Visual Statistical Inference for High Dimension, Small Sample Size Data

Niladri Roy Chowdhury · **Dianne Cook** · **Heike Hofmann** ·
Mahbubul Majumder · **Amy L. Toth**

Abstract Statistical graphics plays an important role in exploratory data analysis, model checking and diagnosis. Recently there were some formal visual methods for determining statistical significance of findings. We often seek to low-dimensional projections in high dimensional data which reveal important aspects of the data. Projection pursuit for classification finds projections that reveal differences between groups. In this paper we are interested in the performance of classification methods when the number of observations is relatively small compared to the number of variables, known as a large p (dimension) small n (sample size) problem using visual statistical inference. We apply projection pursuit for classification to purely noise data and to the data when there is some separation. We use the lineup protocol Buja et al. [2009] to make comparisons among the purely noise data and the data which has some separation.

Keywords statistical graphics · lineup · visualization · projection pursuit

1 Introduction

Many problems needing solutions today require the analysis of data where more variables are measured than samples are taken. This is commonly referred to as high dimensional, low sample size (HDLSS) data (Hall et al. [2005] and Marron et al. [2007]). HDLSS data occur in many application areas like face recognition and gene expression data. Classical statistical methods often fail in this context, because there is insufficient data to be able to estimate properties of matrices, such as the variance-covariance matrix, required by many methods.

Reducing the dimension would seem to be the natural first step in HDLSS data. Principal component analysis (PCA) is the classical approach. PCA requires estimating the eigenvalues (maximum variance) and eigenvectors (direction of maximum variance) of the population variance-covariance based on the sample. With insufficient data this is a Sisyphean task. Just imagine, estimating a line on the foundation of a single point. There are infinitely many lines possible to return. Similarly for classification tasks, finding a low-dimensional representation of the separation between groups is a common first step. Linear discriminant analysis (LDA) is the classical method for this. LDA finds the low-dimensional space where the groups are most separated, by solving an eigen decomposition problem comparing distances between group means with variance around each mean. When there are few sample points differences between groups can be found in many different low-dimensional spaces.

Niladri Roy Chowdhury, Dianne Cook, Heike Hofmann, Mahbubul Majumder
Department of Statistics, Iowa State University, Ames, IA, USA
E-mail: niladrir@iastate.edu, dicook@iastate.edu, hofmann@iastate.edu, mahbub@iastate.edu

Amy L. Toth
Departments of Ecology, Evolution, and Organismal Biology and Entomology, Iowa State University, Ames, IA, USA
E-mail: amytoth@iastate.edu

Marron et al. [2007] describes the estimation issues associated with HDLSS. One of the problems with HDLSS dataset is that not all the measured variables are "important" for understanding the underlying phenomenon of interest. It is important in many applications to reduce the number of dimension of the original data prior to any modeling of the data. There are many established methods of dimension reduction like principal component analysis (PCA), factor analysis, projection pursuit, principal curves, self-organizing maps and many others. Many of the above use linear dimension reduction techniques for normal variables. Others use higher-order dimension reduction methods for datasets which are not realizations of Gaussian distributions. Many advancements in the PCA to handle HDLSS data has been done by Jung et al. [2012] and Yata and Aoshima [2011]. On the other hand, Donoho and Jin [2009] and Donoho and Jin [2008] studies the optimal variable selection and introduces a principle of model selection based on the notion of higher criticism in situations where only a small fraction of the variables are useful and unknown and contributes weakly to the classification decision.

Clearly, though, the issues of working with HDLSS data are not clear to many researchers. In Toth et al. [2010] LDA is used to examine gene expression data of wasps. Figure 1 shows the result. There are 50 different paper wasps divided into 4 types: Foundress (F), Gyne (G), Queen (Q) and Worker (W). There are 14 wasps of type Foundress and 12 each of the other 3 types. The authors, knowing that LDA requires that the number of observations (n) should be larger than (p), first reduced the dimension from 447 to 40 by randomly selected a subset of significantly different oligonucleotides. This is the almost same approach used in Dudoit et al. [2002], by the way, in one of the first studies of classification of gene expression data. What results is a picture of the four groups that suggests big differences in the types of wasps. With visual inference it can be shown that there is no real difference between the groups - what you see is a mirage. Visual inference methods will explain why.
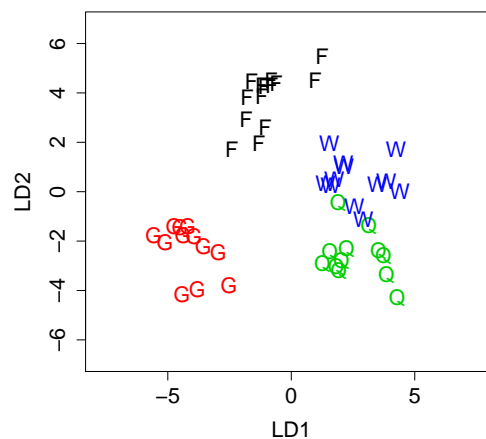

Fig. 1 LD1 versus LD2 from an LDA on a randomly selected subset of 40 significantly different oligos : F, Foundress; G, gyne; Q, queen and W, worker. It can be noticed that the groups F and G are separated. This figure is regenerated to match Figure 2 in Toth et al. [2010].

This paper describes visual inference as applied to dimension reduction for supervised classification problems. In particular we focus on dimension reduction using projection pursuit, and the affect that having large dimension has on the robustness of separation between groups. Small simulation experiments are used to examine the problem in a controlled setting. Visual inference also is used to check the operation of the projection pursuit optimization. The next section explains the methods behind visual inference. Section 3.3 describes theoretically what will happen with dimension reduction of HDLSS data containing two groups. Section 3.2 discusses the experiment designed to examine people's perception of separation in the presence of real separation and purely noise for

2

simulated HDLSS data, and the performance of the optimization algorithm. The wasps data is revisited at the end of the paper.

2 Visual inference methods

Buja et al. [2009], following from Gelman [2004], proposed two protocols that allow the testing of discoveries made from statistical graphics. Among the two protocols explained in Buja et al. [2009] the lineup protocol particularly interests us. The lineup protocol is particularly used for testing significance of findings. Like any statistical test visual test is also associated with a test statistic. But unlike classical hypothesis testing, the test statistic in visual inference is not a real number, but a plot that is appropriately chosen to describe the parameter of interest. The plot of the observed data which is our test statistic is placed randomly among a set of $(m - 1)$ null plots. These type of a plot is known a lineup of m plots. The null plots are generated by a method consistent with the null hypothesis. Human subjects are asked to identify the plot which has the most distinct feature(s). If the human subjects can identify the plot of the observed data, we reject the null hypothesis. When the alternative hypothesis is true, it is expected that the plot of the observed data, the test statistic, will have visible feature(s) inconsistent with the null hypothesis and human subjects will be able to identify the plot of the observed data as different from all the other null plots. For more details see Buja et al. [2009].

Let us consider the following example. The data represents the concentration of a metal in mg/kg for two sites A and B. We want to test whether there exists a significant difference between the concentration levels in the two sites A and B. Let μ_1 denote the mean concentration level in Site A and μ_2 denote the mean concentration level in Site B. To test that, the following null and alternative hypothesis is proposed:

$$H_o : \mu_1 = \mu_2 \qquad \text{versus} \qquad H_a : \mu_1 \neq \mu_2$$

(Technically the problem this data addressed was more interested in testing a one-sided alternative, whether site B has higher concentration than site A, but it is more interesting for this example to consider the two-sided alternative hypothesis.) The test statistic is the plot of the real data. The 19 null plots are generated by assuming that null hypothesis $H_o : \mu_1 = \mu_2$ is true. So the group variable site is permuted to obtain the null plots keeping the other variables fixed. The observed data plot is placed randomly among these 19 plots in a lineup given in Figure 2. The viewer is asked to identify the plot which is most different. If the viewer can identify the plot of the real data, there will be reasons to believe that the observed data plot has a pattern which is absent in the null plots. So the null hypothesis would be rejected. If the viewer cannot identify the observed data plot, we fail to reject the null hypothesis.

If the viewer could identify the plot then there are reasons to believe that there exists a statistically significant difference between the mean concentration levels in site A and site B. So the lineup protocol is the basis of the visual inference while the Rorschach protocol helps viewer understand the extent of randomness.

