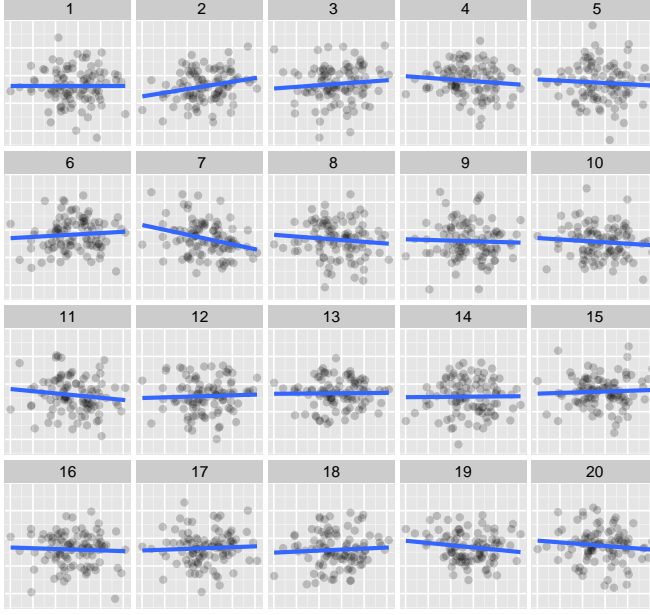


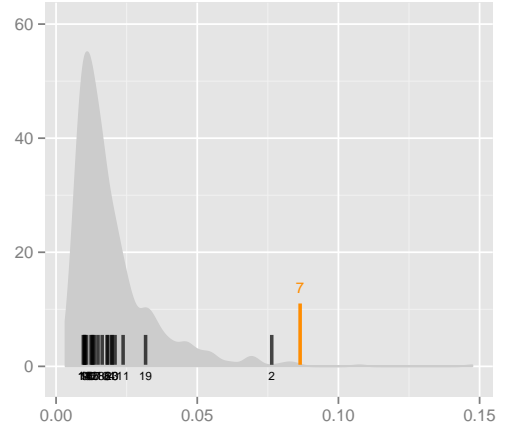(b) Regression based distance metrics.



Figure 5: Illustration of the behavior of a distance metric with a lineup plot in (a) and the distribution of regression based distance metric in (b). A lineup of size $m = 20$ is shown (left) for testing whether there exists a significant linear relationship between $X_1$ and $X_2$. The 19 null plots are obtained by simulating from the null model. The empirical distribution of the distance metric is shown on the right and overlaid by vertical line segments for the true plot and the null plots (in orange and black, respectively).

(a) Lineup: which of these pairs of boxplots shows the biggest vertical difference?
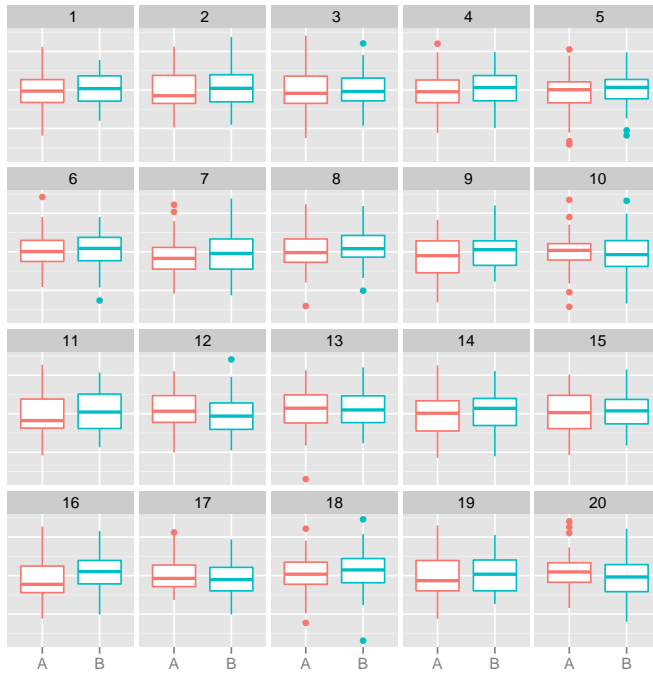
(b) Boxplot based distance
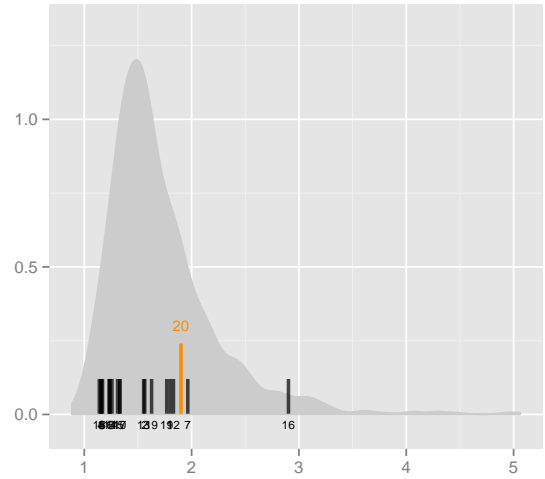


Figure 6: Illustration of the behavior of a distance metric for a more 'difficult' lineup. The lineup is shown in (a), the density plot on the right shows the boxplot based distance metric. Of interest is whether there exists a significant shift between the two groups. The orange line (boxplot distance of the true plot) is among the black lines of the nulls, indicating that the boxes in the true plot show no more difference than a null plot from other null plots.

easier. Similarly, a negative difference indicates that there is at least one null plot which is more extreme compared to the true plot based on the distance metric.

However, this statistic does not imply how many null plots are more extreme than the true plot. So we define,

2. $\gamma$-**Number of Extreme Nulls:** the number of null plots which have larger mean distances than the mean distance of the true plot is noted. Mathematically, for lineup $\ell$, we define the $\gamma$-number as

$$\gamma(\ell) = \sum_{j=1}^{m-1} \mathbb{1}\left(\bar{d}_{\text{null}_j} > \bar{d}_{\text{true}}\right),\tag{2}$$

where $\mathbb{1}(.)$ is a zero/one indicator function.

$\gamma(.)$ takes integer values between 0 and $(m-1)$. Higher values indicate more null plots being more extreme than the true plot, making it harder to identify it from the lineup.

Another choice for comparing lineups would be to use empirical $p$-values from the empirical distribution of the distance metric. While this would enable a comparison of all the metrics for any lineups, this approach is computationally extremely expensive, in particular, because we are generally interested in the extreme values of the empirical distribution, which needs a very large number of simulation runs $N$ for a reliable estimation.

In order to assess how well a distance metric reflects observers' choices, we relate distance metric to the rate at which the data plot in each lineup is being identified. In an ideal scenario, a detection rate of 0.05 corresponds to a $\delta$-difference of zero. With an increase in $\delta$-difference we would expect a simultaneous increase in detection rate. For the evaluation of metrics in the next section, we will fit a logistic regression of detection rate in $\delta$-difference.

# 3    Experimental Results and Analysis of Metrics

A number of experiments employing the lineup protocol were run using the MTurk service (Amazon, 2005-). A complete list and access to each experiment can be found in Hofmann et al. (2013).

Some experiments are used for evaluating the power of visual inference against that of classical tests (Majumder et al., 2013), some for comparison of different designs (Hofmann et al., 2012; Loy et al., 2015), or for targeted conclusions based on visual inference in situations where traditional tests do not perform well (Yin et al., 2013). In each study, subjects were recruited through the MTurk service and were shown a set of lineups. For evaluating the distance metrics we used the data collected on three experiments described in Table 1.

We evaluated the performance of the distance metrics in each of the experiments by comparing the distances to the responses from observers.

## 3.1 Experiment I – Side by Side Boxplots

This study was designed to investigate the power of visual inference in the classical situation of assessing the significance of a co-variate $X$ in a linear model. For the first study, the assumption is that the covariate is discrete (with two levels), while in Experiment II (see section 3.2) we assume the covariate to be continuous. The visual test statistic consists of side-by-side boxplots of the dependent variable against the two levels of covariate $X$. The data for the lineups comes from a model of the form $y_i + \mu + \beta_{x_i} + \varepsilon_i$ where $\mu$ is an overall average, $x_i \in 1, 2$ with $\beta_1 = -\beta_2$ is the effect for each of the two levels of $X$, and $\varepsilon_i \sim N(0, \sigma^2)$, independent for $i = 1, ..., n$. The null generating mechanism is then a simplified model without the covariate, i.e. $\beta_1 = \beta_2 = 0$. Each of the subjects, recruited through the MTurk service, was asked to evaluate ten lineups, and to identify, in each one, the plot that exhibits the largest vertical difference between groups A and B. The type of lineup used in this experiment is shown in Figure 6a.

