Utilizing Distance Metrics on Lineups to Examine What People Read From Data Plots

Niladri Roy Chowdhury, Dianne Cook, Heike Hofmann, Mahbubul Majumder, Yifan Zhao

June 30, 2014

Abstract

Graphics play a crucial role in statistical analysis and data mining. This paper describes metrics developed to assist the use of lineups for making inferential statements. Lineups embed the plot of the data among a set of null plots, and engage a human observer to select the plot that is most different from the rest. If the data plot is selected it corresponds to the rejection of a null hypothesis. Metrics are calculated in association with lineups, to measure the quality of the lineup, and help to understand what people see in the data plots. The null plots represent a finite sample from a null distribution, and the selected sample potentially affects the ease or difficulty of a lineup. Distance metrics are designed to describe how close the true data plot is to the null plots, and how close the null plots are to each other. The distribution of the distance metrics is studied to learn how well this matches to what people detect in the plots, the effect of null generating mechanism and plot choices for particular tasks. The analysis was conducted on data that has already been collected from Amazon Turk studies conducted with lineups for studying an array of data analysis tasks.

1 Introduction

Graphics are an important component of big data analysis, providing a mechanism for discovering unexpected patterns in data. Pioneering research by Gelman [2004], Buja et al. [2009] and Majumder et al. [2013] provide methods to quantify the significance of discoveries made from visualizations. Buja et al. [2009] introduced two protocols which bridge the gulf between traditional statistical inference and exploratory data analysis. These are the Rorschach and the lineup protocols. The Rorschach protocol helps to understand the extent of randomness. The lineup protocol places a statistical plot firmly in the hypothesis testing framework, where a plot of the data is considered to be a test statistic. Unlike the simpler numeric test statistics in classical inference, though, the plot as a test statistic is a complex entity. This plot is compared with a set of null plots, obtained from an appropriate distribution consistent with the null hypothesis. The lineup protocol places the data plot randomly among the obtained null plots, and requires a human observer to examine the plots and identify the most different plot. If this plot is that of the data, this is quantifiable evidence against the null hypothesis. The lineup protocol was formally tested in a head-to-head comparison with the equivalent conventional test by Majumder et al. [2013]. The experiment utilized human subjects from Amazon's Mechanical Turk (Amazon [2010]) and used simulation to control conditions. The results suggest that the visual inference is comparable to conventional tests in a controlled conventional setting. This provides support for its appropriateness for testing in real exploratory situations where no conventional test exists. Interestingly, the power of a visual test increases with the number of observers engaged to evaluate lineups, and the pattern in results suggests that the power will provide results consistent with practical significance (Kirk [1996]).

*** Point 1: Finite comparison vs infinite comparison, measuring how different data plot is from null plots

In traditional hypothesis testing, the sampling distribution of a test statistic is functional and continuous. In the lineup protocol, although conceptually we may have an infinite collection of plots from the null distribution, in practice, we can only evaluate against a finite number of null plots. A human judge has a physical limit on the number of plots they can peruse. This poses one of the issues with using the lineup protocol. Figure 1 illustrates the difference. In traditional inference, the black curve represents the sampling distribution for the t-distribution under the null hypothesis, and the shaded red area shows the rejection region. In visual inference, let us consider that the black curve gives
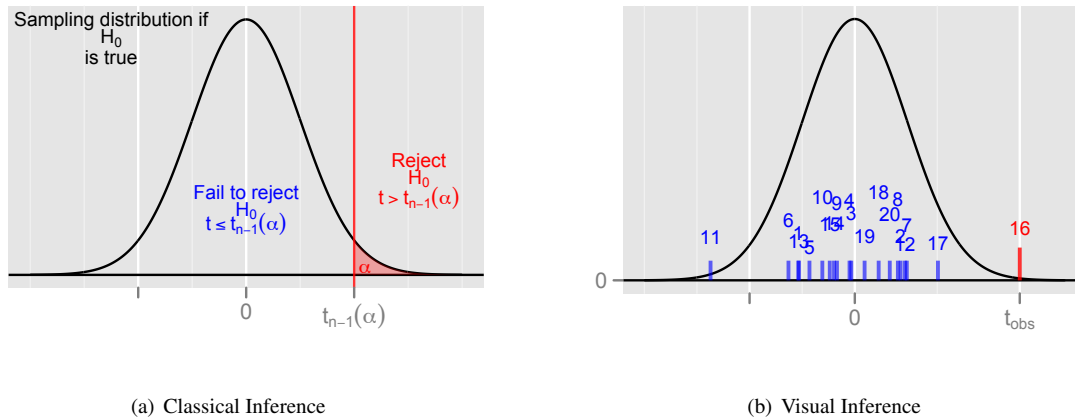
(a) Classical Inference (b) Visual Inference

Figure 1: Visual inference in relation to classical inference: (a) Decision region for classical inference for $H_0 : \mu = \mu_0$ vs $H_a : \mu > \mu_0$ (shaded in red) 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.

the sampling distribution although the sampling distribution is essentially a distribution of null plots. Although the test statistic is not numeric, the true data plot which is the test statistic is represented using red bar and the null plots that are drawn from the null distribution are the blue bars. Effectively, in visual inference the red line is compared only to these finite number of blue lines visually to make a decision, unlike classical inference where we look at the rejection region (Figure 1) to make decisions. Even though the data plot might be extreme, it is possible by randomly selecting from the null distribution, to obtain a null plot that is more extreme, as Tukey suggested [Fernholz, 2003]:

> "There [in Tukey's Data Analysis class] I discovered that [...] a random sample is indeed a "batch of values" which "fail to be utopian" most of the time."

***Point 2: Use metrics to ensure that a range of comparisons is made available to observers

This can be partially solved by having a large number of observers, who each evaluate lineups constructed with different null plots. Having some idea of the type of coverage of the sampling distribution that is provided by the lineups would be useful ahead of engaging observers and evaluating the lineups. Could we say that lineup X is expected to be "difficult" but lineup Y is expected to be "easy" then it may help in determining an appropriate number of observers? A difficult lineup is one where the data plot is similar to the null plots, and an easy lineup is where the data plot has some feature that makes it very different from the null plots. Being able to compute a plot to plot distance metric would be very helpful ahead of running a lineup protocol.

*** Point 3: Metrics might replace human observers, eventually, but as of now, human eye can still beat numbers for finding unexpected patterns. The lineup protocol gives us a chance to evaluate metrics to finding unexpected structures - check out the scagnostics literature

This is a two way process: As metrics are devised to measure the quality of a lineup, the lineup protocol also provides an opportunity to measure the performance of a metric. The human eye can detect patterns in a plot that just cannot be easily quantified numerically, which is why graphics provide an important tool for exploring data and finding the unexpected. Describing plots numerically, is something of an oxymoron, it cannot be universally done. An example in past work are scagnostics [Tukey, 1977, Wilkinson et al., 2005] which were developed to assess the different aspects of scattered points like outliers, shape, trend, density and coherence. If a scatterplot has just one of these structures the scagnostics are descriptive, however, they fail terribly if a plot contains more than one. The goal here is to find some distance measures that can provide some indications of the quality of a lineup, and then to use the results of observer evaluation to determine which metrics best match what people see.

Following up on choices, observers are asked to describe their reasoning. These reasons are used to obtain more information about the rejection: was it some nonlinear dependency, an outlier, clustering, that triggered the detection of the data plot? Good distance metrics may also help relate the descriptive words used with mathematically defined features.

The article is organized as follows. Section 2 discusses the null generating mechanisms. Section 3 defines the distance measures and discusses the choice of the measures. The distribution of the distance measures are studied in Section 4. Section 5 describes the effect of the plot type and the question of interest on the distance measure while Section 6 talks about the distance evaluations. In Section 7, the methods to select the number of bins for the binned distance is described. Section 8 presents a comparison of the distance measures to the performance of human subjects in several experiments conducted by Amazon's Mechanical Turk.

2 Null Generating Mechanism

The lineup protocol embeds the true data plot among a set of null plots. The method of obtaining the data for these null plots is called the null generating mechanism. The null hypothesis directly affects the choice of null generating method. Some examples are:

- Permutation: This is the most common approach, because it requires few assumptions and can be used in a variety of problems. Permutations can be used to break association between two or more variables, and thus is appropriate when the null hypothesis is that there is no association. Consider two variables X_1 and X_2. Either X_1 or X_2 is permuted keeping the other variable fixed. Any association between X_1 and X_2 is broken in the process. The marginal distribution of X_1 and X_2 remains the same while the joint distribution is altered. The method works in situations where one or both the variables are continuous or categorical. Let us consider a case where we have one categorical variable, say, Group and a continuous variable. Let us assume that the variable Group has two levels (say, A and B) and we want to test whether there is any significant difference between the two groups, i.e. $H_o : \mu_A = \mu_B$. To generate the null data, the values of the variable Group are permuted keeping the continuous variable fixed. If there is a difference between the two groups, this difference is broken by the permutation, and any difference observed in the permuted data is consistent with random variation.

- Simulation under the null model: Sometimes there is a model underlying the problem being studied. In this situation simulating from the model will be the null generating mechanism. Assuming that the null hypothesis is true, the model is fitted to the true data. The parameter estimates are obtained from the fitted model and then the data is generated using the parameter estimates. Let us consider that we are interested in testing whether there is any significant linear relationship between two continuous variables X_1 and X_2. Hence we test for $H_o : \beta_1 = 0$ versus $H_a : \beta_1 \neq 0$. Under the null hypothesis, we fit the following model to the data:

$$Y = \beta_0 + \varepsilon$$

where $\varepsilon \sim \text{Normal}(0, \sigma^2)$. The parameter estimates of β_0 and σ^2 are obtained and the null data is generated from $\text{Normal}(\widehat{\beta_0}, \widehat{\sigma}^2)$.

- Simulation from a specific distribution: When the null hypothesis is that the data comes from a specific distribution, this distribution can be used to simulate new samples that generate the null plots. The parameters for the null distribution are obtained from the estimates calculated using the data. For example, suppose we want to test whether data comes from a Normal distribution, $H_o : \text{data} \sim \text{Normal}$ vs. $H_A : \text{data} \nsim \text{Normal}$, then the null data are generated from the Normal distribution with mean and standard deviation equal to the estimated mean and standard deviation from the data.

Other null generating mechanisms might be utilized depending on the null hypothesis underlying a data plot.

3

3 Distance Measures

By calculating the "distance" between plots we may be able to determine if the the actual data plot is detectably different from the null plots, and also to better understand what aspect of the plot people use to make their choice. It is not an easy task to measure the difference between plots. Here we examine several possibilities that match up to plots that have been examined in a series of Amazon Turk studies conducted for other purposes.

A number of different distance measures were examined. We initially started with several candidates that generically measure plot distance, but since revised the candidate list to take into accoutn specific features of the plots that were used in the experiments.

Some of the distance metrics are generic and uses the raw data to calculate the distances. They do not consider any graphical element in the plot which may somehow affect the decision of the subjects in identifying the true plot. A regression line overlaid on a scatterplot may affect the decision of a plot. Same can be done when a boxplot is used to represent the data instead of a dot plot. The distance metrics should take into account the presence of these graphical elements. Distance metrics like distance based on regression line, distance based on boxplots are designed to address this issue.

For all of the distance measures below, let X denote the true dataset with one or two variables. Let Y denote the null dataset obtained from X using any of the above mentioned null generating mechanism.