Majumder et al. describes a comparative study between the visual inference method and the classical inference methods, focusing on plots that might be used in linear modeling. In his work the expected power of the visual test is compared with the power of the uniformly most powerful (UMP) test. The power of the visual test is computed by responses from several large samples of lineup evaluators recruited through Amazon Turk [Amazon, 2010]. The results suggest that the expected power of a visual test is almost as good as the power of UMP test, that visual inference compares favorably with classical testing, in the traditional setting where the classical test performs well. They established properties and efficacy of visual testing procedures in order to use them in situations where traditional test cannot be used. In addition Majumder et al. provide a nice way of making the leap traditional hypothesis testing to visual inference. The table is adapted for the $H_o : \mu_1 = \mu_2$ vs $H_a : \mu_1 \neq \mu_2$ example described and plotted in Figure 2, which can be seen in Table 1.
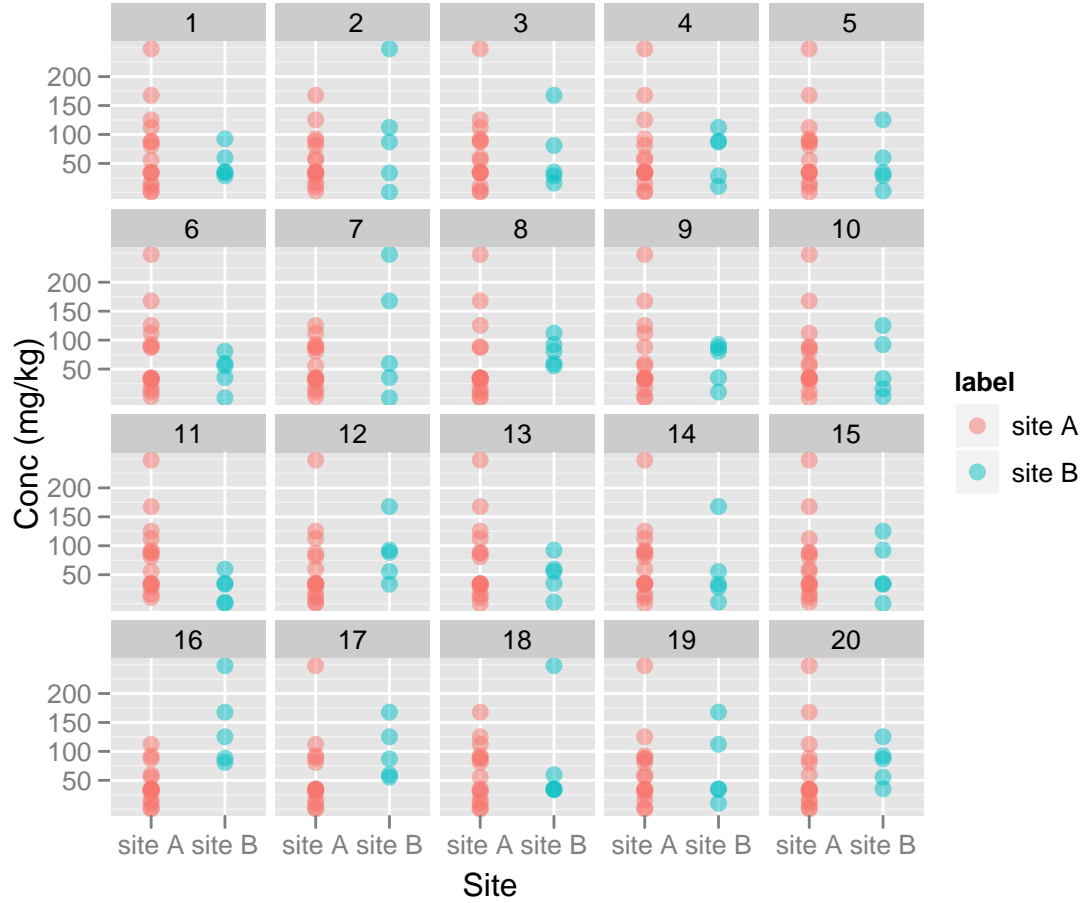
Fig. 2 A typical lineup ($m = 20$) for testing $H_o : \mu_1 = \mu_2$. When the alternative hypothesis is true the observed data plot should have the largest vertical difference between the centers. Can you identify the observed data plot? The solution to the lineup is provided in the Appendix.
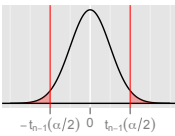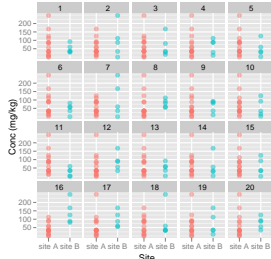
3 Explanation of the methods

3.1 Dimension reduction

In this paper the idea of projection pursuit which was implemented by Friedman and Tukey [1974] is used. Projection pursuit (PP) is a statistical tool to find the most interesting low dimensional projections from high dimensional to low dimensional space that reveals the most details about the structure of the data. In this paper the one and two dimensional projections obtained by projection pursuit on higher dimensions is mainly used. As pointed out in Huber [1985] the most exciting feature of projection pursuit is that it can bypass the curse of dimensionality. Hence it works really well in HDLSS situations. The other methods fail to pick up small features unless the sample size is large. Also projection pursuit are able to ignore noisy and non-informative variables where it is in advantage over minimal spanning trees, multidimensional scaling and most clustering methods. Though these methods work in a HDLSS situation, they cannot ignore the "noise" variables. The major drawback of PP method is its demand of the computer time. But in times of super computer and high speed computing this does not seem that big an issue.

PP method needs a projection index on the basis of which it finds the low dimensional projections of a high dimensional data. Various indices are used for this like Linear Discriminant Analysis (LDA) index, Quadratic

4

Table 1 Comparison of visual inference with traditional hypothesis testing.

	Mathematical Inference	Visual Inference
Hypothesis	$H_o : \mu_1 = \mu_2$ vs $H_a : \mu_1 \neq \mu_2$	$H_o : \mu_1 = \mu_2$ vs $H_a : \mu_1 \neq \mu_2$
	$\downarrow$	$\downarrow$
Test Statistic	$T(y) = \dfrac{\bar{y}_1 - \bar{y}_2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$	$T(y) =$
	$\downarrow$	$\downarrow$
Sampling Distribution	$f_{T(y)}(t);$	$f_{T(y)}(t);$
	$\downarrow$	$\downarrow$
Reject H_o if	observed T is extreme	observed data plot is identifiable

Discriminant Analysis (QDA) index and many others. But in this paper Penalized Discriminant Analysis (PDA) index described in Lee and Cook [2009] is used. PDA index is an improvement over the LDA index in a HDLSS situation.

In this paper the `tourr` package in R (R Development Core Team [2009]) by Wickham and Cook [2010] is used to generate the low dimensional projections. The `tourr` package produces tours of multivariate data. The package also includes functions for creating different types of tours like grand, guided and little tours, which project multivariate data with p dimensions to 1, 2, 3 or d dimensions where $d \leq p$. In this paper the guided tour function is mainly considered. Instead of picking a new projection completely at random, the guided tour function picks projections that are closer to the current projection, so that we eventually converge to a single maximally interesting projection. The `tourr` package comes with different indices like cmass, holes, and LDA index but the PDA (Penalized Discriminant Analysis) index described in Lee and Cook [2009] is used which is specifically designed to handle HDLSS problems.

In this paper, data with p dimensions of noise and no real separation is termed as "p dimensional purely noise data" while data which has k dimensions of real separation and $p - k$ dimensions of purely noise is termed as "p dimensional data with real separation". In this paper k is either 1 or 2.

3.2 Experimental Setup

An experiment is designed to study the ability of human observers to detect the effect of one or two dimensions of real separation in p dimensions of noise. Data is simulated for different values of $p(= 20, 40, 60, 80, 100)$. So two groups of p dimensions of data with 15 observations in each group were generated from $N(0, 1)$. The data from the first group is labeled as group 1 and the data from the second group as group 2. So a $30 \times (p + 1)$ matrix, say, X, is obtained where the first 15 observations are from group 1 and the last 15 observations are from group 2. So X can be written as

$$X^{n \times (p+1)} = (X_1, X_2, \ldots, X_p, \text{Group})$$

where each X_i is a vector of dimension 30 for $i = 1, \ldots, p$. This matrix X excluding the Group variable gives the p-dimensional purely noise data with 2 groups.

To bring in the real separation in the data, 3 is subtracted from the elements of the p-th variable X_p which belongs to group 1 and 3 is added to the elements of the same variable which belongs to group 2. So,

$$X_p = \begin{cases} X_p - 3 & \text{if } X_p \in \text{group 1} \\ X_p + 3 & \text{if } X_p \in \text{group 2} \end{cases}$$

This is done so that the mean of the observations in the two groups in the p-th dimension is separated by 6 units. This modified X matrix works as the data with one dimension of real separation.

Hence two sets of data is obtained, each with p dimensions with $n = 30$ observations in each dimension divided into two groups. One of the datasets is purely noise and the other has some real separation in the form of the p-th dimension. On each of these datasets of p dimension, a projection pursuit optimization with a PDA index is performed to obtain the $d = 1$ dimensional projections.

The above procedure is repeated to obtain the $d = 2$ dimensional projections as well. But instead of labeling the 15 observations as group 1 and group 2, the first 10 observations is assigned as group 1, the second 10 observations as group 2 and the last 10 as group 3. Once again a $30 \times (p + 1)$ matrix, say, X, is obtained where the first 10 observations are from group 1, the second 10 observations are from group 2 and the last 10 observations are from group 3. So, collectively, X can be written as

$$X^{n \times (p+1)} = (X_1, X_2, \ldots, X_p, \text{Group})$$

where each X_i is a vector of dimension 30 for $i = 1, \ldots, p$. This matrix X excluding the Group variable gives the p-dimensional noise data with 3 groups.

As before to bring in the real separation, the means of the 3 groups are adjusted in the last two dimensions i.e. X_{p-1} and X_p. The adjustment is done in the following way:

$$(X_{p-1}, X_p) = \begin{cases} (X_{p-1} - 3, X_p) & \text{if } (X_{p-1}, X_p) \in \text{group 1} \\ (X_{p-1} + 3, X_p) & \text{if } (X_{p-1}, X_p) \in \text{group 2} \\ (X_{p-1}, X_p + \sqrt{27}) & \text{if } (X_{p-1}, X_p) \in \text{group 3} \end{cases}$$

If X_{p-1} versus X_p are plotted in a scatterplot, the points cluster along the vertices of an equilateral triangle of side 6. X works here as the data with real separation in the last two dimensions. In this manner the data with 2 dimensions of real separation divided into 3 groups is obtained. So for obtaining the two dimensional projections two sets of data are obtained - one with purely noise and the other with two dimensions of real separation. Again on each of the above datasets a projection pursuit with a PDA index is performed to obtain the $d = 2$ dimensional projections.

Finally 4 different sets of data are obtained for different values of p. The different combinations of sample size n, dimension p, projection d and presence of noise were generated as shown in Table 2. Three replicates of each level were generated. These produced 60 different "observed data sets".

Table 2 Values of parameters considered for the experiment.

n	projection(d)	separation	dimension (p)
30	1	Yes	20, 40, 60, 80, 100
		No	20, 40, 60, 80, 100
30	2	Yes	20, 40, 60, 80, 100
		No	20, 40, 60, 80, 100

To obtain a lineup with $m = 20$ plots for one dimensional projections for purely noise, the one dimensional projections obtained from the projection pursuit with the PDA index is plotted on the horizontal axis and a fixed value (1.5 in this case) is plotted on the vertical axis colored by the group variable. This plot is called the "observed

data plot" as it is obtained from the observed dataset. The limits of the vertical axis are adjusted so that the points form circular clusters instead of bands of points. The 19 null plots in the lineup were obtained by permuting the group variable so that breaking any dependency between the group variable and the other variables. Then the observed data plot is placed randomly among the 19 null plots to obtain one lineup. Two different statistics Wilk's λ (Johnson and Wichern [2002]) and WBratio which is the ratio of the average distance within clusters to the average distance between clusters (Hennig [2010]) are calculated for each of the 20 plots in the lineup.

