

Visual Statistical Inference for High Dimension, Small Sample Size Data

Niladri Roy Chowdhury · Dianne Cook · Heike Hofmann ·
Mahbubul Majumder · Eun-Kyung Lee · Amy L. Toth

Received: date / Accepted: date

Abstract Statistical graphics play an important role in exploratory data analysis, model checking and diagnosis. In high dimensional data we often seek low-dimensional projections which reveal important aspects of the data. Projection pursuit for classification finds projections that reveal differences between groups. In many contemporary data sets the number of observations is relatively small compared to the number of variables, which is known as a large dimension small sample size (HDLSS) problem. This paper explores the performance of dimension reduction methods for finding good low-dimensional pictures of HDLSS, using new visual inference methods. These methods may be helpful to broaden the understanding of issues related to HDLSS data in the data analysis community. Methods are illustrated using data from a published paper, which erroneously found real separation, and using a simulation study done with Amazon's Mechanical Turk.

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 there are samples taken. This is commonly referred to as high dimensional, low sample size (HDLSS) data e.g. 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 of insufficient data support for estimating parameters, such as the variance-covariance matrix, as required by these 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 possibilities for lines. 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.

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

Eun-Kyung Lee
Department of Statistics, Ewha Womans University, Seoul, Korea
E-mail: lee.eunk@gmail.com

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

LDA finds the low-dimensional space where the groups are most separated, by solving an eigenvalue decomposition problem comparing distances between group means with variance around each mean. When there are few sample points in high dimensions, differences between groups can be found in many different low-dimensional spaces.

Marron et al (2007) describes the estimation issues associated with HDLSS. One of the problems with HDLSS data is that not all the measured variables are “important” for understanding the underlying phenomenon of interest. Advancements in PCA to handle HDLSS data have been done by Jung et al (2012) and Yata and Aoshima (2011). Donoho and Jin (2009) and Donoho and Jin (2008) study optimal variable selection and introduce 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, hence contributes weakly to the classification decision. Penalization (e.g. Witten and Tibshirani, 2011; Lee and Cook, 2009) is another common approach.

The issues of working with HDLSS data are still not clear to many data analysts. For example, in Toth et al (2010) LDA is used to examine gene expression data of wasps containing 447 variables and 50 cases. Figure 1 reproduces the result in this paper. There are 50 different paper wasps divided into 4 types: Foundress (F), Gyne (G), Queen (Q) and Worker (W), 14 wasps of type Foundress and 12 each of the other 3 types. The authors, knowing that LDA requires that the dimension (p) should be smaller than the number of observations (n), first reduced the dimension from 447 to 40 by randomly selecting a subset of significantly different oligonucleotides. LDA produced a $d = 2$ projection of best separation. This is almost the same approach as used in Dudoit et al (2002), 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. There exists no conventional inferential methods which enables us to conclude whether this apparently clear separation is statistically significant or not. Typically data is broken into training and test sets, or cross-validation is conducted to assess the significance of difference, using test set error.

We suspect that new methods for inference on graphics might be helpful for building understanding very generally. Visual statistical inference was first conceptually introduced by Buja et al (2009) and later formalized and validated by Majumder et al (2012). Using visual inference, it can be shown that there is no real difference between the wasp groups - what you see is a mirage.

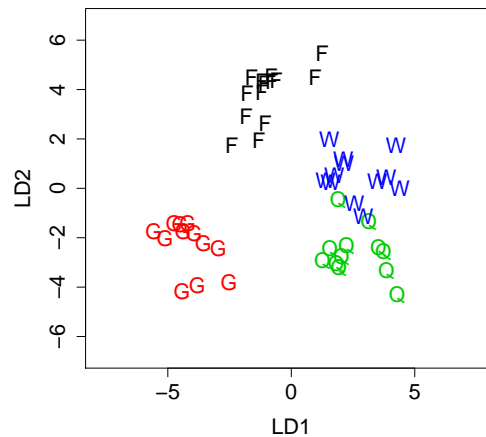


Fig. 1 LD1 versus LD2 from an LDA on a randomly selected subset of 40 significantly different oligos : F, Foundress; G, gyne; Q, queen and W, worker. It can be noticed that the groups F and G are separated. This plot is generated to match Figure 2 in Toth et al (2010).

This paper describes visual statistical inference as applied to dimension reduction for HDLSS. In particular we focus on dimension reduction using projection pursuit, and the effect that having large dimension has on the

robustness of the separation between groups. Small simulation experiments are used to examine the problem in a controlled setting. The next section explains visual inference methods. Section 3 discusses the dimension reduction methods. Section 4 discusses the experiment designed to examine people’s perception of separation in the presence of real separation and “purely noise” for simulated HDLSS data. Amazon’s Mechanical Turk (Amazon, 2010) was used to conduct the experiment. Section 5 discusses the collected data and results. The wasps data is revisited at the end of the paper.

2 Visual inference methods

Buja et al (2009) proposed two protocols, the Rorschach and the lineup protocols. While the Rorschach protocol helps to understand the extent of randomness, the lineup protocol is used for testing significance of findings. This method is called visual statistical inference. Majumder et al (2012) made a head-to-head comparison between visual statistical inference tests and classical tests which yielded comparable results. Unlike classical hypothesis testing, the test statistic in visual inference is not numeric, but a plot that is appropriately chosen to display a distinctive pattern in case that the null hypothesis is not true. The lineup protocol embeds this observed data plot amongst the null plots. Null plots are created consistently with the null hypothesis. Human subjects are asked to identify the plot with the most distinct feature(s). If the human subjects choose the plot of the observed data, this is regarded as evidence against the null hypothesis and with enough support, the null hypothesis is rejected. 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 the other null plots. A comparison of visual testing with conventional testing is shown in Table 1.

For example, suppose we have data that 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. Thus the null and alternative hypothesis would be

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

We choose as the visual test statistic, $V(y)$, to be a dot plot of the observed data for two groups as shown in Table 1. The null plots are generated by assuming that H_o is true. This is achieved by randomly permuting the values of group variable site. The observed data plot is placed randomly among 19 null plots to obtain a lineup shown in Figure 2. If H_o is not true, a vertical displacement of the two groups should be a distinctive feature in the test statistic that is not present in the null plots. The lineup is shown to a human observer who is asked to identify the plot that is most different than the others in the lineup. If the observer can identify the plot of the real data, there will be reason to believe that the observed data plot has a pattern which is absent in the null plots leading to a rejection of the null hypothesis. If the viewer cannot identify the observed data plot, we fail to reject the null hypothesis.

Majumder et al (2012) describes the methods of obtaining the power of the visual test. For their simulation experiments they recruited human observers from Amazon Mechanical Turk (Amazon, 2010) and estimated power from human evaluations of lineups. They also obtained the power theoretically with an assumption supported by the experimental data. The result suggests that the power of visual statistical inference is comparable to conventional tests in a setting of testing the parameters of linear regression models. The subject specific power of the visual test is also estimated from the multiple responses data from each human observer.