For each lineup, the detection rate is calculated based on the number of evaluations and data identifications by subjects and related to its $\delta$-difference and $\gamma$-number of extreme null plots using both the distance based on boxplots ($d_{BX}$) and the binned distance ($d_{BN}$, using 8 bins in y direction and 2 in x).

These values are plotted in Figure 7. We see that as $\delta$-difference increases, detection rate generally increases. The solid lines in Figure 7a are fits from logistic regression models using a quadratic effect for distance. The fitted lines come very close to the non-parametric

(a) Scatterplots of detection rate versus δ-difference
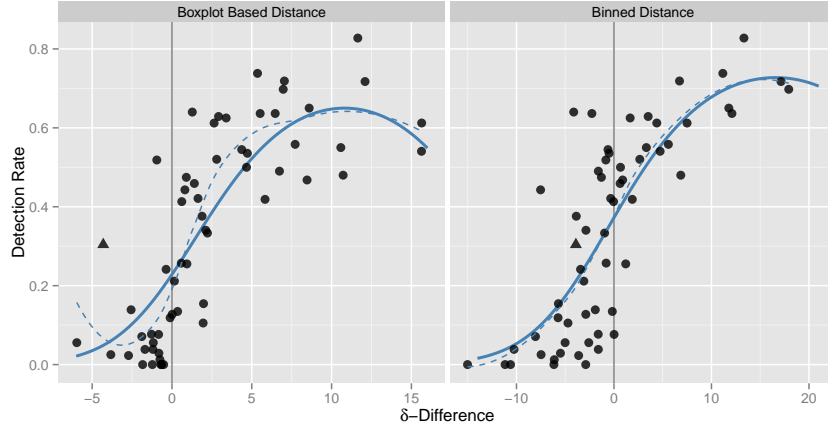


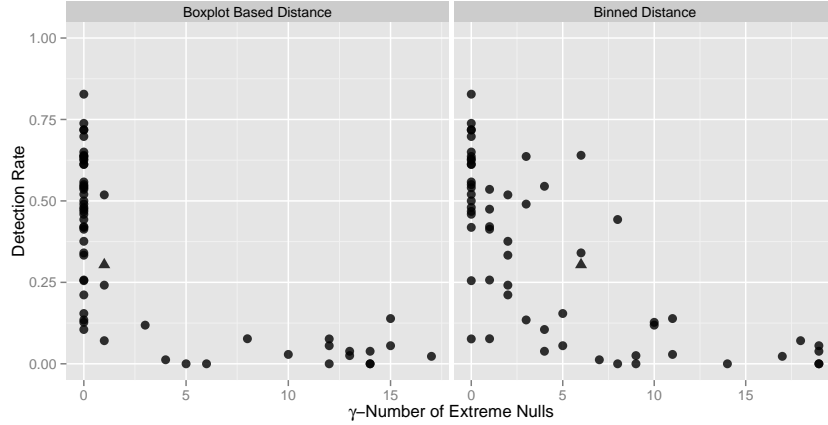(b) Scatterplots of detection rate versus γ-number of extreme nulls



Figure 7: Comparison of distance metrics for side-by-side boxplots. Detection Rate (a) and the number of plots greater than the observed (b) are plotted against the difference based on the boxplot and binned distance. The vertical line represents the difference equal to zero when there is at least one null plot similar to the observed plot. The detection rate increases with the difference. As the number of plots with distance greater than the observed increases, the detection rate decreases. The triangle represents a lineup which has a high detection rate but a negative δ-difference. This particular lineup is examined in Figure 8.

smooth shown by the dashed line. Qualitatively, both boxplot based distance and binned distance show very similar fits for detection rate. Based on the logistic regressions, boxplot distance is fitting detection rate a bit better (AIC: 798.7) than binned distance (AIC: 826).

Figure 7b shows the relationship between detection rate and the γ-number of extreme

null plots. As this number increases, it gets harder for subjects to identify the data plot. It is interesting to see that for some lineups subjects are able to pick the data plot even if there are one or two more extreme null plots.

Though distance based on boxplots works better a bit better, binned distance does a decent job in this case. According to the binned distance, there are a few lineups, in which the data plot is identified in more than 60% of all evaluations despite a negative $\delta$-difference (see also Figure 7). It should be noted that the binned distance does not take any graphical elements of the plot (such as the box or whiskers) into account but calculates distance solely based on the data. So outliers may have an effect on the binned distance but might not affect the distance based on the boxplots.

If participants base their choice on graphical elements, binned distance might therefore not be able to adequately reflect this. In order to investigate on what participants base their choice, we are going to have a closer look at an individual lineup.

From Figure 7 we see that for some of the lineups the detection rate is higher than its $\delta$-difference would suggest. One such lineup is marked using a triangle in Figure 7. Figure 8 shows the lineup and the distribution of distance metrics for a closer look at what might observers lead to pick the data plot as different. The grey numbers at the bottom right of each of the plots in the lineup shows a summary of how often a plot was picked by a participant in 168 evaluations.

The observed data plot is Plot #20, which has been picked most often by participants, but there are other plots that are being chosen quite frequently. Plot #16 is the plot with the largest boxplot distance. This is reflected by the large number of times this plot has been singled out by observers.

Maybe surprisingly, plots Plots #19 and #15, which have relatively large differences between the quartiles, are not being chosen by many participants. Instead, observers seem to focus on the difference in interquartile ranges (i.e. the height of the box in a boxplot). Plots #1, #13 and #16 exhibit a large interquartile difference and are being picked often by observers.

The time subjects take to respond to a lineup is another measure that can be used to evaluate their difficulty. Due to the presence of some huge outliers, we decided to use the

(a) Lineup of side-by-side boxplots.
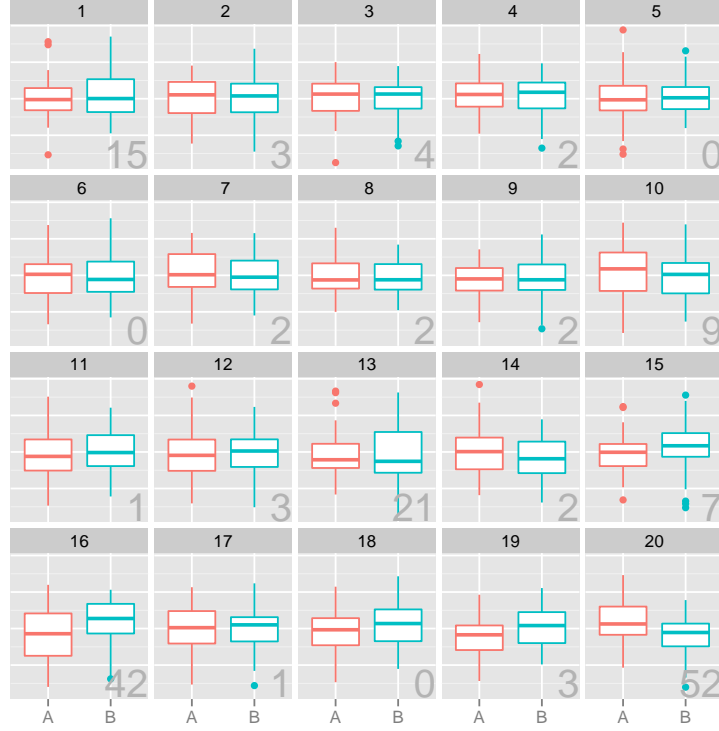


(b) Boxplot based distance    (c) Binned (2,8) distance    (d) Binned (2,2) distance
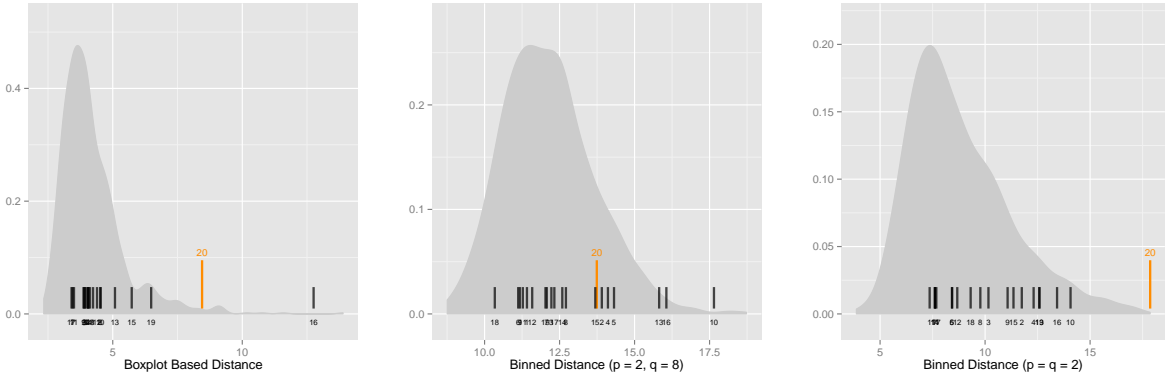


