MACHINE LEARNING METHODS IN HDLSS SETTINGS

Xi Yang

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Statistics and Operations Research.

Chapel Hill
2021

Approved by:

J.S Marron

Jan Hannig

Katherine A. Hoadley

Mariana Olvera-Cravioto

Yufeng Liu

ABSTRACT

Xi Yang: MACHINE LEARNING METHODS IN HDLSS SETTINGS
(Under the direction of J.S Marron, Jan Hannig and Katherine A. Hoadley)

During the exploration of high dimension-low-sample-size (HDLSS) data in different fields such as genetics, finance, computer science, etc, various machine learning methods have been developed. This dissertation includes the invention of novel methods and the improvement of current methods, which are evaluated using cancer genetics data.

The statistical significance of the difference between subgroups is a central question in the setting of HDLSS data. The Direction Projection Permutation (DiProPerm) hypothesis test provides an answer to this that is directly connected to a visual analysis of the data. However, under some circumstances, the DiProPerm test can be less powerful and accurate when measuring the significance of the test pairs. In this dissertation, we first introduce a new permutation method. This increases the power of the test in high signal situations. Furthermore, the simulated null test statistics tend to be more reasonable and uni-modal. Then, our theoretical analysis provides an adjustment to the inference for both permutation schemes. This enables us to exploit the improved power available. We also add confidence measures that reflect the Monte Carlo uncertainty in that test, which is seen to be very useful for the comparison of results across different contexts.

Another important goal of this dissertation is to understand the drivers of Angle-Based Joint and Individual Variation Explained (AJIVE). An important open problem is a statistical inference on the AJIVE loadings to determine which are significant features of the analysis. Jackstraw is a method that generally aims to find the statistically significant drivers associated with unobserved latent variables. In this dissertation, we develop a method based on similar ideas in the richer context of AJIVE to give a precise estimation.

Genetic data sets are used to evaluate the above-proposed machine learning methods, which also give results of independent interest to biologists.

To

ACKNOWLEDGEMENTS

First, I am extremely grateful to my supervisors, Dr. Marron, Dr. Hannig, and Dr. Hoadley for their advice, support, encouragement, and patience during my Ph.D. study. Without them, I can never complete my Ph.D. program and have a such wonderful academic research life. I cannot express enough thanks to Dr. Marron for his patience in teaching me writing, presentation, statistical consulting skills besides academic research. I would like to thank all the members of my Ph.D. committee. Thank you very much for your time. I would like to express my deep and sincere gratitude to my parents and grandparents for their tremendous understanding and encouragement in the past 20 more years. Finally, my special thanks go to my friends for their company, friendship, and empathy.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

Introduction

In the age of big data, valuable information can be extracted using modern data analysis methods. Big data is a currently fashionable topic in many fields. People gather different types of information such as graphs, texts, sound, etc, and transform these into data. Then statisticians use various techniques to analyze and visualize such data. For example, researchers transfer the massive online reviews or comments into big data with large sample size and then use statistical methods to analyze the corresponding sentiment. Another example is recommendation systems. Researchers combine click rates, web open rates, etc. to analyze the needs of different customers and recommend products to the corresponding customer to make higher profits.

Big data has become a major trend in statistics. The above examples refer to big data with a large sample size. Another type of big data has a large number of features. For a single sample, we can observe millions or even billions of features using various modern techniques. Our research in this dissertation focuses on a particular part of big data, where the sample size may not be very large, but the dimension is very large. We call such settings High Dimension Low Sample Size (HDLSS) data. Classical mathematical analysis of statistical methods often use asymptotic tools, such as taking the limit as the sample size goes to infinity. However, relatively few methods focus on the data with extremely large dimensions, especially when the feature size is larger than the sample size. Many important data sets, especially in bioinformatics have high dimensions, but a low sample size. In this dissertation, we explore some important aspects of statistical analysis in such settings.