- Binned Distance: 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. (i, j)-th cell represents the j-th bin of X_2 corresponding to the i-th bin of X_1. Let $C(X_1, X_2)$ be defined as a $p \times q$ matrix. Each (i, j)-th element of the matrix represents the number of points falling in the (i, j)-th cell, where $i = 1, \ldots, p$, $j = 1, \ldots, q$. The Binned distance is then defined as

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

This method also works for data with two categorical variables and data with one continuous and one categorical variable. For the categorical variable, it is sensible to pick the number of bins 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 exact cut-offs. This is particularly problematic for small number of points. As a remedy to that, we considered using kernel density estimates instead of point frequencies. But the results were not promising. Hausdorff distance (Huttenlocher et al. [1993]) was also studied as an option. Although the results were promising, Hausdorff distance was computationally intensive and hence was not considered.

The remaining distance measures are different from the ones above, in that they are used for specific plot types and cannot be used for any type of data. The following distance measures uses the graphical element to calculate the distances.

- Distance for univariate data: Let X be a continuous variable. Then the distance metric is given by

$$
d^2_{\mathsf{uni}}(X, Y) := ||m(X) - m(Y)||^2 = \sum_{i=1}^{4} ((m(X))_i - (m(Y))_i)^2
$$

where $m(.)$ is a vector of the mean, the standard deviation, the skewness and the kurtosis of the variable. This distance metric works for univariate distributions using only the graphical elements in the plot.

- Distance based on boxplots : Let X_1 be a categorical variable representing the groups in the data and X_2 be a continuous variable. Then the distance metric is given by

$$
d^2_{\mathsf{box}}(X, Y) := ||d_q(X) - d_q(Y)||^2 = \sum_{i=1}^{3} ((d_q(X))_i - (d_q(Y))_i)^2
$$

4
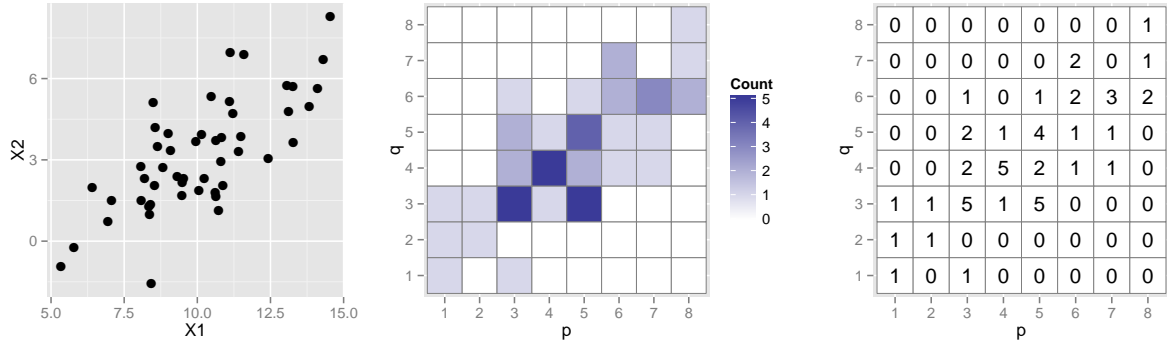
(a) Dataset X with two variables X_1 and X_2
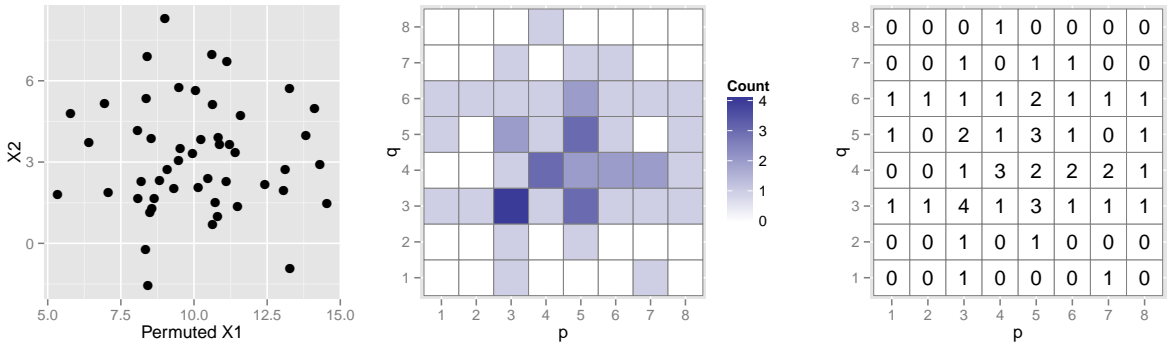
(b) Dataset Y with permuted X_1 and original X_2

Figure 2: Illustration of binned distance, for data with strong association (a), and the same data where one variable has been permuted (b). The scatterplot of the data is shown (left) along with the binned view of the data (center) and the number of points in each cell (right). Binned distance is the euclidean distance of these counts. The binned distance between these plots is 6.4807.

where $d_q(.)$ is a vector giving the absolute difference of the first quartile, median and the third quartile of X_2 between the two groups in X_1. This distance measure works specifically for the boxplots using only the graphical elements. This is based on the assumption that after the boxplots have already been constructed, the subjects only look at the difference in the boxes to make the distinction.

- Distance based on the regression line: Let X_1 and X_2 be two continuous variables. X_1 and X_2 are plotted in a scatterplot and assume that the scatterplot is binned vertically into b bins. In each vertical bin, a linear regression model is fitted and the regression coefficients i.e. the estimated intercept and the estimated slope are noted. The distance metric based on the regression coefficients is given by

$$d_{\text{reg}}^2(X,Y) := \text{tr}(B(X)-B(Y))'(B(X)-B(Y)) = \sum_{i=1}^{b}((b_0(X))_i - (b_0(Y))_i)^2 + \sum_{i=1}^{b}((b_1(X))_i - (b_1(Y))_i)^2$$

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 skewness of the data.

- Distance based on separation: 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 (Hennig [2010]). Two versions are used in this paper. Let us define,

(i) $s_g(.)$ be a vector of cluster wise minimum distance between a point in the cluster to the points in other clusters for g clusters. The distance metric based on separation is defined as

$$d_{\text{minsep}}^2(X,Y) := ||s_g(X) - s_g(Y)||^2 = \sum_{i=1}^{g}((s_g(X))_i - (s_g(Y))_i)^2$$

(ii) $s_g(.)$ be a vector of cluster wise average distances of all the points in the cluster to all point of other clusters for g clusters. The distance metric based on separation is defined as

$$d_{\text{avesep}}^2(X,Y) := ||s_g(X) - s_g(Y)||^2 = \sum_{i=1}^{g}((s_g(X))_i - (s_g(Y))_i)^2$$

Figure 3 illustrates the difference between the two methods of separation. Here the two methods are applied on two dimensional projections of a dataset with 60 dimensions and 30 observations with three classes. The same is done on the projections of the same dataset with permuted classes. The average separation calculates the euclidean distance between each point in one cluster to all the points in the other two clusters. For example, for the original data, the distance between the points in the green cluster and the other two clusters are calculated and then the average of these distances represent the average distance for the green cluster. Similarly, the average distances for the other two clusters are calculated. The minimum distance also calculates the euclidean distance between each point in one cluster to all the points in the other two clusters. But instead of taking the average, it looks at the minimum distance. For the original data, the minimum distance between any point in the green cluster and the other two clusters is the distance between the point and a point in the red cluster.

The last three distance measures are also dependent on the question of interest and hence can be changed accordingly. But, in general, this should not be a problem because the question which is typically asked is "Which plot among these is different ?".
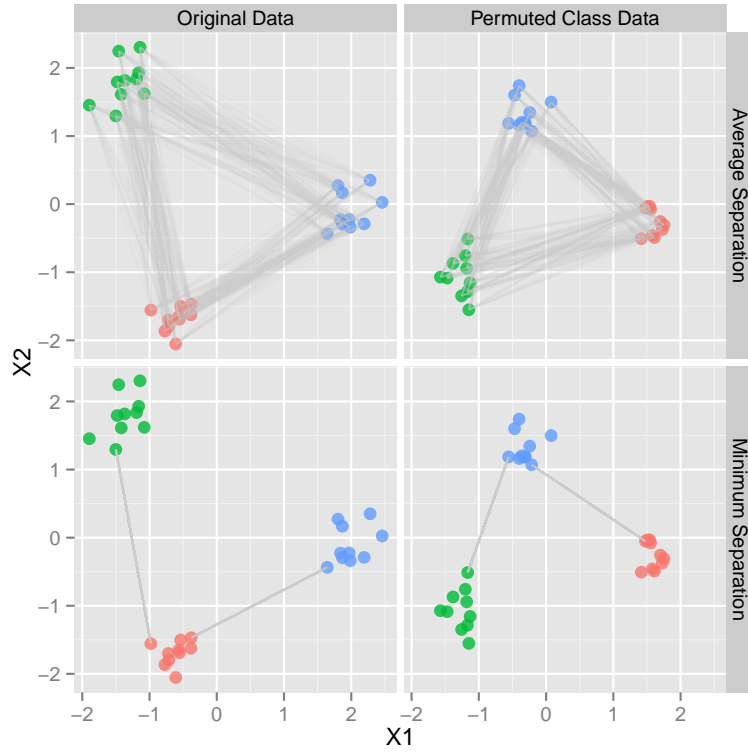
Figure 3: Two dimensional projections with 3 classes for a particular data and a data with classes being permuted. Two different separation distances are used. Average distance calculates the distance between all the points in a cluster to all the points in the other clusters and takes the average of these distances. The lines show the distance between the points. On the other hand, minimum separation calculates the minimum distance of the points in a cluster to all the points in the other clusters. The line shows the minimum distances.

4 Distance Metric Distribution

The empirical distribution of the distance measures is obtained by calculating the distances between the null plots among themselves. One null data is generated from the true data set using the null generating mechanism. Assuming this null data to be the "true" data set, a number of null data sets are obtained from this null data and the distances between these datasets are calculated. One single distance value is obtained by averaging all these distances. This process is repeated a large number of times, say, N where N is a large number of the order 10^3 or 10^4. Finally N mean distances or average distances are obtained which gives the empirical distribution of the distance.

The empirical distribution of the distance works as the t-distribution in the classical setting. In the classical setting, the test statistics follows a t-distribution under the null hypothesis. The observed test statistic is then compared to this distribution, as shown in Figure 1. In visual inference, the mean distances of the null plots gives the empirical distribution. The mean distance of the true plot from the null plots in the lineup acts as the observed test statistic. Unlike the t-distribution, the empirical distribution is generally skewed.

The mean distance between the true plot and the null plots in a lineup of size $m = 20$ is calculated by averaging over the distances between the true plot and each of the $(m-1)$ null plots. The mean distances for the $(m-1)$ null plots in the lineup are calculated by taking the mean of the distances of the particular null plot and the other $(m-2)$ null plots. The mean distances for the true dataset and the null datasets are plotted on the empirical distribution. If the mean distance of the true plot is larger than any of the null plots, the lineup would be regarded as "easy". Otherwise, it is a "difficult" lineup.

The empirical distribution of the distance based on regression is shown in Figure 4. To generate this distribution, $N = 1000$ and $m = 20$ was used. Figure 4(a) 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 4(b) shows the empirical distribution of the distance with the mean distances for the true plot (in orange) and the null plots (in black) for the particular. The true plot is easy to be identified in the lineup (Figure 4(a)). It can also be seen in Figure 4(b) as the orange line is extreme compared to the black lines.