Figure 8: Illustration of the behavior of the three distance metrics. The lineup is shown in (a) and distributions of the different metrics based on this lineup is shown in the other plots: boxplot based distance in (b), binned distance with 2 and 8 bins on x and y axis in (c) and binned distance with 2 bins along both axes in (d). Grey numbers on the lineup show the counts of subject choices. The lineup corresponds to the point marked with a triangle in difference vs. detection rate plot in Figure 7.

median of time taken for each lineup. These values are plotted against $\delta$-difference for both distance measures as shown in Figure 9. Both $\delta$-difference exhibit a similar pattern: the time to respond peaks at a $\delta$-difference of zero. This indicates, that the situation where two or more plots exhibits similarly extreme features is the most difficult for observers to judge. With negative $\delta$-differences at least one null plot is more extreme, and observers seem to be able to make their choice quickly (probably for the extreme null plot). When $\delta$-difference is positive, the median time to respond decreases rapidly as $\delta$-difference increases. Hence, subjects are able to make their choice of a plot more quickly if the true plot is extreme compared to the null plots.

Note that taking the log of time to respond is an alternative to using the median. Qualitatively, the relationships to distance metrics are very similar for median and log time taken, but median time is much easier to interpret.
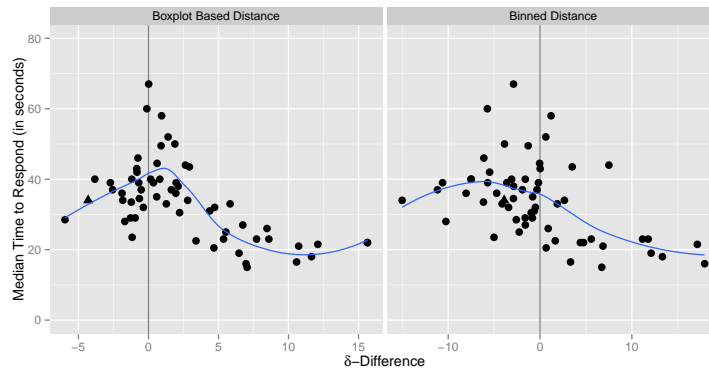


Figure 9: Comparison of distance metrics for side-by-side boxplots. Median time to respond is plotted against $\delta$-difference based on boxplot and binned(2,8) distance. The vertical line represents a $\delta$-difference equal to zero when there is at least one null plot similar to the observed plot. Median time to respond decreases as $\delta$-difference increases. The triangle marks again the lineup examined in detail in Figure 8.

## 3.2   Experiment II – Scatterplots with an Overlaid Regression Line

The question of interest of experiment II is very similar to the one in experiment I: again, the focus is to investigate the power of visual methods in the framework of normal models.

In contrast to experiment I, we are interested in the significance of a continuous variable $X$. The test statistic therefore is a scatterplot of the dependent variable and $X$ overlaid by a regression line. As mentioned in the introduction and further discussed in section 5.2 of Majumder et al. (2013), $X$ is assumed to be standard normal, and dependent data $Y$ is also simulated from a normal distribution for various correlation settings. Null data correspondingly is simulated from $N(X\widehat{\beta}, \widehat{\sigma}^2)$. Figures 1 and 5 show examples of the type of lineup used in the study. Subjects recruited from MTurk were shown a set of ten lineups and asked to identify the plot with the steepest slope in each.

For each lineup in this experiment, distances between the plots were computed using both regression based distance ($d_{RG}$) and binned distances ($d_{BN}$) with a small number of bins. For each lineup the proportion of data identifications was calculated from participants' responses and plotted against $\delta$-difference and the $\gamma$-number of extreme null plots, as shown in Figure 10.

As $\delta$-difference increases, average detection rate increases, i.e. subjects do better in easier lineups than hard ones. The solid lines show fits of logistic regressions of detection rate in $\delta$-difference based on regression (AIC: 746.9) and based on binned distance (AIC: 2150.7). Both fits come reasonably close to the dashed lines of a non-parametric loess smooth, indicating that they explain most of the relationship between $\delta$-difference and detection rate. The regression based distance works well in capturing the complexity of the lineups. A few lineups have a $\delta$-difference close to zero, marked by the vertical line – for those lineups detection rates of lineups flip from being close to zero to close to one within a very short interval.

For binned distance the situation is quite different. Although detection rate increases with difference, the detection rate is already quite high for lineups with negative differences. This is a classic scenario where the distance does not capture all the features on which observers base their choice: here, a graphical element –the line– affects the response, and shifts detection rates horizontally. While this makes an absolute comparison of distances across different types of plots impossible, we can still use binned distance as a relative measure to judge difficulty.

Figure 10b shows detection rate against the $\gamma$-number of extreme null plots. As the

(a) $\delta$-difference



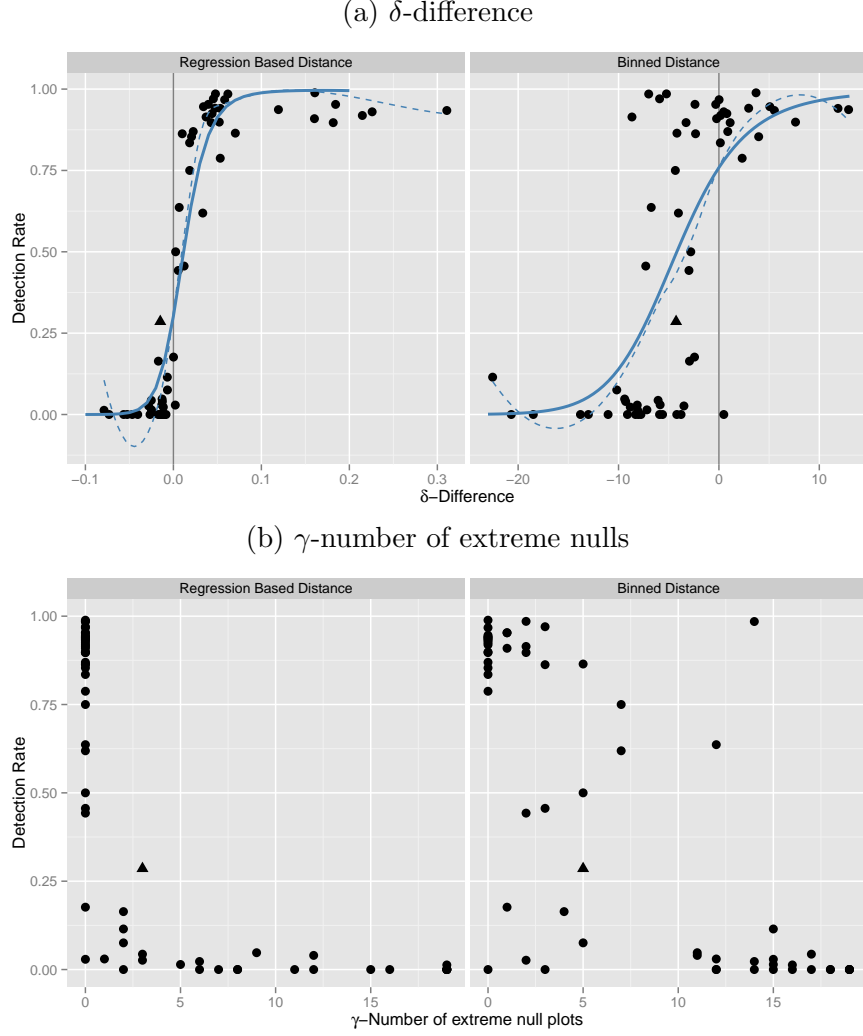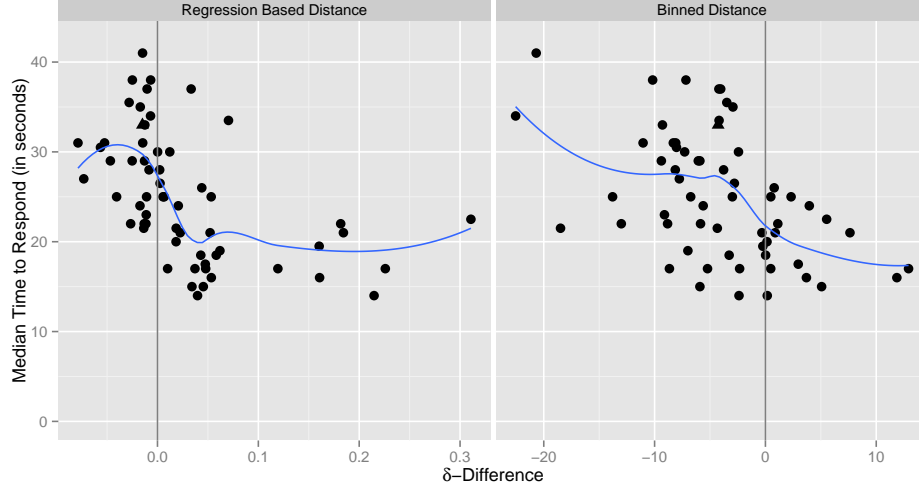(b) $\gamma$-number of extreme nulls