We first focus on classification problems in the HDLSS context. The traditional t-test, often called the A/B test in industry only deals with data with a single or small amount of features. When the feature size is larger than the sample size, there are relatively few available statistical methods. The Direction Projection Permutation (DiProPerm) hypothesis test aims to solve such problems in HDLSS settings. In particular, it gives insights into high dimensional visualizations, because

it is directly linked to what is seen there. DiProPerm is especially valuable in the comparison of different hypotheses. Is one set of hypotheses more significant than the other and how to quantify such a difference? For example, when doing a genetic analysis, which animals are closer to human beings, dolphins or gorillas? This type of comparison or difference is one of the important parts of this dissertation studied in Chapter 3. DiProPerm comparison between different hypotheses can sometimes be uncertain due to permutation variation. This is addressed by developing novel confidence intervals to assess the variation.

Another major challenge is feature selection. Feature selection is an important topic in both statistics and machine learning, such as in regression or decision trees. It has the potential to both increase the model accuracy and can save time/cost for stakeholders. A traditional solution such as the enumeration method can be very slow to compute in the HDLSS setting due to the extremely large number of features. Besides, traditional solutions mostly assume the sample size is larger than the number of features, ie. dimension. However, feature selection is essential in HDLSS settings. The Jackstraw methodology focuses on finding a subset of features that are closely related to the latent information regarding PCA. This dissertation focuses on finding a subset of features that may contribute to interesting underlying aspects. Such aspects are extracted using a data integration technique called Angle-based Joint and Individual Variation Explained (AJIVE). A fundamental contribution of this dissertation is to adapt the Jackstraw methodology to joint and also individual features derived from AJIVE in Chapter 4.

CHAPTER 2

Background and Literature Review

In this chapter, we review papers related to this dissertation. Section 2.1 gives a literature review of DiProPerm for high himensional hypothesis tests. We will carefully investigate it in Chapter 3 showing some surprising problems in the high signal case for which some solutions are developed. Section 2.2 gives a literature review of Angle-Based Joint and Individual Variation Explained (AJIVE). A hypothesis test based statistical significance of variables driving systematic variation in high-dimensional data (Jackstraw method) is discussed in Section 2.3. AJIVE and Jackstraw form the basis of the new method proposed in Chapter 4.

2.1 The DiProPerm Test

In this section, we give an introduction to the DiProPerm Test (Wei et al. (2016)). The motivation is given in Section 2.1.1 and the algorithm is given in Section 2.1.2. A real data example is given in Section 2.1.3. The DiProPerm Test is evaluated and improved in Chapter 3.

2.1.1 Motivation

During the exploration of breast cancer data, previously well-determined subtypes (Sørlie et al. (2001)) have been playing a prominent role. In studies such as The Cancer Genome Atlas (TCGA) (**?**), an interesting question is how different are these subtypes? A useful way to quantify this is to evaluate hypothesis tests for differences between subtypes. In Figure 2.1, the input data is from a gene expression matrix from a TCGA breast cancer study Ciriello et al. (2015), with 1038 cases and 20249 genes. These are two views from a principal component analysis (PCA) (see Jolliffe (2011) for a good introduction). Colors and symbols are used to contrast subtypes. The panel on the right is a standard PCA plot, the first principle component (PC1) score is on the vertical axis and the second principal component (PC2) score is on the horizontal axis. The plot on the

left shows the 1-dimensional distribution of PC1 scores. The symbols are shown as a jitter plot with random heights on the vertical axis and scores are shown on the horizontal axis. The colored curves are kernel density estimates, i.e. smooth histograms. The Basal subtype (red ◁) is clearly different from the others, while LumA (blue +), LumB (cyan ×) and Her2 (magenta *) appear to be relatively closer to each other, but with some differences between. For example, the Her2 (magenta *) cases appear to be somewhat intermediate in these views.