The above procedure was repeated for the one dimensional projections for data with one dimension of real separation. Figure 3 gives the lineup for the one dimensional projections with $p = 20$ dimensions where 19 dimensions are purely noise data and 1 dimension has real separation. This is an example of a lineup that is used in the experiment.
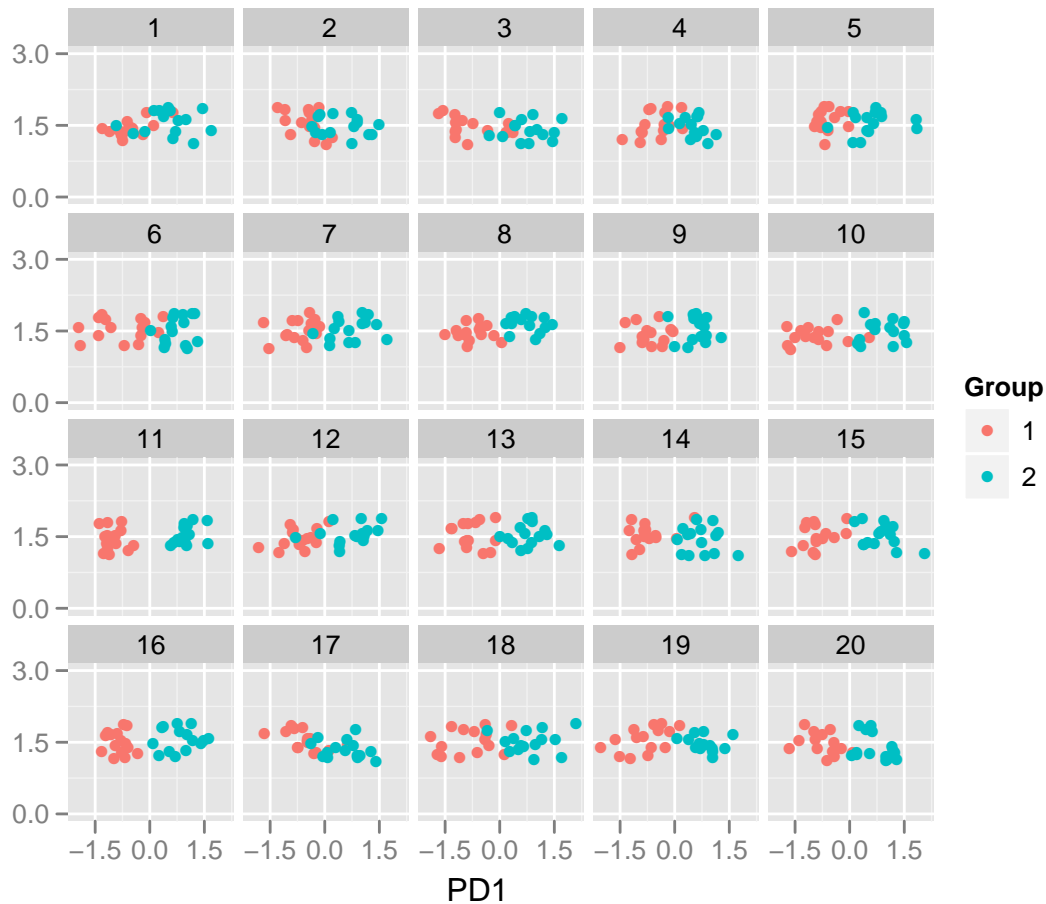

Fig. 3 lineup ($m = 20$) with 19 dimensions of purely noise and 1 dimension of real separation. So the lineup is generated from a data with 1 dimension of real separation. The subjects were asked to identify the plot with the most separated colors. Can you identify the observed data plot? The solution to the lineup is provided in the Appendix.

Similar procedure was followed to obtain the lineup for two dimensional projections for purely noise and data with two dimensions of real separation. But in this case the first two projections are plotted on the horizontal and vertical axis respectively colored by the group variable to get a scatterplot. This again gives the "observed data plot". To obtain the null plots, the group variable is permuted and the projection pursuit is performed on the permuted dataset. The "observed data" is placed randomly among the 19 null plots to obtain a lineup of

$m = 20$ plots. Figure 4 gives the lineup for the two dimensional projections for $p = 100$ dimensions where all the 100 dimensions are purely noise data. As before the Wilk's λ (Johnson and Wichern [2002]) and WBratio are calculated for each of the 20 plots in the lineup.
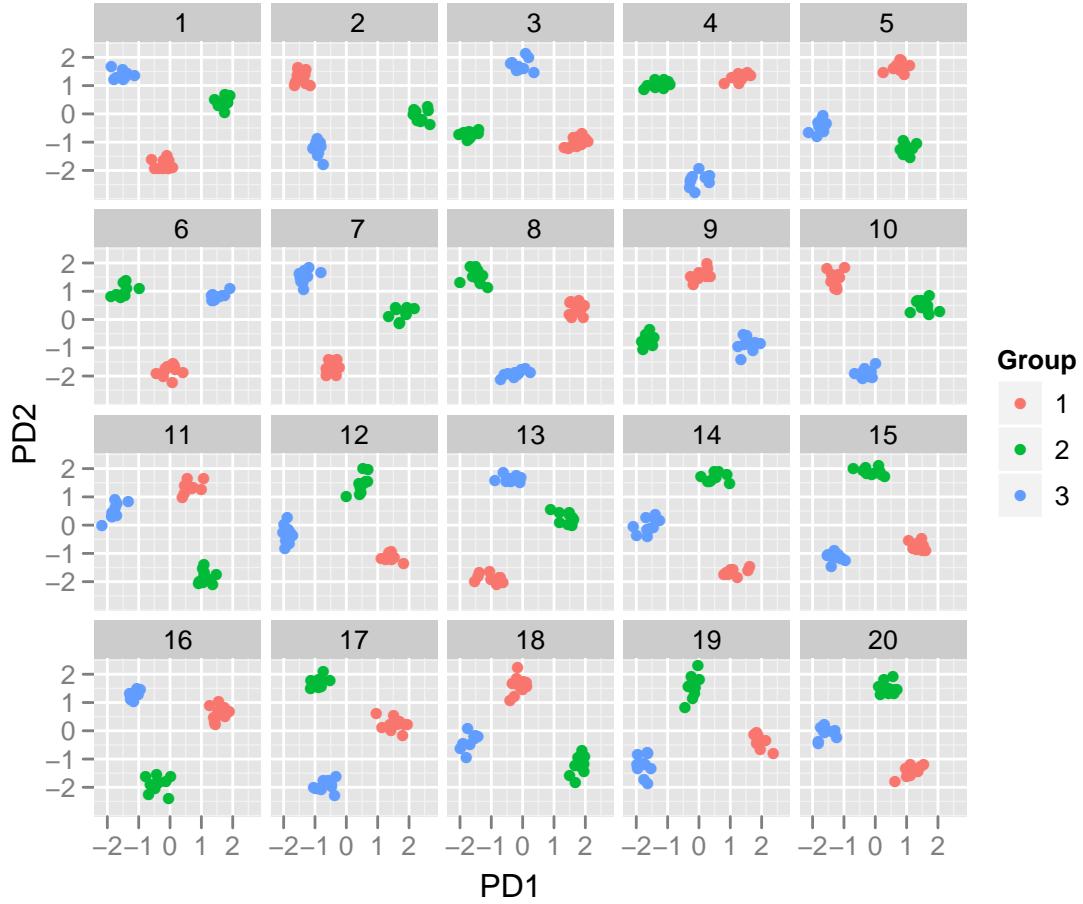

Fig. 4 lineup ($m = 20$) with 100 dimensions of purely noise. In this case the lineup is generated from a data with 100 dimensions of purely noise. The subjects were asked to identify the plot with the most separation between the colored groups. Can you identify the observed data plot? The solution is provided in the Appendix.

Subjects for the experiment were recruited through Amazon Mechanical Turk [Amazon, 2010]. Amazon [2010] is a platform to get feedback on this kind of experiments. Each subject was shown a sequence of 10 lineups. Subjects also had the freedom of answering more than 10 lineups. They were asked to identify the plot which has the most separation between the colored groups. Their response was recorded along with a reason for their choice of the plot and the level of confidence they have in their decision. Gender, age, educational qualification and location of each subject were also noted. In total, 1137 lineups were evaluated by 103 subjects from different locations.

3.3 Theory

To decide on the different dimensions, the distribution of the absolute difference of the means of the two groups of a purely noise data is looked at. In this case one dimension of purely noise is considered which is divided into two groups. The means of the observations in each group is calculated. Since we are interested in the projections, the absolute differences of the means is considered.

Let us denote X_{ij} as the j-th observation in the i-th group where $j = 1, \ldots, n_i$. Essentially in our case there are only $i = 2$ groups. X_{ij}s are purely noise obtained from a standard normal distribution. The difference between the means of the purely noise data for the two groups, group 1 and group 2 is given by $\bar{X}_{1.} - \bar{X}_{2.}$ and

$$\bar{X}_{1.} - \bar{X}_{2.} \sim \text{Normal}(0, 1/n_1 + 1/n_2)$$

where $n_1 = n_2 = 15$. So we have

$$\bar{X}_{1.} - \bar{X}_{2.} \sim \text{Normal}(0, 2/15)$$

Let us define

$$Y = |\bar{X}_{1.} - \bar{X}_{2.}|$$

where $Y \sim$ Half Normal with scale parameter $\sigma = \sqrt{1/n_1 + 1/n_2} = \sqrt{2/15}$.

