18.338 project: Hypothesis testing in high dimensions

Jiahai Feng

December 9, 2021

Abstract

Traditionally, statistics are applied in regimes where the the number of samples n is much greater than the number of dimensions of the data p. In this context, statistics such as the sample mean and the sample variance are well behaved. In the regime where both n and p are large, a lot of our intuition breaks down. This report aims to explore hypothesis testing in limit where both n and p are both large, and provide empirical results using real world datasets.

1 Introduction

Conventional statistical analysis often operates in the regime where the dimension of the features p is much smaller than the number of data points n. Often, the asymptotics are analyzed in the limit where $n \to \infty$ and p is held constant. In machine learning, however, it is common for p to be the same order of magnitude, and in some cases larger than n. For example, Table 1 shows the values of p and n for a few popular image classification datasets in machine learning.

	MNIST	ImageNet	CIFAR-10
n	6000	$\sim 1,000$	6000
p	784	$\sim 400,000$	3072

Table 1: n, the number of data points per class and p, the dimension of features for 3 popular image classification datsets. For color images with 3 image channels, p will be the number of pixels times 3.

This is one motivation for studying statistics in the regime where $p, n \to \infty$, $p/n \to c$ for some $c \in (0, \infty)$. Many decision and hypothesis testing tools

fail in this regime. For example, covariance estimation becomes very tricky, because we only have $\Theta(pn)$ observations to estimate a covariance matrix with $\Theta(p^2)$ entries. This report investigates specifically the behavior of a binary hypothesis test, the linear discriminant analysis (LDA) in this regime. We'll first give a quick overview of how LDA works, and then summarize the random matrix analysis of LDA as presented in [2, 3]. This analysis makes heavy use of resolvents, which are quite different from the tools we learnt in 18.338. Therefore we also attempt to give a flavor of how these techniques work. Lastly, we'll show empirical results testing the random matrix predictions for LDA.

2 Linear Discriminant Analysis

Suppose we have a distribution $p_x(\cdot)$ from which we draw n i.i.d. samples $x_1, \ldots, x_n \in \mathbb{R}^p$. Let $\mathbf{X} = [x_1, \ldots, x_n] \in \mathbb{R}^{p \times n}$. Now, suppose p_x is a gaussian mixture with two components, so that each sample can be drawn from either of two normal distributions. Specifically, let $\mathcal{H}$ denote the latent unobserved random variable representing which hypothesis $\mathbf{x}$ is drawn from.

$$\mathcal{H}_0 : \mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}_0, \mathbf{C}_0)$$
$$\mathcal{H}_1 : \mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}_1, \mathbf{C}_1)$$

For LDA, we assume that we have access to labelled data from the two normal distributions, $\mathbf{X}^{(l)} \in \mathbb{R}^{p \times n_l}$, for $l \in \{0, 1\}$, where n_0 and n_1 are the number of data points from each of the two hypotheses. Then, LDA estimates the means of the two hypothesis classes:

$$\hat{\boldsymbol{\mu}}_l = \frac{1}{n_l} \sum_{i=1}^{n_l} \mathbf{x}_i^{(l)}, \qquad l \in \{0, 1\}.$$

Crucially, LDA makes the assumption that the two hypothesis have the same covariance matrix, even if it is untrue in reality. It then, for a regularizer $\gamma \geq 0$, makes the estimations

$$\hat{\mathbf{C}}_l^{(\gamma)} = \frac{1}{n_l - 1} \sum_{i=1}^{n_l} (\mathbf{x}_i^{(l)} - \hat{\boldsymbol{\mu}}_l)(\mathbf{x}_i^{(l)} - \hat{\boldsymbol{\mu}}_l)^T + \gamma \mathbf{I}_p$$

to obtain one covariance matrix estimate:

$$\hat{\mathbf{C}}^{(\gamma)} = \frac{n_0 - 1}{n - 2} \hat{\mathbf{C}}_0^{(\gamma)} + \frac{n_1 - 1}{n - 2} \hat{\mathbf{C}}_1^{(\gamma)}.$$