3 Explanation of dimension reduction methods

In this work projection pursuit (e.g. Friedman and Tukey, 1974) is used for dimension reduction. Projection pursuit (PP) finds the most interesting low dimensional projection of high dimensional data by maximizing some criterion

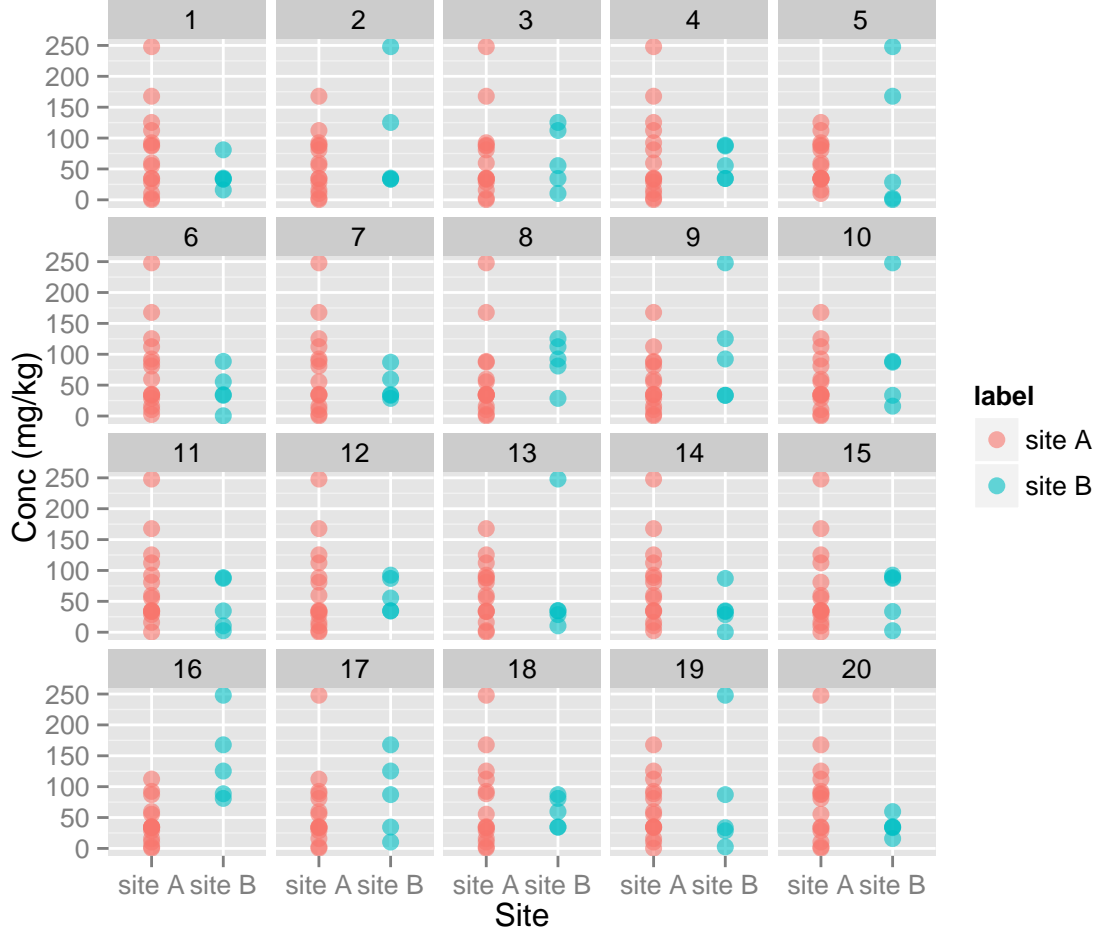


Fig. 2 A typical lineup ($m = 20$) for testing $H_o : \mu_1 = \mu_2$. When the alternative hypothesis is true, the observed data plot should have the largest vertical difference between the centers. Can you identify the observed data plot? The solution to the lineup is provided in the Appendix.

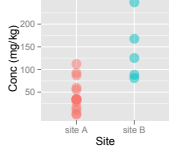
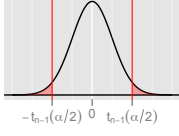
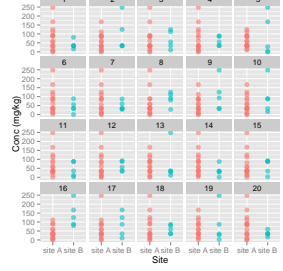
of interest, e.g. variance or clustering or group separation. As pointed out in Huber (1985) the most exciting feature of projection pursuit is that it can bypass the curse of dimensionality.

In classification problems, linear discriminant analysis (LDA) can be used to find a low-dimensional space where the groups are most separated. This corresponds to using the LDA index (Lee and Cook, 2009) in projection pursuit. Let \mathbf{X}_{ij} be the p -dimensional vector of the j th observation in the i th class, $i = 1, \dots, g$, $j = 1, \dots, n_i$, g is the number of classes, n_i is the number of observations in class i , and $n = \sum_{i=1}^g n_i$. Let $\bar{\mathbf{X}}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} \mathbf{X}_{ij}$ be the i th class mean and $\bar{\mathbf{X}}_{..} = \frac{1}{n} \sum_{i=1}^g \sum_{j=1}^{n_i} \mathbf{X}_{ij}$ be the total mean. The LDA PP index is

$$I_{LDA}(\mathbf{A}) = \begin{cases} 1 - \frac{|\mathbf{A}^T \mathbf{W} \mathbf{A}|}{|\mathbf{A}^T (\mathbf{W} + \mathbf{B}) \mathbf{A}|} & \text{for } |\mathbf{A}^T (\mathbf{W} + \mathbf{B}) \mathbf{A}| \neq 0 \\ 0 & \text{for } |\mathbf{A}^T (\mathbf{W} + \mathbf{B}) \mathbf{A}| = 0 \end{cases} \quad (1)$$

where \mathbf{A} is an orthogonal projection onto a k -dimensional space and

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

$$\mathbf{B} = \sum_{i=1}^g n_i (\bar{\mathbf{X}}_{i.} - \bar{\mathbf{X}}_{..}) (\bar{\mathbf{X}}_{i.} - \bar{\mathbf{X}}_{..})^T : \text{between-class sums of squares,}$$

$$\mathbf{W} = \sum_{i=1}^g \sum_{j=1}^{n_i} (\mathbf{X}_{ij} - \bar{\mathbf{X}}_{i.}) (\mathbf{X}_{ij} - \bar{\mathbf{X}}_{i.})^T : \text{within-class sums of squares.}$$

For HDLSS data, the penalized discriminant analysis (PDA) index (Lee and Cook, 2009) is more robust. Let \mathbf{X}_{ij}^* be the standardized vector of \mathbf{X}_{ij} . Then

$$\mathbf{B}^s = \sum_{i=1}^g n_i (\bar{\mathbf{X}}_{i.}^* - \bar{\mathbf{X}}_{..}^*) (\bar{\mathbf{X}}_{i.}^* - \bar{\mathbf{X}}_{..}^*)^T : \text{between-class sums of squares of the standardized data}$$

$$\mathbf{W}^s = \sum_{i=1}^g \sum_{j=1}^{n_i} (\mathbf{X}_{ij}^* - \bar{\mathbf{X}}_{i.}^*) (\mathbf{X}_{ij}^* - \bar{\mathbf{X}}_{i.}^*)^T : \text{within-class sums of squares of the standardized data}$$

where $\bar{\mathbf{X}}_{i.}^*$ is the i th class mean of the standardized data and $\bar{\mathbf{X}}_{..}^*$ is the total mean of the standardized data, which is 0. The PDA index is defined as

$$I_{PDA}(\mathbf{A}, \lambda) = 1 - \frac{|\mathbf{A}^T \{ (1 - \lambda) \mathbf{W}^s + n \lambda \mathbf{I}_p \} \mathbf{A}|}{|\mathbf{A}^T \{ (1 - \lambda) (\mathbf{B}^s + \mathbf{W}^s) + n \lambda \mathbf{I}_p \} \mathbf{A}|} \quad (2)$$

where \mathbf{A} is an orthonormal projection onto a k -dimensional space and $\lambda \in [0, 1)$ is a predetermined parameter. Penalized LDA (Witten and Tibshirani, 2011) is a similar approach.

These indices are available for projection pursuit using the `tourr` package (?) in R (R Development Core Team, 2009). 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 $d \leq p$. The guided tour function is used here. The guided tour will converge to a maximally interesting projection, here that is one where groups show the biggest separation. For this paper we used $d = 1$ or 2.