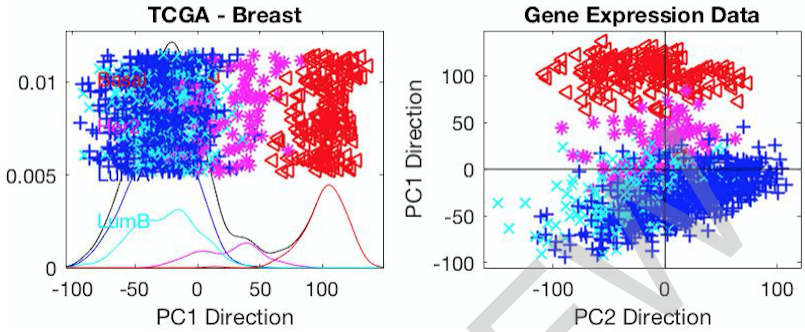
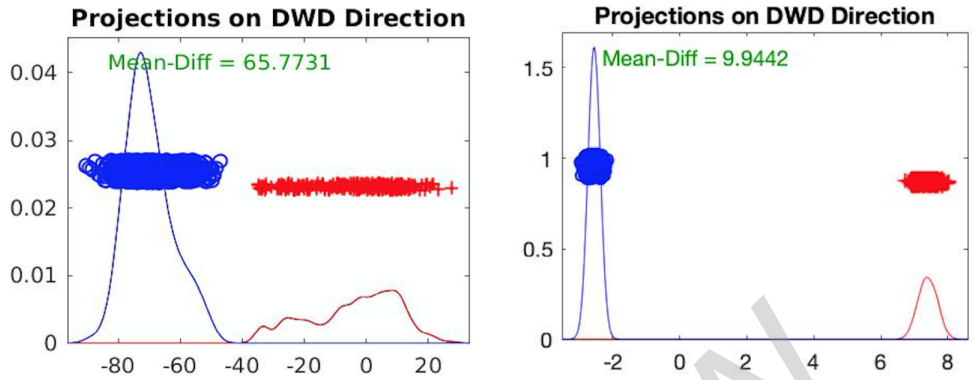


Figure 2.1: PCA views of TCGA breast cancer gene expression with subtypes: Basal: (red ◁); Her2: (magenta *); LumA: (blue +); LumB: (cyan ×). Left: distribution of each subtype projected on the PC1 direction with colors and symbols used to contrast subtypes; right: scatter plot of raw data projected on PC1 (vertical axis) and PC2 (horizontal axis) directions. This suggests Basal is much different from the others, which are somewhat different from each other.

While PCA views such as Figure 2.1 often show interesting structure in data, it is important to keep in mind that the first principal component only reflects the direction of maximum variation. Other aspects such as subtype differences may not appear clearly in the few directions studied in a PCA scatter plot. This is apparent in the left panel of Figure 2.2, which shows the projection onto a direction that much more clearly separates the union of (Basal & Her2) (red ◁ & magenta * in Figure 2.1 and blue ◯ in Figure 2.2) from the (LumA & LumB) (blue + & cyan × in Figure 2.1 and red + in Figure 2.2), than what can be seen in Figure 2.1. This clear difference between subtypes is less visible in Figure 2.1 because there is less total variation in this direction as indicated by the horizontal axis.

Figure 2.2 shows how a projection direction for distinguishing classes can provide good visual separation. A common choice for separating classes is the support vector machine (SVM) Vapnik (1995), but as noted in Marron et al. (2007) that suffers from a data piling problem in the case of High Dimension Low Sample Size (HDLSS) data, which results in poor visualization and gener-

alizability. A much clearer visual impression of subtype differences comes from the projection on the Distance Weighted Discrimination (DWD) direction Marron et al. (2007) trained on subclasses as in Figure 2.2. While such graphics are suggestive, they can be deceptive. This is illustrated