The expectation and the variance of Y can be calculated:

$$E(Y) = \sigma\sqrt{2/\pi}$$

$$Var(Y) = \sigma^2(1 - 2/\pi)$$

Let us consider p dimensions of purely noise data, each dimension being divided into $i = 2$ groups with $n_i = 15$ observations. Let us define

$$Y_m = |\bar{X}_{m1.} - \bar{X}_{m2.}|$$

where X_{mij} is the j-th observation in the i-group for the m-th dimension. The sum of the absolute difference between the means is obtained for all the p dimensions. So, define

$$Y = \sum_{m=1}^{p} Y_m = \sum_{m=1}^{p} |\bar{X}_{m1.} - \bar{X}_{m2.}|$$

Assuming independence among the dimensions of purely noise data,

$$E(Y) = p\sigma\sqrt{2/\pi}$$

$$Var(Y) = p\sigma^2(1 - 2/\pi)$$

To check for the optimization procedure one dimension of real separation is considered in the data. To bring in the real separation, we adjust the mean of X_{ij}. We subtract C from the mean of the first 15 observations (group 1) and add C to the mean of the last 15 observations (group 2). So we define

$$X_{ij}^* \sim \text{Normal}(-C, 1) \qquad \text{for} \qquad j = 1, \ldots, n_1$$

$$X_{ij}^* \sim \text{Normal}(C, 1) \qquad \text{for} \qquad j = 1, \ldots, n_2$$

Hence we have

$$\bar{X}_{1.}^* - \bar{X}_{2.}^* \sim \text{Normal}(2C, 1/n_1 + 1/n_2)$$

where $n_1 = n_2 = 15$. Let us define

$$Y^* = |\bar{X}_{1.}^* - \bar{X}_{2.}^*|$$

where $Y^* \sim$ Folded Normal Distribution with scale parameter $\sigma = \sqrt{1/n_1 + 1/n_2} = \sqrt{2/15}$.

The expectation and the variance of Y^* can be calculated :

$$E(Y^*) = \sigma\sqrt{2/\pi}\exp(-2C^2/\sigma^2) + 2C[1 - \Phi(-2C/\sigma)]$$

$$Var(Y^*) = 4C^2 + \sigma^2 - (E(Y^*))^2$$

Now the means of X_{mij}s are adjusted in the p-th dimension. So we define Z as the sum of the absolute differences of the mean with one dimension of real separation as

$$Z = \sum_{m=1}^{p-1} |\bar{X}_{m1.} - \bar{X}_{m2.}| + Y_p$$

where Y_p follows a Folded Normal Distribution.

Again assuming independence among the dimensions of purely noise data and data with real separation,

$$E(Z) = (p-1)\sigma\sqrt{2/\pi} + \sigma\sqrt{2/\pi}\exp(-2C^2/\sigma^2) + 2C[1 - \Phi(-2C/\sigma)]$$

$$Var(Z) = (p-1)\sigma^2(1 - 2/\pi) + 4C^2 + \sigma^2 - \left(\sigma\sqrt{2/\pi}\exp(-2C^2/\sigma^2) + 2C[1 - \Phi(-2C/\sigma)]\right)^2$$

In our case, $C = 3$ and $\sigma^2 = 2/15$. Therefore,

$$\exp(-2C^2/\sigma^2) \approx 0 \qquad \text{and} \qquad \Phi(-2C/\sigma) \approx 0$$

Hence,

$$E(Z) = (p-1)\sigma\sqrt{2/\pi} + 6$$

$$Var(Z) = (p-1)\sigma^2(1 - 2/\pi) + \sigma^2$$

As the value of p increases for a fixed n, the spread of Y increases with a factor of the dimension p as the spread of Z increases as well. The means of both Y and Z also increase with a factor of p but the expected value of the difference between Y and Z stays constant and is independent of the number of dimensions (p).

$$E(Z - Y) = (p-1)\sigma\sqrt{2/\pi} + 6 - p\sigma\sqrt{2/\pi} = 6 - \sigma\sqrt{2/\pi}$$

Figure 5 shows the phenomenon described above. In Figure 5 notice that as the number of dimensions (p) increases from 20 to 100, the dark blue region of the distribution of Y which is greater than the 5th percentile of Z increases as the variability in Y and Z increases as a factor of the dimension (p).

The next step is to find the number of dimension for a specific value of the common region between the two distributions. To allow some error, the value of the number of dimension p is considered for which

$$P[Y > Z_\alpha] = \delta$$

where Z_α is the α-th percentile of Z. So for a given value of δ, the value of p is calculated for which the above condition holds true for $\alpha = 0.05$.

The above described procedure is done for 5 different values of $\delta = 0.0000001, 0.01, 0.05, 0.1$ and 0.2. The above procedure is repeated 100 times to get 100 values of each dimension p corresponding to each δ. Table 3 shows the summaries of the number of dimensions for each value of δ.
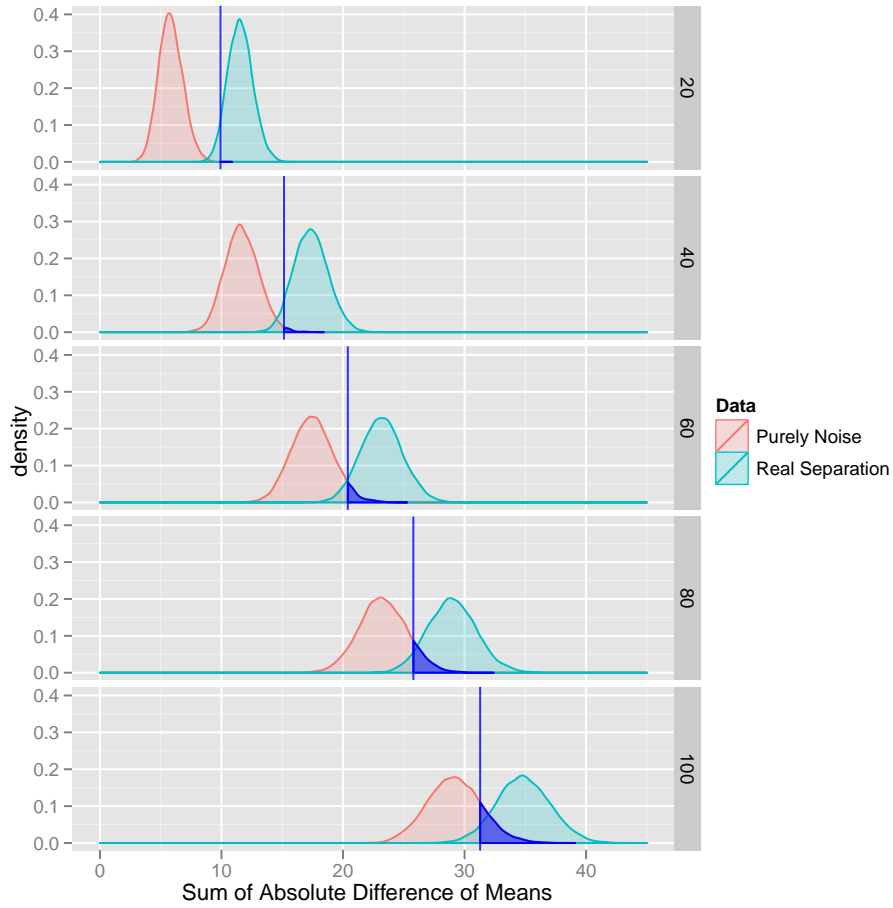
Fig. 5 Plot showing the distribution of the sum of absolute difference of means for data with purely noise (Y) and real separation (Z) for different values of the dimension. The distribution of Y and Z are colored in red and blue respectively with the dark blue line showing the 5th percentile of Z. The dark blue area shows the area of Y which is greater than the 5th percentile of Z. It can be noticed that the area of the dark blue region increases as the number of dimensions increases. This indicates that the probability of obtaining the sum of absolute difference for data with purely noise similar to the difference for data with real separation increases with p.

Table 3 Numerical summaries of dimension p for each value of δ. Notice that with the increase in the dark blue region δ, the median number of dimensions required to obtain the region also increases. This also means for a large value of p the difference between the groups for data with purely noise is close to the difference between the groups for data with real separation.

δ	Median	5th percentile	95th percentile
0.0000001	24	19	28
0.01	41	38	44
0.02	61	56	64
0.1	77	72	81
0.2	106	99	112

11

4 Results

4.1 Data cleaning

Amazon Mechanical turk subjects are paid for their responses. Since the process is not monitored manually, some subjects does not make an honest effort to find the observed data plot but just picks plots randomly. To counter this problem, each subject responded to a comparatively easy lineup (a lineup with $p = 10$ dimensions with some real separation). The subjects who failed to give a correct response to this lineup were removed from the study. If the response in this lineup is correct, we remove the response of this lineup but keep all the other responses. In this way we finally retain the response of 101 subjects.

4.2 Effect of experimental factors

It would be interesting to look at the proportion of subjects identifying correctly the observed data plot for data with real separation versus data with purely noise, for 1D projections versus 2D projections and for the different dimensions. Figure 6 shows the proportion of the subjects who successfully picked the plot of the observed data for each replicate of each dimension. Notice that when there is some real separation, the rate of success in making the correct response overall is higher than the rate of success when the data is purely noise and there is no real separation. Also there is a difference in terms of the variability among the replicates in each dimension. The rate of success is not very different for 1D and 2D projections both in case of data with real separation and also for noise data. Also with the different values of the dimension p, the rate of success in picking the correct plot is different. The rate is higher for dimension $p = 20$ than when the dimension is $p = 40$. Strangely enough it can be noticed that the rate of success for $p = 100$ is higher than the rate of success for $p = 80$ for both 1D and 2D projections.