4 Experiment

4.1 Experimental design

The goal is to determine how well real separation is distinguishable from random noise. To achieve this, the experiment is set up with factors, real separation or purely noise, data dimension and projection dimension. Real separation is achieved by setting 1 or 2 variables with real separation among a number of purely noise variables. Sample size is fixed to keep the experiment manageable. Also mean shift is kept fixed. The levels of the factors used in the experiment are given in Table 2. Three replicates at each level are generated. These produced 60 different “observed data sets”, and thus, 60 different lineups.

Table 2 Levels of the factors used for the experiment.

n	projection(d)	separation	dimension (p)	replicates
30	1	Yes	20, 40, 60, 80, 100	3
		No	20, 40, 60, 80, 100	3
30	2	Yes	20, 40, 60, 80, 100	3
		No	20, 40, 60, 80, 100	3

Figure 3 shows the distribution of absolute difference of means for data with and without separation for different dimensions. The distributions of data with and without separation are shown in red and blue respectively. The area of the distribution of pure noise which is above the 5th percentile of the distribution of data with separation is shown in dark blue and it can be seen that the dark blue region increases with dimension (p). This indicates that as dimension increases, the distributions of data with or without separation gets closer. Hence the probability of obtaining separation among groups for purely noise data increases with higher dimensions. Fixing the area of the dark blue region and calculating the dimensions to obtain the required region provides the choice of levels of dimension used in the experiment. More details are provided in Section 7.

4.2 Simulation process

Two groups of p dimensions of data with 15 observations in each group are 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, \mathbf{X} , is obtained where the first 15 observations are from group 1 and the last 15 observations are from group 2. The data for both the groups is purely noise, having no dependence on the groups. So \mathbf{X} can be written as

$$\mathbf{X}^{n \times (p+1)} = (\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_p, \text{Group})$$

where each \mathbf{X}_i is a vector of dimension 30 for $i = 1, \dots, p$. This matrix \mathbf{X} excluding the Group variable gives the p -dimensional pure noise data or data with no separation.

To introduce real separation in the data, values for p -th variable \mathbf{X}_p are shifted apart by 6 units between the two groups:

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

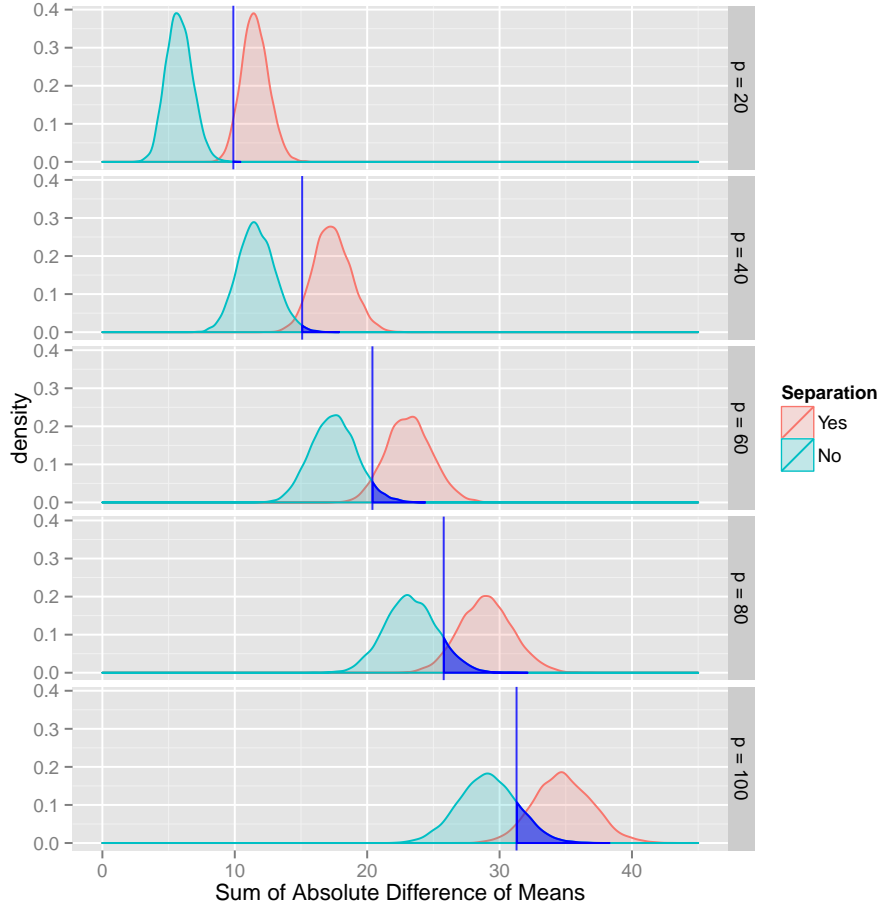


Fig. 3 Plot showing the distribution of the sum of absolute difference of means for data with and without separation for different dimensions. The distributions of data with real separation (V) and purely noise data (U) are shown in red and blue respectively with the dark blue line showing the 5th percentile of V. The dark blue area shows the area of U which is greater than the 5th percentile of V. The dark blue region (δ) increases as dimension (p) increases.

Hence two sets of data are obtained, each with p dimensions with $n = 30$ observations in each dimension divided into two groups. One of the datasets is pure noise and the other has some real separation in 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 projection of best separation, yielding $\mathbf{Y} = \mathbf{XA}$.

The above procedure is effectively the same for $d = 2$ dimensional projections with few key differences. The first 10 observations is assigned as group 1, the second 10 observations as group 2 and the last 10 as group 3. So, collectively, \mathbf{X} can be written as

$$\mathbf{X}^{n \times (p+1)} = (\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_p, \text{Group})$$

where each \mathbf{X}_i is a vector of dimension 30 for $i = 1, \dots, p$. This matrix \mathbf{X} excluding the Group variable gives the p -dimensional noise data or data with no separation with 3 groups.

To introduce real separation, the means of the 3 groups are adjusted in the last two dimensions i.e. \mathbf{X}_{p-1} and \mathbf{X}_p . The adjustment is done in the following way:

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

If \mathbf{X}_{p-1} versus \mathbf{X}_p are plotted in a scatterplot, the points cluster along the vertices of an equilateral triangle of side 6. Hence the data with 2 dimensions of real separation divided into 3 groups is obtained. A projection pursuit with a PDA index is performed to obtain the $d = 2$ -dimensional projection of best separation, yielding $\mathbf{Y} = \mathbf{XA}$.

4.3 Producing lineups

Two different visual test statistics, $V_1(\mathbf{Y})$ and $V_2(\mathbf{Y})$ are used in this paper, for representing 1D and 2D data. $V_1(\mathbf{Y})$ is a horizontal jittered dot plot, with color representing groups. $V_2(\mathbf{Y})$ is a scatterplot with color representing groups. Symbols are kept constant for uniformity of appearance. Figure 4 shows the two different visual test statistics.

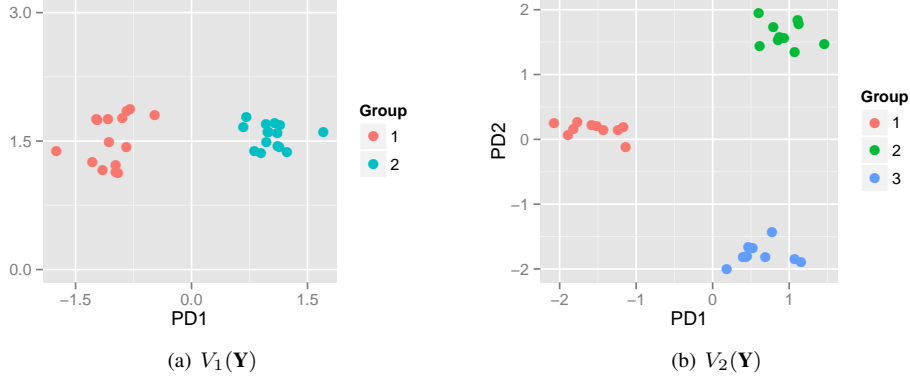


Fig. 4 The visual test statistics $V_1(\mathbf{Y})$ and $V_2(\mathbf{Y})$ used. $V_1(\mathbf{Y})$ is a horizontal jittered dot plot while $V_2(\mathbf{Y})$ is a scatterplot of the first and second dimensional projections, with color representing groups in both cases.