(a)   (b)
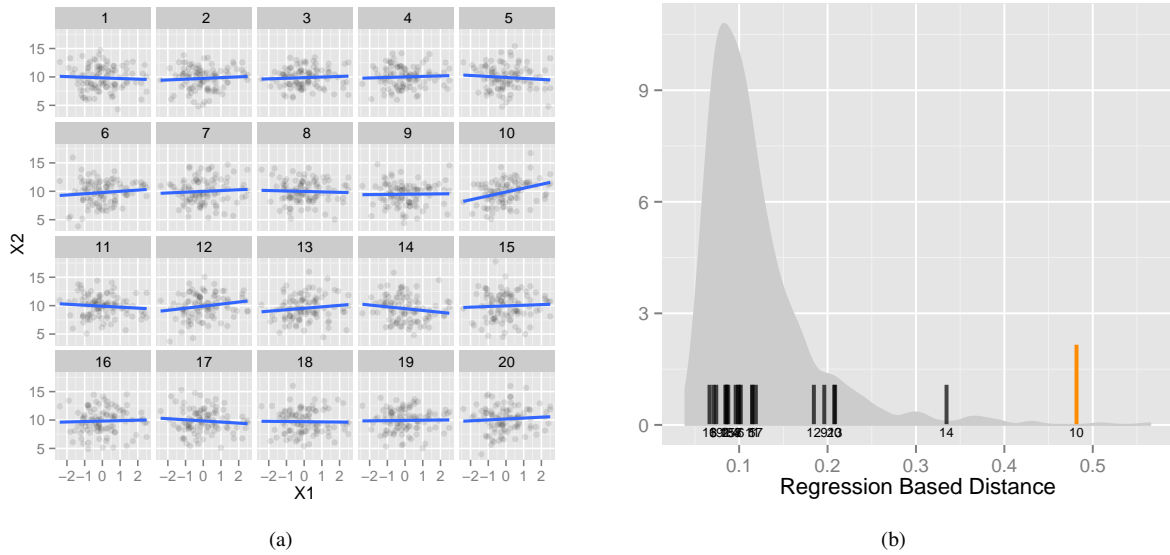
Figure 4: (a) Lineup Plot ($m = 20$) for testing whether there exists a significant linear relationship between X_1 and X_2. The 19 null plots are obtained by simulating from the null model. (b) The chart on the right shows the empirical distribution of the distance based on regression parameters. The distance of the true plot is shown in orange while the distance for the null plots are shown in black.

Figure 5(a) shows the lineup plot for $m = 20$ for testing whether there exists a significant difference between the two groups A and B. The 19 null plots are generated by permuting the group variable keeping the other variable fixed.
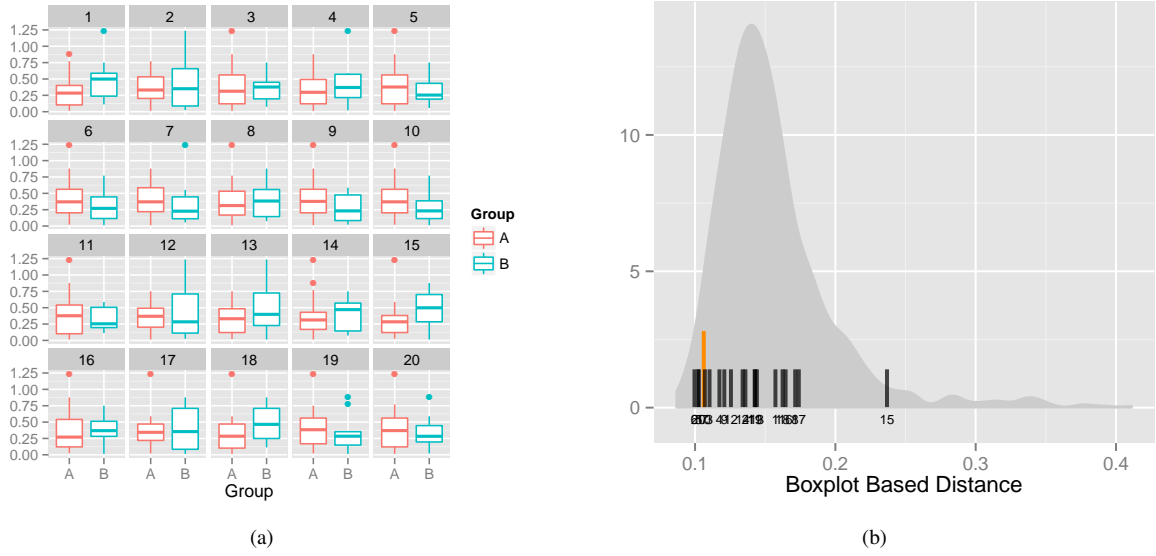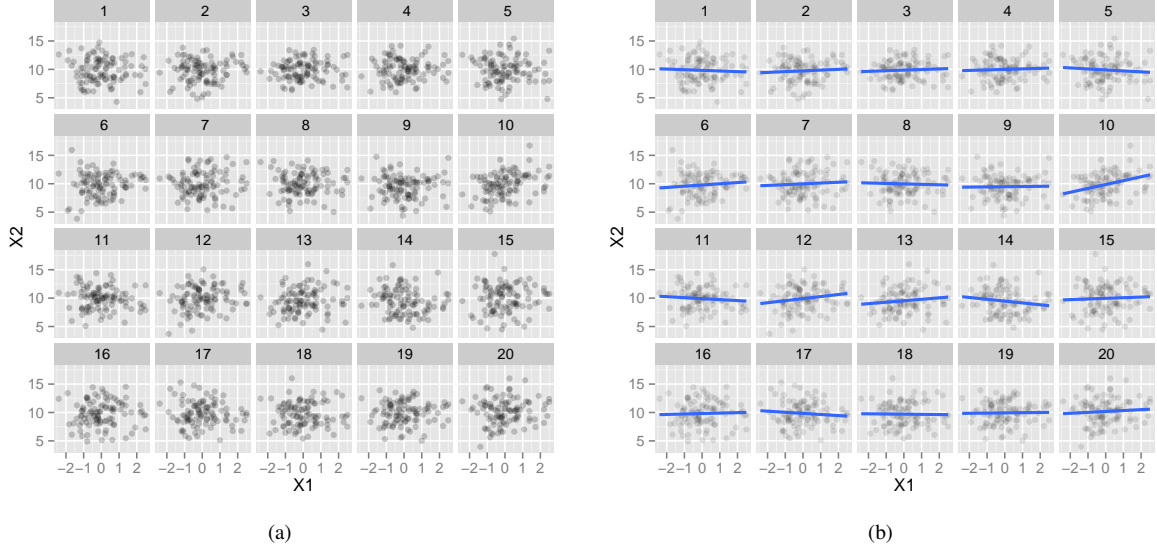
Figure 5: (a) Lineup Plot ($m = 20$) for testing whether there exists a significant difference between the two groups A and B. The 19 null plots are obtained by permuting the group variable while keeping the continuous variable fixed. (b) The chart on the right shows the empirical distribution of the distance based on boxplots. The distance of the true plot is shown in orange while the distance for the null plots are shown in black.

Figure 5(b) 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). The true plot is hard to be identified from the lineup which is also evident in the distribution since many black lines are to the right of the orange line.

5 Effect of Plot Type and Question of Interest

Previous studies have suggested that the type of plot used in the lineup have an effect on the response of the subjects [Zhao et al., 2012]. For example the subjects find it easier to identify the true plot for a large sample data when a box plot is used in the lineup instead of a dot plot. Similarly the distance metric should also be altered according to the plot type. The distance metric should account for the additional information provided by the graphical elements in the lineup. The graphical elements, like the presence of a box or a regression line overlaid on a scatterplot may influence the response of the subject. Figure 5 illustrates this idea.

Figure 6(a) shows a lineup of scatterplots with 100 points between two variables X_1 and X_2. Figure 6(b), on the other hand, gives a lineup of the same scatterplots with the regression line overlaid. Showing Figure 6(a), if the subjects are asked to identify the plot which has the steepest slope, then the subjects probably will face some difficulty in identifying the true plot. But in Figure 6(b), the regression line overlaid makes it easier for the subjects to identify the true plot. A different distance metric should be used in each case to correctly measure the quality of the lineup.

The question asked to the subjects plays an important role to identify the true plot in the lineup. A minor change in the question can change the response of the subject. In Figure 5(a), if the subjects are asked to identify the plot in which the green group has a larger vertical difference than the red group, the subjects should pick Plot 6. If the subjects are asked which plot has the largest vertical difference between the two groups, the subjects should pick Plot 15. A distance metric should also take into account the question of interest. But, in general, the question of interest is which plot among these is different.

Figure 6: (a) Lineup Plot ($m = 20$) for testing whether there exists a significant difference between the two groups. The 19 null plots are obtained by permuting the group variable while keeping the other variable fixed. (b) The chart on the right shows the empirical distribution of the distance based on boxplots. The distance of the true plot is shown in orange while the distance for the null plots are shown in black.

6 Metric Evaluation

For a lineup of size $m = 20$, the distance for the true plot is compared to the 19 null plots. This comparison can sometimes complicate things. A logical solution can be to look at one statistic for one lineup. Such a statistic can be defined as the difference between the mean distance of the true plot and maximum of the mean distances for the null plots. Hence we define,

1. Difference: the difference between the mean distance for the true plot and the maximum of the mean distances for the null plots. Mathematically,

$$\delta_{\text{lineup}} = \bar{d}_{\text{true}} - \max_{j} \bar{d}_{\text{null}_j}$$

   for $j = 1, \ldots, (m-1)$. A positive difference would indicate that the mean distance of the true plot is greater than the maximum of the mean distances of the null plots. Hence the true plot is extreme compared to all the null plots. Similarly a negative difference indicates that there is at least one null plot which is extreme compared to the true plot based on the distance.

   The issue with this statistic is that δ_{lineup} indicates an "easy" or "difficult" lineup only on the basis of whether it is positive or negative, although it may be really close to 0. The statistic does not imply how many null plots are more extreme than the true plot. So we define,

2. Larger than the true plot: the number of null plots which have larger mean distances than the mean distance of the true plot is noted. Mathematically,

$$\gamma_{\text{lineup}} = \sum_{j=1}^{m-1} a_j$$

   where

$$a_j = \begin{cases} 1 & \text{if } \bar{d}_{\text{null}_j} > \bar{d}_{\text{true}}, \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

10

γ_{lineup} takes all values between 0 and $(m - 1)$. A large value of this measure would indicate that there are a number of null plots more extreme than the true plot and hence it is hard to identify the true plot in the lineup.

7 Selection of the Number of Bins

Binned distance works for any type of data and for any null generating mechanism. It does not take into account the graphical elements in the plot, and the raw data is used. 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 choice of number of bins or the bin size highly affects the distance. A wrong choice may produce erroneous or conflicting results. Hence the choice of the number of bins is important.

The choice of number of bins or bin sizes is investigated with different types of data. Different null generating mechanisms are also used for the same data type. Null datasets are obtained for a true data using a null generating mechanism and hence a lineup is constructed. Mean binned distance is 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 δ_{lineup} is calculated for each combination. Table 1 and Table 2 shows the type of data, the observed plot, the null generating mechanism, a typical null plot, the difference δ_{lineup} and also the maximum value of δ_{lineup}, the x-bin and y-bin for which the maximum was obtained. The minimum δ_{lineup} is also reported to get an idea of the range of values.

