Using Visual Statistical Inference to Better Understand Random Class Separations in High Dimension, Low Sample Size Data

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

Abstract Statistical graphics play an important role in exploratory data analysis, model checking and diagnosis. With high dimensional data, this often means plotting low-dimensional projections, for example, in classification tasks projection pursuit is used to find low-dimensional projections that reveal differences between labelled groups. In many contemporary data sets the number of observations is relatively small compared to the number of variables, which is known as a high dimension low sample size (HDLSS) problem. This paper explores the use of visual inference on understanding low-dimensional pictures of HDLSS data. Visual inference helps to quantify the significance of findings made from graphics. This approach 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 in microarray data, and with a simulation study conducted using Amazon's Mechanical Turk.

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 (see for e.g. Hall et al (2005)). HDLSS data occur in many application areas like face recognition, spectroscopy and gene expression analysis. Classical statistical methods often fail in this context, because of insufficient data to support for parameter estimation.

Reducing the dimension, using principal component analysis (PCA), would be a classical first step in the analysis of HDLSS data. 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

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

Mahbubul Majumder
Department of Mathematics, University of Nebraska, Omaha, NE, USA
E-mail: mmajumder@unomaha.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

a Sisyphean task. Just imagine, estimating a line on the foundation of a single point - there are infinitely many possibilities for lines. For classification tasks, finding a low-dimensional space where the classes are separated is a common first step. Linear discriminant analysis (LDA) is the classical approach. LDA solves an eigenvalue decomposition problem comparing distances between group means with variance around each mean. Estimating the variance-covariance is problematic when there are few points. In addition, when there are few sample points in high dimensions, differences between groups can be found in many different low-dimensional spaces, simply because of the sparseness of space.

Marron et al (2007) describes the estimation issues associated with HDLSS. 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 for problems where only a small fraction of the variables are useful and unknown. Penalization is another common approach to handle HDLSS, and has been applied to classification problems (e.g. Witten and Tibshirani, 2011; Lee and Cook, 2010). Estimates of the variance-covariance are obtained by an interpolation with the identity matrix, effectively reducing the importance of some variables.

So, although substantial research has produced many new approaches to creating better models and estimation for HDLSS data, the major issues are still not clear to many data analysts. For example, Toth et al (2010) make a common mistake of seeing structure where none exists. Figure 1 reproduces the result in this paper. They use LDA to examine gene expression data of wasps containing 447 variables and 50 cases. 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 2D projection ($d = 2$) 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 method that enables us to conclude whether this apparently clear separation is statistically significant or not. For prediction, typically data is broken into training and test sets, or cross-validation is conducted to assess the significance of difference, using test set error. This approach does not work well for visualization.
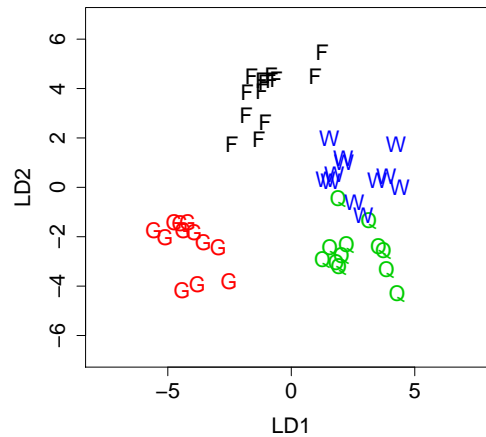

Fig. 1 LD1 versus LD2 from an LDA on a randomly selected subset of 40 significantly different oligos : F, Foundress; G, gyne; Q, queen and W, worker. It can be noticed that the groups F and G are separated. This plot is generated to match Figure 2 in Toth et al (2010).

We propose 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), formalized and validated

by Majumder et al (2013). Using visual inference, it can be shown that there is no real difference between the wasp groups - what you see is a mirage.

Visual inference may also be useful in related applications, such as checking algorithms that produce visual results. We used the approach described in this paper to check the optimization algorithm of projection pursuit in the `tourr` package (Wickham et al, 2011) in `R` (R Core Team, 2013). The optimization procedure was new, and we suspected that it was being sensitive to the order of data values, and returning projections of purely noise data that we thought were surprisingly distinct from noise. Visual inference was able to temper our concerns.

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 high 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 describes Amazon's Mechanical Turk (Amazon, 2010) which was used to conduct the experiment. The application of visual inference methods on the wasp data (Toth et al (2010)) is described in Section 5. Section 6 discusses the experiment designed to examine people's perception of separation in the presence of real separation and "purely noise" for simulated HDLSS data. Section 7 discusses the collected data and results.

2 Visual inference methods

Buja et al (2009) proposed two protocols, the Rorschach and the lineup. While the Rorschach protocol helps to understand the extent of randomness, the lineup protocol is used for testing significance of findings. These methods together are called visual statistical inference. Majumder et al (2013) made a head-to-head comparison between visual statistical inference tests and classical tests which showed that the lineup protocol performs similarly to the classical tests. 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 false. The lineup protocol of size m embeds the observed data plot amongst (m - 1) null plots. Null plots are created by a mechanism consistent with the null hypothesis. Human subjects are asked to identify the plot in the lineup with the most distinct feature(s). 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. If the subjects choose the plot of the observed data, this is evidence against the null hypothesis and with enough support, the null hypothesis is rejected.
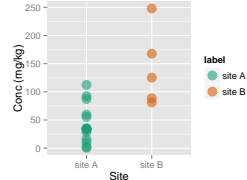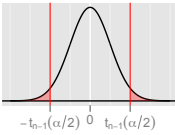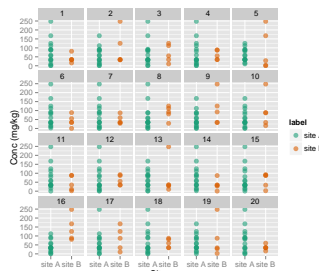
An illustration of the lineup protocol in contrast to the conventional test is shown in Table 1. Both start with the same hypothesis but the test statistic in a conventional setting is the parameter estimate divided by its standard error. In visual inference the test statistic is a plot of the observed data. In this case, a dot plot is used since the variable of interest is continuous with two groups. In a conventional test the value of the test statistic is compared with all possible values of the sampling distribution, the distribution of the test statistic if the null hypothesis is true. If it is extreme, then the null hypothesis is rejected. In visual inference, the plot of the data is compared with a finite number of samples drawn from the sampling distribution. If the actual data plot is selected as the most different, then the null hypothesis is rejected.