Figure 10: Comparison of distance metrics for scatterplots with a regression line overlaid. Detection rate is plotted against (a) $\delta$-difference and (b) $\gamma$-number of extreme null plots based on regression and binned distances. The vertical line represents the difference equal to zero when there is at least one null plot with an identical difference measure to the data plot. Detection rate increases on average with $\delta$-difference. As the $\gamma$-number of extreme null plots increases, detection rate decreases. The triangle represents a lineup which has high detection rate but negative difference. This particular lineup is examined in Figure 12.

number of extreme plots increases, detection rate decreases on average — indicating that identifying the data plot from a lineup becomes harder when there are more extreme plots in the lineup. For a few lineups, almost all evaluations led to an identification of the data plot although there was one null plot with an extreme feature. From Figure 10a, we see

Figure 11: Comparison of distance metrics for scatterplots with a regression line over laid. Median time to respond is plotted against $\delta$-difference based on regression based and binned(2,2) distance. The vertical line represents a $\delta$-difference equal to zero, i.e. there is at least one null plot similar to the observed plot. The median time to response decreases as $\delta$-difference increases. The triangle represents a lineup which is examined in Figure 12.

that this difference is marginal in most cases, though.

Figure 11 shows the relationship between the median time taken to respond and $\delta$-difference for both the distances. It can be clearly seen that there is a strong negative association: as $\delta$-difference increases, the subjects take less time to respond. Similar to the previous example of section 3.1 time to respond peaks at a $\delta$-difference close to zero.

Although the regression based distance seems to efficiently identify the quality of the lineup, there is one lineup (marked by a solid triangle in Figure 10) with a negative $\delta$-difference which was nevertheless identified by observers reasonably successfully. Figure 12 shows the lineup and the corresponding distributions of distance metrics.

The lineup in Figure 12 is a difficult one as suggested by the distribution of the distance metrics based on regression. For the data plot (in panel #10) the conventional $p$-value for testing the slope equal to zero is 0.085. However, the signal in the plot is strong enough, to make around 28% of all subjects pick this plot.

The binned distance with 2 bins on each axes does not let the data plot stand out in any way, however, the binned distance using the optimal number of bins (8 on the x-axis and 2 on the y-axis) by the optimal number of bins selection method identifies the data

(a) Lineup of scatterplots with overlaid regression lines.

(b) Regression based distance

(c) Binned (2,2) distance

(d) Binned (8,2) distance

(e) $p$-values
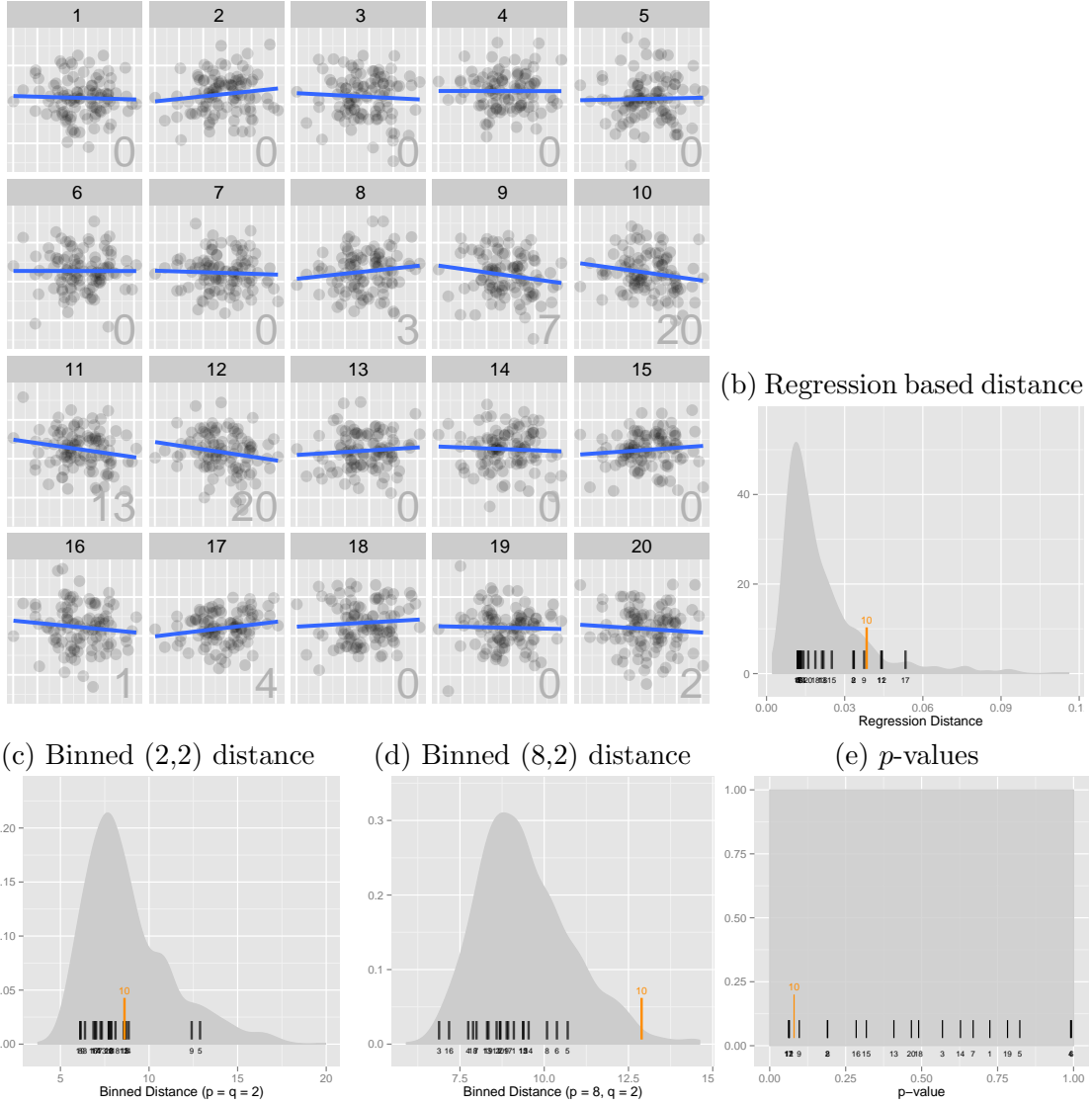
Figure 12: Illustration of the behavior of different distance metrics. The lineup is shown in (a) and the distributions of different distance metrics using this lineup are shown in plots (b)–(d): regression based distances in (b), binned distance with 2 bins on each axes in (c), and binned distance with 8 and 2 bins in x and y axis (d). In (e), the distribution of the conventional $p$-values are plotted with $p$-values for the lineups marked on the distribution. Grey numbers on the lineup show the counts of subject choices. The lineup corresponds to the point marked with a triangle in difference vs. detection rate plot in Figure 10.

plot as different from the others.

The number of picks lines up best with the difference measure based on the slope (regression without intercept). Plots #10, 12, 11, 9, and 7 all have a relatively steep slope, and are also the plots that were picked the most often. Again, a distance measure derived directly from one of the graphical elements in the plot leads to the best assessment of the choices made by human evaluators.