To obtain the null plots in a lineup, the group variable is permuted in order to break any dependence between the group variable and the other variables. Projections are obtained and plotted in the same way as the test statistic. The test statistic, which is the observed data plot, is placed randomly among the 19 null plots. To maintain the same orientation of the two groups in the 1D projection lineup, the mean of the projections for each group is calculated for each plot in the lineup and the group with the lower mean is considered to be group 1 and the other group 2. Figure 5 shows an example lineup having treatment levels $p = 20$, separation = Yes and $d = 1$. Similarly, Figure 6 shows an example lineup for $p = 100$, separation = No and $d = 2$.

A statistic measuring the ratio of the average distance within clusters to the average distance between clusters (Hennig, 2010), called WBratio, is calculated for each plot in the lineup of both 1D and 2D projections. An additional statistic Wilk's λ (e.g. Johnson and Wichern, 2002) is calculated for 2D projections. To account for the occasional lack of convergence of the projection pursuit optimization, 30 null plots are generated. The 19 null plots which have the smallest Wilk's λ values are used for the lineup.

4.4 Data collection

Subjects for the experiment were recruited through Amazon Mechanical Turk (Amazon, 2010). Each subject was shown a sequence of 10 lineups. Subjects also had the freedom to process 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 across the globe.

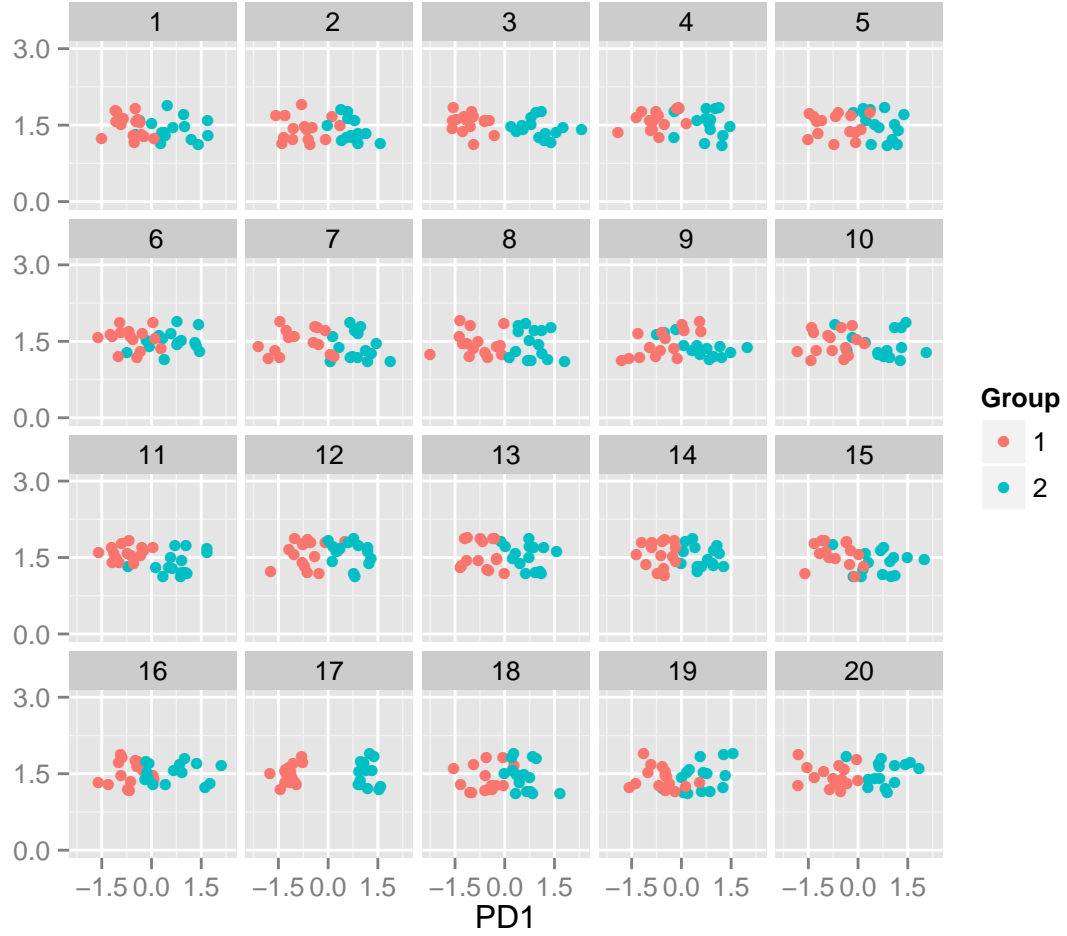


Fig. 5 Lineup ($m = 20$) from treatment with $p = 20$, separation = Yes and $d = 1$. The subjects were asked to identify the plot with the most separated colors. Can you identify the observed data plot? The solution to the lineup is provided in the Appendix.

5 Results

5.1 Data cleaning

Amazon Mechanical Turk subjects are paid for their responses. Since the process is not manually monitored, it is possible that some subjects do not make an honest effort to find the observed data plot. To counter this problem, each subject is given a very easy lineup (a lineup with $p = 10$ dimensions with some real separation). The subjects who failed to give a correct response to this lineup are removed from the study. If their response in this lineup is correct, we remove the subject's response for this lineup but keep all their responses for all the other lineups evaluated.

5.2 Effect of experimental factors

We would expect that subjects are correct more often when there is real separation and that as dimension increases, correctness decreases. Figure 7 examines this. The proportion correct is plotted against the different levels of dimension faceted by the levels of separation and projection. The 3 different dots for each level shows the 3 lineup