For example, suppose we have two sample data on the concentration of a metal in mg/kg for sites A and B of sizes n_1 and n_2 respectively. We want to test whether there exists a difference between the concentration levels in the two sites A and B. To test for statistically significant difference between the two populations from which the data was sampled, 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

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

The conventional test would be a two-sample t-test with test statistic $T(y)$ described in Table 1. One way to plot this data is a side-by-side dotplot. Let this be the visual test statistic $V(y)$. Null plots are generated assuming that H_o is true. Here, this is achieved by randomly permuting the site label. The observed data plot is placed randomly among the null plots to obtain a lineup. In this example, $m = 20$ is the size of the lineup - there are 19 null

Table 1 Comparison of visual inference with traditional hypothesis testing. Starting with the same hypothesis, the test statistic in a conventional setting is a real number while in visual inference it is a plot of the observed data. In conventional testing the value of the test statistic is compared with all possible values of the sampling distribution. H_o is rejected if it is extreme. In visual inference, the plot of the data is compared with a finite number of samples drawn from the null distribution. If the actual data plot is identifiable, then the null hypothesis is rejected.

	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	$T(y) = \dfrac{\overline{y}_1 - \overline{y}_2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$	$V(y) =$
Sampling Distribution	$f_{T(y)}(t);$	$f_{V(y)}(t);$
Reject H_o if	observed T is extreme	observed data plot is identifiable

plots. If H_o is not true, the dots of one group should be vertically shifted relative to the other group. If the human 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. Under the null hypothesis, each observer has a $1/m$ chance of picking the observed plot from a lineup of size m. Hence $1/m$ is the minimal value at which we can set the Type I error, α, consistent with $\alpha = 0.05$ if $m = 20$. Majumder et al (2013) provides more detailed discussion about this. For this problem, visual inference enables the handling of the small sample size and non-normality of the population. However, in general the setting for visual inference would be problems where no conventional test exists.

Majumder et al (2013) describes the methods of obtaining the power of the visual test, by combining results from multiple users. For their simulation experiments human observers were recruited through Amazon's Mechanical Turk (Amazon, 2010). Power of the visual test used in their simulation was also calculated theoretically. Their results suggest 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 can also be estimated from the multiple responses data from each human observer, which might help quantify individual visual skills.

3 Dimension reduction

Projection pursuit (e.g. Friedman and Tukey, 1974) is used for dimension reduction in our studies. Projection pursuit (PP) finds the most interesting low dimensional projection of high dimensional data by maximizing some
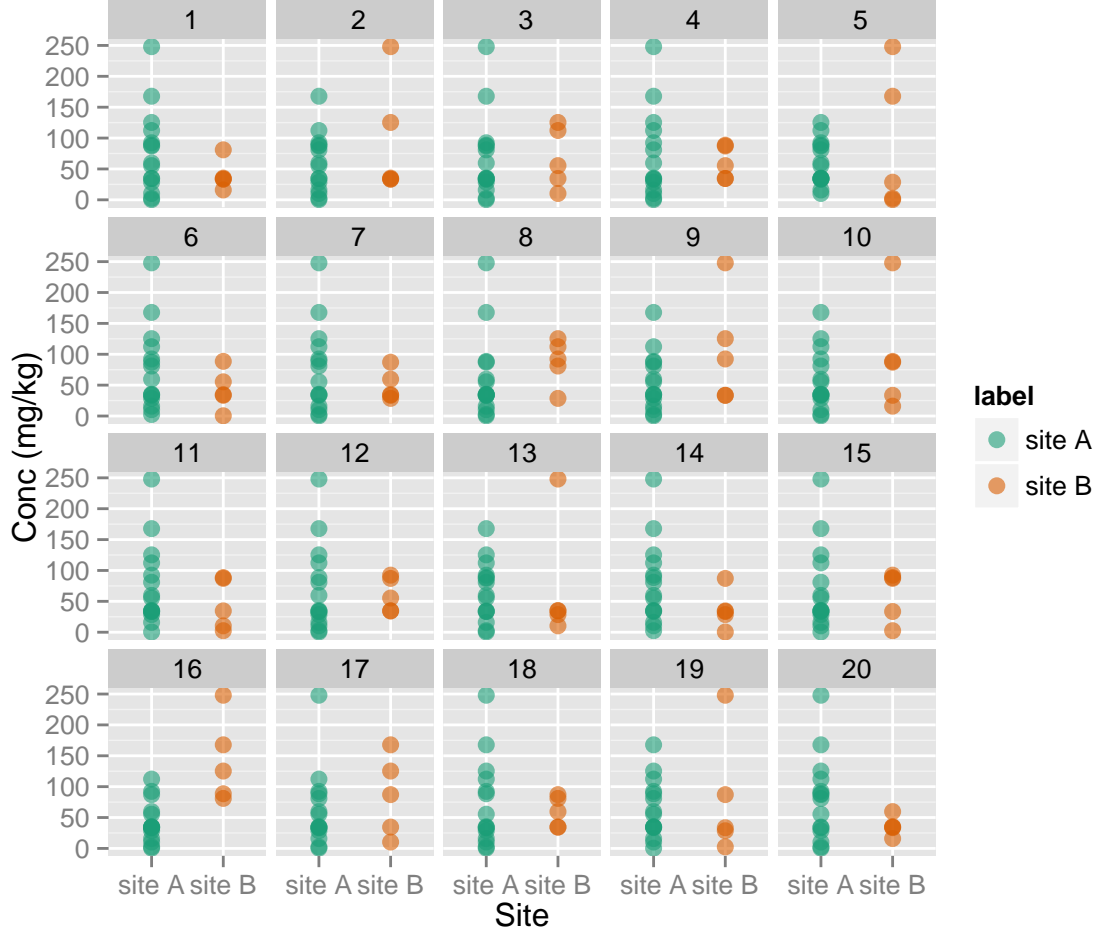
Fig. 2 A typical lineup ($m = 20$) for testing $H_o : \mu_1 = \mu_2$. When the alternative hypothesis is true, the observed data plot should have the largest vertical difference between the centers. Can you identify the observed data plot? The solution to the lineup is provided in the Appendix.

criterion 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, 2010) in projection pursuit. Let $\mathbf{X_{ij}}$ be the p-dimensional vector of the jth observation in the ith class, $i = 1, \ldots, g, j = 1, \ldots, n_i, g$ is the number of classes, n_i is the number of observations in class i, and $n = \sum_{i=1}^{g} n_i$. Let $\overline{\mathbf{X}}_{i.} = \frac{1}{n_i} \sum_{j=1}^{n_i} \mathbf{X}_{ij}$ be the ith class mean and $\overline{\overline{\mathbf{X}}}_{..} = \frac{1}{n} \sum_{i=1}^{g} \sum_{j=1}^{n_i} \mathbf{X}_{ij}$ be the total mean. The LDA PP index is

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

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

5

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

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

For HDLSS data, the penalized discriminant analysis (PDA) index (Lee and Cook, 2010) is more robust. Let $\mathbf{X}_{ij}^*$ be the standardized vector of $\mathbf{X}_{ij}$. Then

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

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

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

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

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 (Wickham et al, 2011) in R (R Core Team, 2013). 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 a projection where groups show the biggest separation. For this paper we used $d = 1$ or 2.