## 3.3   Experiment III – Large $p$, Small $n$ Data

The motivation behind this experiment is to study the effect of high dimensions on separability in data. Scenarios with pure noise and some real separation in two or three groups were investigated. A projection pursuit with Penalized Discriminant Analysis Index (Lee and Cook, 2010) was used to obtain one (for two groups) or two (for three groups) dimensional projections. Depending on the number of groups, either a jittered dotplot or a scatterplot was used as test statistic in a lineup setting. The null plots are obtained by permuting the group variable and plotting the two dimensional projections obtained from a projection pursuit with PDA index. The subjects were shown these lineups and were asked to identify the plot with the most separated colored groups.

The distances between the plots in this experiment were computed using the distance based on minimum separation and average separation of the clusters and also the binned distance. The number of bins used for the lineups with one dimensional projections is larger (10 in this case) but for the lineups with two dimensional projections, the number of bins used is 5. The proportion of correct response is plotted against $\delta$-difference and $\gamma$-number of extreme nulls for both distances. Figure 13 shows the results.

In Figure 13, the detection rate is plotted against the difference for distance based on minimum separation, average separation and the binned distance. The vertical line shows a difference equal to zero. It can be seen that as the difference increases, the detection rate increases and all distances do a reasonably good job in capturing the response of the subjects. In terms of the logistic regression fit to the data, average separation is a bit ahead according to AIC (AIC: 349.6) compared to minimal separation (AIC: 422.8) and binned distance (AIC: 543.4). Dunn separation, motivated from cognitive perception, comes in at
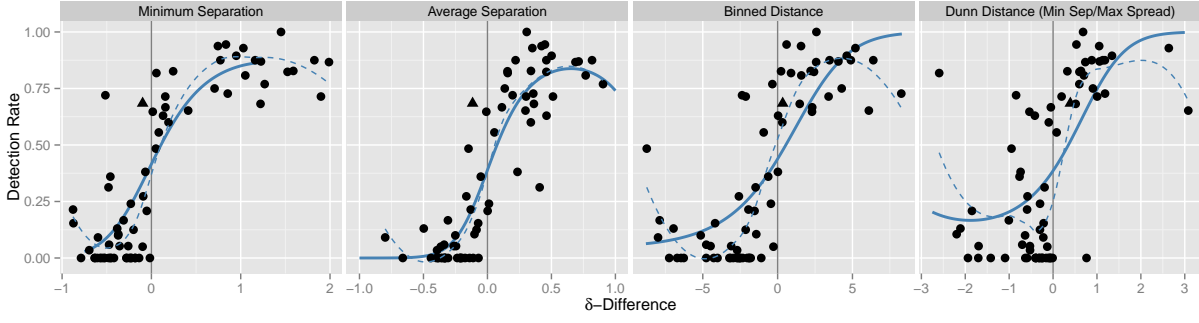
(b) $\gamma$-number of extreme null plots
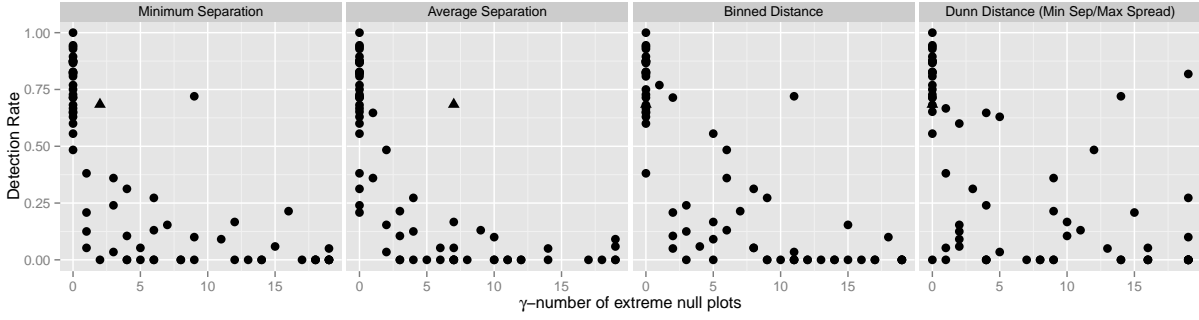


Figure 13: Comparison of distance metrics for the scatterplot with clusters. Detection rate is plotted against (a) $\delta$-difference and (b) against $\gamma$-number of extreme nulls, using distances based on minimum separation, average separation, binned and Dunn's distance. The vertical line represents the difference equal to zero when there is at least one null plot similar to the observed plot. Solid blue line represents the fitted logistic regression model and the dashed blue line shows a loess smoother. Detection rate generally increases with $\delta$-difference. As the $\gamma$-number of extreme null plots increases, detection rate decreases. The triangle represents a lineup with high detection rate and negative difference based on the average separation distance. This is examined in Figure 15.

a maybe surprising last place (AIC: 697.3).

In (b) it can be seen that as there are more extreme null plots compared to the observed plot, the subjects find it difficult to pick the observed plot. For a few lineups, a large number of the subjects identify the observed plot although there is more extreme null plots.

Figure 14 shows the relationship between the median time observers take to respond and $\delta$-difference for the three different distances. It can be clearly seen that there is a
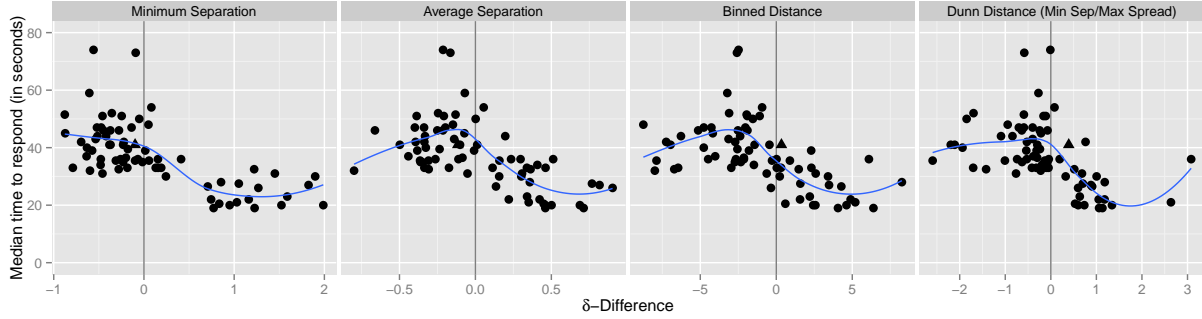
Figure 14: Plot showing the median time to respond by the subjects against the difference based on the minimum separation distance, average separation and binned distance. The vertical line represents the difference equal to 0 when there is at least one null plot similar to the observed plot. The median time decreases as the difference increases.

strong negative association; as the difference increases, subjects take less time to respond. For both average separation and binned distance we see a peak in response time, i.e. for large negative $\delta$-difference the median time to respond decreases again.

Figure 15 shows the lineup in a high dimension, low sample size setting. The number of dimensions used is 100 and two of the dimensions have some separation. Plot #20 shows the two-dimensional projections of the original data. This plot is, indeed, chosen in 13 out of 19 evaluations. As the true plot does have real separation, it is to be expected that subjects would be able to identify the plot. The distance based on average separation yields a negative difference showing that the lineup is difficult, while the distance based on minimum separation yields a positive difference. The distance metrics identify different characteristics in a plot. The average separation looks at the average of the distances of the points in a cluster to the points in other clusters. Dunn separation performs very well in this example and correctly identifies the data panel as the panel with the strongest signal. While the binned distance shows plot #20 as the one with the largest distance, this should be taken with care, because binned distance is, unlike the other two distances, not rotation invariant. In the case of plot #20 the large distance merely indicates the difference in the arrangement of the clusters rather than their separation.

(a) Lineup of scatterplots of three groups. Which plot shows the best separation?