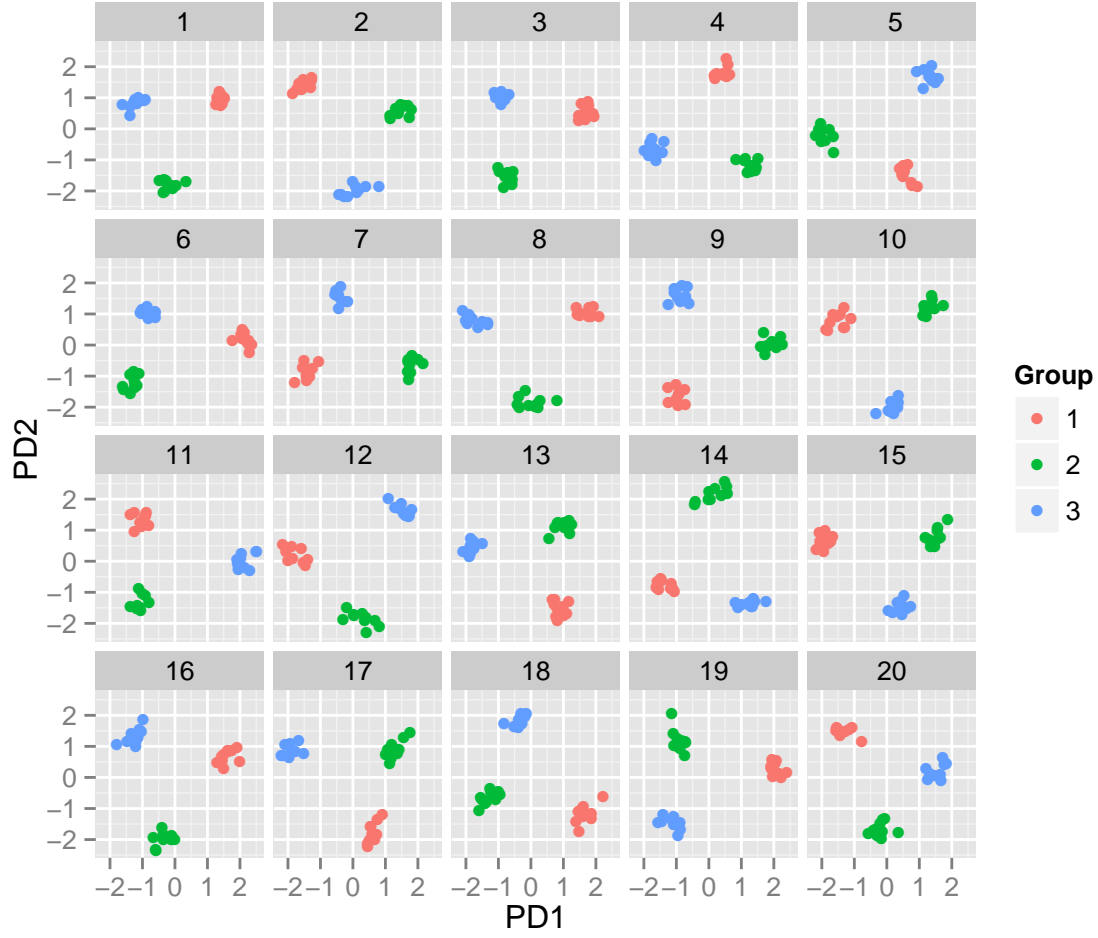


Fig. 6 Lineup ($m = 20$) from treatment with $p = 100$, separation = No and $d = 2$. The subjects were asked to identify the plot with the most separation between the colored groups. Can you identify the observed data plot? The solution is provided in the Appendix.

replicates for each level combination of the factors. The line represents the fitted fixed effects from a logistic regression model. The success rate is higher for low dimension, and decreases as dimension increases, for real separation for both 1D and 2D projections. For noise data, the success rate is flat across dimension. There also appears to be increasing variance as dimension increases.

Table 3 shows the estimates of the parameters from the fixed effects logistic regression model, the standard errors and the corresponding p -values. We 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 suggests that the difference between 1D and 2D projections is not significant at the 5% level of significance. One of our concerns with the 2D projections is that the rotation of the group was not adjusted, and that this might diminish the subjects ability to identify the observed data plot. The lack of significant difference between 1D and 2D results suggest that the lack of adjustment was not important. The slope of the covariate dimension is different when there is separation or not. When the observed data plot has real separation, the ability to detect it decreases with dimension.

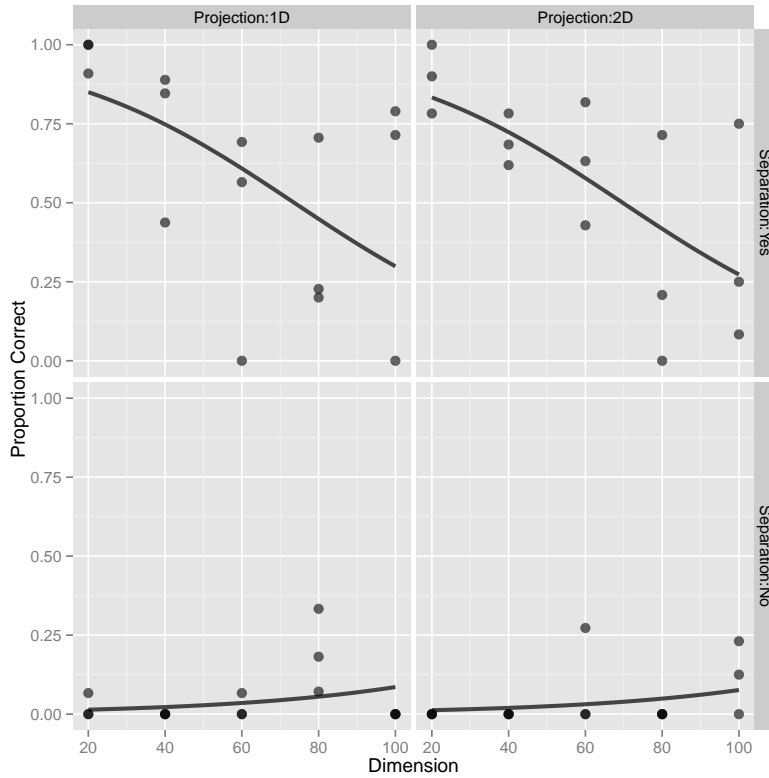


Fig. 7 Proportion of correct response against dimension faceted by projection and separation. The three points represents the three replicates for each treatment level. A fixed effects logistic regression model is overlaid on the points. It can be seen that the success rate decreases as dimension increases for data with real separation. When the data is purely noise data, the success rate is flat across dimensions. The success rate does not change with projection.

Table 3 Table showing the estimate, the standard error and the p -value of the parameters used in logistic regression model. The covariates dimension and separation are highly significant while projection is not significant at 5% level of significance. The interaction term between dimension and separation is also significant.

Parameters	Estimate	Std. Error	p -value
Intercept	2.381	0.278	0.000
dimension(p)	-0.032	0.004	0.000
separation = No	-7.097	0.911	0.000
projection = 2D	-0.127	0.181	0.483
separation:dimension	0.056	0.011	0.000

5.3 Time taken to respond

We would expect that the amount of time taken to respond will increase with the difficulty of identifying the observed data plot in a lineup. Figure 8 shows the distribution of the time taken in seconds to respond (on a log scale) by dimension and projection. Color indicated separation or not. A loess smoother is added. Notice that, when the data has some real separation (red), as the dimension increases, subjects take more time to respond to the lineups. But when the data is purely noise (blue), the increase of dimension does not have any effect on the time. This suggests that as the number of dimensions increases, it becomes harder 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 dimension. 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.

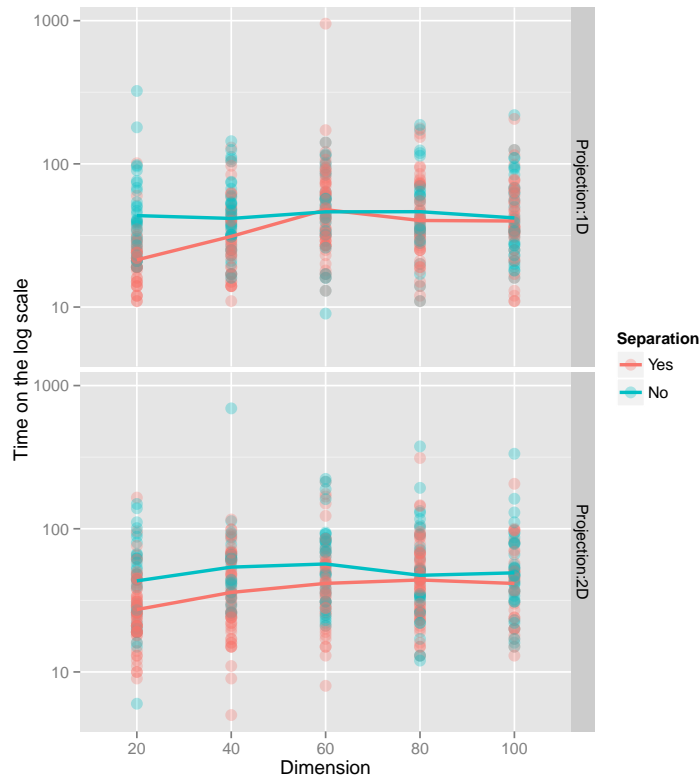


Fig. 8 Time taken in seconds to respond on log scale against dimension colored by separation and faceted by projection. A loess curve is fitted through the points for each separation and projection. Time taken to respond is higher when the data has no separation. Also as dimension increases, the time to answer increase when there is real separation.