LDA then uses the estimated parameters as the guess for the true parameters, and performs hypothesis testing. In both the non-parametric and the Bayesian setting, it is useful to consider the log-likelihood ratio

$$\log \frac{p(\mathbf{x}|\mathcal{H}_0)}{p(\mathbf{x}|\mathcal{H}_1)},$$

which, under the assumptions that LDA makes, reduces to:

$$T_{\mathrm{LDA}}^{(\gamma)}(\mathbf{x}) = (\mathbf{x} - \hat{\boldsymbol{\mu}})^T [\hat{\mathbf{C}}^{(\gamma)}]^{-1} (\hat{\boldsymbol{\mu}}_0 - \hat{\boldsymbol{\mu}}_1),$$

where $\hat{\boldsymbol{\mu}} = \frac{1}{2}(\hat{\boldsymbol{\mu}}_0 + \hat{\boldsymbol{\mu}}_1)$. Then, both the optimal Neyman-Pearson test and Bayesian prediction will take the form:

$$T_{\mathrm{LDA}}^{(\gamma)}(\mathbf{x}) \underset{\mathcal{H}_1}{\overset{\mathcal{H}_0}{\gtrless}} \xi$$

for some $\xi \in \mathbb{R}$.

Thus, $T_{\mathrm{LDA}}(\mathbf{x})$ is the quantity that is critical to analyzing LDA.

3 Theoretical Analysis

3.1 Background

First, we introduce the notion of a deterministic equivalent. $\overline{\mathbf{Q}} \in \mathbb{R}^{n \times n}$ is a deterministic equivalent for the symmetric random matrix $\mathbf{Q} \in \mathbb{R}^{n \times n}$, if for arbitrary deterministic matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ and vectors $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$ of unit operator and Euclidean norms, we have, as $n \to \infty$,

$$\frac{1}{n} \operatorname{tr} \mathbf{A}(\mathbf{Q} - \overline{\mathbf{Q}}) \to 0, \qquad \mathbf{a}^T(\mathbf{Q} - \overline{\mathbf{Q}})\mathbf{b} \to 0$$

Deterministic equivalents are useful because quantities like the spectral density can be expressed in terms of them, via the resolvent. For a matrix $\mathbf{M} \in \mathbb{R}^{n \times n}$, the resolvent is defined:

$$\mathbf{Q}_{\mathbf{M}}(z) = (\mathbf{M} - z\mathbf{I}_n)^{-1}.$$

Define also the Stieljes Transform $m_\mu : \mathbb{C} \setminus \mathrm{supp}(\mu) \mapsto \mathbb{C}$ of a probability measure μ with support $\mathrm{supp}(\mu) \subset \mathbb{R}$ as:

$$m_\mu(z) = \int \frac{1}{t - z} \mu(dt).$$

Then, it follows that if $\mu_{\mathbf{M}}(x) = \frac{1}{n}\sum_{i=1}^{n}\delta_{\lambda_i(\mathbf{M})}(x)$ is the measure describing the density of eigenvalues of $\mathbf{M}$,

$$m_{\mu_M}(z) = \frac{1}{n}\operatorname{tr}\mathbf{Q}_{\mathbf{M}}(z)$$

Thus, with the help of the inverse Stielfjes Transform, finding a deterministic equivalent $\overline{\mathbf{Q}}$ for the resolvent $\mathbf{Q}_{\mathbf{M}}$ is sufficient to retrieve the spectral density.

As a familiar example, the Marcenko-Pastur law, written in the language of resolvents, looks like [2]:

Let $\mathbf{X} \in \mathbb{R}^{p\times n}$ have i.i.d. columns $\mathbf{x}_i$ such that $\mathbf{x}_i$ has independent zero-mean, unit-variance entries, and denote $\mathbf{Q}(z) = (\frac{1}{n}\mathbf{X}\mathbf{X}^T - z\mathbf{I}_p)^{-1}$ the resolvent of $\frac{1}{n}\mathbf{X}\mathbf{X}^T$. Then, as $n, p \to \infty$ with $p/n \to c \in (0,\infty)$, $\mathbf{Q}(z)$ has a deterministic equivalent $\overline{\mathbf{Q}}(z) = m(z)\mathbf{I}_p$, where

$$zcm^2(z) - (1-c-z)m(z) + 1 = 0$$

$m(z)$ can be obtained by solving the quadratic equation, and upon taking the inverse Stielfjes transform will result in the familiar Marcenko-Pastur distribution.

3.2 LDA

We return to the LDA setting. We look at the limit where $n_l/n \to c_l$ and $p/n \to c$. We may without loss of generality assume that $\|\mathbf{C}\| = 1$. It turns out that to avoid having an asymptotically trivial or impossible decision task, we require $\|\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1\| = O(1)$ as well. Under these settings, it was shown in [3] that

$$T^{(\gamma)}_{\mathrm{LDA}}(\mathbf{x}) = (\mathbf{x} - \hat{\boldsymbol{\mu}})^T[\hat{\mathbf{C}}^{(\gamma)}]^{-1}(\hat{\boldsymbol{\mu}}_0 - \hat{\boldsymbol{\mu}}_1)$$

satisfies a central limit law. The mean and variance of the limiting gaussian distribution can be found via deterministic equivalents. The process of deriving them is technically involved. While the derivation is instructive, we will not reproduce it here, and instead refer interested readers to [2]. Instead, we provide the main results.