(b) Binned (5,5) Distance
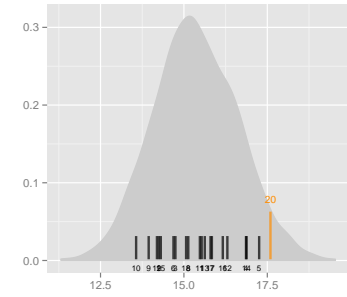


(c) Minimum Separation
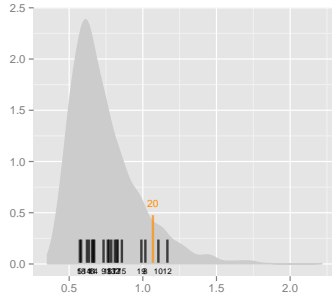
(d) Average Separation
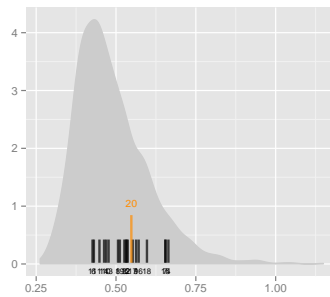
(e) Dunn Separation
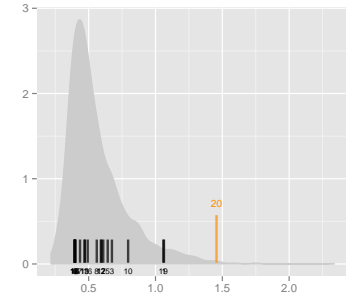


Figure 15: Illustration of the behavior of different distance metrics. The lineup is shown in (a) and the distributions of different distance metrics are shown in the other plots: binned distance with 6 and 4 bins in x and y axis respectively in (b), distance based on minimum separation in (c), distance based on average separation in (d) and distance based on Dunn separation in (e). Grey numbers on the lineup show the counts of subject choices. The reference distributions are based on 100 null sets of 19 nulls each, yielding 1,800 distances.

# 4 Conclusion

Distance metrics are compared to the response of human subjects on lineups. What we see in each of the experiments is that the general approach of the binned distance, while mostly performing decently, is usually out-done by a distance that is more tailored to the question of interest or takes the graphical elements into account. These were a special collection of experiments where the task was very focused, but in practice the question is generic and we would expect the binned distance to be most broadly applicable.

The distance metrics clearly can be used to help in future experiments with lineups, to roughly divide them into three groups, easy, moderate, difficult. Subjects are typically given a block of ten lineups to evaluate. Having the difficulty rating will help provide a reasonable set for each subject.

It is also clear that some distance metrics better match with subjects' choices. Boxplot du=istance better matched subject choices in experiment I. Regression based distance matched better than binned distance for the subject choices in experiment II. By the $\gamma$-number minimum separation and average separation matched better than the other distance metrics for experiment III.

None of the distance metrics perfectly matches subject choices. This suggests that there is a lot of scope for exploring new ways to numerically characterize structure in plots. The metrics we have described still are based on the data, rather than the graphical elements in the plot. Utilizing the graphical elements, e.g. length of line relative the plot size to build metrics might provide a close fit, and also help to understand more precisely what viewers are visually responding to in the plot.

## SUPPLEMENTARY MATERIAL

**Software:** R-package `nullabor` containing code to create lineups and calculate the distance measures described in the article. Available on CRAN (R Development Core Team, 2015), with development versions at https://github.com/dicook/nullabor.

**Reproducibility:** All the code and anonymized data used in this analysis is available at https://github.com/niladrir/metrics-paper.

[Experiments:] Add the link to the experiments web site

# References

Amazon (2005-), "Mechanical Turk," https://www.mturk.com/mturk/welcome, Accessed: 2016-08-17.

Anscombe, A. J. (1972), "Graphs in Statistical Analysis," *The American Statistician*, 27:1, 17–21.

Baddeley, A. J. (1992), "An Error Metric for Binary Images," in *Robust Computer Vision: Quality of Vision Algorithms*, eds. Forstner, W. and Ruwiedel, S., Karlsruhe: Wichmann, pp. 59–78.

Bhattacharyya, A. (1946), "On a measure of divergence between two multinomial populations," *Sankhyā: The Indian Journal of Statistics*, 401–406.

Buja, A., Cook, D., Hofmann, H., Lawrence, M., Lee, E., Swayne, D., and Wickham, H. (2009), "Statistical Inference for Exploratory Data Analysis and Model Diagnostics," *Royal Society Philosophical Transactions A*, 367, 4361–4383.

Dunn, J. C. (1973), "A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters," *Journal of Cybernetics*, 3, 32–57.

Halkidi, M., Batistakis, Y., and Vazirgiannis, M. (2001), "On Clustering Validation Techniques," *Journal for Intelligent Information Systems*, 17, 107–145.

Hamming, R. W. (1950), "Error detecting and error correcting codes," *Bell System technical journal*, 29, 147–160.

Hannig, J., Lee, T. C. M., and Park, C. (2013), "Metrics for SiZer map comparison," *Stat*, 2, 49–60.

Hennig, C. (2015), *fpc: Flexible Procedures for Clustering*, R package version 2.1-10.

Hofmann, H., Follett, L., Majumder, M., and Cook, D. (2012), "Graphical tests for power comparison of competing designs," *Visualization and Computer Graphics, IEEE Transactions on*, 18, 2441–2448.

Hofmann, H., Majumder, M., and Cook, D. (2013), *Experiments for Visual Inference*, http://www.public.iastate.edu/~hofmann/experiments.html, Accessed: 2015-08-31.

Hofmann, H., Wickham, H., and Kafadar, K. (2015), "Letter-value plots: Boxplots for large data," *Journal of Computational and Graphical Statistics*, submitted.

Huttenlocher, D., Klanderman, G., and Rucklidge, W. J. (1993), "Comparing Images Using the Hausdorff Distance," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15:9.

Lee, E.-K. and Cook, D. (2010), "A projection pursuit index for large p small n data," *Statistics and Computing*, 20, 381–392.

Loy, A., Follett, L., and Hofmann, H. (2015), "Variations of Q-Q Plots – the Power of our Eyes!" *The American Statistician*, 2015, 1–36.

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

Marron, J. S. and Tsybakov, A. B. (1995), "Visual Error Criteria for Qualitative Smoothing," *Journal of the American Statistical Association*, 90, 499–10.

R Development Core Team (2015), https://cran.r-project.org.

Roy Chowdhury, N., Cook, D., Hofmann, H., Majumder, M., Lee, E.-K., and Toth, A. L. (2015), "Using visual statistical inference to better understand random class separations in high dimension, low sample size data," *Computational Statistics*, 30, 293–316.

Shapiro, S. S. and Wilk, M. B. (1965), "An analysis of variance test for normality (complete samples)," *Biometrika*, 591–611.

Stephens, M. A. (1974), "EDF statistics for goodness of fit and some comparisons," *Journal of the American Statistical Association*, 69, 730–737.

Tukey, J. W. (1977), *Exploratory Data Analysis*, Reading, MA: Addison-Wesley Publishing Company.

Vander Plas, S. and Hofmann, H. (2015), "Clusters beat Trend!? Testing feature hierarchy in statistical graphics," *Journal of Computational and Graphical Statistics*, submitted.

Wickham, H. (2009), *ggplot2: Elegant Graphics for Data Analysis*, useR, Springer.

Xie, Y. (2015), *Dynamic Documents with R and knitr*, Boca Raton, Florida: Chapman and Hall/CRC, 2nd ed., iSBN 978-1498716963.

Yin, T., Majumder, M., Roy Chowdhury, N., Cook, D., Shoemaker, R., and Graham, M. (2013), "Visual Mining Methods for RNA-Seq Data: Data Structure, Dispersion Estimation and Significance Testing," *Journal of Data Mining in Genomics & Proteomics*, 4.

Zhao, Y., Cook, D., Hofmann, H., Majumder, M., and Roy Chowdhury, N. (2013), "Mind Reading Using an Eyetracker to See How People Are Looking at Lineups," *International Journal of Intelligent Technologies and Applied Statistics*, 6, 393–413.

# Appendix

## Details of Distance Metrics

For all of these distance measures, let $X$ denote the true dataset with one or two variables. Let $Y$ denote the null dataset obtained by using an appropriate null generating mechanism.