Suppose each of K independent subjects responded to multiple lineups and responses are considered to be a binary random variable Y_{lk} where

$$Y_{lk} \sim \text{Ber}(p_{lk})$$

where $Y_{lk} = 1$ if the k-th subject correctly identifies the observed data plot in the l-th lineup, $1 \leq k \leq K, 1 \leq l \leq L$ and 0 otherwise. The main interest is to predict the proportion of successes by dimension. Treating dimensions as a continuous variable, a fixed effects logistic regression model is used for $P(Y_{lk} = 1) = p_{lk} = E(Y_{lk})$ with the different parameters of the lineup as covariate. The model can be fit as:

$$g(p_{lk}) = \theta + \alpha_{lk} + \beta_{lk} + \gamma x_{lk} + \alpha_{lk}\gamma x_{lk} \tag{1}$$

where θ is the intercept, α_{lk} is the effect of noise on the l-th lineup for the k-th subject, β_{lk} is the effect of projection on the l-th lineup for the k-th subject and x_{lk} is the dimension of the l-th lineup for the k-th subject. An interaction between the effect of noise and dimension is also included in the model. γ is the corresponding slope parameter and $g(.)$ denotes the logit link function $g(\pi) = \log(\pi)$ - $\log(1 - \pi)$; $0 \leq \pi \leq 1$. The logit link function is inverted to estimate the lineup specific probability of successful evaluation as

$$\hat{p}_{lk} = g^{-1}(\hat{\theta} + \hat{\alpha}_{lk} + \hat{\beta}_{lk} + \hat{\alpha}_{lk}\hat{\gamma}x_{lk}) \tag{2}$$

Model 1 is fitted to the data and the estimated proportions of successful evaluation for each lineup is obtained. The line in Figure 6 shows the estimated proportion of successful evaluation for the different levels of the parameters while the points gives the observed success rate. Notice that as the number of dimensions p in the lineups with purely noise increases, the probability of success in picking the true plot decreases. This is understandable since as the number of dimensions of noise increases keeping the number of real separation fixed to one or two, the distance between the groups in the observed data plot get closer to the distance between the groups in a null plot making it difficult for the subjects to pick the observed data plot. Table 4 shows the estimates of the parameters, the standard errors and the corresponding p-values.

In Table 4 observe that the p-values corresponding to dimension and presence of real separation is very highly significant. The interaction term is also significant. But the p-value corresponding to the projection is high, which
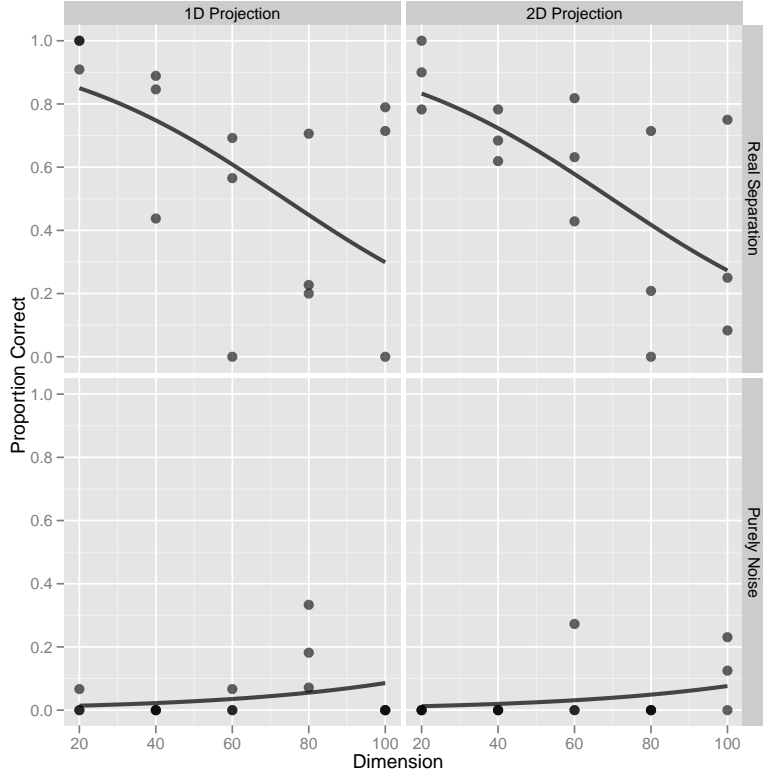
Fig. 6 Proportion of correct responses for 1D projection and 2D projection and data with real separation and purely noise are shown. The three points represents the three replicates for each treatment level. A fixed effects logistic regression model is overlaid on the points. It can be seen that the proportion of correct response decreases as the number of dimensions (p) increases from 20 to 100 for data with real separation. When the data is purely noise data, the proportion of correct response is very low which is understandable as the subjects are not expected to identify the plot.

Table 4 Table showing the estimate, the standard error and the p-value of the parameters used in Model 1. Notice that the covariates dimension and noise are highly significant while projection is not significant at 5% level of significance. Also the interaction term between dimension and noise is also significant.

Parameters	Estimate	Std. Error	p-value
Intercept	2.381	0.278	0.000
dimension	−0.032	0.004	0.000
purely noise	−7.097	0.911	0.000
2D projections	−0.127	0.181	0.483
purely noise:2D projections	0.056	0.011	0.000

suggests that projection is not significant at 5% level of significance which is indicated in the previous section as well.

The estimated models for the different treatment levels are provided.

$$g(\hat{p}_{lk}) = \begin{cases} 2.381 - 0.032x_{lk} \text{ for data with real separation and 1D projections} \\ -4.716 + 0.024x_{lk} \text{ for data with purely noise and 1D projections} \\ 2.254 - 0.032x_{lk} \text{ for data with real separation and 2D projections} \\ -4.843 + 0.024x_{lk} \text{ for data with purely noise and 2D projections} \end{cases}$$

The fitted models for the different treatment levels suggest that the slope of the covariate dimension is different for purely noise data and data with real separation for a fixed level of projection. On the other hand, there is no difference between 1D projections and 2D projections in the fitted models, when the level of the presence of noise is fixed.

4.3 Subject-wise responses

To look at the subject specific probability of successful evaluation Model 1 is modified to bring in a random effect for the individuals. Using the setup of Model 1, we have

$$g(p_{lk}) = \theta + \alpha_{lk} + \beta_{lk} + \gamma x_{lk} + \alpha_{lk}\gamma x_{lk} + u_k \qquad (3)$$

where u_k is a random effect specific to individual k, $k = 1 \leq k \leq K$ and $u_k \sim N(0, \sigma_u^2)$ independently. As before the link function is inverted to obtain the probability of successful evaluations.

Using *lme4* package in R, a mixed effects logistic regression model is fitted as specified in Model 3. The thin black lines in Figure 7 shows the estimated probability of successful evaluation of each individual with the blue line showing the fixed effect. It can be noticed that some of the subjects has a very high probability of success for both one and two dimensional projections. Also as the number of dimensions increases, the probability of successful evaluation decreases. Most of the individuals perform similarly when the data has no real separation. Also there is a higher variability in the individual responses for data with real separation but not much when the data is purely noise.
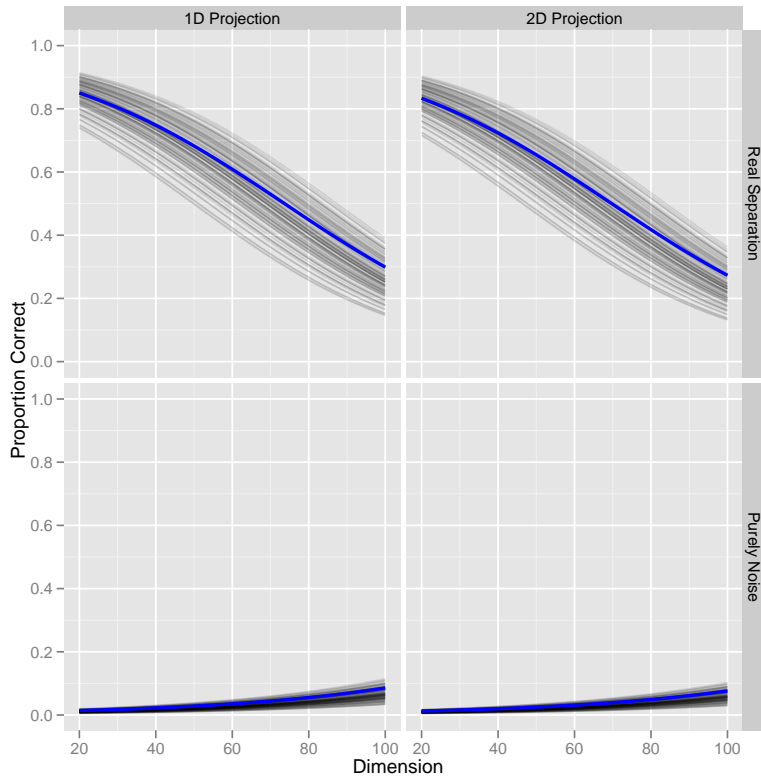

Fig. 7 Plot showing subject wise proportions for each subject estimated using Model 3. The thick blue line shows the overall average proportion for all the subjects. It can be noticed that some of subjects have higher proportion of correct response than the others.

14

4.4 Does rotation in the 2D projections affect the responses?

The 1D projections were plotted in a lineup making the adjustment that the group with the lower values of projection are always considered to be group 1 and the group with larger values of the projection are considered to be group 2, as can be seen in Figure 3. So the orientation of the groups does not affect the response of the subjects. But in case of 2D projections, the groups are rotated in a different way in each plot in the lineup (Figure 4) which may influence the response of the subjects. It would be interesting to test if the rotation in the 2D projections actually effect the response. Since the 1D projections are adjusted, the lineups of the 1D projections are used as a control and the responses for the 1D and 2D projections are compared. Figure 6 shows that there is not much of a difference between the 1D and 2D projections both in presence and absence of real separation. This is also verified in Table 4 where the effect of the projections is not significant at 5% level.