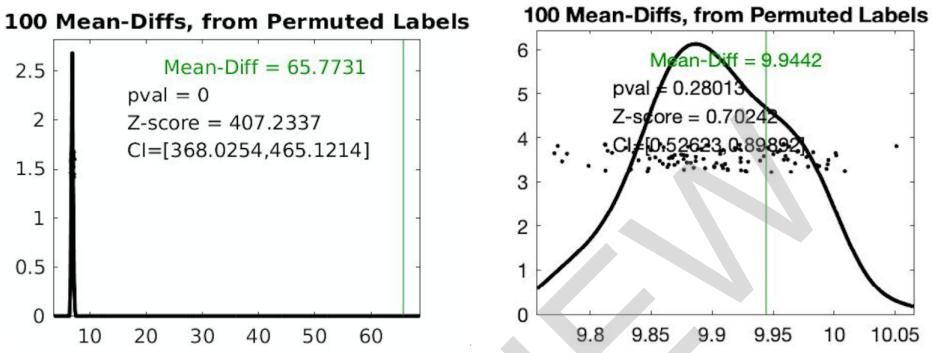
(a) Basal&Her2 vs. LumA&LumB (b) Normal(0,1) vs. Normal(0,1)

Figure 2.2: Distribution of data projected on DWD directions using new symbols and colors: class +1 (blue ◯) and class -1 (red +). Left panel: class +1: Basal&Her2 ($n_1 = 266$) class -1: LumA&LumB ($n_2 = 772$) and $d = 20249$. Right panel: both classes are the same number of independent samples from the same d dimensional standard normal. This shows that visual impression of class difference may not be reliable since the relative distance in the left panel appears to be much smaller than in the right, which is surprising since the 2 classes in the right panel are from the same distribution.

in the right panel, which shows red and blue data sets both simulated from the standard normal distribution, i.e. there is no underlying distributional difference between the classes. For straight forward comparison with the left panel, the simulated data set uses the same dimensions $d = 20249$, and sample sizes $n_1 = 266$, $n_2 = 772$. The distance between the 2 distributions in the left panel appears to be relatively smaller than that in the right panel. Since the right panel has both classes drawn from the same distribution, there is no statistically significant difference despite the apparent strong visual difference (this is an effect of very high dimensionality). Hence it is very important to do the needed statistical inference provided by a formal hypothesis test in such visualizations.

In the classical setting, where the sample size is larger than the dimension, many methods have been developed to test the equality of two distributions. However, few methods are designed for HDLSS data. The classical methods are nearest neighbor tests Bickel et al. (1983), Henze et al. (1988), Schilling (1986). A more recent method is the energy test Székely et al. (2004), which is based on the Euclidean distance to find the nearest neighbor coincidences. The classical Hotelling

T^2 test is useful when testing the equality of means, but not computable due to the singular covariance matrix of HDLSS data. Bai and Saranadasa (1996), Chen et al. (2010), Srivastava and Du (2008) solved this problem by using a diagonalized version instead. Another method that uses the traditional Hotelling T^2 statistic is to project the HDLSS data onto a low dimensional space (Lopes et al. (2011)). However, the above methods have the assumption of equal covariances. The DiProPerm test makes no such assumption. Figure 3.9 shows how the formal DiProPerm hypothesis



(a) Basal&Her2 vs. LumA&LumB (b) Normal(0,1) vs. Normal(0,1)

Figure 2.3: Black curves: estimated null distributions from permuted differences shown as black dots; green vertical line: the observed mean difference between class +1 and class -1. The permutation p-value reveals a strongly significant difference in panel (a) but insignificance in panel (b) as expected, despite the visual impression from Figure 2.2.

test, which quantitatively analyzes such differences is implemented. The DiProPerm test of Wei et al. (2016) directly targets what we visually see, which is different from the above methods. As noted in Section 2.1.2, this is done with a summary statistic such as the difference of group means. These summaries are highlighted in Figure 3.9 as the horizontal coordinate of the vertical green lines. Significance is assessed using a permutation null distribution, where the permuted statistics (from permutations as stated in step 3 in Section 2.1.2) are shown as black dots, whose distribution is indicated using a black kernel density estimate. In the left panel of Figure 3.9, the black dots are strongly piling up in a relatively narrow neighborhood, which makes them hard to see and thus the kernel density estimate is relatively narrow on the scale. The vertical green line is relatively far from the black dots and the kernel density estimate, which indicates the test pair is rather strongly significantly different. The empirical p-value is the proportion of the black dots to the right of