4 Amazon Turk experiments

Amazon's Mechanical Turk (Amazon, 2010) is a service that enables researchers to employ people to do tasks which computers perform poorly. In exchange for their efforts, the subjects are paid, not substantially, but on the scale of the minimum wage of the USA. For visual inference studies, subjects are typically given a block of ten lineups to evaluate during a job. From this block, one lineup is typically used as a filter, and the remaining lineups produce data for the studies. Because a subject evaluates more than one lineup, and a lineup is evaluated by more than one subject, we obtain some replication in the results upon which to estimate variation. The one filter lineup, in which the observed data plot is markedly different from the nulls, is necessary because Turkers are not manually monitored, and a few attempt to maximize financial gain without taking the exercise seriously.

For this paper, two Turk studies were run, one for the wasps data, and the other for the simulation study. Turkers are redirected from Amazon to a website which describes the study in detail, provides some practice trials, collects demographic details, and responses. The website of the simulation study is `http:/mahbub.stat.iastate.edu/feedback_turk7/homepage.html`.

5 Wasps application

We return 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:

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 a real difference between the expression levels for the types of wasps then the observed data plot should be detectable in the lineup. Figure 3 shows a lineup. Three different lineups were created using this procedure.
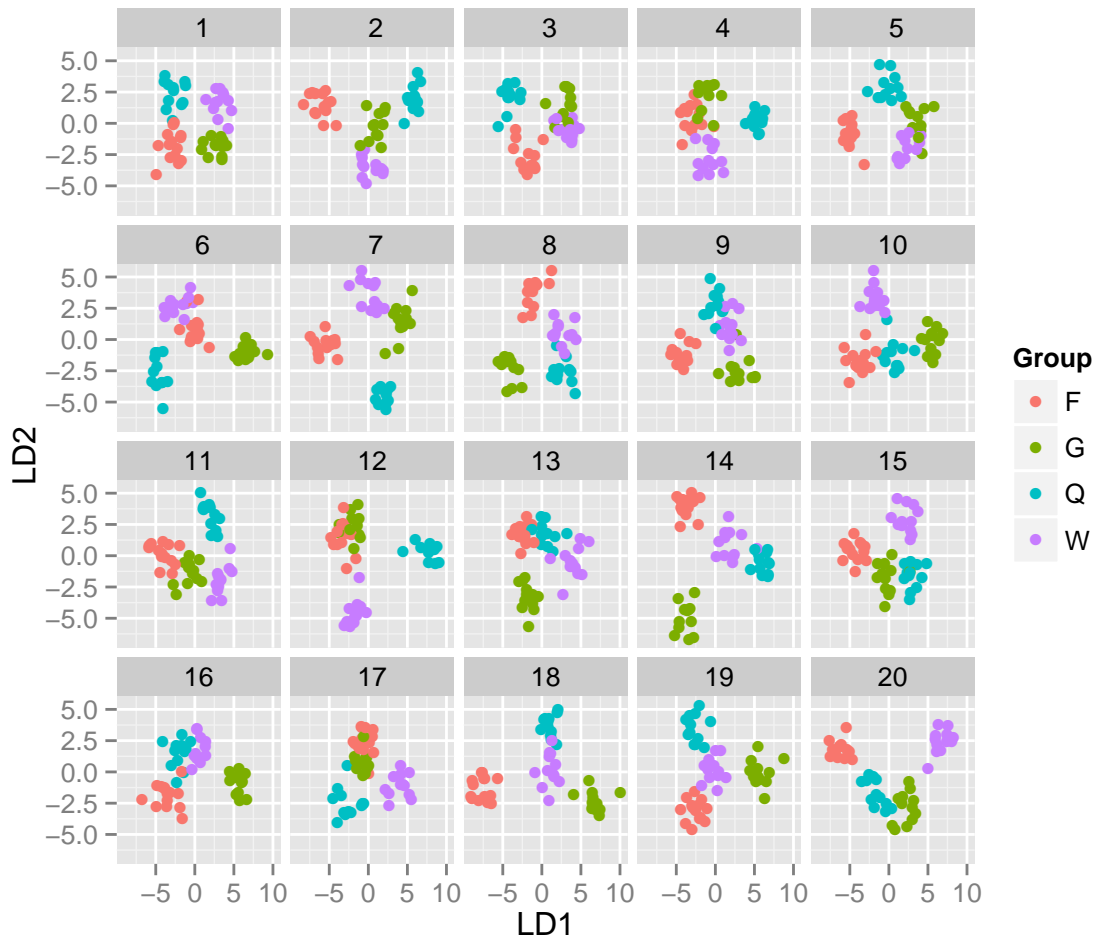

Fig. 3 Lineup of the wasps data. One plot shows the observed data and the remaining 19 show null data where the wasp type labels were randomly assigned. Each plot was produced by conducting LDA on the 40D data to produce the 2D projection with best separation. Which plot shows the most separation between the 4 groups? The solution is provided in the Appendix.

In addition, three more lineups are made containing only null plots, with one plot randomly chosen to act as an observed data plot. These lineups were shown to the subjects recruited from Amazon Turk. A total of 116 subjects evaluated the 6 lineups. Table 2 shows the results. The detection rate for 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 very often detected 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 (2013). 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 (Toth et al (2010)) is not real. It is purely the effect of high dimensionality.

Table 2 Results of the Turk study on the wasps data. Detection rate for each lineup is shown, with the number of subjects, and p-value associated. The detection 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 (Toth et al (2010)) is consistent with pure noise induced by the high dimensions.

Data	Replicate	Num Subjects	Detection rate	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

The probability of separation by chance between two groups in purely noise data, given a fixed sample size and dimension, was quantified by Ripley (1996) (Proposition 3.1). Figure 4 illustrates this result for 1D projections, for different p, where sample size is fixed at $n = 30$. When $p = 2$, P(separation$|n = 30, p = 2$) = 0 and it reaches 1 when $p = 25$. For data of the size of the wasps, $n = 50$ and $p = 38$, the probability of obtaining separation with only two groups is 1, so we would expect that there would certainly be separation between four groups in 2D.
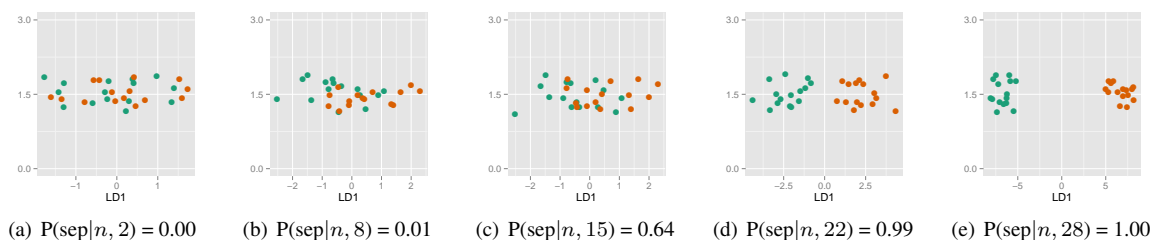