Let l take either 0 or 1, and $\mathbf{x} \sim \mathcal{H}_l$. Then,

$$\mathbb{E}[T^{(\gamma)}_{\mathrm{LDA}}(\mathbf{x})] = \frac{(-1)^l}{2}(\boldsymbol{\mu}_0-\boldsymbol{\mu}_1)^T\overline{\mathbf{Q}}^{\circ}(-\gamma)(\boldsymbol{\mu}_0-\boldsymbol{\mu}_1) - \frac{1}{2}g_0(-\gamma) + \frac{1}{2}g_1(-\gamma) + o(1)$$

where $\overline{\mathbf{Q}}^{\circ}$ is the deterministic equivalent for $\mathbf{Q}^{\circ}$, which is the high rank part of $[\hat{\mathbf{C}}^{(\gamma)}]^{-1}$. The precise meaning of that statement is beyond the scope of this report. We present here an expression for $\overline{\mathbf{Q}}^{\circ}$:

$$\overline{\mathbf{Q}}^{\circ}(z) = -\frac{1}{z}\left(\mathbf{I}_p + \sum_{l=0}^{1} c_l \tilde{g}_l(z)\mathbf{C}_l\right)^{-1},$$

where $g_l(z), \tilde{g}_l(z)$ satisfy

$$g_l(z) = \frac{1}{n}\operatorname{tr}\mathbf{C}_l\overline{\mathbf{Q}}^{\circ}(z), \qquad \tilde{g}_l(z) = \frac{1}{z}\frac{1}{1+g_l(z)}.$$

The variance is given by:

$$\begin{aligned}\operatorname{var}[T_{\mathrm{LDA}}^{(\gamma)}(\mathbf{x})] = (\boldsymbol{\mu}_0-\boldsymbol{\mu}_1)^T\overline{\mathbf{Q}^{\circ}\mathbf{C}_l\mathbf{Q}^{\circ}}(\boldsymbol{\mu}_0-\boldsymbol{\mu}_1) + \frac{1}{n_0}\operatorname{tr}\mathbf{C}_0\overline{\mathbf{Q}^{\circ}\mathbf{C}_l\mathbf{Q}^{\circ}} \\ + \frac{1}{n_1}\operatorname{tr}\mathbf{C}_1\overline{\mathbf{Q}^{\circ}\mathbf{C}_l\mathbf{Q}^{\circ}} + o(1),\end{aligned}$$

where the deterministic equivalent $\overline{\mathbf{Q}^{\circ}\mathbf{C}_l\mathbf{Q}^{\circ}}$ was derived in [1] to be:

$$\overline{\mathbf{Q}^{\circ}\mathbf{C}_l\mathbf{Q}^{\circ}} = \overline{\mathbf{Q}}^{\circ}\mathbf{C}_l\overline{\mathbf{Q}}^{\circ} + \overline{\mathbf{Q}}^{\circ}(R_{0l}\mathbf{C}_0 + R_{1l}\mathbf{C}_1)\overline{\mathbf{Q}}^{\circ},$$

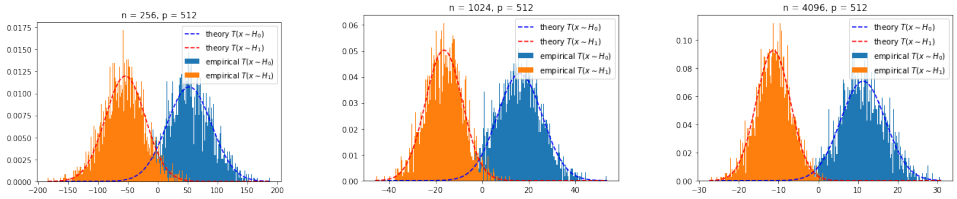
where $R_{ij} = \frac{c_i}{c_j}[(\mathbf{I}_2-\mathbf{S})^{-1}\mathbf{S}]_{i+1,j+1}, [\mathbf{S}]_{i+1,j+1} = c_j\gamma^2\tilde{g}_i(-\gamma)^2\frac{1}{n}\operatorname{tr}\mathbf{C}_i\overline{\mathbf{Q}}^{\circ}C_j\overline{\mathbf{Q}}^{\circ}$.