5.4 What affects decisions?

Figure 9 examines the subjects choices in detail. 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 that the plot was chosen. Red represents the observed data plot. Two separate figures are made for the two projections. The top three rows correspond to data containing real separation between the groups, and for the bottom three rows all of the data is purely noise. Columns indicate dimension (p). Replicates are in different rows. The taller the pin the more often that particular plot is chosen from the lineup. We asked subjects to pick the plot where the groups are most separated, and this is effectively what they picked. The plot in each lineup with the smallest WBratio tend to have the highest frequency. This is more obvious when there is real separation, and also when dimension is 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 though, the choices are a little surprising, for example separation = No, dimension = 60, rep = 2, projection = 2D. Investigating these lineups further may reveal why this is. (See supplementary material).

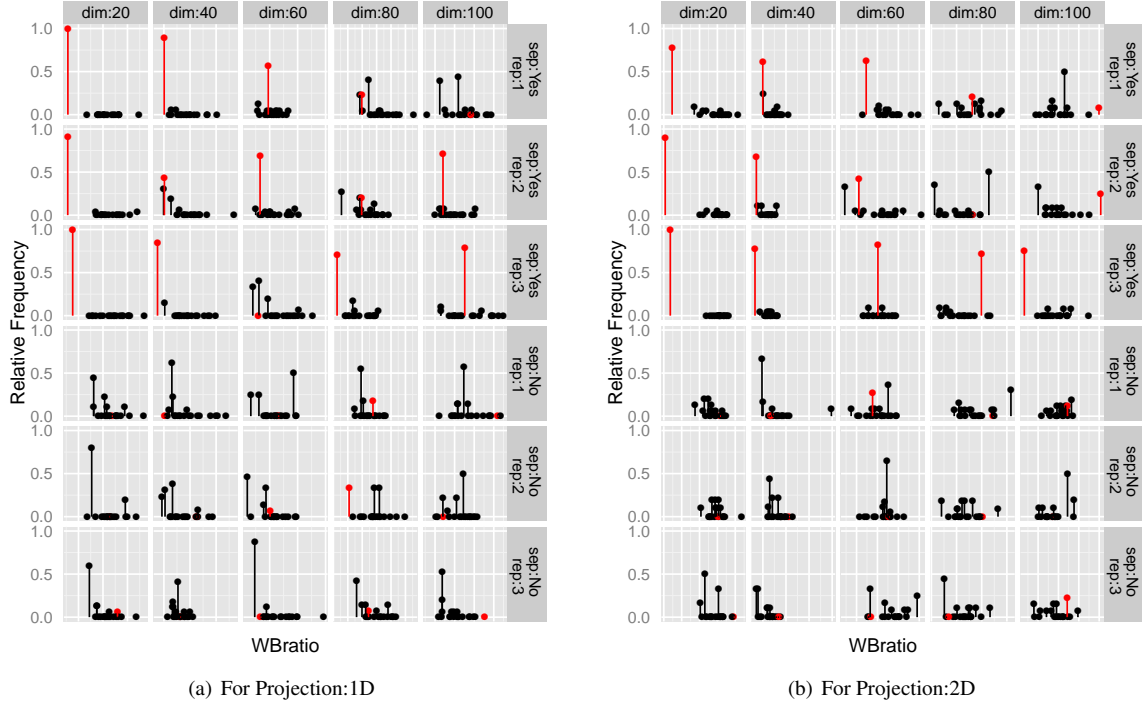


Fig. 9 Comparing the choices that subjects make for each lineup. Relative frequency of plots chosen against a measure of the distance between means, WBratio, the smaller the value the more separated are the groups. Each cell here shows the data for one of the lineups used in the experiment, 60 in total, and each “pin” represents a plot in the lineup, 20 for each lineup. Red indicates the observed data plot. Subjects are asked to pick the plot in the lineup where the groups are 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 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. Figure (a) shows 1D projections and Figure (b) shows for 2D projections. There is not much difference between the two figures.

5.5 How do the null plots affect choices?

We have learned that subjects tend to pick the plot in the lineup that exhibits the most separation. Because visual inference only allows for a finite (small) number of comparisons against the sampling distribution, the influence of the null plots in the lineup on the observer’s choice is important. If any null plot has a strong signal, subjects may choose this plot over the observed data plot. To gauge the influence of the null plots, we calculate the ratio between the minimum WBratio of the null plots and the WBratio for the observed data plot for each lineup. Figure 10 examines the influence of the null plots on this pick. The proportion correct and mean time taken in seconds are plotted against ratio. The vertical line is a reference line where the ratio is 1: the observed data plot has the same signal as the most extreme null plot. The points to the right of the line should indicate easier lineups and those to the left indicate more difficult lineups in the sense that the null plots have more signal than the observed data plot. We can see that as the ease increases, the success rate increases. Also as the ratio increases, time taken to choose decreases, suggesting easier lineups. More details on the measurement of the influence are available in Roy Chowdhury et al (2012)).

6 Wasps data, revisited.

We return now to the motivating example. Figure 1 suggested that the expression patterns of the wasp groups are different. The question of interest is “Is this separation real?”. This can be investigated by testing the hypothesis:

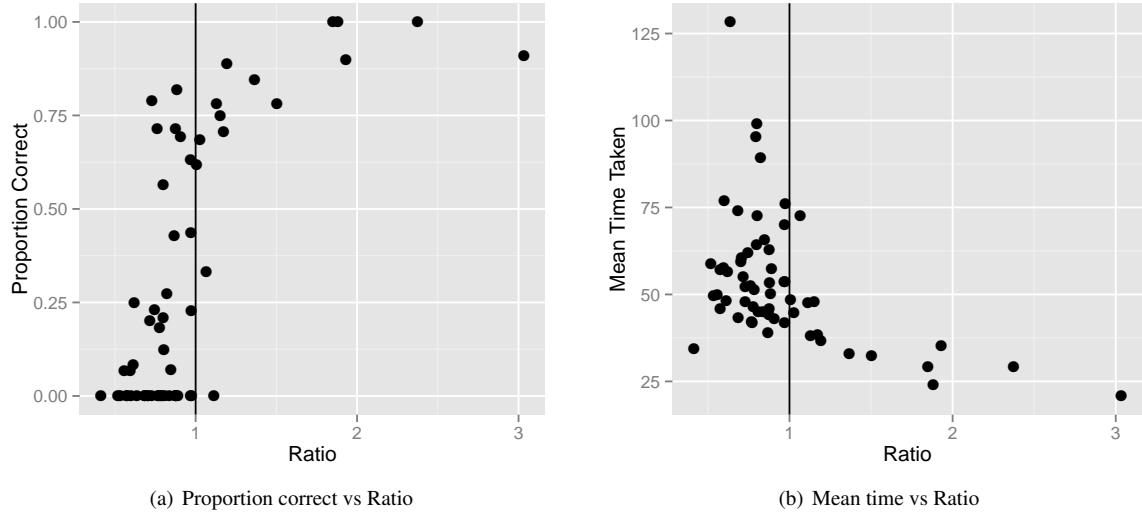


Fig. 10 (a) Proportion of correct response against the ratio of the minimum WBratio of the null plots and the WBratio of the observed data plot for each lineup. (b) Mean time taken to respond in seconds against the ratio. The vertical line represents ratio equal to 1 when WBratio of the observed data plot is equal to the minimum WBratio of the null plots. The points left to the 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. As the ratio increases, the success rate increases and the mean time taken decreases indicating that the subjects have an easier time in identifying the observed data plot.

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.

A lineup is made of the wasp data obtained from Toth et al (2010) to test H_o where the null plots are 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 detected in the lineup. Figure 11 shows a lineup.