4.5 Time taken to respond

The amount of time taken to respond has a relationship with the difficulty in identifying the observed data plot in the lineup. So the distribution of the time taken to respond by the subjects is considered for the different parameters. Figure 8 shows the distribution of the time taken to respond on the logarithm scale by the subjects for the dimensions for the different projections. The colors suggest whether the data has some real separation or the data is purely noise with a loess smoother fitted through the points. Notice that as the dimensions increases, the subjects takes more time to respond to the lineups when the data has some real separation. But when the data is purely noise, the increase of dimension does not have any effect on the time. This suggests that as the number of dimension of noise increases with fixed number of real separation, it becomes extremely hard to spot the observed data plot among the null plots. On the other hand, the difficulty of spotting the observed data plot for a data with purely noise does not vary with dimensions. It can also be seen that the time taken when the data is purely noise is overall higher than the time taken when the data has some real separation. There is not much of a difference between the time taken for 1D and 2D projections.

4.6 What affects the decision of the people?

For some of the lineups, we noticed that more subjects than expected were able to identify the observed data plot, even though there was no real separation, all of the plots showed purely noise data. Figure 9 examines subjects choices in depth. The relative frequency of picks of each plot in the lineup is plotted against a measure of distance separation between groups, the WBratio. Each cell of this figure shows data from one of the lineups used in the study, 60 in total. Each "pin" represents a plot in a lineup, so each cell here has 20 pins, indicating the frequency the plot was chosen. Red represents the observed data plot. The top three rows correspond to data containing real separation between the groups, and the bottom three rows all of the data was purely noise. Columns indicate dimension (p) and projection dimension (d). Replicates are in different rows. The taller the pin the more often that particular plot was chosen. We asked subjects to pick the plot where the groups were most separated, and this is effectively what they picked. The plot in each lineup with the smallest WBratio tended to have the highest frequency. This is more obvious when there was real separation, and also when dimension was small, but it is also seen in the lineups containing pure noise data. This is reassuring – that subjects did well at detecting the biggest difference. For some of the lineups different choices were made, and these are a little surprising. Investigating these lineups further may reveal why this is.

4.7 How does the null plots affect the response?

In classical inference the test statistic can be compared with an infinite number of possible values from the sampling distribution under the null hypothesis. But in visual inference the subjects can compare the test statistic (the observed data plot) with only a finite number of null plots which makes the null plots an important factor in
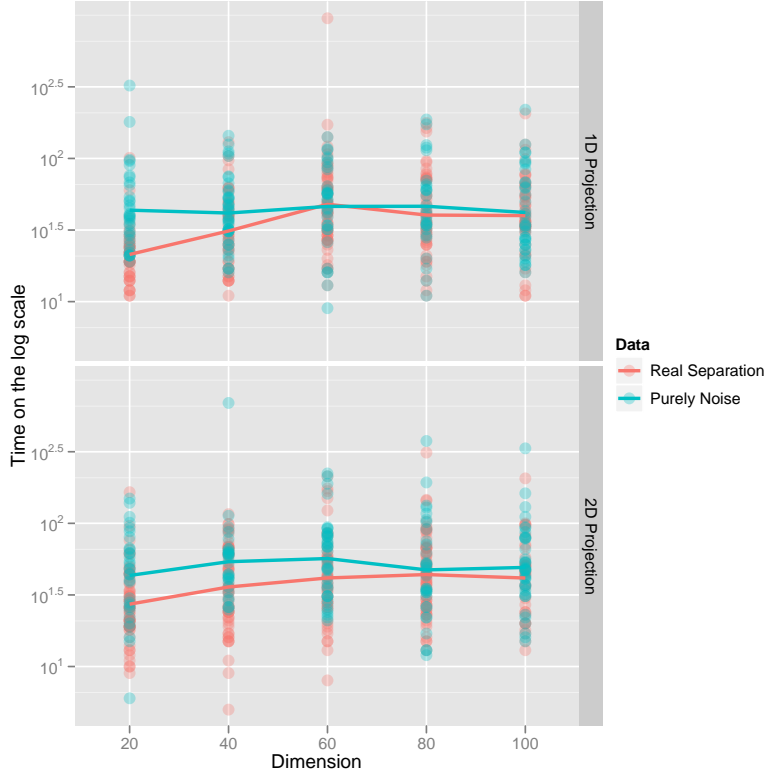
15

Fig. 8 Plot showing the time taken to respond on log scale by each dimension colored by the type of the data. A loess curve is fit through the points. It can be observed that the time taken to respond is higher when the data is purely noise than when there is some real separation. Also it can be noticed that as the number of dimensions increases, the difference between the time taken to respond for purely noise data and real separation decreases.

decision making. It is mentioned in Section 4.2 that the probability of correct evaluation decreases for dimension $p = 80$ and increases for dimension $p = 100$ which is opposite to what we should expect. To investigate this the difference between the minimum WBratio of the null plots and the WBratio for the observed data plot for each lineup is considered. Since the values of WBratio vary for each lineup depending on the presence or absence of real separation, the difference is scaled by dividing the difference by the WBratio of the observed data plot of the lineup. To put the differences on a logarithmic scale, the differences are adjusted by adding a constant term. Figure 10 shows the relationship between the logarithm of the adjusted difference and the probability of successful evaluation.

Figure 10 shows that for data with real separation dimension has an effect on the difficulty level but for data with noise there is no effect of dimension. For some lineups with noise data the observed data plot has a lower WBratio value than the null plots. It can be observed from Figure 10 that the subjects could pick the observed data plot if it has the lowest WBratio value. The unexpected performance of the subjects when the dimension is $p = 80$ can be explained by the fact that the one or more of the null plots have lower WBratio value than the observed data plot.

5 Distance Increasing with p for Fixed n

Consider two dimensions of data with purely noise, each dimension having 30 observations divided into 2 groups with 15 observations in each group. If the projection pursuit is performed on the above data and the one dimensional projections are plotted with different color for the groups against a fixed constant. It can be noticed that there
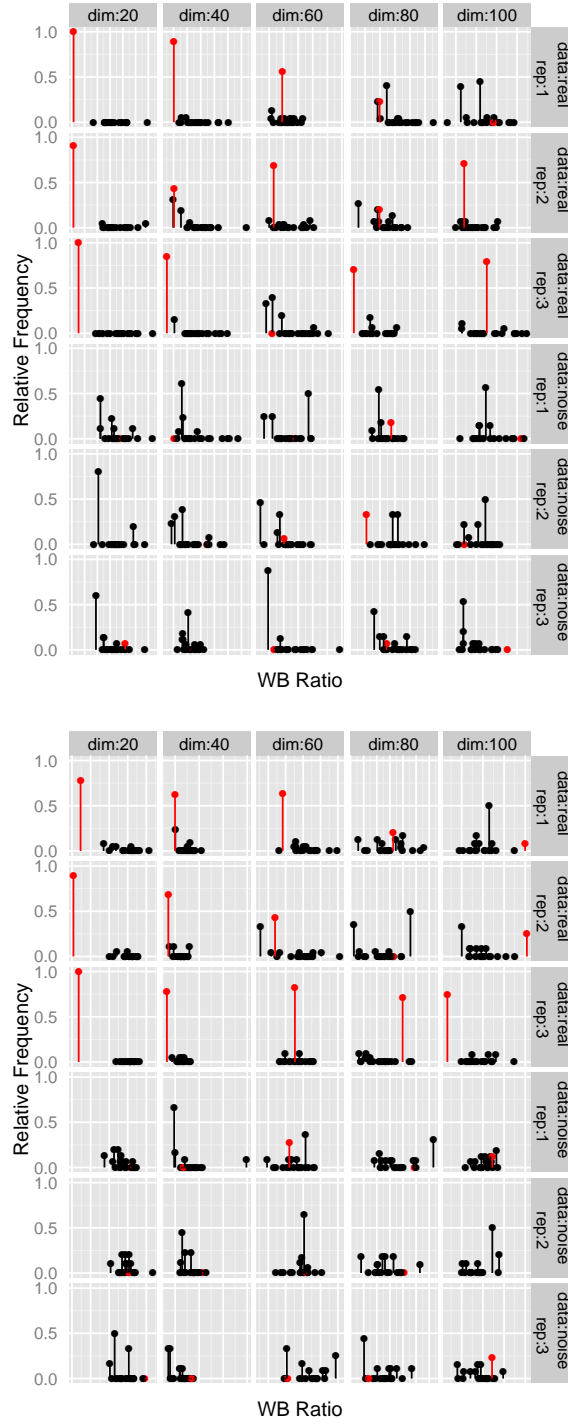
Fig. 9 Comparing the choices that subjects make for each lineup. Relative frequency of plots chosen against a measure of the distance between means, WBratio, the smaller the value the more separated are the groups. Each cell here shows the data for one of the lineups used in the experiment, 60 in total, and each "pin" represents a plot in the lineup, 20 for each lineup. Red indicates the observed data plot. Subjects were asked to pick the plot in the lineup where the groups were the most separated, so we would expect that more subjects would pick the plots with the smallest WBratio. In general, this happens, the tallest pins are in the left of each cell. The top three rows show the results for the data with real separation, so the observed data plot (red) is typically the pin on the very left of the cell, less so for the higher dimensions which are the cells at right. Also the figure on the left is for 1D projections and on the right is for 2D projections. We do not see much difference between the two figures.
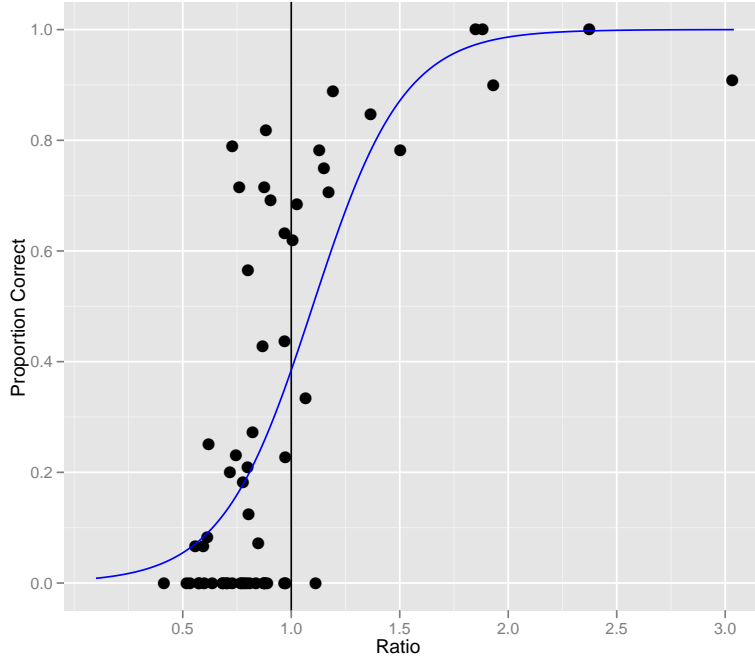
17