This concludes the theoretical analysis of LDA. Let us summarize what the results are. Starting from a gaussian mixture model with two components, we draw labelled samples that constitute the training data. From the training data, LDA estimates the means and the covariances of the two components, assuming crucially that the two components have the same covariance, even if that may not be true. To test a new data point $\mathbf{x}$, LDA computes the quantity $T_{\mathrm{LDA}}^{(\gamma)}(\mathbf{x})$ and makes a decision based on its threshold value. The theoretical analysis in this section predicts that the distribution of $T_{\mathrm{LDA}}^{(\gamma)}(\mathbf{x})$ has a central limit to a gaussian with the above moments.

4 Empirical Results

Experimentally, one way to verify the limiting distribution is to take a finite value of p and n, perform LDA, and then see the distribution of $T_{\mathrm{LDA}}^{(\gamma)}$ converges to the central limit. The results are shown in Figure 1.

The code generating these plots was adapted from [2].



(a) $n = 256, p = 512$ (b) $n = 1024, p = 512$ (c) $n = 4096, p = 512$

Figure 1: $T^{(\gamma)}_{\text{LDA}}$ scores for synthetically generated GMM data

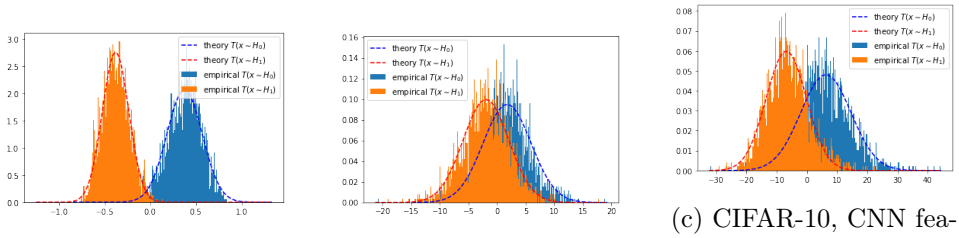
Another interesting experiment we can do is to perform LDA on real world data. Image classification datasets consist of many images with a corresponding classification, such as whether the image contains a dog or a cat. We might imagine that the set of images containing a dog forms a distribution in the high dimensional space of images, and the set of images containing a cat forms another distribution. These distributions are probably not gaussian, but we might hope that because of some form of universality, we will obtain similar distributions for $T^{(\gamma)}_{\text{LDA}}$ if we perform LDA on these datasets. [2] found that for Fashion-MNIST and Kannada-MNIST datasets, there is a remarkably close fit with the theoretical LDA predictions.

We recreate the plot for the MNIST and CIFAR-10 datasets in Figures 2a and 2b. The theoretical limiting distribution of T appears to fit remarkably well again, even for the more complex CIFAR-10 dataset.

We also trained, on CIFAR-10, a simple convolutional neural network with two convolutional layers, each with 50 channels, 3x3 kernels and stride 2, and one fully connected layer. We take the penultimate output as features, and ran LDA on these features. We can see that the two hypotheses' distributions have separated more in Figure 2c than in Figure 2b, which is what we expect the features to be able to achieve. The features are non-linear, complicated transformations of the initial dataset, and we still see in Figure 2c that the distribution is still well predicted by the central limit theorem.

5 Conclusion

In the $p, n \to \infty$ regime, random matrix methods could be used to prove the central limit property of the $T^{(\gamma)}_{\text{LDA}}$ score from LDA. This distribution has been empirically shown to be robust, even on real world data that are not gaussian mixtures. Further work could investigate failure cases for the central limit property. It also could be useful to study intermediate outputs



(a) MNIST dataset (b) CIFAR-10 dataset (c) CIFAR-10, CNN features

Figure 2: $T^{(\gamma)}_{\text{LDA}}$ scores for real world image classification datasets

of deep neural networks by looking at the means and covariances, since this analysis seems to suggest that those two moments capture a lot of what is going on. This line of inquiry is reminscent of recent work on neural tangent kernels, with the key difference that the gaussian distributions in NTK analysis arise from randomness in the initialization, whereas in this context it comes from the input distribution.

References

[1] F. Benaych-Georges and R. Couillet. Spectral analysis of the gram matrix of mixture models. *ESAIM: Probability and Statistics*, 20:217–237, 2016.

[2] R. Couillet and L. Zhenyu. *Random Matrix Methods for Machine Learning*. 2021.

[3] K. Elkhalil, A. Kammoun, R. Couillet, T. Y. Al-Naffouri, and M.-S. Alouini. A large dimensional study of regularized discriminant analysis. *IEEE Transactions on Signal Processing*, 68:2464–2479, 2020.