(a) P(sep$|n, 2$) = 0.00   (b) P(sep$|n, 8$) = 0.01   (c) P(sep$|n, 15$) = 0.64   (d) P(sep$|n, 22$) = 0.99   (e) P(sep$|n, 28$) = 1.00

Fig. 4 Plots showing example 1D projections for $p = 2, 8, 15, 22, 28$ and $n = 30$ for purely noise data. The probability of obtaining a projection where groups are separated is calculated and displayed below the plots. Of course, once a projection is computed the groups are either separated or not - the event occurred or didn't - and we can see that the last two plots display separated groups. The difference between the groups increases as p increases, and the likelihood of obtaining a projection with separation increases.

In the original paper (Toth et al, 2010), the dimensionality was reduced from 389 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 the penalized discriminant analysis (PDA) index (Lee and Cook, 2010), 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).

6 Follow-up simulation experiment

6.1 Experimental design

The goal is to determine how well people can detect the presence of real separation as distinguishable from random noise. To achieve this, the experiment is set up with several factors: real separation or pure noise, data dimension and projection dimension. Real separation is achieved by setting 1 or 2 variables with real separation among a number of noise variables. Sample size is fixed to keep the experiment manageable. Also mean difference is kept fixed. The levels of the factors used in the experiment are given in Table 3. Three replicates at each level are generated. These produced 60 different "observed data sets", and thus, 60 different lineups.

Table 3 Levels of the factors used for the simulation 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

6.2 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, yielding a $30 \times (p+1)$ matrix, $\mathbf{X}$, 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

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

where each $\mathbf{X}_i$ is a vector of dimension 30 for $i = 1, \ldots, p$. 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:

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

and then standardized to have unit variance again. On each dataset of p dimension, a projection pursuit optimization with the PDA index is performed to obtain the 1D projection of best separation, yielding $\mathbf{Y} = \mathbf{XA}$.

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

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

where each $\mathbf{X}_i$ is a vector of dimension 30 for $i = 1, \ldots, p$. 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:

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

and standardized. 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 2D projection of best separation, yielding $\mathbf{Y} = \mathbf{XA}$.

6.3 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 5 shows the two different visual test statistics.
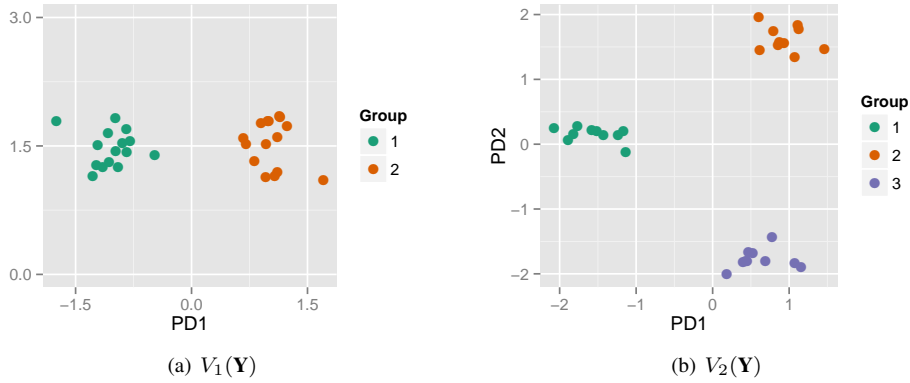

(a) $V_1(\mathbf{Y})$ (b) $V_2(\mathbf{Y})$

Fig. 5 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 6 shows an example lineup having treatment levels $p = 20$, separation = Yes and $d = 1$. Similarly, Figure 7 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, 2014), 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.
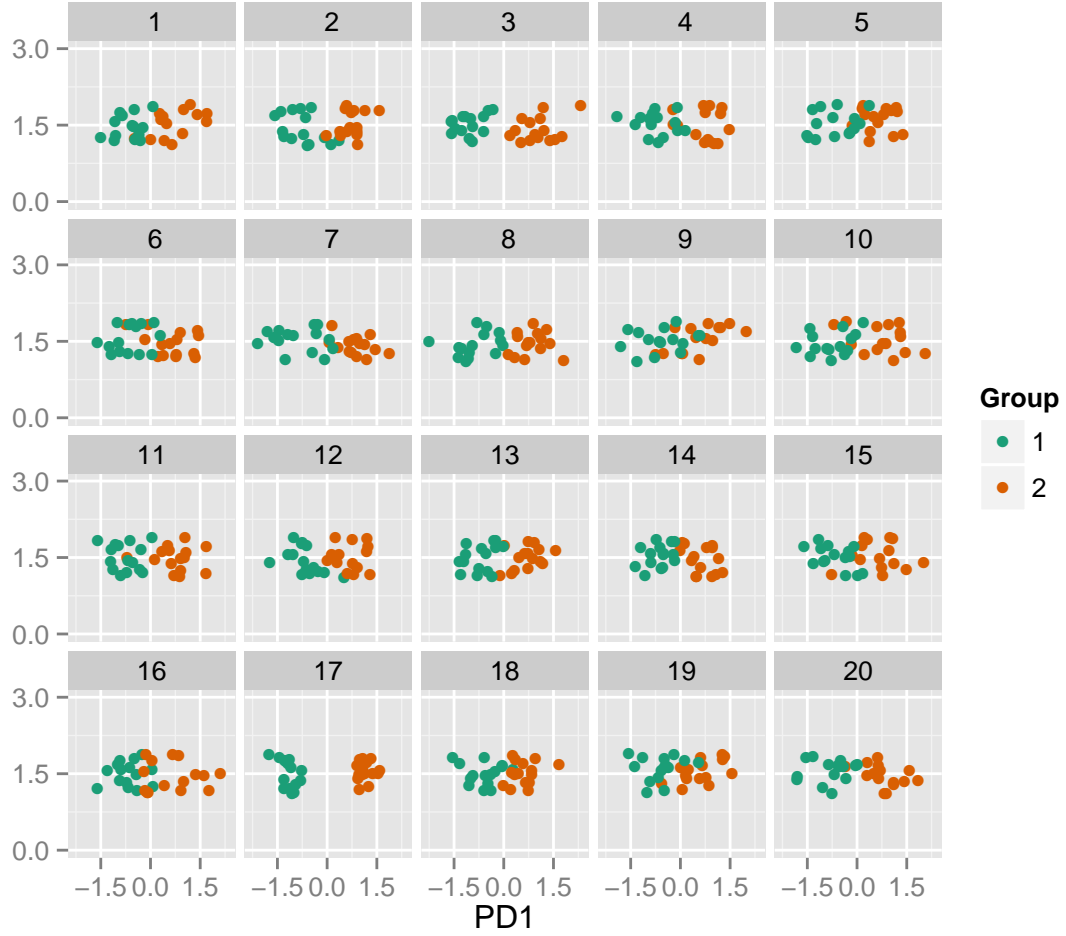
Fig. 6 Lineup ($m = 20$) from treatment with $p = 20$, separation = Yes and $d = 1$. The subjects were asked to identify the plot with the most separated colors. Can you identify the observed data plot? The solution to the lineup is provided in the Appendix.