Fig. 10 Proportion of successful evaluation is plotted against the ratio of the minimum WBratio of the null plots and the WBratio of the observed data plot for each lineup. The vertical line represents the ratio when the WBratio of the observed data plot is equal to the minimum WBratio of the null plots. The points left to the zero line indicates a difficult lineup in the sense that at least one of the null plots had a lower WBratio value than the observed data plot. The blue line shows a logistic regression model which indicates that there is a positive effect of the ratio on the proportion correct.

is an overlap of all the colors in the plot. This signifies that there is no real group and hence the colors overlap. But as the number of dimensions increases for fixed sample size n, it can be noticed the colors starts separating out and the clusters are formed. Figure 11 shows the one dimensional projections for $p = 2$, $p = 20$, $p = 50$ and $p = 100$.

Similarly, consider 3 dimensions of data with purely noise, each dimension having 30 observations divided into 3 groups with 10 observations in each group. The projection pursuit is performed as discussed above on the data and the two dimensional projections are plotted in a scatterplot. Similarly, in this case the 3 colors starts separating out as the number of dimensions increases. Figure 11 also shows the two dimensional projections for $p = 3$, $p = 20$, $p = 50$ and $p = 100$.

It can be observed that even if the data is purely noise, the groups separate out as the number of dimensions increases for a fixed sample size. It would be interesting to find the optimal number of dimensions for which the LDA index will be unreliable given a fixed sample size n. Ripley [1996] proposes a combinatorial method to compute the probability that n observations with p dimensions divided into 2 groups are linearly separable. It quantifies that the probability that n randomly chosen observations from a continuous distribution in p-dimensional space are linearly separable when randomly divided into two groups:

$$\text{Prob}(n, p) = \frac{1}{2^{n-1}} \sum_{i=0}^{\min(n-1, p)} \binom{n-1}{i}$$

The above probability is being calculated for $n = 30$ and $n = 50$ which is represented in Figure 12. It can be noticed that the probability starts from 0 and increases steeply and eventually becomes 1. So for a fixed sample size, the probability of observing linearly separable groups becomes 1 as the number of dimensions increases
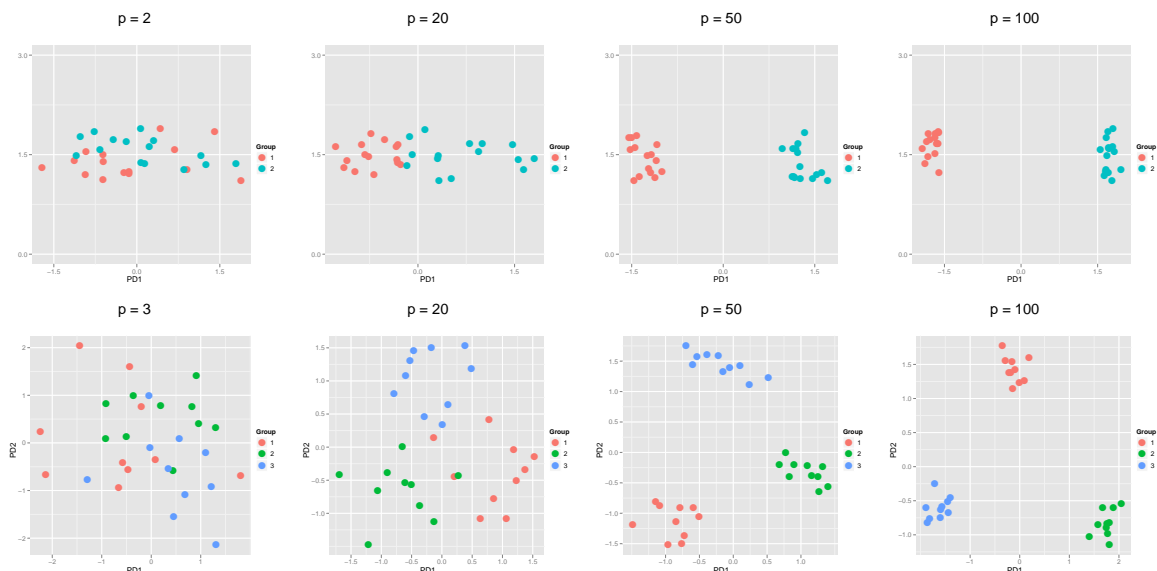
18

Fig. 11 Plots showing one and two dimensional projections for 4 different number of dimensions $p = 3$, $p = 20$, $p = 50$ and $p = 100$. Note that the groups look more and more separated as the number of dimensions increases when actually there is no real separation and the data is purely noise.

although the data is purely noise. The number of dimensions after which the LDA method becomes unreliable for $n = 30$ is 25 and for $n = 50$, it is 38.

6 Wasps data, revisited.

As seen earlier in Figure 1 it looks as if the expression patterns of the wasp groups are separated, most clearly that two of the groups are separated from the other two. The interesting question is "Is this separation real?". This can be investigated by testing the hypothesis:

H_o: There is NO difference in the expression levels between the types of wasp.
H_a: At least one of the types of wasps has different expression levels.

To test the above hypothesis, a lineup was made of the wasp data obtained from Toth et al. [2010] where the null plots were made by permuting the wasp type label, and re-doing the LDA. If there is real difference between the expression levels for the types of wasps then the observed data plot should be detectable in the lineup. Figure 13 shows a lineup.

This was replicated three times, to provide three different lineups, where the null plots changed but the observed data plot was the real wasps data. In addition, three more lineups were made that contained purely null plots. An Amazon Turk was conducted to evaluate the lineups. A total of 116 subjects evaluated the lineups. Table 5 shows the results. Success rate in detecting the plot of the wasp data was 0! This compares with that of purely noise data, with the exception of one of the replicates of the purely noise data lineup, where the simulated real data had more separation than any other plot in the lineup, and subjects picked up on this. The p-value was calculated according to the procedure given by Majumder et al.. The large p-values indicate that there is no statistically significant evidence to reject the null hypothesis. Thus we have to conclude that the separation in the wasp data is not real. It is purely the effect of high dimensionality.

In the original paper [Toth et al., 2010], the dimensionality was reduced from much higher, by choosing the genes that showed the greatest separation. So the problem of high dimensionality is actually even worse for these data. In general, reducing the data dimensions so that the sample size is bigger than dimension is not,
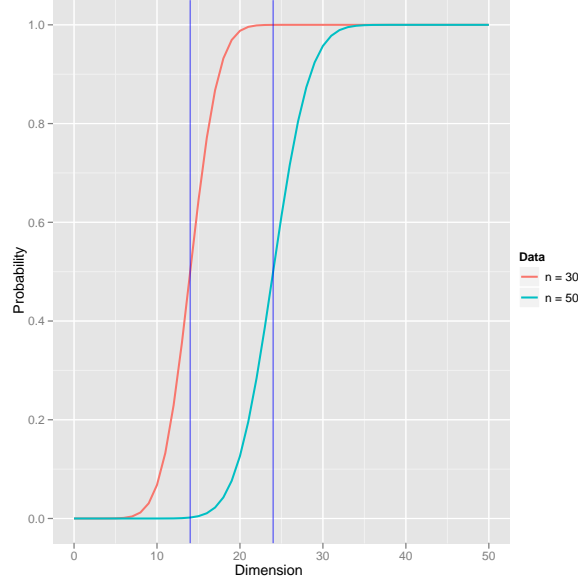
Fig. 12 Probability that $n = 30$ and $n = 50$ randomly chosen observations with p dimensions randomly divided into two groups are linearly separable. For $n = 30$, when $p = 14$, the probability of obtaining linearly separable groups is 0.5 and for $n = 50$, the probability is 0.5 when $p = 24$ which are represented by the blue vertical lines.

Table 5 Results of the Turk study on the wasps data. Proportion of successful evaluations of each lineup is shown, number of subjects, and p-value associated with the result. Notice that the most success came from oneof the pure noise lineups, which occurred because the plot with the most difference between groups happened to be the one that was randomly generated as the "real" data. Averaging the p-values for each set of lineups, for the wasps was 1.0, and for the pure noise is 0.67 suggests that the apparent separation in the wasp data is consistent with pure noise induced by the high dimensions.

Data	Replicate	Num Subjects	Prop Correct	p-value
	1	25	0.0000	1.0000
Wasps	2	13	0.0000	1.0000
	3	27	0.0000	1.0000
	1	19	0.2632	0.0002
Purely noise	2	18	0.0000	1.0000
	3	14	0.0000	1.0000

on its own, sufficient. It is important, even, with so few cases to do cross-validation, or break the sample into training and test sets before conducting analysis. LDA is known also to be a problem for HDLSS data, because it requires estimating more parameters than the available data allows. A better prospect for dimension reduction is the penalized discriminant analysis (PDA) index [Lee and Cook, 2009], which helps adjust for the over-estimation. Other results and the overall conclusions in Toth et al. [2010] are not affected by the inadequacy revealed by this visual inference analysis. A similar LDA performed on wasp gene expression data with a much higher sample size in [Toth et al., 2007] did not suffer from the HDLSS problem; we determined that there were robust separations between the groups based on those data (results not shown).