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 mechanisms are also selected for the same reason. In Table 1 the first four observed data plots corresponds to the datasets described by Francis Anscombe in [Anscombe, 1972] but with large number of data points. Although the datasets have the same pattern, the datasets do not follow the properties of Anscombe's quartet. The fifth dataset is a data with 3 distinct clusters. In Table 2, the first dataset shows a categorical data. The second and the third data are non-linear and linear association with the presence of outliers. The fourth and fifth datasets are the residual plots with curved pattern and non-constant variance pattern. The sixth data is a spiral data while the seventh one is a data with contamination.

The differences, δ_{lineup}, are represented in a tile plot where each tile gives the difference for each combination. The dark blue shows higher values while the white shows lower values. It can be seen that the 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 relationship, small number of bins should be preferred over large number of bins. Also when outlier is present in the data, larger number of bins is preferred at least in one axis.

It is important to mention at this point that Table 1 and Table 2 is not meant to provide any guidelines for the selection of number of bins. The Tables only show that the 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 data before using the binned distance.

8 Results

The performance of the distance metrics was evaluated with comparing the distances with the response of the subjects. A number of experiments were done in Amazon Mechanical Turk [Amazon, 2010]. Subjects were recruited through Amazon Mechanical Turk [Amazon, 2010] and were shown a sequence of lineups. In each experiment, they were asked specific questions. Their responses were recorded along with other demographic informations. The details about the design of experiments can be found in Majumder et al. [2013] and Roy Chowdhury et al. [2013].

8.1 Turk Experiment – Side by Side Boxplots

In this experiment, all the lineups generated had a side by side boxplot as the test statistic. Assuming that the null hypothesis is true, the null plots were generated by assuming that there is no difference between the two distributions. The subjects were shown a few lineups and were asked to identify the plot which has the largest vertical difference between group 1 and group 2. Figure 7 gives such a lineup.

Table 1: Preferable number of bins for different types of observed data to calculate the binned distance.
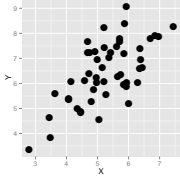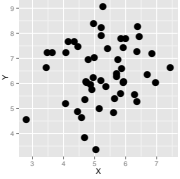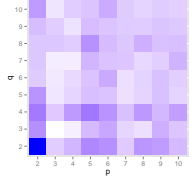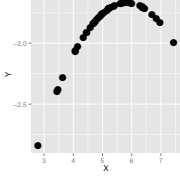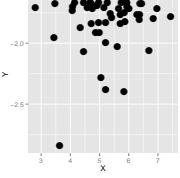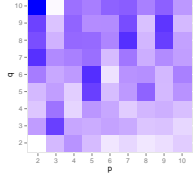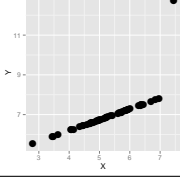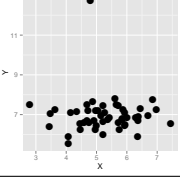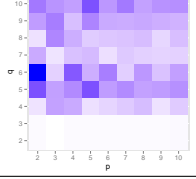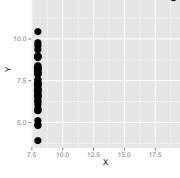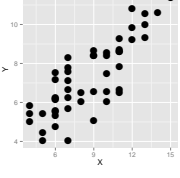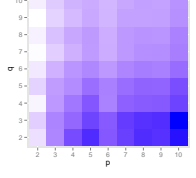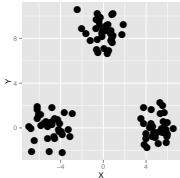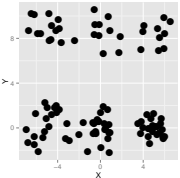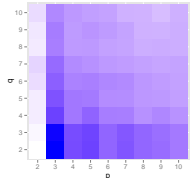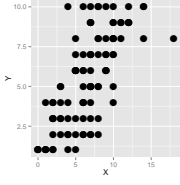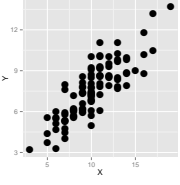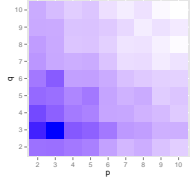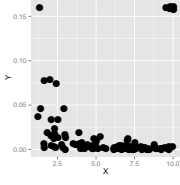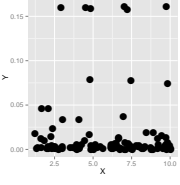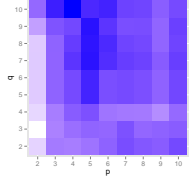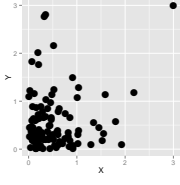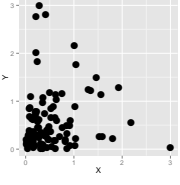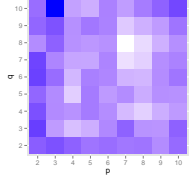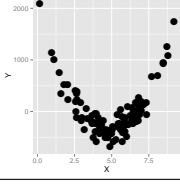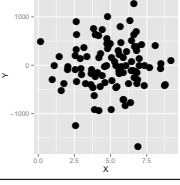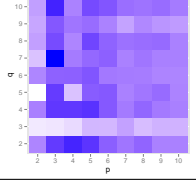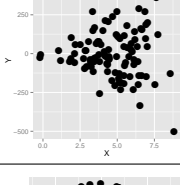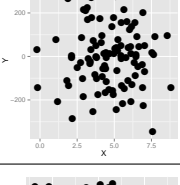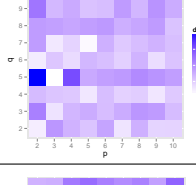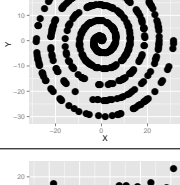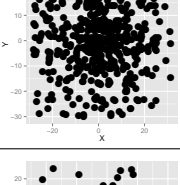
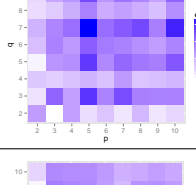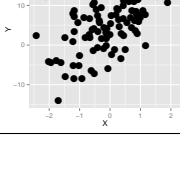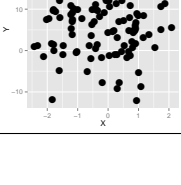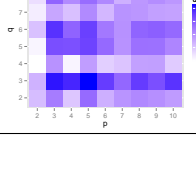| Type of Data	Observed Plot	Null Generating Mechanism	A typical null plot	Difference	(x-bin, y-bin, Max; Min)
Linear association		Permutation			(2, 2, 5.7 ; - 2.5)
Nonlinear relationship		Permutation			(2, 10, 6.2 ; - 0.0)
Linear relation with outliers		Permutation			(2, 6, 16.7 ; - 0.4)
Same values with one outlier		Simulation from a *Poi(9)* distribution			(10, 3, 34.3 ; - 0.1)
Clusters		Permutation			(3, 2, 27.6 ; - 5.7)

12

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

| Type of Data	Observed Plot	Null Generating Mechanism	A typical null plot	Difference	(x-bin, y-bin, Max; Min)
Categorical		Simulation from a Normal distribution			(3, 3, 30.7; 6.2)
Nonlinear relation with outliers		Permutation			(4, 10, 3.9; -3.4)
Linear relation-ship with outlier		Permutation			(3, 10, 0.3; -7.1)
Residual Plot		Simulation from the null model			(3, 7, 17.8; -4.5)
Residual Plot		Simulation from the null model			(2, 5, 4.8; -4.4)
Spiral data		Permutation			(5, 7, 23.6; -11.9)
Contaminated data		Permutation			(5, 3, 8.1; -2.5)
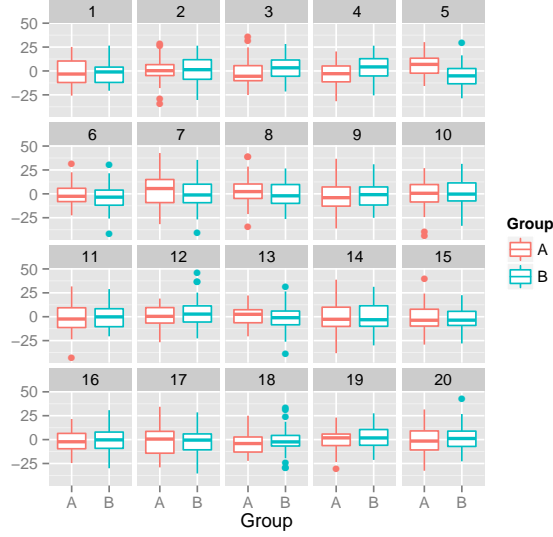
Figure 7: An example lineup from Turk Experiment 1. The lineup has $m = 20$ plots of which one is the observed data plot and the remaining $m - 1$ are the null plots generated assuming that the null hypothesis is true. Subjects were asked to identify the plot which has the largest vertical difference between the two groups. Can you identify the observed plot ?

The response of the subjects were noted and the proportion of correct response was calculated for each lineup. The distances between the plots in each lineup were computed using both the distance based on boxplots (d_{box}) and the binned distance (d_{bin}). The mean distance for the true plot and the null plots were calculated and δ_{lineup} and γ_{lineup} are obtained. The proportion of correct response was plotted against each of the two statistics. Figure 8 shows the detection rate against the difference for d_{box} and d_{bin} and the number of null plots greater than the observed plot for the two distance measures.

In Figure 8, the detection rate is plotted against the difference. The red vertical line represents difference equal to 0 indicating that the mean distance of the true plot is equal to the maximum of the mean distance of the null plots i.e. the mean distance of the true plot is equal to at least one of the mean distance of the null plots. It can be seen that as the difference increases, the detection rate increases. So the subjects do better in the easier lineups than the hard ones. The binned distance was calculated using 8 bins on both the axes. Figure 8 also shows the relation between detection rate and the number of null plots larger than the true plot. 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. It is interesting to see that the subjects can pick the observed plot with one or two extreme null plots.

Though the distance based on the boxplots works better, the binned distance does a decent job in this case. According to the binned distance, there are a few lineups which has a negative difference but the proportion correct is above 60%, which can be also be seen in Figure 8. It should be noted that the binned distance does not take into account the graphical elements of the plot (e.g. boxplot) and calculates the distance solely based on the data. So an outlier may have a huge effect on the binned distance but does not effect the distance based on the boxplots. Hence it is advisable to use a distance based on the graphical elements since that is exactly what the subjects look at in the lineup.

The time taken to respond by the subjects is another measure of difficulty of the lineups. Due the presence of some huge outliers, the mean time taken by the subjects for each lineup is looked at and plotted against the difference for both the distance measures. Figure 9 shows the plots. It can be clearly seen that when the difference is below 0, there is no real trend in the mean time and there is a huge variability, indicating that the time taken depends on the subjects. But when the difference is above 0, the mean time decreases rapidly as the difference increases. Hence the subjects can pick the true plot quickly if the true plot is extreme compared to the null plots.