1. **Binned Distance (BN):** Let $X_1$ and $X_2$ be two continuous variables. Let $X_1$ be divided into $p$ bins and $X_2$ divided into $q$ bins. We define the cell count matrix $C(X_1, X_2)$ as a $p \times q$ matrix consisting of the binned frequency of the joint distribution of $X_1$ and $X_2$, i.e. let the $(i,j)$-th element of the matrix be the cell count of the number of points in interval $i$ of $X_1$ and interval $j$ of $X_2$. The BN distance is defined as

$$
\begin{aligned}
d_{BN}^2(X,Y) \quad &:= \quad ||C_X(X_1, X_2) - C_Y(X_1, X_2)||^2 \\
&= \quad \sum_{i=1}^{p} \sum_{j=1}^{q} (C_X(X_{1i}, X_{2j}) - C_Y(X_{1i}, X_{2j}))^2.
\end{aligned}
\tag{3}
$$

2. **Distance based on boxplots (BX):** Let $X_1$ be a categorical variable with $J$ groups and $X_2$ be a continuous variable. For each of the levels in $X_1$ find the corresponding $X_2$ values, and calculate first quartile, median and the third quartile. Identify the minimum and maximum value of each of these statistics across all $J$ groups. Define $d_q(.)$ as a three dimensional vector giving the absolute maximal difference of the first quartile, median and the third quartile of $X_2$ between the groups in $X_1$.

   Then the distance metric is given by

$$
d_{BX}^2(X,Y) := ||d_q(X) - d_q(Y)||^2 = \sum_{i=1}^{3} (d_q(X)_i - d_q(Y)_i)^2.
$$

3. **Distance based on the regression line (RG):** Let $X_1$ and $X_2$ be two continuous variables plotted in a scatterplot. We assume that the scatterplot is binned horizontally into $b$ bins to allow for a description of a piece-wise linear relationship between the variables. In each vertical slice, a linear regression model is fitted and the regression coefficients. i.e. intercept and slope are estimated. The distance metric

33

based on the regression coefficients is given as

$$d^2_{RG}(X,Y) := \operatorname{tr}(B(X) - B(Y))'(B(X) - B(Y))$$

$$= \sum_{i=1}^{b}((b_0(X))_i - (b_0(Y))_i)^2 + \sum_{i=1}^{b}((b_1(X))_i - (b_1(Y))_i)^2 \qquad (4)$$

where $b_0$ and $b_1$ denote the vector of the intercept and slope respectively while $b$ is the number of bins. $B(.)$ is a $b \times 2$ matrix of the regression coefficients where each row represent the intercept and the slope obtained from each bin. The number of bins have a significant effect on the distance measure. It can be seen that it works best for smaller number of bins like 1 or 2. With larger number of bins (i.e. smaller bin sizes), the regression coefficients are affected by the variability in the data and the signal to noise ratio in the data becomes too low for a reliable detection. Variations might include using slope alone, or absolute value of slope. Note that we assume that $X_1$ and $X_2$ are on the same scale. This does not change the regression or the significance of its parameters, but it does matter for the distance measure. By assuming variables on the same scale, we implicitly expect that the scatterplot has an aspect ratio of 1, which is typically the case in a lineup. This way, a deviation along the x axis is perceived to be about the same as a deviation along the y axis.

4. **Distance based on separation between multiple groups (MS, AS, DS, CM):**
   Let $X_1$ and $X_2$ be two continuous variable. Let $X_3$ be a categorical variable providing the groups associated with each variable. $X_1$ and $X_2$ are plotted in a scatterplot colored by the group variable $X_3$. The separation can be described in a number of different ways:

   (i) **Minimum separation:** let $s_m(.)$ be a vector of the minimum distance of a point in the cluster to a point in any of the other $g - 1$ clusters. The distance metric based on minimal separation is an average of these intercluster minima, defined as

   $$d^2_{MS}(X,Y) := ||s_m(X) - s_m(Y)||^2 = \sum_{i=1}^{g-1}((s_m(X))_i - (s_m(Y))_i)^2.$$

   (ii) **Average point separation:** let $s_a(.)$ be a vector of cluster wise average distances of all the points in a cluster to all points in any of the other $g - 1$ clusters.

The distance metric based on average separation is defined as

$$d_{AS}^2(X,Y) := ||s_a(X) - s_a(Y)||^2 = \sum_{i=1}^{g} ((s_a(X))_i - (s_a(Y))_i)^2.$$

(iii) **Dunn separation:** the Dunn index (Dunn, 1973; Halkidi et al., 2001) $s_d(.)$ is defined as the ratio of the minimal separation between clusters and maximal diameter of any of the clusters. The distance metric based on the Dunn index is defined as

$$d_{DS}^2(X,Y) := (s_d(X) - s_d(Y))^2.$$

The Dunn index is a member of the Dunn index family, defined as ratios of measures of cluster separation and cluster extent. This makes an analysis of variance approach to clustering, such as Ward's method part of this family as well. Minimal separation between clusters and maximal diameter of a cluster are two very concrete measures. This way, the Dunn index stays close to a visual assessment of the clustering from a cognitive perspective. However, because it is based on two extreme statistics, the Dunn index is highly susceptible to outliers.
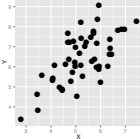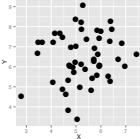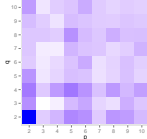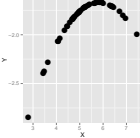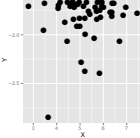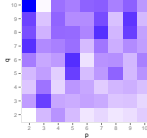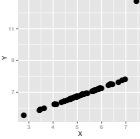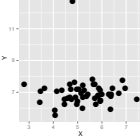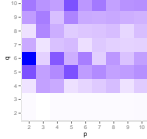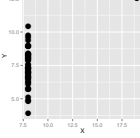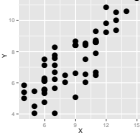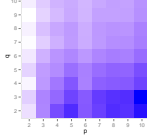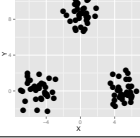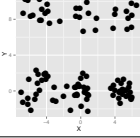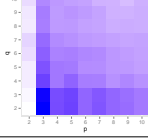
We also examined the distance between the means for each cluster (CM) as another alternative, which would more closely match Fisher's linear discriminant. As we might expect, this did not match how observers read the separation. In practice, many possible metrics could be used to measure the separation, such as those readily available in the `fpc` package (Hennig, 2015). Figure 3 illustrates the different distance metrics for cluster separation.

## Effect of the Number of Bins on Distance

Binned distance works for any type of data and for any null generating mechanism. It is based on the raw data but does not take into account the graphical elements in the plot. Binned distance can be used in situations where no distance measure is known for the particular plot type and hence it can be regarded as universal. But the the number of bin highly affect the distance. An unfortunate choice may lead to hard to interpret or conflicting results.
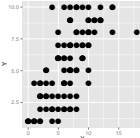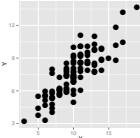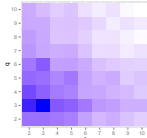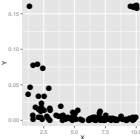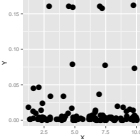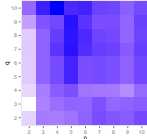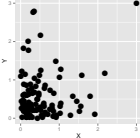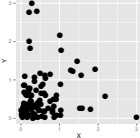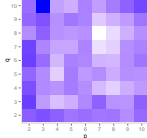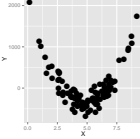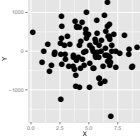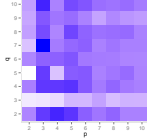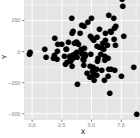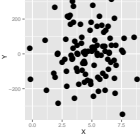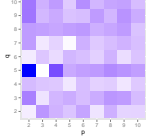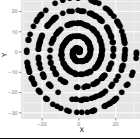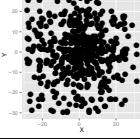
We investigate the choice of the number of bins using different types of data and different null generating mechanisms. Null datasets are obtained for a true data set using a null generating mechanism and hence a lineup is constructed. As described in section sec:dists mean binned distances are calculated between the true data and the null datasets and also among the null datasets. The number of bins for the binned distance are varied from 2 to 10 on both $x$ and $y$ direction and $\delta$-difference is determined for each combination. T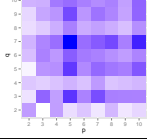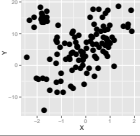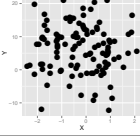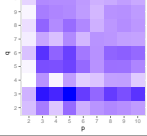ables 2 and 3 show the type of data, the observed plot, the null generating mechanism (NGM), a typical null plot, $\delta$-difference and the maximally observed value of $\delta$, together with the corresponding number of bins in $x$ and $y$ direction. Similarly, the minimal $\delta$ value and its parameters are collected to get an idea of the range of values.