7 Conclusions

The purpose of this paper has been to apply the visual inference methods for dimension reduction using projection pursuit. The visual inference method was also used to check whether the optimization procedure in projection pursuit. It was found that the optimization procedure in the projection pursuit works well. The observers had more
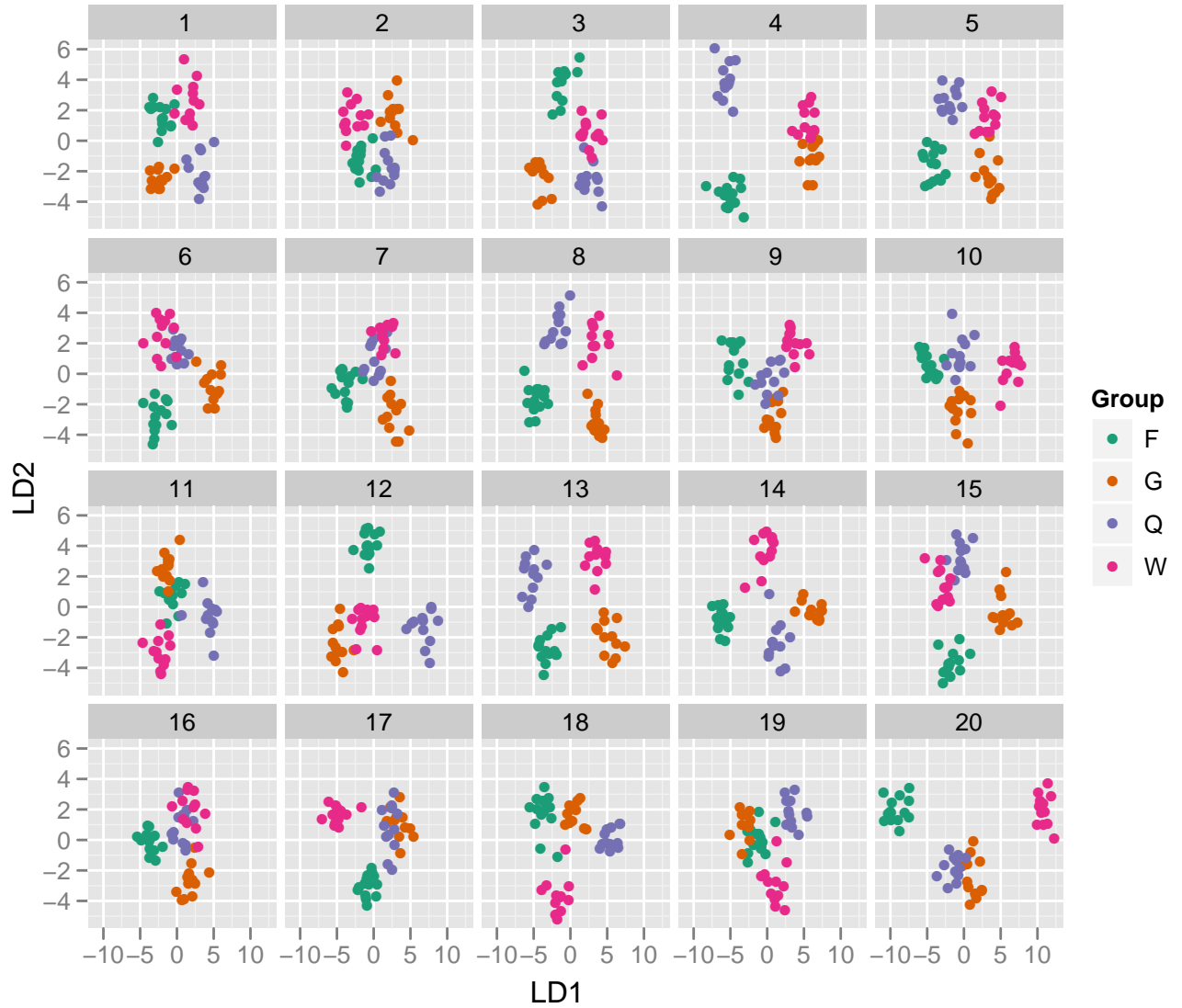
Fig. 13 Lineup showing LD1 versus LD2 from an LDA on a randomly selected subset of 40 significantly different oligos. F, Foundress; G, gyne; Q, queen and W, worker. The observed data plot is placed randomly among the 19 null plots. Which plot shows the most separation between the 4 groups? The solution is provided in the Appendix.

difficulty picking the observed data plot when it was purely noise data and the proportion of correct response was higher for data with real separation.

It was also noticed that the projection does not have a significant effect on the responses. The subjects were equally successful in identifying the separation in 1D or 2D projections. It was also noticed that for data with real separation, as the number of dimensions of purely noise increases keeping the dimensions of real separation fixed at 1 or 2, the proportion of successful evaluations decreases.

Subject specific proportions of each subject was also computed. The performance of the subjects has a large variation. Though some of the subjects do very badly, but there were some subjects who performed exceedingly well. There were no restrictions placed on the ability of the Turk workers. It was also noticed that some subjects

chose the plot which has the largest separation in the groups though it may be in the null plots and not the observed data plot.

Rotation of the groups in the 2D projections were a concern. We believed that the rotation may affect the decision of the subjects. But it was noticed that the performance of the subjects was not significantly different for the 1D and 2D projections which indicates that the rotation in the lineups with 2D projection does not affect the response of the subjects.

The amount of time taken to respond was higher for data with purely noise than data with real separation. But as the number of dimensions increases in the data with real separation, the amount of time taken is closer to the time taken for data with purely noise.

The difficulty of a lineup also has a significant positive effect on the proportion of successful response. The difficulty of a lineup depends on the null plots obtained corresponding to the observed plot. In this paper the difficulty was calculated on the basis of the measure WBratio which was calculated for each plot in each lineup. For more details on this, see Roy Chowdhury et al. [2012].

It was also shown that as the number of dimension increases, the separation between the groups increases even for data with purely noise. For a fixed sample size, the optimal number of dimension which can be used for a data with purely noise with 2 groups to perform a LDA was also given. It would be interesting to calculate an interval for the number of dimension at which the group starts separating for a data with purely noise using a visual method. Future work may explore this.

The visual inference method was also used to test whether the separation obtained among the groups is real or fake. It was noticed that only by making the number of dimensions less than the sample size may not always solve the problem of obtaining fake separation. It is recommended to divide into test and training sets and perform the LDA on the training set and then model it on the test set. It was also shown that the PDA performs better than the LDA in a large p situation.

8 Appendix

- The solution to the lineup at Figure 2 is Plot 16.
- The solution to the lineup at Figure 3 is Plot 11.
- The solution to the lineup at Figure 4 is Plot 15.
- The solution to the lineup at Figure 13 is Plot 3.

References

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

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.

D. Donoho and J. Jin. Higher Criticism Thresholding: Optimal Feature Selection when Useful Features are Rare and Weak. *Proceedings of the National Academy of Sciences of the United States of America*, 105:14790–14795, 2008.

D. Donoho and J. Jin. Feature Selection by Higher Criticism Thresholding achieves the Optimal Phase Diagram. *Philosophical Transactions of the Royal Society A*, 367:4449–4470, 2009.

S. Dudoit, J. Fridlyand, and T. Speed. Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data. *Journal of American Statistical Association*, 97:457:77 – 87, 2002.

J. H. Friedman and J. W. Tukey. A Projection Pursuit Algorithm for Exploratory Data Analysis. *IEEE Transactions on Computers*, c-23:881 – 890, 1974.

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

P. Hall, J.S. Marron, and A. Neeman. Geometric Representation of High Dimension, Low Sample Size Data. *Journal of Royal Statistical Society B*, 67:427 – 444, 2005.

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

P. J. Huber. Projection Pursuit. *The Annals of Statistics*, 13:435 – 475, 1985.

R. A. Johnson and D. W. Wichern. *Applied Multivariate Statistical Analysis (5th ed)*. Prentice-Hall, Englewood Cliffs, NJ, 2002.

S. Jung, A. Sen, and J. S. Marron. Boundary Behavior in High Dimension, Low Sample Size asymptotics of Pca. *Journal of Multivariate Analysis*, 109:190–203, 2012.

E.-K. Lee and D. Cook. A Projection Pursuit Index for Large p Small n Data. *Statistics and Computing*, page http://www.springerlink.com/content/g47n0n342761838m/?p=d2ff5a7b69eb45ef8abf7ef3aba69557&pi=3, 2009.

M. Majumder, H. Hofmann, and D. Cook. Validation of Visual Statistical Inference, Applied to Linear Models. *Journal of American Statistical Association*, Under Revision.

J. S. Marron, M. J. Todd, and J. Ahn. Distance Weighted Discrimination. *Journal of American Statistical Association*, 480:1267–1271, 2007.

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2009. URL http://www.R-project.org. ISBN 3-900051-07-0.

B. D. Ripley. *Pattern Recognition and Neural Networks*. Cambridge University Press, 1996.

N. Roy Chowdhury, D. Cook, H. Hofmann, and M. Majumder. Where's Waldo: Looking Closely at a Lineup. Technical Report 2, Iowa State University, Department of Statistics, 2012.

A. Toth, K. Varala, T. Newman, F. Miguez, S. Hutchison, D. Willoughby, J. Simons, M. Egholm, J. Hunt, M. Hudson, and G. Robinson. Wasp Gene Expression Supports an Evolutionary Link between Maternal Behavior and Eusociality. *Science*, 318:441 – 444, 2007.

A. Toth, K. Varala, M. Henshaw, S. Rodriguez-Zas, M. Hudson, and G. Robinson. Brain Transcriptomic Analysis in Paper Wasps Identifies Genes Associated with Behaviour across Social Insect Lineages. *Proceedings of the Royal Society of Biological Sciences - B*, 277:2139 – 2148, 2010.

H. Wickham and D. Cook. tourr: Implements Tour Methods in Pure R Code. http://www.R-project.org, 2010.

K. Yata and M. Aoshima. Effective PCA for High Dimension, Low Sample Size Data with Noise Reduction via Geometric Representations. *Journal of Multivariate Analysis*, 105:193–215, 2011.