It can be noticed in Figure 8 that for some of the lineups, the detection rate is high but the difference using distance
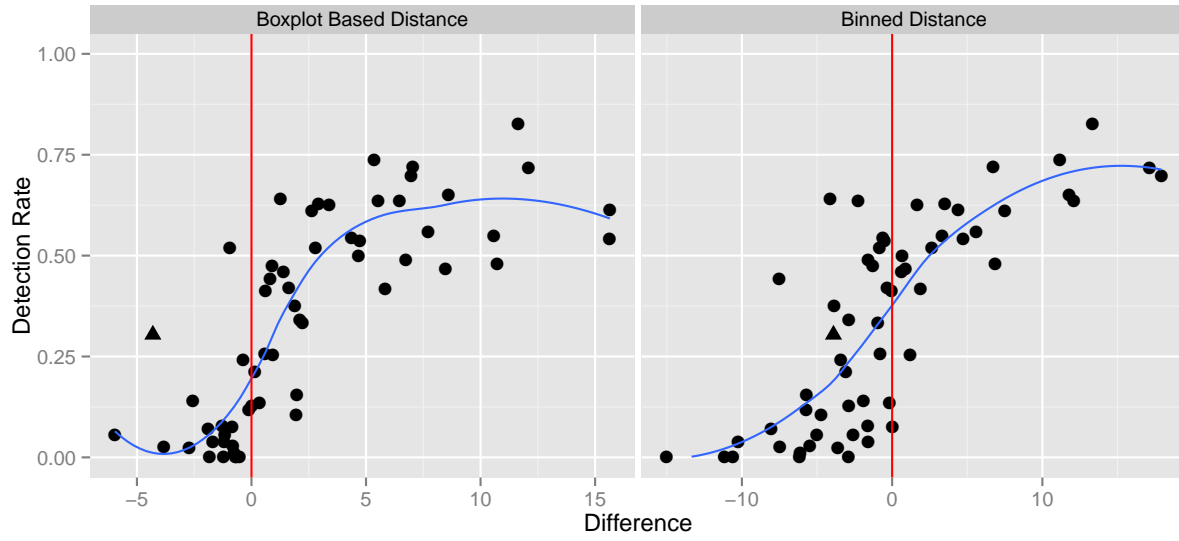
(a)
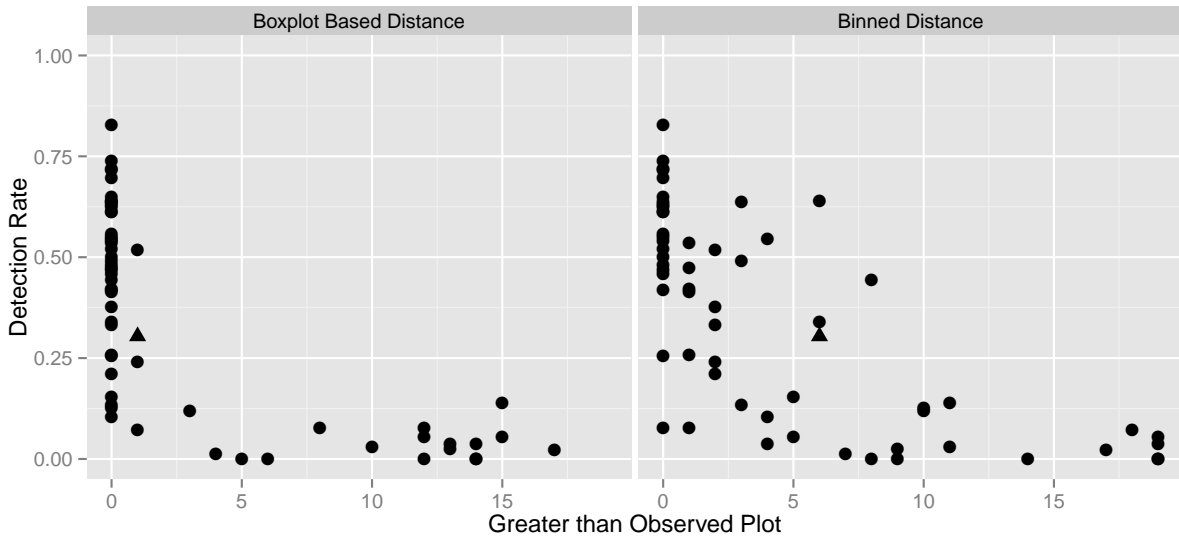
(b)

Figure 8: Plot showing the detection rate against the difference and the number of null plots greater than the observed plot based on the boxplot distance and the binned distance . In (a) the vertical line represents the difference equal to 0 when there is at least one null plot similar to the observed plot. The detection rate increases with the difference. In (b) the detection rate decreases as the number of null plots greater than the observed plot increases.
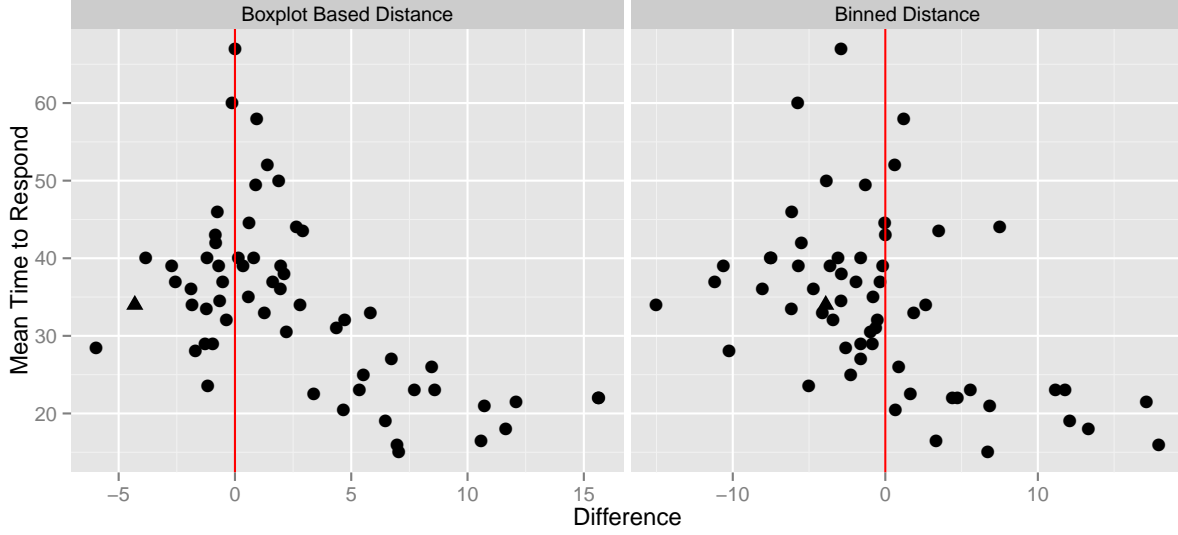
Figure 9: Plot showing the mean time to respond by the subjects against the difference based on the boxplot distance and binned distance. The vertical line represents the difference equal to 0 when there is at least one null plot similar to the observed plot. The mean time decreases with the difference.

metric is negative suggesting that the lineup is difficult. One such lineup is marked using a triangle in Figure 8. It would be interesting to look into the lineup closely to identify what made the people pick the actual plot as different. Figure 10 shows the lineup and the distribution of the distance metrics.

The lineup in Figure 10 is a lineup of side-by-side boxplots. The observed data plot is Plot 20 but there are other candidates who can be picked easily. Plot 19 and Plot 16 seems to have large differences between the quartiles. Specifically in Plot 16, the difference between the first quartiles for the two groups is very large but the differences between the medians and the third quartiles are small. The huge difference of the first quartiles may have affected the huge mean distance of Plot 16 from all the other plots.

8.2   Turk Experiment – Scatterplots with an Overlaid Regression Line

In this experiment, the test statistic is a scatterplot with the regression line overlaid. Assuming that the null hypothesis is true, the null plots are generated by assuming that there is no significant linear relationship between the two variables. The subjects were shown a few lineups and were asked to identify the plot which has the steepest slope. Figure 11 gives such a lineup.

The distances between the plots in this experiment were computed using both the distance based on regression line (d_{reg}) and the binned distance (d_{bin}) with a small number of bins. The proportion of correct response for each lineup was calculated from the response of the subjects and plotted against δ_{lineup} and γ_{lineup}. Figure 12 shows the results for the distance based on the regression line and the binned distance against δ_{lineup}.

Figure 12 shows the detection rate against the difference. The vertical line represents difference equal to 0. It can be seen that as the difference increases, the detection rate increases. So the subjects do better in the easier lineups than the hard ones. The distance based on regression works well in capturing the complexity of the lineups. The difference is positive for lineups with large differences and negative for lineups with small difference. A few lineups have difference close to zero for which the detection rate is close to 50%.

But the binned distance fails terribly. Although the detection rate increases with difference, the detection rate is high for values with negative difference. This is a classic scenario where a graphical element affects the response. The presence of the overlaid regression line on almost transparent points of the scatterplot affected the response of the subjects. One other reason may be the use of the same number of bins (2×2) in this case for all the lineups. The
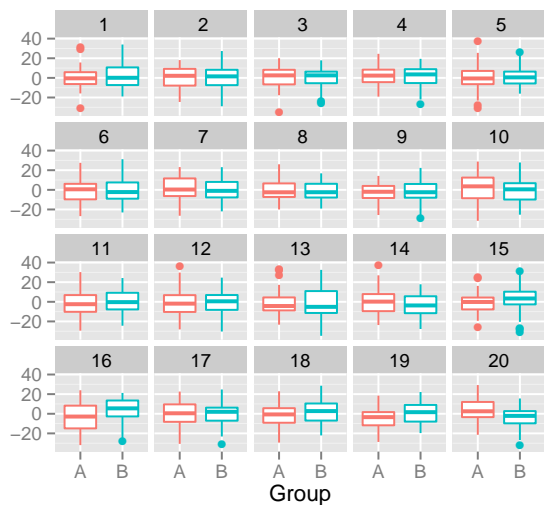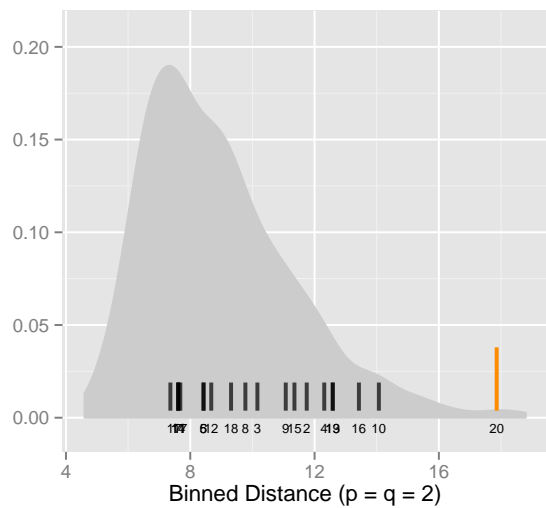
(a)
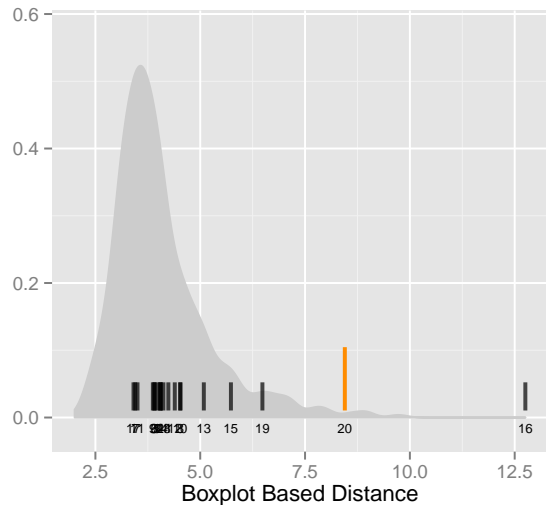
(b)
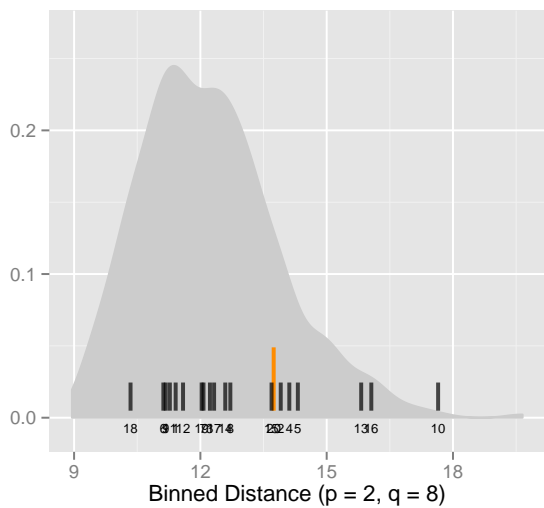
(c)