Table 2: Preferable number of bins for different types of observed data to calculate the binned distance.

| Data | Observed Plot | NGM | Typical Null | Difference | (x-bin, y-bin) Min, Max |
|------|--------------|-----|-------------|-----------|--------------------------|
| Linear association |  | Permutation |  |  | (2,2) -2.5, 5.7 |
| Nonlinear relationship |  | Permutation |  |  | (2,10) 0.0, 6.2 |
| Linear with outliers |  | Permutation |  |  | (2,6) -0.4, 16.7 |
| Same $x$ with one outlier |  | Simulation from Poi(9) |  |  | (10, 3) -0.1, 34.3 |
| Clusters |  | Permutation |  |  | (3,2) -5.7, 37.6 |

The rationale behind selecting different types of data is to investigate how the optimal number of bins or bin sizes varies with different types of data. The different null generating

Table 3: Preferable number of bins for different types of observed data to calculate the binned distance.

| Data | Observed Plot | NGM | Typical Null | Difference | (x-bin, y-bin) Min, Max |
|---|---|---|---|---|---|
| Categorical |  | Simulation from $N(\mu, \sigma^2)$ |  |  | (3,3) 6.2, 30.7 |
| Nonlinear with outliers |  | Permutation |  |  | (4, 10) -3.4, 3.9 |
| Skewed with outliers |  | Permutation |  |  | (3,10) -7.1, 0.3 |
| Residual |  | Simulation from null model |  |  | (3,7) -4.5, 17.8 |
| Residual |  | Simulation from null model |  |  | (2,5) -4.4, 4.8 |
| Spiral |  | Permutation |  |  | (5,7) -11.9, 23.6 |
| Contaminat |  | Permutation |  |  | (5,3) -2.5, 8.1 |

mechanisms are also selected for the same reason. In Table 2 the first four observed data plots corresponds in spirit to the datasets described by Francis Anscombe (Anscombe, 1972), using a larger number of data points. The fifth dataset is a data set with 3 distinct clusters. In Table 3, the first dataset shows a categorical dataset. The second and the third data are non-linear association and skewed with the presence of outliers. The fourth and fifth datasets are residual plots with a curved pattern and non-constant variance pattern. The sixth data is a spiral data while the seventh one is a data with contamination.

The $\delta$-differences are represented in a tile plot where each tile gives the difference for each combination. Darker hues correspond to larger differences values. It can be seen that these tile plots look different for the different datasets. Hence the optimal number of bins varies from data to data. No specific pattern is evident in the plot. But overall it can be seen that for strong linear relationships, small number of bins should be preferred over large number of bins. Also when outliers are present in the data, larger number of bins are preferred on at least one of the axes.

It is important to mention at this point that Tables 2 and 3 is not meant to provide any guidelines for the selection of number of bins. The Tables only show that binned distance is highly affected by the number of bins and the type of data. It is advisable to find the optimal number of bins for a given dataset before using binned distance.

## 4.1   Discussion of the Effect of Plot Type and Question of Interest

Previous studies have suggested that the type of plot used in the lineup affects the response of the subjects (Hofmann et al., 2012; Zhao et al., 2013). For example, subjects more often identify the true plot for large data with side-by-side box plots than dot plots. Similarly, the use of aesthetics, such as colors or shapes, and graphical elements, such as a trend line, also influence an observer's decision (Vander Plas and Hofmann, 2015). In order to account for this, the distance metric has to be adjusted for the plot type and graphical elements. Figure 16 illustrates this idea.

Figure 16a shows a lineup of scatterplots with 100 points between two variables $X_1$ and $X_2$. Figure 16b, on the other hand, gives a lineup of the same scatterplots with the regression line overlaid. The overlaid regression lines are likely to help observers in

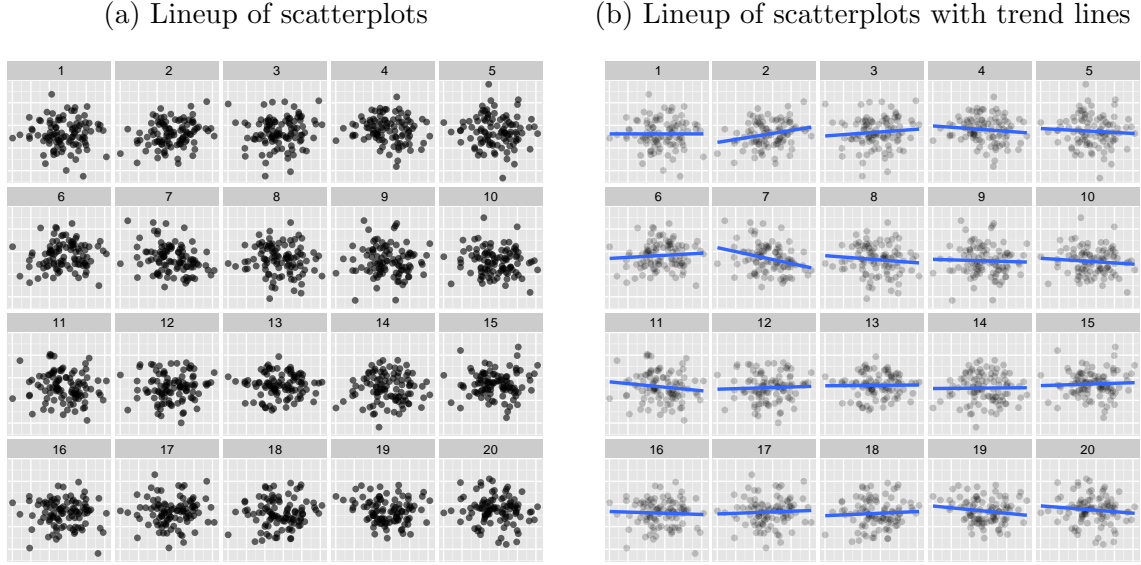(a) Lineup of scatterplots      (b) Lineup of scatterplots with trend lines

Figure 16: Comparison of two lineups: scatterplots in (a) and scatterplots with a regression line overlaid in (b). The raw data is same for the lineups. The human subjects are shown the lineups and asked to identify the plot with the steepest slope. The presence of the regression line directs the decision of observers.

identifying the panel with the steepest slope. The choice of distance metric depends on what kind of comparisons those distance measures are supposed to allow. Internally, when the same design is used, all measures are consistent – and allow for a relative comparison: an easier lineup will show a bigger positive distance between the true plot and the nulls than a more difficult one. The distance measure does not necessarily allow for an absolute assessment of difficulty. If a distance metric is additionally supposed to assist in planning an experiment where different types of plots are considered or different graphical elements on the same type of plot, the distance metric has to be able to take these differences into account.

The question posed to observers plays an important role in the decision they make. A minor change in the question can change the response of the subject. For example, if subjects are asked to identify a plot from one in the lineups in Figure 16, in which there is the strongest positive relationship between the variables, plot 2 exhibits the strongest signal. If the question asks, instead, for the plot with the largest slope, plot 7 would be the obvious pick. A distance metric should also take into account the question of interest,

and the metrics that we have described in the previous section do this for the MTurk studies done to date where quite specific questions were used. In practice, the question of interest very broad to enable detection of any type of significant structure in a plot, and is generically, "which plot in the lineup is the most different from the others", in which case the binned distance should be optimal.

## Description of MTurk Data Collection

TO BE FILLED IN Participants from Amazon Mechanical Turk were recruited for the experiments. These participants were shown lineups and were asked to identify the plot in the lineups which, according to their judgement, closely corresponds to the question they were asked. They were also asked to provide a reason for their choice and the confidence in their response in a categorical scale. Other than these data, gender, age, education and geographic location were also collected.

Each participant was shown at least 10 lineups of varying difficulty, with a fair share of "easy" lineups. These were randomly chosen from a larger pool of lineups having different difficulty levels. The difficulty level of the lineups were controlled by varying the parameters like slope and variability. These are described by greater details in Majumder et al. (2013) and (Roy Chowdhury et al., 2015). None of the participants judged the same lineup twice.

Amazon Mechanical Turk workers were paid for their efforts on the scale of minimum wages in USA. To minimize their efforts, some workers may try to randomly pick a plot in the lineup to get the job done. To avoid this issue, a very easy lineup was randomly placed in this block of 10 lineups, in which every participant should identify the actual plot correctly. During data cleaning, the response from only those participants were considered who identified the actual plot correctly in this easy lineup. The response from the other lineups for these participants were only considered while the response for this easy lineup was not considered.