6.4 Data collection

Subjects for the experiment were recruited through Amazon's Mechanical Turk (Amazon, 2010). Each subject was shown a block of ten 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 level and the geographic location of each subject were also noted. In total, 1137 lineups were evaluated by 103 subjects, from different locations across the globe.

Each subject was given a very easy lineup (a lineup with $p = 10$ dimensions with some real separation) in the block of ten. Data from subjects who failed to give a correct response to this lineup are removed from the study. If their response to this lineup was correct, data for this lineup was removed but responses for the remaining nine were kept for analysis. This produced 66 lineups evaluated by 101 subjects for analysis.
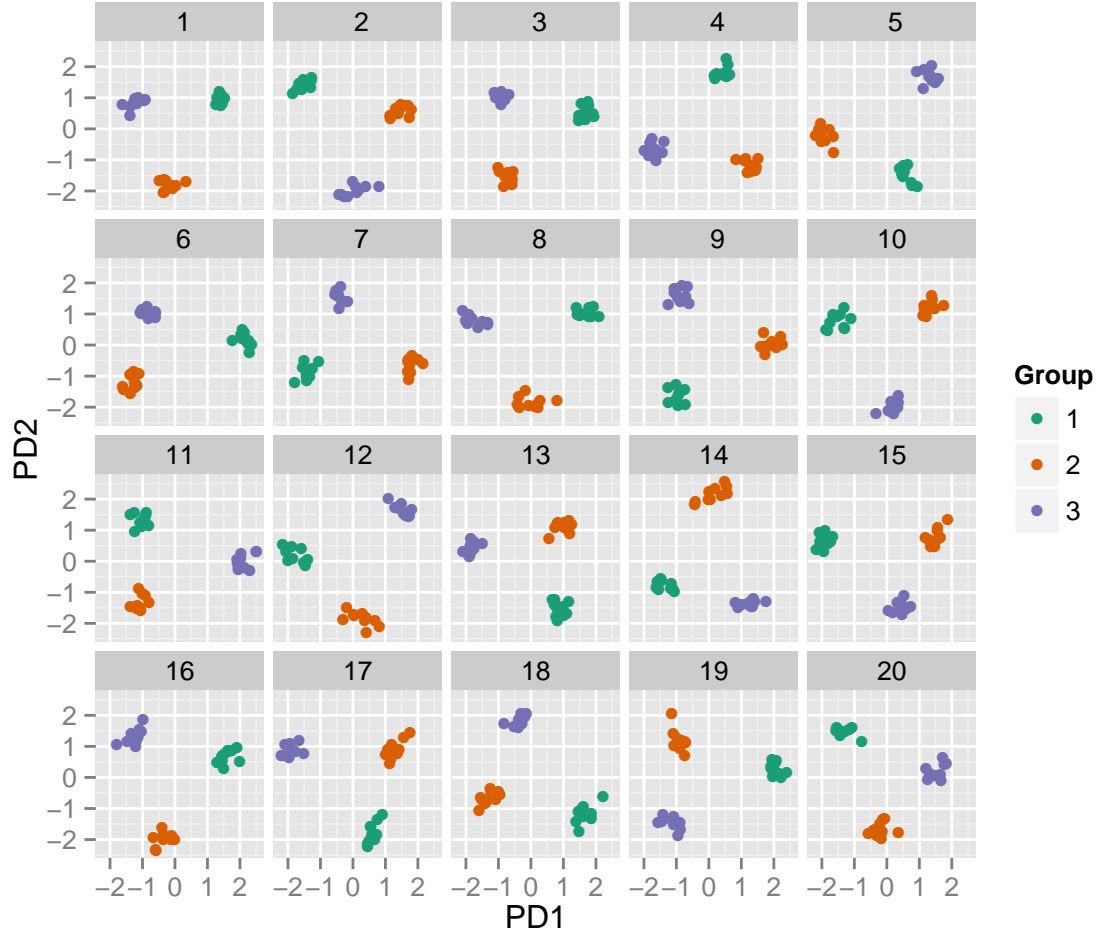
Fig. 7 Lineup ($m = 20$) from treatment with $p = 100$, separation = No and $d = 2$. The subjects were asked to identify the plot with the most separation between the colored groups. Can you identify the observed data plot? The solution is provided in the Appendix.

7 Results

7.1 Effect of experimental factors on detection rate

We would expect that subjects detect the observed data plot more often when there is real separation but that the detection rate diminishes as dimension increases. This is indeed the case, as illustrated by Figure 8. The detection rate is plotted against data dimension (p), faceted by the levels of separation and projection. The three dots for each p represent the three replicates for each treatment. In a few cases, the dots overlap as the detection rate is same for the replicates. For example, when projection = 1D, dimension = 100 and separation = No, the detection rate is 0 for all the replicates and hence a single dot is shown. Alpha-blending is used on the plots so these ties are darker dots. The line shows the fixed effects from a logistic regression model fitted to the data. The detection rate is higher for small p, and on average decreases as p increases, with both 1D and 2D projections for real separation. For purely noise data, the detection rate is effectively flat across p, always less than 0.1 on average. Interestingly, the detection rate is higher for data where separation exists than for pure noise, even at $p = 100$, at 0.25 vs 0.05. For real separation, there also appears to be increasing variance as p increases, which is intriguing, too: it might