(d)

Figure 10: Illustration of the behavior of the different distance metrics. The lineup is shown in (a) and the distributions of different distance 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 in both axes in (d). The lineup corresponds to the point marked with a triangle in difference vs. detection rate plot in Figure 8.
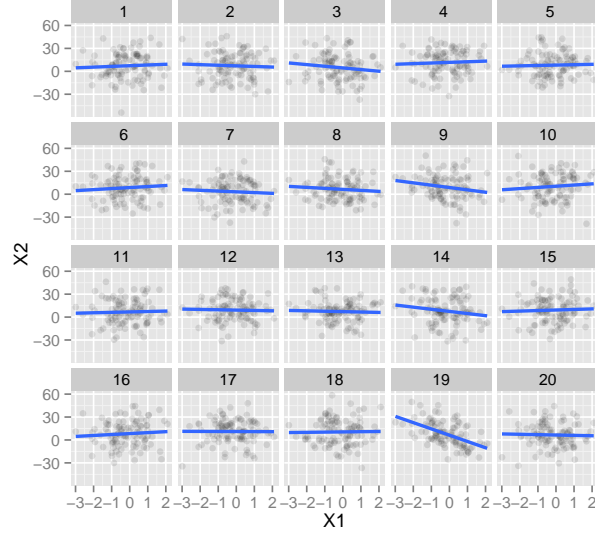
17

Figure 11: An example lineup from Turk Experiment 2. In this lineup, one of the plots is the observed plot and the other 19 plots are the null plots generated assuming that the null hypothesis $H_o : \beta = 0$ is true. Subjects were asked to identify the plot with the steepest slope. Can you identify the observed plot ?

relationship may improve if different number of bins can be used for different lineups.

Figure 12 also shows the detection rate against the number of null plots greater than the observed plot. As the number of plots greater than the observed plot increases, the detection rate decreases. Hence 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, almost all the subjects identify the observed plot although there is one more extreme null plot. Though from Figure 12, it can be seen that the extremeness is marginal in most cases.

Figure 8.2 shows the relationship between the mean time taken to respond and the difference for both the distances. It can be clearly seen that there is a strong negative association showing that as the difference increases, the subjects take less time to respond. Also the variability of the mean time is higher for smaller difference. In case of binned distance, the relationship is negative though the variability is higher for the above mentioned reasons.

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 12) which had a negative difference although people identified the true plot with reasonable success. Figure 14 shows the lineup and the distributions of different distance metrics.

The lineup in Figure 14 is a difficult one as suggested by the distribution of the distance metrics based on regression. Although around 28% of the people identified the true plot correctly, the conventional p-value for testing the slope equal to 0 is 0.085, which shows that the relationship is not significant. The binned distance with 2 bins on each axes also shows the same. 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 true plot as different from the others.

8.3 Turk Experiment – Large p, Small n Data

The motivation behind this experiment is to study the effect of large dimensions in a data with complete noise and some real separation. Data was simulated with different dimensions and fixed sample size. Data was divided into two or three groups. A projection pursuit with Penalized Discriminant Analysis Index was used and the one and two dimensional projections were obtained. The one or two dimensional projections were then plotted which resulted in the observed data plot. To generate the null data, the group variable in the data was permuted and the projection pursuit was applied. The subjects were shown these lineups and were asked to identify the plot with the most separated colored groups. Figure 15 gives an example of such a lineup with two dimensional projections with 3 colored groups.
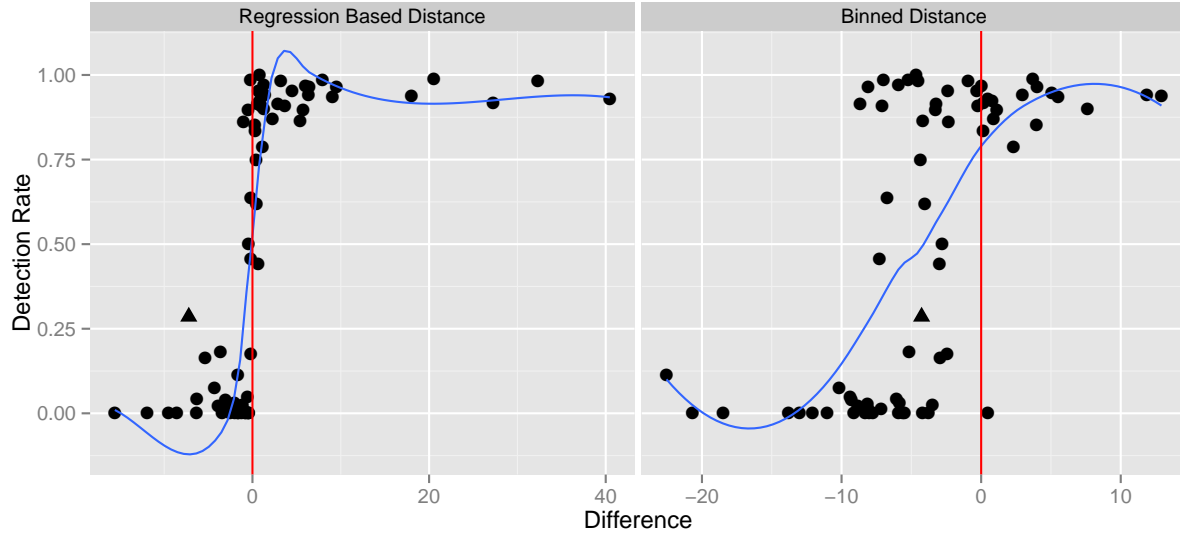
(a)
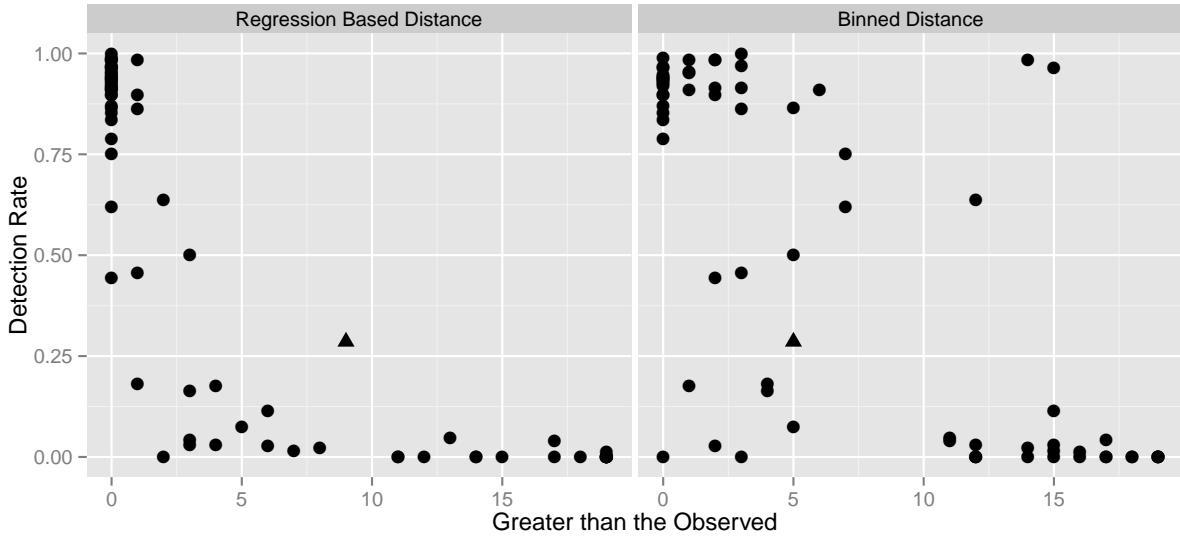
(b)

Figure 12: Plot showing the detection rate against the difference and the number of null plots greater than the observed based on the regression distance and based on the binned distance. In (a) the vertical line represents the difference equal to 0 when there is at least one null plot similar to the observed plot. The detection rate increases with the difference. In (b) the detection rate decreases as the number of null plots greater than the observed plot increases.
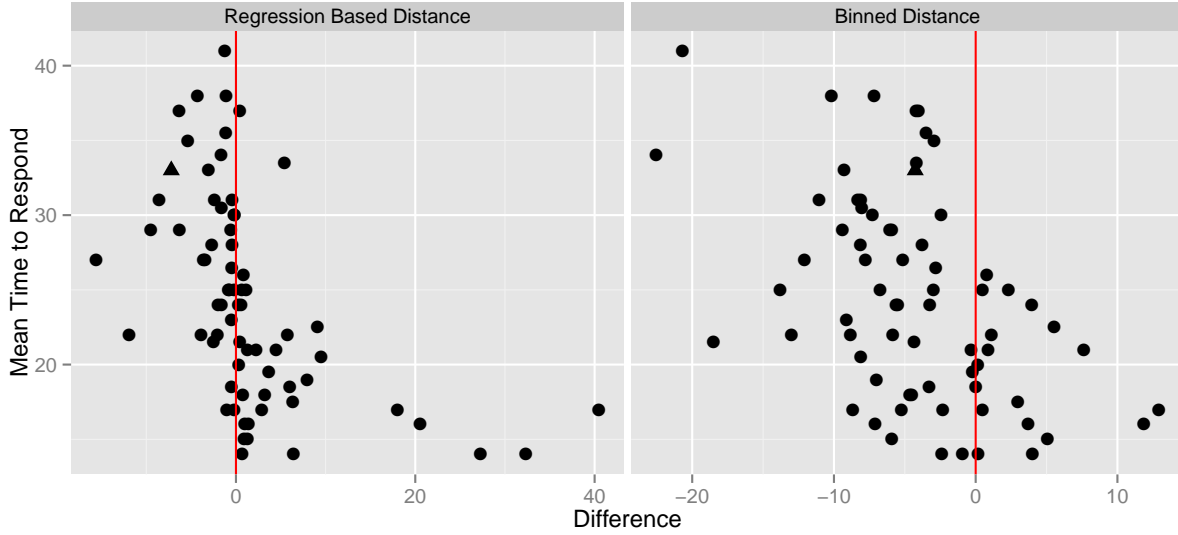
Figure 13: Plot showing the mean time to respond by the subjects against the difference based on the regression distance and binned distance. The vertical line represents the difference equal to 0 when there is at least one null plot similar to the observed plot. The mean time decreases with the difference.