the green line, i.e. $\frac{\#\{\text{black dots on the right of the green line}\}}{\#\{\text{black dots}\}}$. In many cases, the empirical p-values are zeros which makes the comparison of the significance of different test pairs difficult. Thus, we need another measurement of significance for test pairs, such as the z-score. The z-score is the difference of the observed mean difference and the mean of the black dots divided by the standard deviation of the black dots, i.e. z-score $= \frac{C-\bar{C}}{S}$, where C is the observed mean difference, shown as the x coordinate of the vertical green line; $\bar{C}$ is the mean of the black dots; S is the standard deviation of the black dots. The 0 empirical p-value and large z-score (as well as the 90% confidence interval for the z-score, shown as 'CI' in Figure 3.9) is consistent with what we visually see. However, in the right panel of Figure 3.9, the black dots are more clear and the green vertical line is in the middle of the black dots (and the kernel density estimate), with a relatively large p-value and small z-score, which indicates this test pair is not significantly different. The results shown in Figure 3.9 are as expected that different tumor subtypes are drawn from different distributions, while the 2 different groups of standard normal samples are drawn from the same distribution, i.e. there is no significant difference.

2.1.1.1 DiProPerm diagnostic graph

Figure 2.4 gives an example of the DiProPerm diagnostic graph using the GE data from TCGA breast cancer data and testing the subtype pair: Basal vs Her2, which we will also study later in Section 3.1. The top left panel shows the data projected on the original direction, with colors and symbols representing the classes. The green text indicates the observed test statistic is the mean-difference statistic, which is 109.5. This gives a visual impression of the separation of the data and the goal of DiProPerm is to asses a significant difference, which is done to the panel to the right. The top right panel shows the permutation null distribution where the black dots are the permuted statistics and the black curve is the kernel density estimate of the black dots. The empirical p-value is 0 and the z-score is 41.6, indicating a significantly different. The x coordinate of the vertical green line represents the observed test statistic. The position of this line indicates the significance of the difference which is observed in the left panel. When it is in the middle of the distribution, it is not significant. When it is to the right, it is significant. The bottom 2 panels show 2 random permutations with colors representing the permuted labels and symbols representing the original labels. The black text on the top indicates the corresponding permuted test statistics are

26.5 and 24.0 respectively, which are 2 of the black dots in the upper right. These are diagnostic providing a visual impression of the impact of the permutation on the statistic.