12

be indicative of the increase in unexplained variance associated with the reduction to low dimensions from the increasingly higher p, and the sparsity of space.
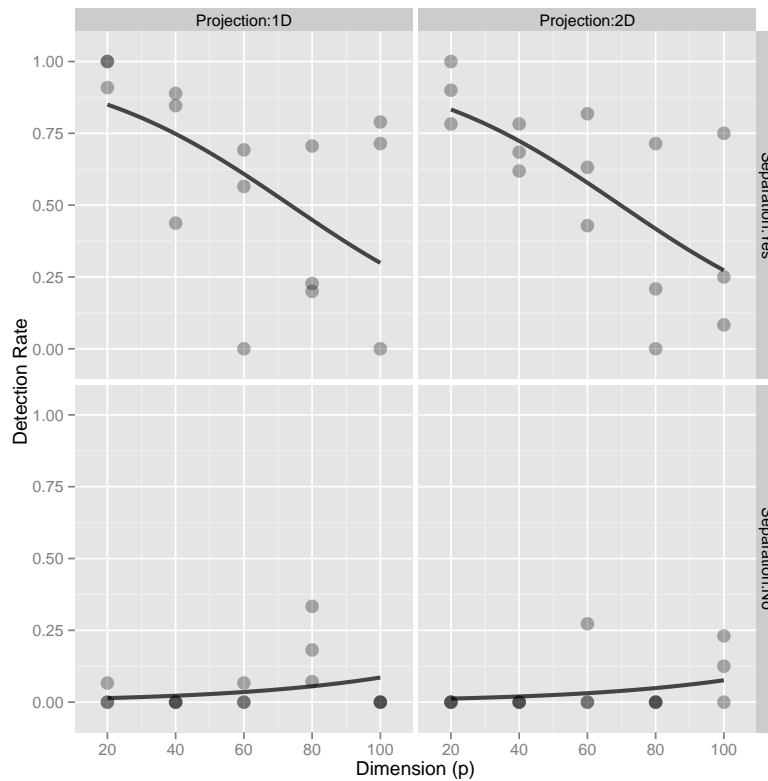

Fig. 8 Detection rate by dimension, faceted by projection and separation. The three points represents the three replicates for each treatment level. A fixed effects logistic regression model is overlaid on the points. It can be seen that the detection rate decreases as p increases for data with real separation. When the data is purely noise data, the detection rate is flat across dimensions. Detection rate does not change with projection. Even with $p = 100$ subjects more often detected separation than would be expected by chance.

Table 4 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. However, the p-value corresponding to the projection is large, which suggests that the difference between 1D and 2D projections is not significant. 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 rotation was not important. The interaction term is also significant, which says that the detection rate changes in the presence or absence of separation – detectability of the observed data plot decreases with dimension when there really is separation.

The effect of demographic variables (age, gender and education) on the detection rate was also studied, and found to be insignificant.

7.2 Time taken to respond under different treatments

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 9 shows the time taken in seconds to respond (on a log scale) by p, facetted by

Table 4 Table summarizing results of experiment. Columns correspond to the estimate, the standard error and the p-value of the parameters used in logistic regression model. As dimension (p) increases, detection of separation decreases. Subjects can detect the separation if it exists even when $p = 100$. Subjects were equally good in 1D or 2D projections.

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

projection. Color indicates presence or absence of separation. The line shows the trend over dimension. Bootstrap resampling bands (Buja and Wolgang (2005)) are overlaid, which effectively provide confidence bands for the curves. Notice that, when the data has some real separation (green), as the dimension increases on average, subjects take more time to respond to the lineups. But when the data is purely noise (brown), the increase of dimension does not have any effect on the time taken. 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. The bootstrap resampling bands suggest that there is only significantly reduced time taken when $p = 20$ or 40.
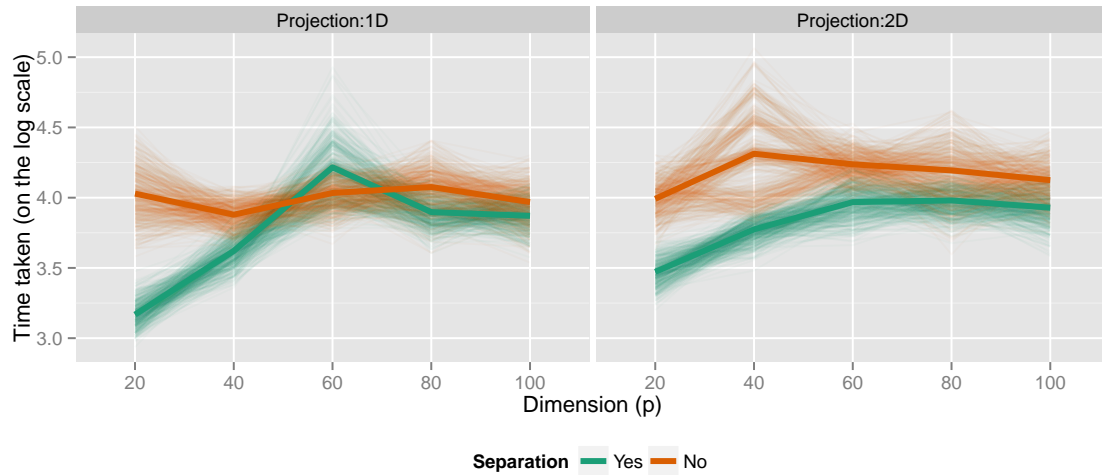

Fig. 9 Time taken in seconds to respond on log scale against dimension colored by separation and faceted by projection. A line shows the trend over dimension for each separation within projection. Bootstrap resampling bands are drawn for each colored lines. Time taken to respond is higher when the data has no separation. Also as dimension increases, the time to answer when there is separation is equal to the time taken when there is no separation.

7.3 What affects decisions?

Figure 10 examines the subjects choices in detail. The relative frequency of picks of each plot in the lineup is plotted against a measure of average separation between groups. 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 1D and 2D 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 largest average separation tends 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.
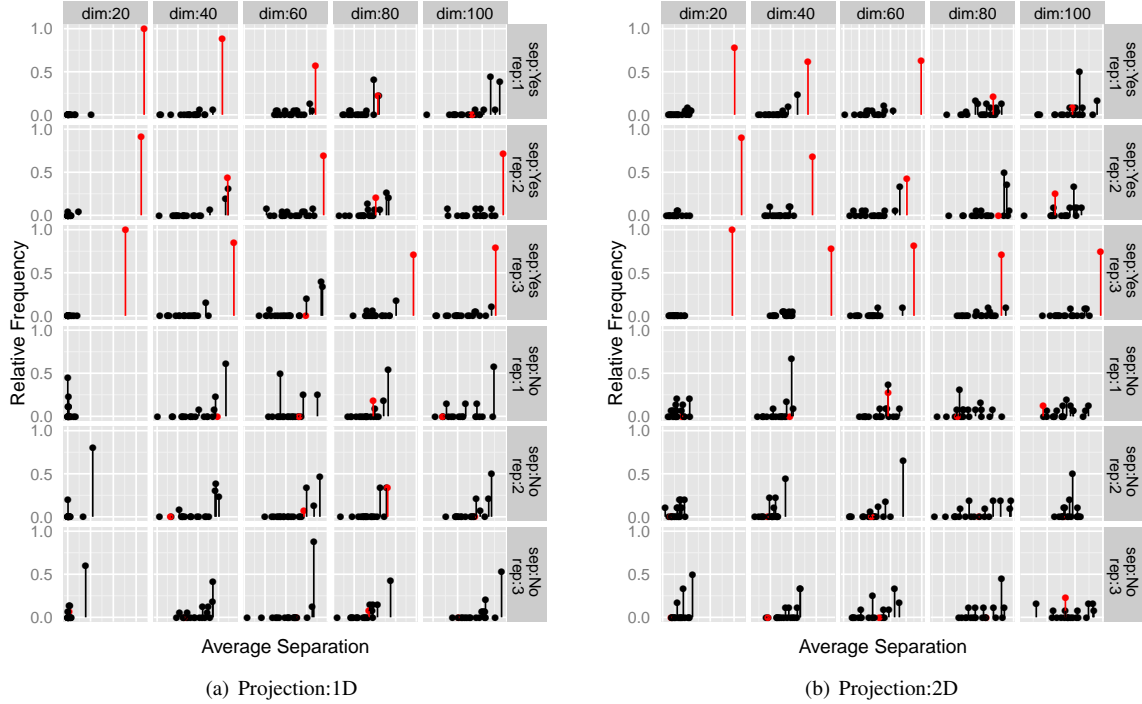