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 δ_{lineup} and γ_{lineup} for both the distances. Figure 16 shows the results.

In Figure 16, the detection rate is plotted against the difference for distance based on minimum separation, average separation and the binned distance. The red vertical line shows difference equal to 0. It can be seen that as the difference increases, the detection rate increases and both the distances do a good job in capturing the response of the subjects. 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 17 shows the relationship between the mean time taken to respond and the difference for the three different distances. It can be clearly seen that there is a strong negative association showing that as the difference increases, the subjects take lesser time to respond. Also the variability of the mean time is higher for smaller difference. In case of binned distance, the relationship is negative though the variability is higher for all differences.

Figure 18 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. The null plots are obtained by permuting the group variable and plotting the two dimensional projections obtained from a projection pursuit with PDA index. Since the true plot has real separation, it is expected that the 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 identifies 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. The presence of an outlier point in the opposite side of the other clusters affects this distance considerably. On the other hand, the minimum separation looks at the minimum of the distances. Hence it is not affected by the outlier point.
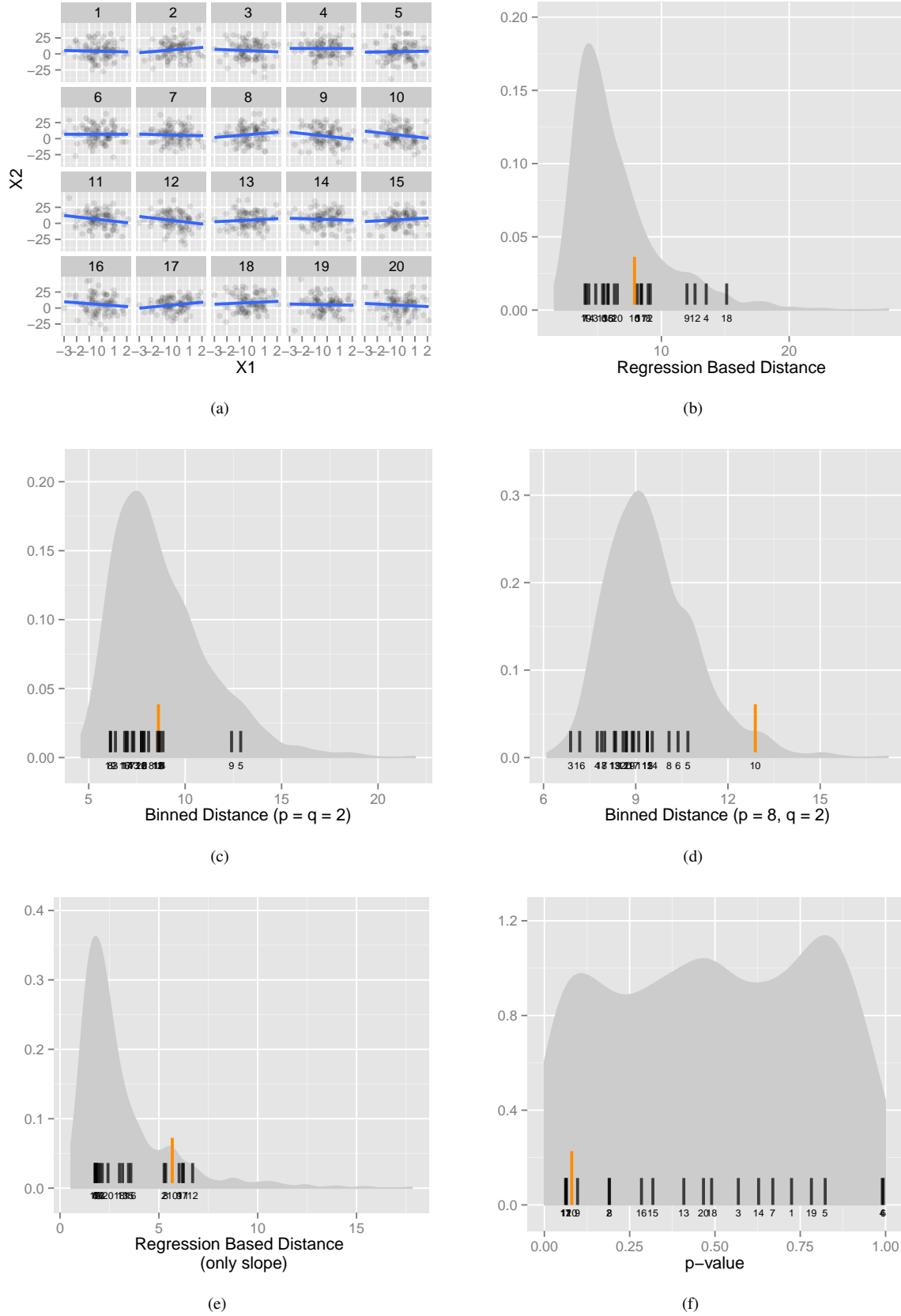
Figure 14: Illustration of the behavior of different distance metrics. The lineup is shown in (a) and the distributions of different distance metrics using this lineup is shown in the other plots ((b) - (e)): regression based distance in (b), binned distance with 2 bins on each axes in (c), binned distance with 8 and 2 bins in x and y axis respectively in (d) and regression distance based only on slope in (e). In (f), the distribution of the conventional p-values are plotted with p-values for the lineups marked on the distribution. The lineup corresponds to the point marked with a triangle in difference vs. detection rate plot in Figure 12.
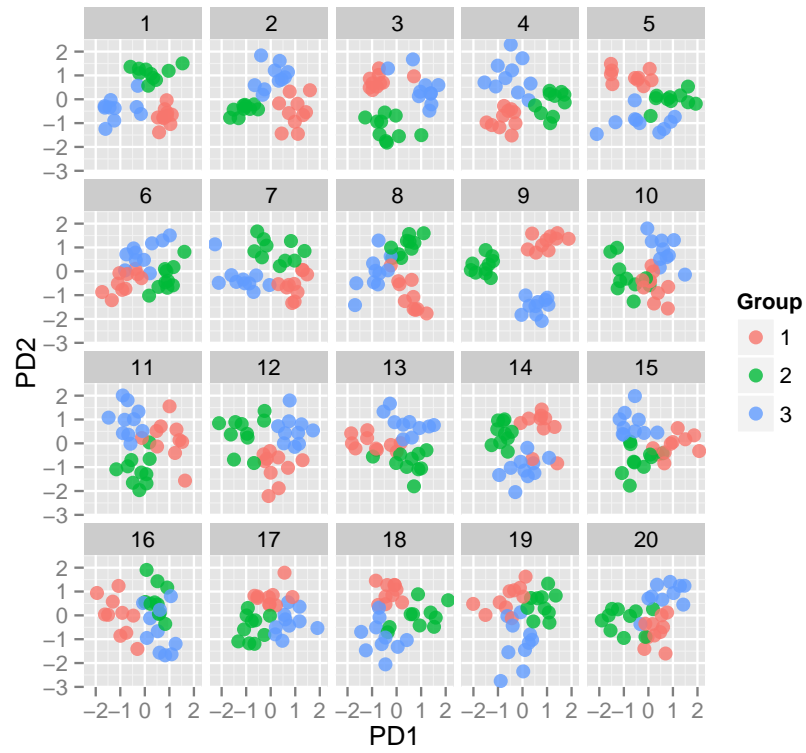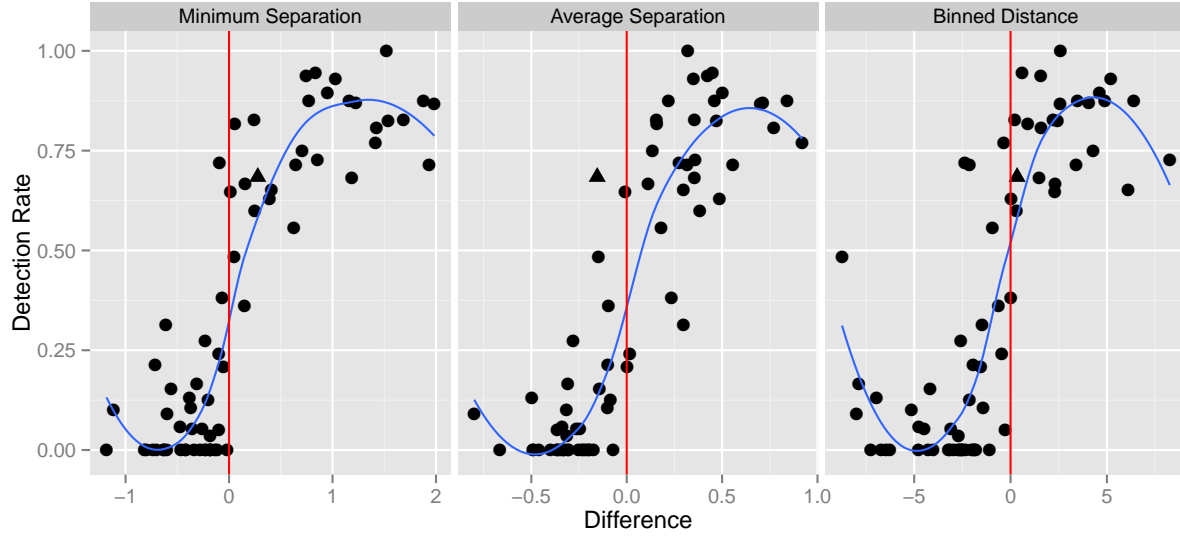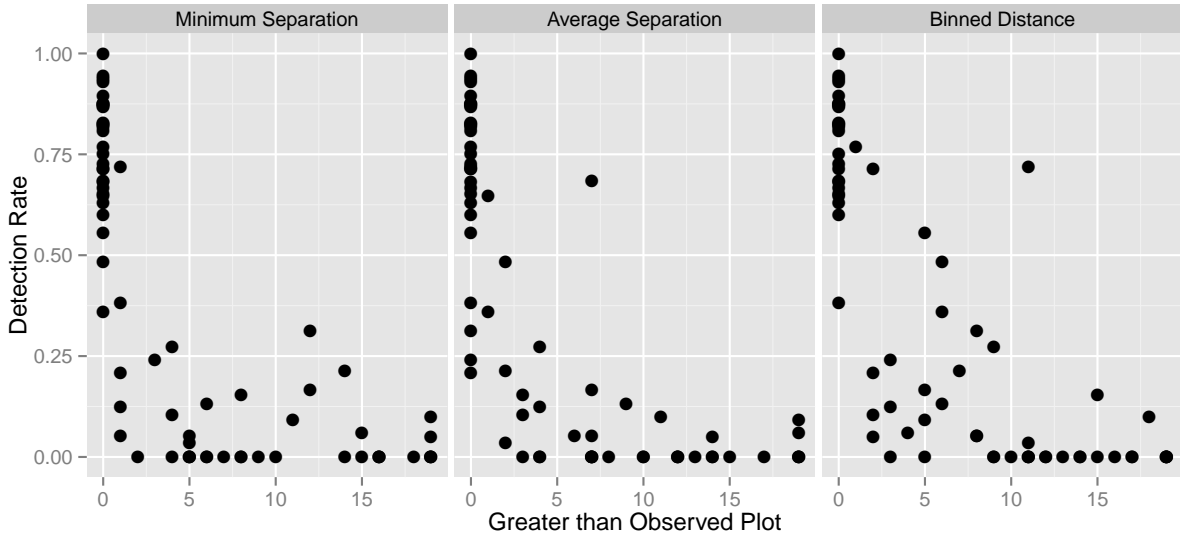
Figure 15: An example lineup from Large p, Small n Turk Experiment. Here two dimensional projections of the PDA index are plotted for a data with separation having $p = 20$ dimensions and $n = 30$ observations. The subjects were asked to identify the plot with the most separated colors. Can you identify the observed data plot?