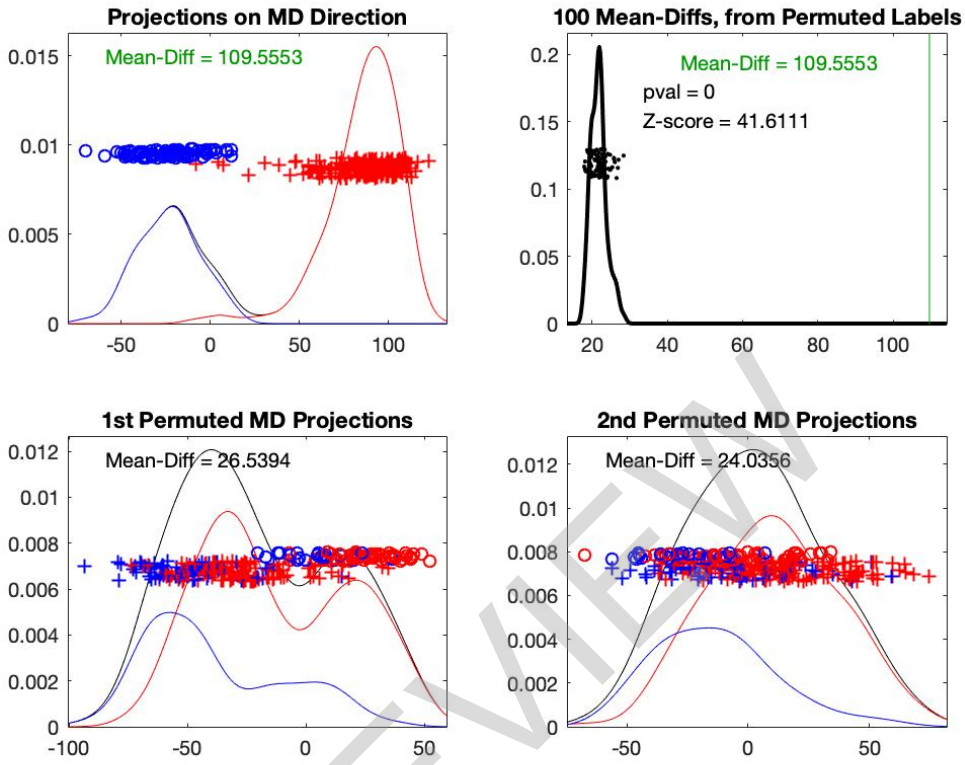


Figure 2.4: The DiProPerm diagnostic graph. Top left: data projected on the original direction with class 1 (blue circles) and class 2 (red pluses). Top right: black dots are the permuted statistics; the black curve is the kernel density estimate of the black dots; the green vertical line is the observed test statistics between class +1 and class -1. Bottom 2 panels: 2 random permutations with colors representing the permuted labels and symbols representing the original labels. From the black texts on the top right panel, the test pair is significant since the p-value is 0 and the z-score is 41.6111.

2.1.2 Algorithm

As stated in Wei et al. (2016), the aim is to test whether the 2 independent random samples of real-valued random vectors are from the same distribution. The mean difference test statistic, which is the difference between the mean of projections of each class onto the chosen projection direction, is recommended. DiProPerm is a three-step procedure:

1 Direction: choose a direction vector which is trained to best (in some sense) separate the class labels e.g. mean difference (MD) direction, DWD direction, SVM direction, etc.

2 Projection: project the 2 samples onto this direction, as shown in Figure 2.2: the blue dots (and curve) are samples (and a kernel density estimate) respectively from the projection of the first class and the red are from the projection of the second class. Then we calculate the mean difference test statistic, which is the x coordinate of the vertical green line in Figure 3.9, labeled here as C, to measure the distance between the mean of the projections of the two classes.

3 Permutation: assess the significance from a permutation test. To be specific, (a) pool the two samples and randomly permute the class labels; (b) take the normal vector to the binary linear classifier retrained on the permuted class labels; (c) project data onto this direction and re-calculate the univariate two-sample statistic. Doing this step N times (e.g N=100 or 1000) gives a sample representing the null distribution: $\{C_i\}_{i=1,\ldots,N}$, which are the black dots in Figure 3.9.

Finally, we assess the significance using the empirical p-value: $\frac{\#\{\text{black dots on the right of the green line}\}}{\#\{\text{black dots}\}}$ and the z-score: $\frac{C-\bar{C}}{S}$, where $\bar{C}$ is the sample mean of $\{C_i\}_{i=1,\ldots,N}$, and S is the sample standard deviation of $\{C_i\}_{i=1,\ldots,N}$.

This method is studied deeply and an improved method is proposed in Chapter 3.

2.1.3 Real Data Application: TCGA Lung Adenocarcinoma Data

In this section, we explore TCGA Lung Adenocarcinoma Data (Network et al. (2014)) using the DiProPerm test. This data involves 2 distinct types of measurements: one is the gene expression (GE) data and the second data type is the copy number region (CNR) data. These are both measured on the same patient samples. The GE data has $d_{GE} = 24776$ genes and $n = 402$ samples; the CNR data has $d_{CNR} = 806$ copy number regions and $n = 402$ samples. The 402 samples are grouped into 3 subtypes: 147 TRU (blue) cases, 115 ProxProlif (red) cases and 140 ProxInflam (green) cases. We use DiProPerm to test each of the three pairwise subtype differences, as well as each subtype versus the rest for both GE and CNR data.

Figure 2.5 shows the z-scores for several tests. The upper panel shows the results from GE and the lower panel gives the results from the CNR. The six different hypotheses appear in the six columns. The columns on the left are pairwise tests and the columns on the right are the 3