| (a) Projection:1D | (b) Projection:2D |

Fig. 10 Comparing the choices that subjects make for each lineup. Relative frequency of plots chosen against a measure of the average separation between groups, the larger 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 largest average separation. In general, this happens, the tallest pins are in the right 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.

7.4 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 first calculate the average separation between the clusters in each plot of the lineup. For each lineup we then calculate the difference between the maximum average separation of the null plots and the average separation for the observed data plot for each lineup. Figure 11 examines the influence of the null plots on this pick. The detection rate and mean time

15

taken in seconds are plotted against difference. The vertical line is a reference line where the difference is 0 – the value at which 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 difference increases, the detection rate increases and also time taken to choose decreases, suggesting easier lineups. More details on the measurement of the influence of the null plots are available in Roy Chowdhury et al (2012).
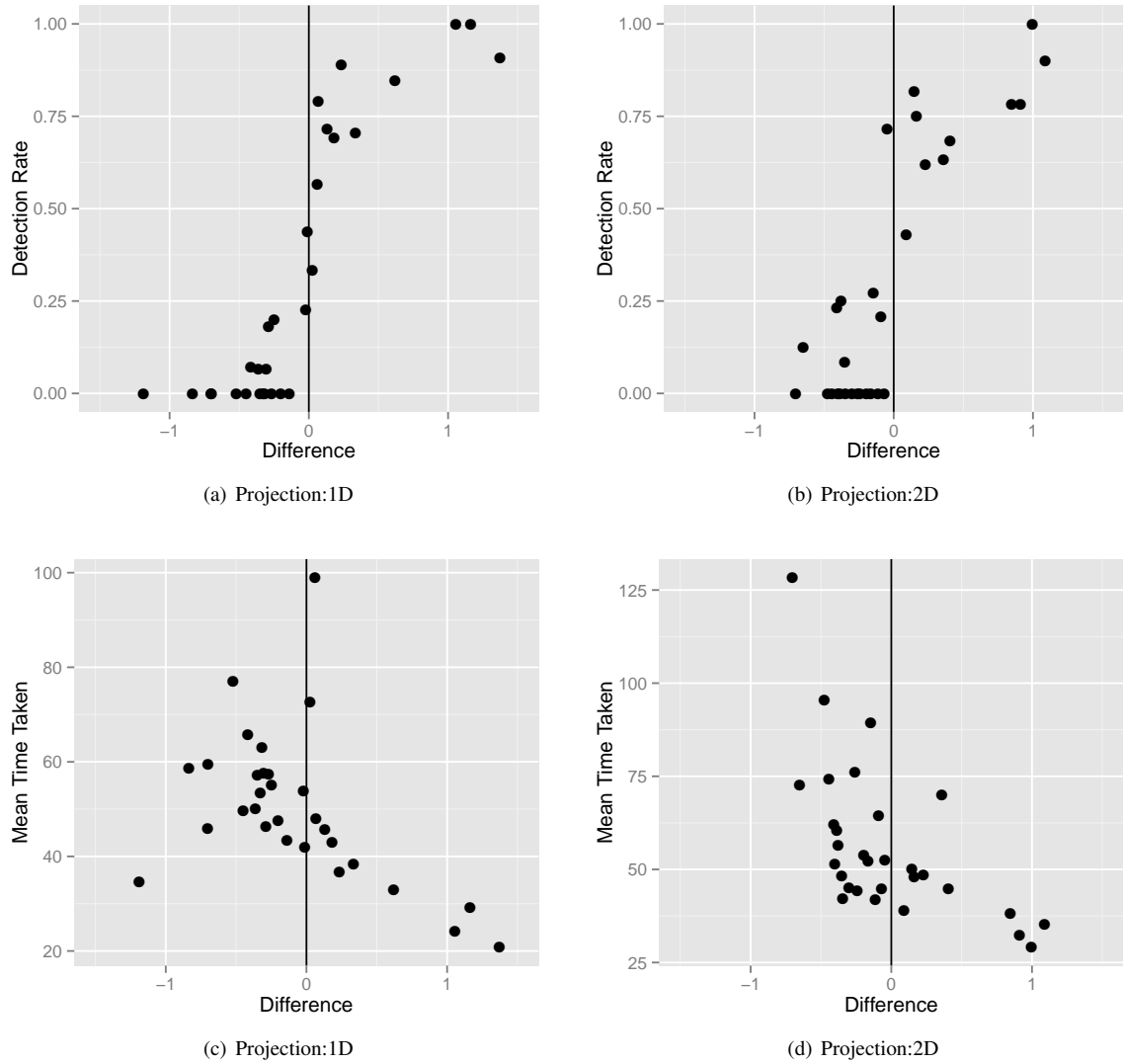

(a) Projection:1D

(b) Projection:2D

(c) Projection:1D

(d) Projection:2D

Fig. 11 Detection rate and mean time taken to respond in seconds are plotted against the difference for 1D and 2D projections separately. The difference is between maximum separation of all the null plots and separation of the observed data plot for each lineup for 1D projections but for 2D projections the difference is based on the average separation between the groups. The vertical line represents difference equal to 1 when the average separation of the observed data plot is equal to the maximum average separation of the null plots for 2D projection. The points left to the line indicates a difficult lineup in the sense that at least one of the null plots had a lower average separation value than the observed data plot. (a) and (b) As difference increases, detection rate increases. (c) and (d) As difference increases, mean time taken decreases indicating that the subjects have an easier time in identifying the observed data plot.

8 Conclusions

The results of applying visual inference procedures to classification problems on HDLSS data suggest that visual inference may be effective for improving the understanding of the emptiness of space in this type of data. With visual inference we saw that people can visually detect real separation as different from noise up to a reasonably high dimension, for 1D and 2D projections. Visual inference provides a calibration for reading the separation.

