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

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

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

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 based on the regression coefficients is given as

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

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 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

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

   (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 **??** illustrates the different distance metrics for cluster separation.
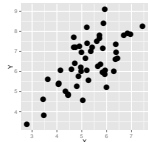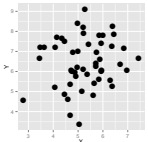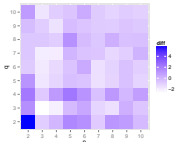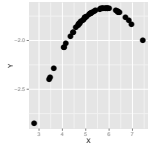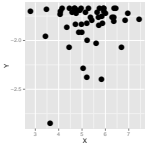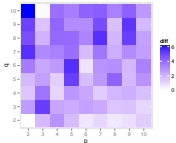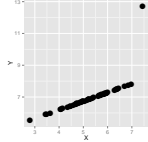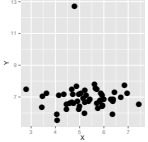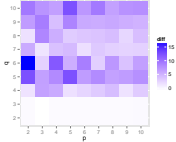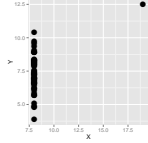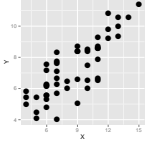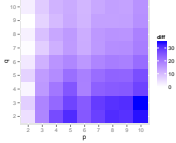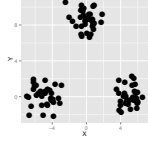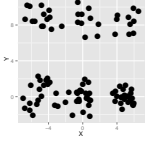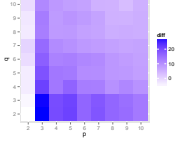
Effect of the Number of Bins on Distance

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 δ-difference is determined for each combination. Tables 1 and 2 show the type of data, the observed plot, the null generating mechanism (NGM), a typical null plot, δ-difference and the maximally observed value of δ, together with the
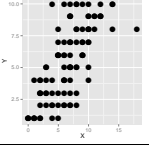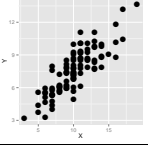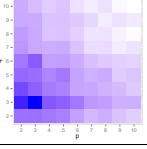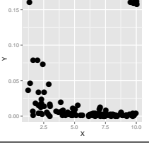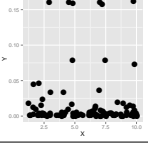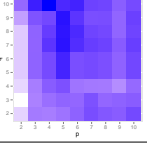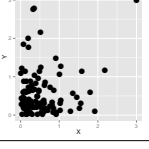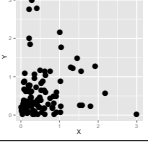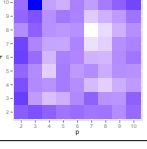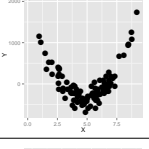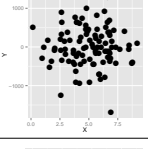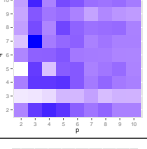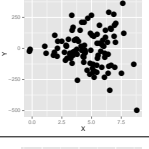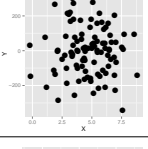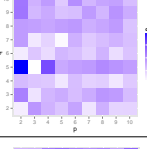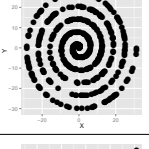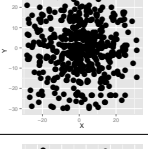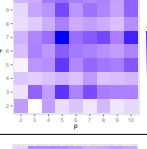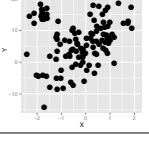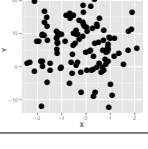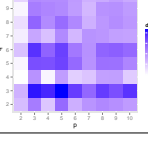
corresponding number of bins in x and y direction. Similarly, the minimal δ value and its parameters are collected to get an idea of the range of values.

Table 1: 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 mechanisms are also selected for the same reason. In Table 1 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 2, 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.

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
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

The δ-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 1 and 2 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.

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, 2016). In order to account for this, the distance metric has to be adjusted for the plot type and graphical elements. Figure 1 illustrates this idea.

Figure 1a shows a lineup of scatterplots with 100 points between two variables X_1 and X_2. Figure 1b, 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 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

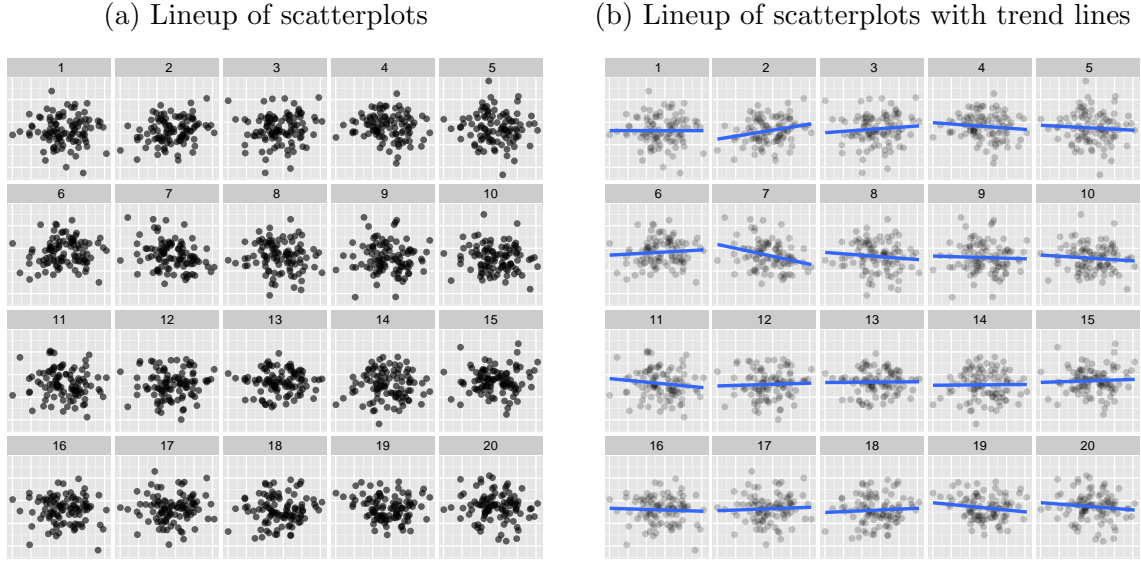(a) Lineup of scatterplots   (b) Lineup of scatterplots with trend lines

Figure 1: 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.

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 1, 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

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.

References

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

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.

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.

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.

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.

Vander Plas, S. and Hofmann, H. (2016), "Clusters beat Trend!? Testing feature hierarchy in statistical graphics," *Journal of Computational and Graphical Statistics*, To appear.

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.