This is replicated three times, to provide three different lineups, where the null plots are changed but the observed data plot is the real wasps data. In addition, three more lineups are made that contained only null plots, with one plot chosen randomly to act as an observed data plot. These lineups were included in the Amazon Turk experiment. A total of 116 subjects evaluated the 6 lineups. Table 4 shows the results. Success rate in detecting the plot of the wasp data is 0! This is worse than that of purely noise data. You will notice that for one of the purely noise lineups, subjects were very often correctly picked the (random) observed data plot. This happened because the randomly generated observed data plot actually had more separation than any other plot in that lineup. This is the nature of randomness, but makes for interesting results here. The p -value is calculated according to the procedure given by Majumder et al (2012). The large p -values indicate that there is no statistically significant evidence upon which we 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.

The probability of separation by chance in purely noise data given a fixed sample size and dimension, was quantified by Ripley (1996) (Proposition 3.1). Figure 12 illustrates this probability with 1D projections having sample size $n = 30$, for different p . When $p = 2$, $P(\text{separation}|n = 30, p = 2) = 0$ and it reaches 1 when $p = 25$. For data of the size of the wasps, $n = 50$, the probability of obtaining separation with 2 groups is 1 when $p = 38$, so it would be much less than this if we could calculate the probability for 4 groups, in 2D.

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, 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

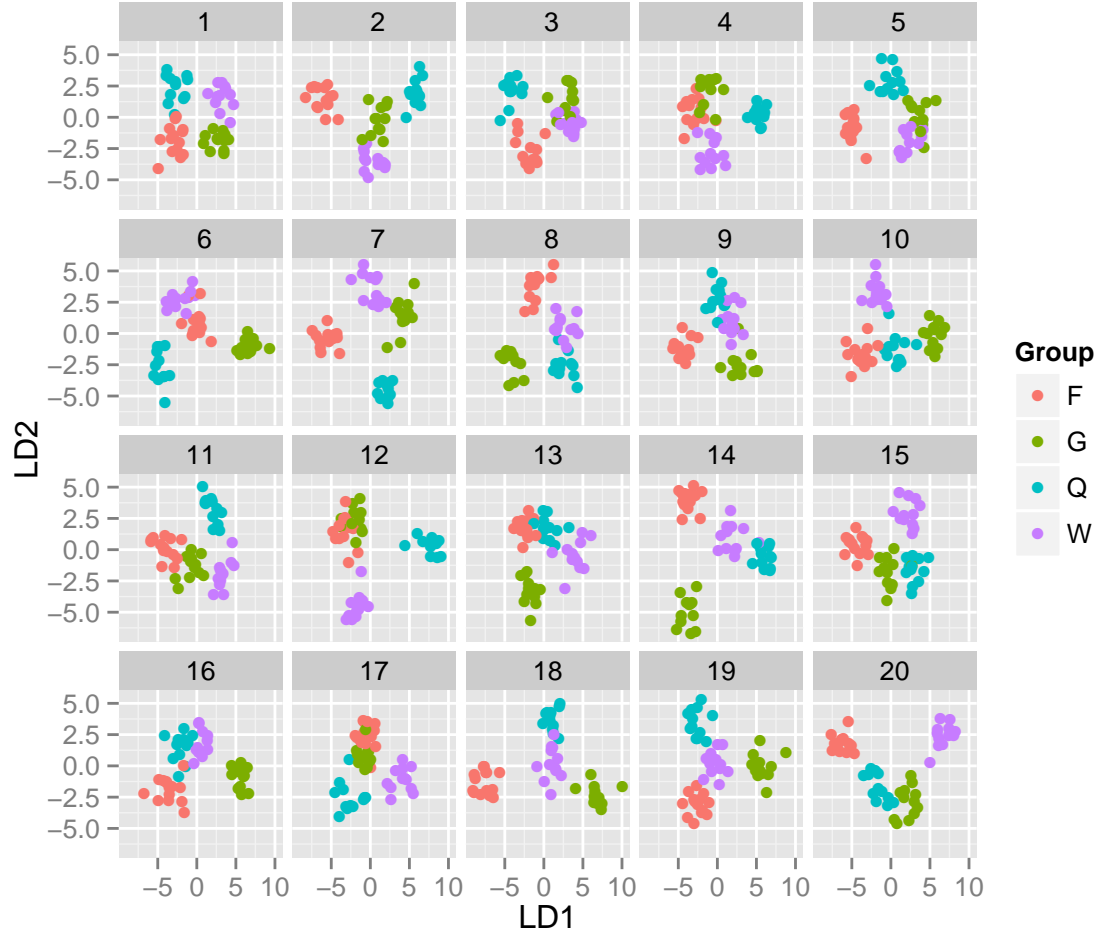


Fig. 11 Lineup showing LD1 versus LD2 from an LDA on a randomly selected subset of 40 significantly different oligos. F, Foundress; G, gyne; Q, queen and W, worker. The observed data plot is placed randomly among the 19 null plots. Which plot shows the most separation between the 4 groups? The solution is provided in the Appendix.

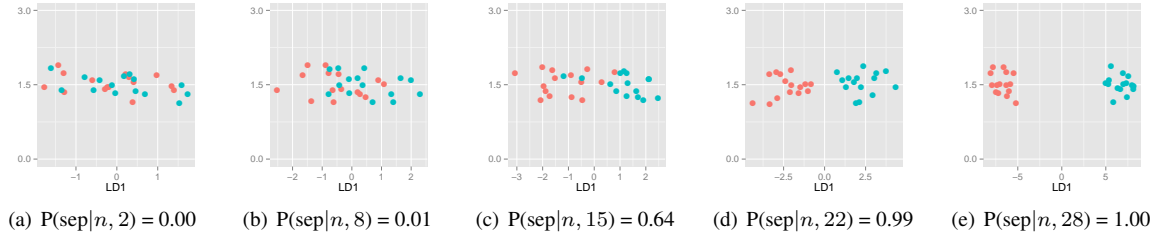


Fig. 12 Plots showing one dimensional projections for 4 different values of dimension $p = 2, p = 8, p = 15, p = 22$ and $p = 28$ for a fixed sample size of $n = 30$. The 1D projections separate out as dimension increases although the data has no real separation.

Table 4 Results of the Turk study on the wasps data. Proportion correct for each lineup is shown, with the number of subjects, and p -value associated. The success rate is highest for one of the purely noise lineups, which occurred because the plot with the most difference between groups happened to be the one that is randomly generated as the “real” data. Averaging the p -values for each set of lineups, for the wasps is 1.0, and for the pure noise, is 0.67 suggesting 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
Wasps	1	25	0.0000	1.0000
	2	13	0.0000	1.0000
	3	27	0.0000	1.0000
Purely noise	1	19	0.2632	0.0002
	2	18	0.0000	1.0000
	3	14	0.0000	1.0000

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

This paper examined how visual inference works for HDLSS data. From the results it is clear that people can detect real separation from pure noise up to a reasonably high dimension, for 1D and 2D projections. Visual inference can be used to improve the understanding of the emptiness of space in HDLSS data. That is, a 1D or 2D projection of HDLSS data, as provided by LDA, for example, will likely look separated purely due to this emptiness. Visual inference makes this clearer by placing the data along with comparisons where it is known that they were produced with no real difference between classes.

There are several natural next steps for this research. One is to examine the possibility of using visual inference to obtain confidence bands for the value of p , where separation is certain, for fixed sample size and dimension, particularly if a component of real separation is included. Another direction is to build metrics to quantify the difficulty of a lineup and the influence that null plots have on identifying the data plot.

Acknowledgement

This work was funded by National Science Foundation grant DMS 1007697.