We also learned from visual inference, although didn't discuss this in the paper, that the projection pursuit optimization procedure in `tourr` package is performing correctly. It is possible that visual inference might be used to calibrate results of similar algorithms, where the optimization is used to yield visual products, like multi-dimensional scaling, PCA, independent component analysis (ICA) (Comon (1994)) and local linear embeddings (Roweis and Saul (2000)).

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. It may be useful to incorporate the approaches into the statistics curriculum, particularly elementary statistics and applications areas such as gene expression analysis, to improve understanding of randomness.

Acknowledgement

References

Amazon (2010) Mechanical Turk. URL http://aws.amazon.com/mturk/

Buja A, Wolgang R (2005) Calibration for Simultaneity: (Re)Sampling Methods for Simultaneous Inference with Applications to Functional Estimation and Functional Data. Tech. rep., URL http://stat.wharton.upenn.edu/~buja/PAPERS/paper-sim.pdf

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

Comon P (1994) Independent Component Analysis: a new concept? Signal Processing 36(3):287 – 314

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 (2014) fpc : Flexible Procedures for Clustering URL http://CRAN.R-project.org/package=fpc, r package version 2.1-7

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 (2010) A Projection Pursuit Index for Large p Small n Data. Statistics and Computing 20:3:381–392

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

Marron JS, Todd MJ, Ahn J (2007) Distance Weighted Discrimination. Journal of American Statistical Association 480:1267–1271

R Core Team (2013) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, URL http://www.R-project.org/

Ripley BD (1996) Pattern Recognition and Neural Networks. Cambridge University Press, New York, NY, USA

Roweis S, Saul L (2000) Nonlinear Dimensionality Reduction by Locally Linear Embedding. Science 290:2323 – 2326

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, URL http://www.stat.iastate.edu/preprint/articles/2012-02.pdf

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 (2009) ggplot2: Elegant Graphics for Data Analysis. Springer New York, URL http://had.co.nz/ggplot2/book

Wickham H, Cook D, Hofmann H, Buja A (2011) tourr: An R Package for Exploring Multivariate Data with Projections. Journal of Statistical Software 40(2):1–18, URL http://www.jstatsoft.org/v40/i02/

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 3 is Plot 8.
- The solution to the lineup at Figure 6 is Plot 17.
- The solution to the lineup at Figure 7 is Plot 20.

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 $j = 1, \ldots, n; i = 1, ..., g$. 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 $\overline{X}_{1.} - \overline{X}_{2.}$ and

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

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

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

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

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:

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

and by independence it follows that

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

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

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

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

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:

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

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

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,

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

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

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).

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

Two p-dimensional datasets are generated with 30 observations in each dimension. The datasets are then divided into two groups with 15 observations in each group. For one set, data is obtained from random noise and hence there is no real separation between the two groups. But for the other set, one dimension among these p is adjusted so that the data have some real separation between the groups in that dimension. The absolute difference of the means for each group in each of these p dimensions is considered for both datasets. The absolute difference is considered as we are concerned with projections. These absolute differences between the groups are then summed over all the dimensions to obtain the absolute difference of means for the data. This process is repeated 1000 times. These 1000 sum of absolute differences are then plotted for the different values of p.
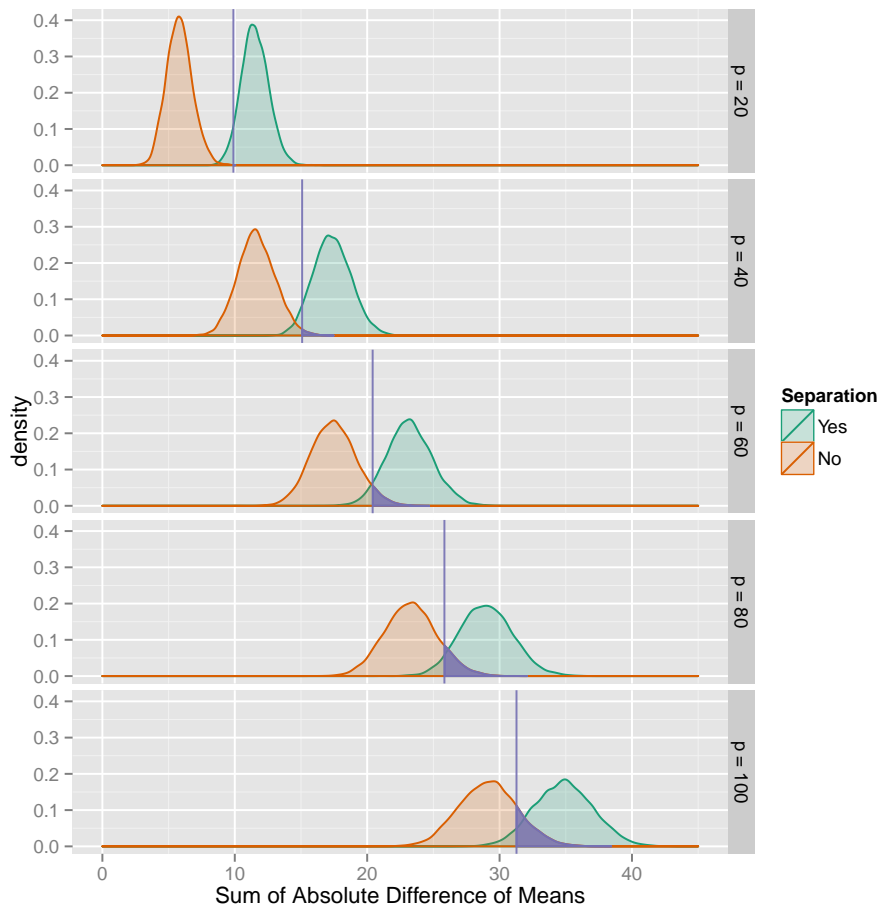

Fig. 12 Plot showing the distribution of the sum of absolute difference of means for data with and without separation for different dimensions. The distributions of data with real separation (V) and purely noise data (U) are shown in brown and green respectively with the dark purple line showing the 5th percentile of V. The dark purple area shows the area of U which is greater than the 5th percentile of V. The dark purple region (δ) increases as dimension (p) increases.

Figure 12 shows the distribution of sum 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 brown and green

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 purple. Hence a 5% error is allowed and let the area of the distribution of U greater than the 5th percentile of V be denoted by δ. Mathematically,

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

where V_α is the 100α-th percentile of V, where $\alpha = 0.05$. It can be seen that the dark purple region increases with dimension (p). This indicates that as dimension increases, the distributions of data with or without separation gets closer. Hence it gets harder to detect real differences with higher dimensions. Fixing the area of the dark purple region (δ) and calculating the dimensions to obtain the required region provides the choice of levels of dimension used in the experiment.

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