(a)

(b)

Figure 16: Plot showing the detection rate in (a) and the number of plots greater than the observed in (b) against the difference based on the minimum separation, 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 detection rate increases with the difference. As the number of plots greater than the observed increases, the detection rate decreases.
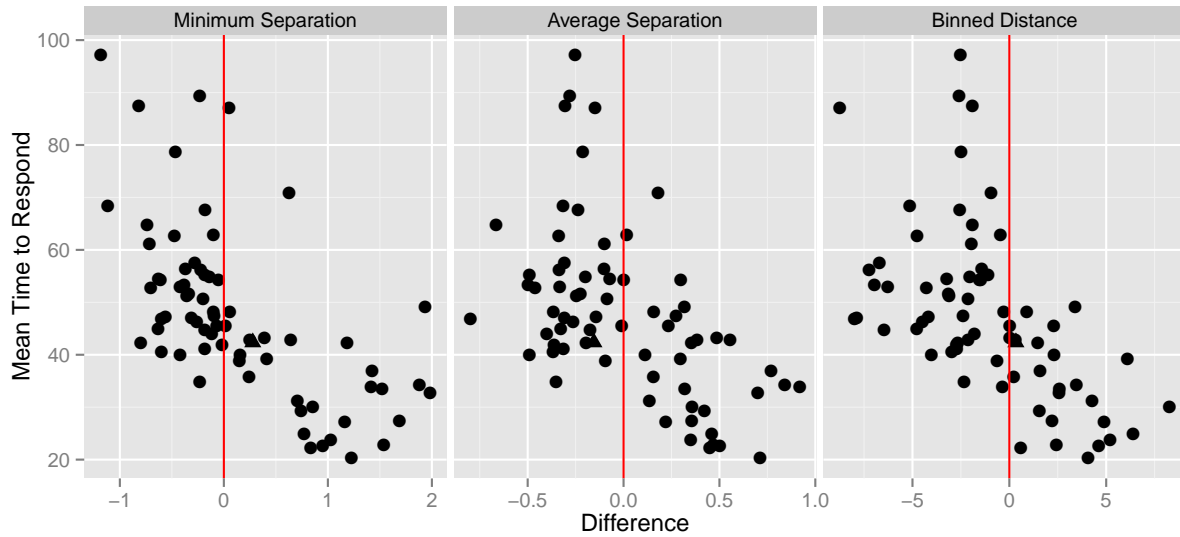
Figure 17: Plot showing the mean time to respond by the subjects against the difference based on the minimum separation distance, average separation and binned distance. The vertical line represents the difference equal to 0 when there is at least one null plot similar to the observed plot. The mean time decreases as the difference increases.

9 Conclusion

1. Distance metrics are compared to the response of human subjects on lineups. They are comparable to a certain extent except in certain situations where they disagree. There seems to be various reasons behind the disagreement. When people look at a lineup, they may identify a plot as different from the others due to various reasons. But the distance metrics are constructed such that it takes into specific properties of the plot.

2. Distance metrics can be used to measure the quality of a lineup before showing the lineups to human subjects. Hence the distance metrics allows us to provide a range of lineups to the human subjects to evaluate.

3. In classical inference, the test statistic under null hypothesis follows a certain distribution. Similarly the null plots in visual inference can also be assumed to be random samples from a sampling distribution. Though theoretically this is true, practically it is impossible to investigate such a distribution. The distribution of the distance metrics approximates such a sampling distribution for a given distance metric. The value of the distance metric for the actual plot can be compared to all the other plots using such a distribution.

4. The reason of choice can provide a way of evaluating the performance of a distance metric. For example, for a lineup of scatterplots with regression line overlaid, if the choice of reason for majority is steepest slope, the regression based distance may work better than the binned distance. Similarly if the reason of choice is presence of outliers, the binned distance with large number of bins on both axes may be the best distance metric. This can be a probable future work.

The lineup protocol places a statistical plot in the hypothesis testing framework. The null plots in the lineup has an instrumental effect on the response of the subjects since there are only a finite number of null plots which the subjects compare the observed data plot to. A 'bad' set of null plots makes it difficult to identify the observed data plot. The quality of the lineup is measured by describing plots numerically using a set of distance metrics.
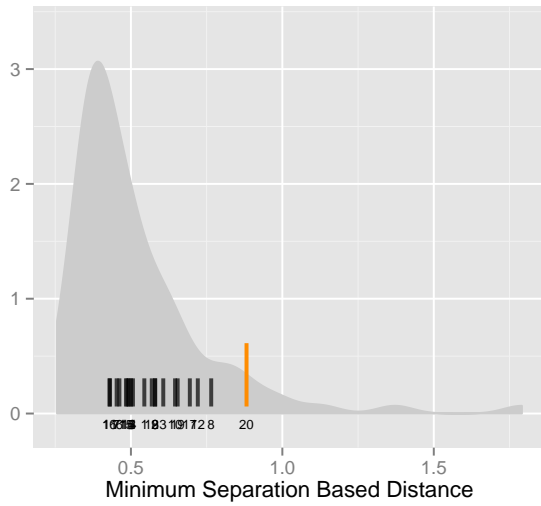
A number of existing distance metrics are studied. Most of these metrics use the raw data to calculate the distances. A number of distance measures are suggested which takes the graphical elements into account. The graphical elements
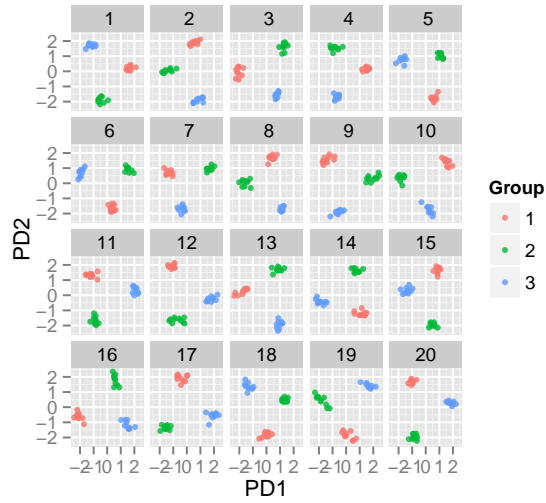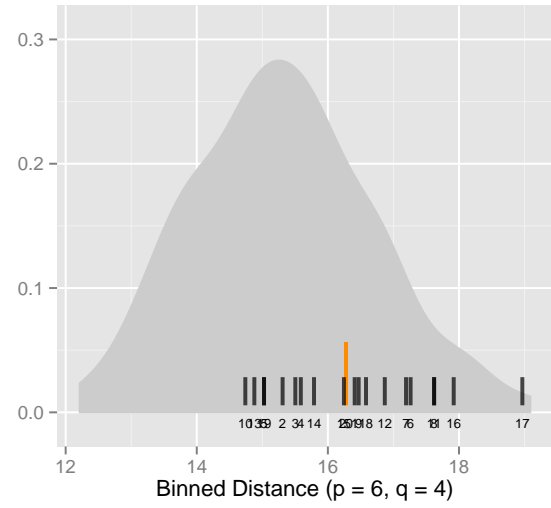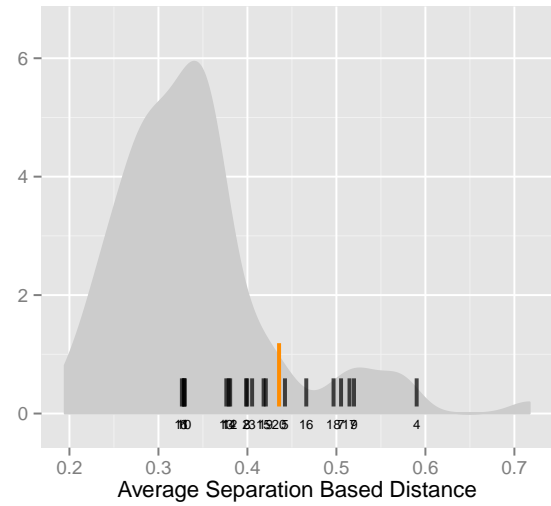
Figure 18: Illustration of the behavior of different distance metrics. The lineup is shown in (a) and the distributions of different distance metrics are shown in the other plots: binned distance with 6 and 4 bins in x and y axis respectively in (b), distance based on average separation in (c) and distance based on minimum separation in (d).

in a plot of a lineup affects the response of the subjects. So considering the graphical elements in the distance metric calculation seems logical.

Unlike classical inference, the test statistic in visual inference is not a number, it is a plot. Hence it is not possible to obtain the distribution of the test statistics in visual inference. The empirical distribution of the distance metrics relates to the t-distribution followed by the test statistic in the classical inference framework. The distance for the observed data plot is compared to the empirical distribution and also the null plots obtained in the lineup. This provides a good idea about how extreme the observed plot is compared to the nulls.

Comparing the observed data plot to the null plots may sometime complicate things. A single measure of the quality of a lineup is easy to interpret. Two measures are developed: the first one being the difference between the mean distance of the observed data plot and the maximum of the mean distances of the null plots and the second being the number of null plots which are more extreme than the true plot.

References

Amazon. Mechanical Turk, 2010. URL `http://aws.amazon.com/mturk/`.

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

A. Buja, D. Cook, H. Hofmann, M. Lawrence, E. Lee, D. Swayne, and H. Wickham. Statistical Inference for Exploratory Data Analysis and Model Diagnostics. *Royal Society Philosophical Transactions A*, 367(1906):4361–4383, 2009.

L. Fernholz. Remembering John W. Tukey. *Statistical Science*, 18(3):pp. 336–340, 2003. ISSN 08834237. URL `http://www.jstor.org/stable/3182751`.

A. Gelman. Exploratory Data Analysis for Complex Models. *Journal of Computational and Graphical Statistics*, 13(4):755–779, 2004.

C. Hennig. fpc : Flexible Procedures for Clustering. *R package version 2*, 2010.

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

R. E. Kirk. Practical significance: A concept whose time has come. *Educational and psychological measurement*, 56.5:746 – 759, 1996.

M. Majumder, H. Hofmann, and D. Cook. Validation of Visual Statistical Inference, Applied to Linear Models. *Journal of American Statistical Association*, 108(503):942–956, 2013.

N. Roy Chowdhury, D. Cook, H. Hofmann, M. Majumder, E. Lee, and Amy Toth. Visual Statistical Inference for High Dimension, Small Sample Size Data. *Computational Statistics: Submitted*, 2013.

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

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

Leland Wilkinson, Anushka Anand, and Robert L Grossman. Graph-theoretic scagnostics. In *INFOVIS*, volume 5, page 21, 2005.

Y. Zhao, D. Cook, H. Hofmann, M. Majumder, and N. Roy Chowdhury. Mind Reading Using an Eyetracker to See How People Are Looking at Lineups. *The American Statistician: Submitted*, 2012.