References

- Amazon (2010) Mechanical Turk. URL <http://aws.amazon.com/mturk/>
- Buja A, Cook D, Hofmann H, Lawrence M, Lee E, Swayne D, Wickham H (2009) Statistical Inference for Exploratory Data Analysis and Model Diagnostics. Royal Society Philosophical Transactions A 367(1906):4361–4383
- Donoho D, Jin J (2008) 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 105:14,790–14,795
- Donoho D, Jin J (2009) Feature Selection by Higher Criticism Thresholding achieves the Optimal Phase Diagram. Philosophical Transactions of the Royal Society A 367:4449–4470
- Dudoit S, Fridlyand J, Speed T (2002) Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data. Journal of American Statistical Association 97:457:77 – 87

- Friedman JH, Tukey JW (1974) A Projection Pursuit Algorithm for Exploratory Data Analysis. *IEEE Transactions on Computers* c-23:881 – 890
- Hall P, Marron J, Neeman A (2005) Geometric Representation of High Dimension, Low Sample Size Data. *Journal of Royal Statistical Society B* 67:427 – 444
- Hennig C (2010) *fpc* : Flexible Procedures for Clustering. R package version 2
- Huber PJ (1985) Projection Pursuit. *The Annals of Statistics* 13:435 – 475
- Johnson RA, Wichern DW (2002) *Applied Multivariate Statistical Analysis* (5th ed). Prentice-Hall, Englewood Cliffs, NJ
- Jung S, Sen A, Marron JS (2012) Boundary Behavior in High Dimension, Low Sample Size asymptotics of Pca. *Journal of Multivariate Analysis* 109:190–203
- Lee EK, Cook D (2009) A Projection Pursuit Index for Large p Small n Data. *Statistics and Computing* p <http://www.springerlink.com/content/g47n0n342761838m/?p=d2ff5a7b69eb45ef8abf7ef3aba69557&pi=3>
- Majumder M, Hofmann H, Cook D (2012) Validation of Visual Statistical Inference, Applied to Linear Models. *Journal of American Statistical Association* Under revision
- Marron JS, Todd MJ, Ahn J (2007) Distance Weighted Discrimination. *Journal of American Statistical Association* 102:1267–1271
- R Development Core Team (2009) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, URL <http://www.R-project.org>, ISBN 3-900051-07-0
- Ripley BD (1996) *Pattern Recognition and Neural Networks*. Cambridge University Press
- Roy Chowdhury N, Cook D, Hofmann H, Majumder M (2012) Where's Waldo: Looking Closely at a Lineup. Tech. Rep. 2, Iowa State University, Department of Statistics
- Toth A, Varala K, Newman T, Miguez F, Hutchison S, Willoughby D, Simons J, Egholm M, Hunt J, Hudson M, Robinson G (2007) Wasp Gene Expression Supports an Evolutionary Link between Maternal Behavior and Eusociality. *Science* 318:441 – 444
- Toth A, Varala K, Henshaw M, Rodriguez-Zas S, Hudson M, Robinson G (2010) 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
- Wickham H, Cook D (2010) *tourr*: Implements Tour Methods in Pure R Code. <http://www.R-project.org>
- Witten D, Tibshirani R (2011) Penalized Classification using Fisher's Linear Discriminant. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)* 73(5):753 – 772
- Yata K, Aoshima M (2011) Effective PCA for High Dimension, Low Sample Size Data with Noise Reduction via Geometric Representations. *Journal of Multivariate Analysis* 105:193–215

Appendix

Solution

- The solution to the lineup at Figure 2 is Plot 16.
- The solution to the lineup at Figure 5 is Plot 17.
- The solution to the lineup at Figure 6 is Plot 20.
- The solution to the lineup at Figure 11 is Plot 8.

Choice of dimensions

The experiment is set up with the 3 factors separation, dimension and projection dimension. To decide on the levels of dimension to use, we considered the distribution of the absolute difference of the sample group means,

for data with two groups, no separation and projection dimension $d = 1$. The same levels are used for data with 3 groups, $d = 2$, and for data with separation.

Let \mathbf{X}_{ij} denote the j -th observation in the i -th group where $j = 1, \dots, n; i = 1, \dots, g$. The \mathbf{X}_{ij} 's are random noise, generated by drawing samples from a standard normal distribution. For this experiment, $g = 2$ and $n = 15$. The difference between the group means is given by $\bar{\mathbf{X}}_{1.} - \bar{\mathbf{X}}_{2.}$ and

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

Let $U = |\bar{\mathbf{X}}_{1.} - \bar{\mathbf{X}}_{2.}|$ where $U \sim \text{Half Normal}$ with scale parameter $\sigma = \sqrt{2/15}$. The expectation and the variance of U are $E(U) = \sigma\sqrt{2/\pi}$ and $\text{Var}(U) = \sigma^2(1 - 2/\pi)$, respectively.

For p dimensions, consider p independent samples from the same distribution, denoted as

$$\mathbf{U}_m = |\bar{\mathbf{X}}_{m1.} - \bar{\mathbf{X}}_{m2.}|, \quad m = 1, \dots, p$$

where \mathbf{X}_{mij} is the j -th observation in the i -th group for the m -th dimension. The difference between the two group means projected into one dimension, is the sum over p dimensions of the absolute difference between the means:

$$U = \sum_{m=1}^p \mathbf{U}_m = \sum_{m=1}^p |\bar{\mathbf{X}}_{m1.} - \bar{\mathbf{X}}_{m2.}|$$

and by independence it follows that

$$E(U) = p\sigma\sqrt{2/\pi}, \quad \text{Var}(U) = p\sigma^2(1 - 2/\pi)$$

Thus we expect to find this amount of separation between the projected sample means, for data sampled from populations with the same means.

Now consider data where there is some separation (equal to $2c$) between the population means:

$$\mathbf{Z}_{1j} \sim \text{Normal}(-c, 1)$$

$$\mathbf{Z}_{2j} \sim \text{Normal}(c, 1)$$

giving $\bar{\mathbf{Z}}_{1.} - \bar{\mathbf{Z}}_{2.} \sim \text{Normal}(2c, 2/15)$. Then define $Z = |\bar{\mathbf{Z}}_{1.} - \bar{\mathbf{Z}}_{2.}|$ where $Z \sim \text{Folded Normal Distribution}$ with scale parameter $\sigma = \sqrt{2/15}$. The expectation and the variance of Z can be calculated to be:

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

$$\text{Var}(Z) = 4c^2 + \sigma^2 - (E(Z))^2$$

Suppose that only one of the p dimensions is simulated from this distribution, and all of the rest are simulated from populations having identical means. Define V as the sum of the absolute differences of the mean with one dimension of real separation as

$$V = \sum_{m=1}^{p-1} \mathbf{U}_m + \mathbf{Z}$$

Then, by independence, it follows that:

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

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

In this experiment, $c = 3$ and $\sigma^2 = 2/15$. Therefore,

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

Hence,

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

$$\text{Var}(V) = (p - 1)\sigma^2(1 - 2/\pi) + \sigma^2$$

As dimension p increases for a fixed n , the spread of both U and V increases by a factor of p . The means of U and V also increase with a factor of p but the expected value of the difference between U and V stays constant and is independent of dimension (p).

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

Figure 3 illustrates this phenomenon. The distributions of data with separation (V) and without separation (U) are shown in red and blue respectively. As dimension (p) increases, the spread of both the distributions increases while the means are equally apart. As a result, the common region between the distributions increases, indicating the chance of obtaining a random separation for purely noise data increases with dimension(p). A 5% error is allowed and the area of the distribution of U greater than the 5th percentile of V is considered and we call this δ . In Figure 3, δ is indicated by the dark blue region. Mathematically,

$$P[U > V_\alpha] = \delta$$

where V_α is the 100α -th percentile of V , where $\alpha = 0.05$. Dimension p is calculated for fixed values of δ . The various values of δ are chosen such that the distributions has no separation ($\delta \approx 0$) or has 1%, 5%, 10% and 20% common region. For each value of δ , the procedure is repeated 100 times and Table 5 shows the summaries of the dimension (p) for each value of δ .

Table 5 Numerical summaries of dimension p for each value of δ . As the common region δ increases, the median dimension required to obtain the region increases.

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