18.338 - High Dimensional Hypothesis Testing

Jiahai Feng

High dimensional asymptotics

Conventional hypothesis testing:

- Holds number of features *p* constant, takes number of samples *n* to infinity

Machine learning applications:

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

High dimensional asymptotics

$$p, n \to \infty,\ p/n \to c \text{ for some } c \in (0, \infty)$$

Binary Hypothesis Testing

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

Linear Discriminant Analysis

Estimates log-likelihood ratio:

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

Estimate sample mean and variance:

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

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

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

Linear Discriminant Analysis

Log-likelihood ratio reduces to:

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

Optimal decision:

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

Deterministic equivalents

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 $n \to \infty$,

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

Resolvent

Let $\mathbf{M} \in \mathbb{R}^{n \times n}$

The resolvent:

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

Stieljes Transform

Stieljes Transform of measure μ

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

It follows that:

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

Deterministic equivalents

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 $n \to \infty$,

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

Marcenko-Pastur in Resolvents

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

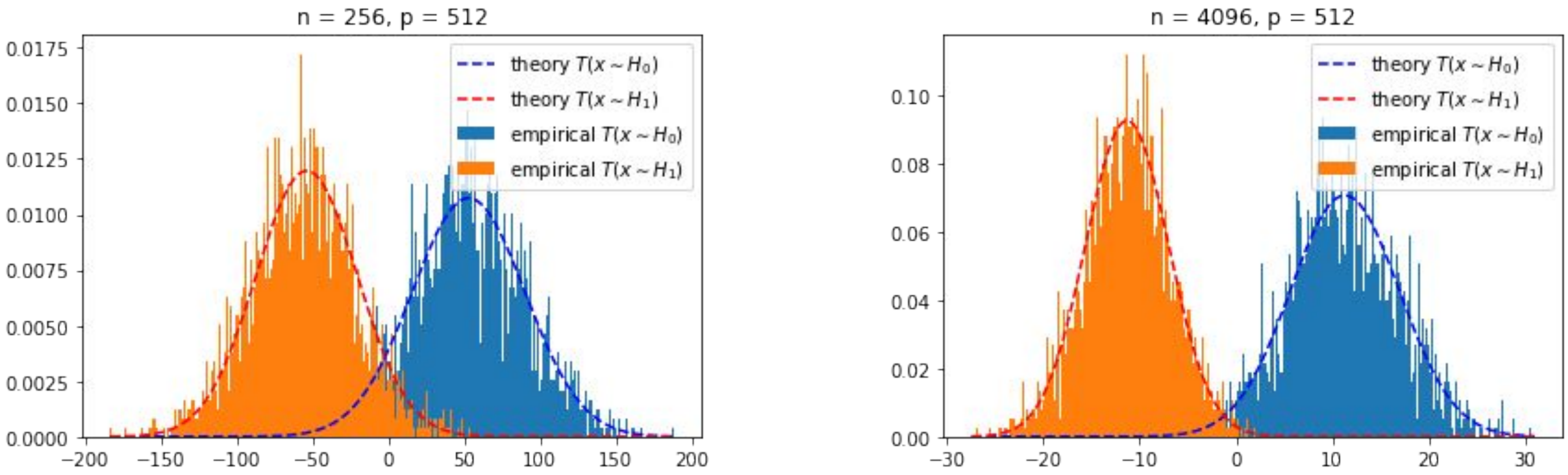
$$zcm^2(z) - (1 - c - z)m(z) + 1 = 0$$

RMT analysis of LDA

- It was shown that the LDA statistic has a central limit behavior, and tends towards a gaussian
- Moments of this gaussian, or at least of a deterministic equivalent of it, can be computed from n, p, and the underlying means and covariances of the mixture model

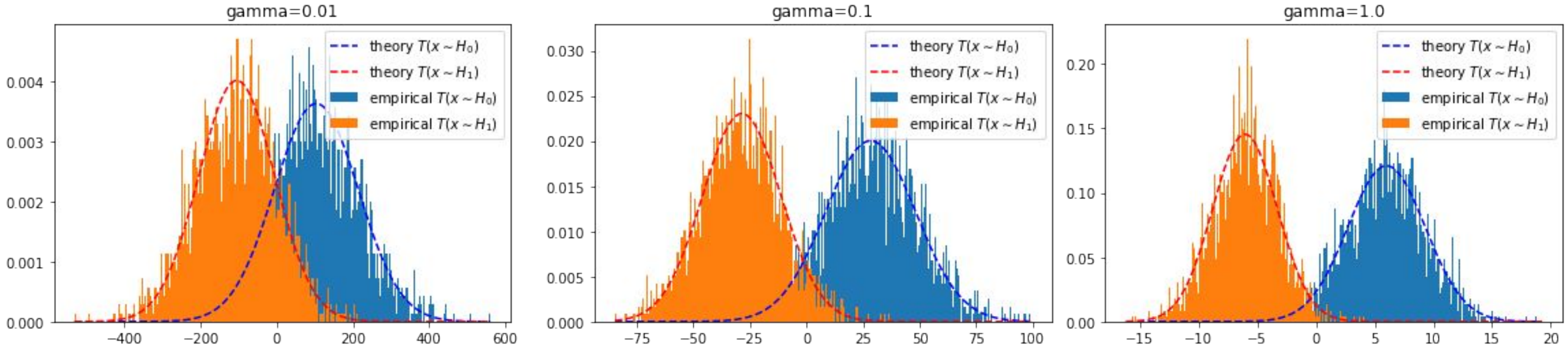
Experiments: Gaussian Mixtures

Asymmetric because mixtures have different covariances

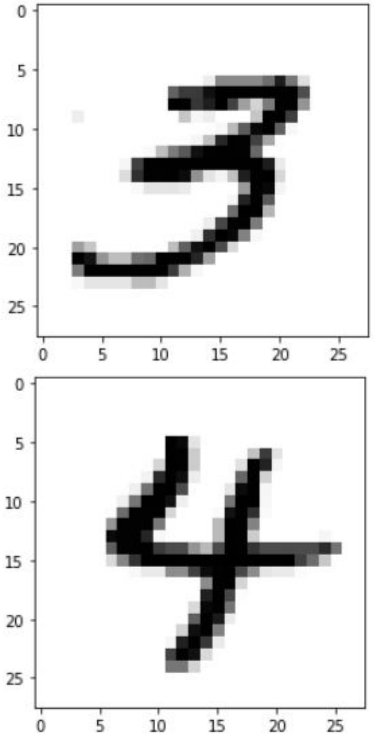


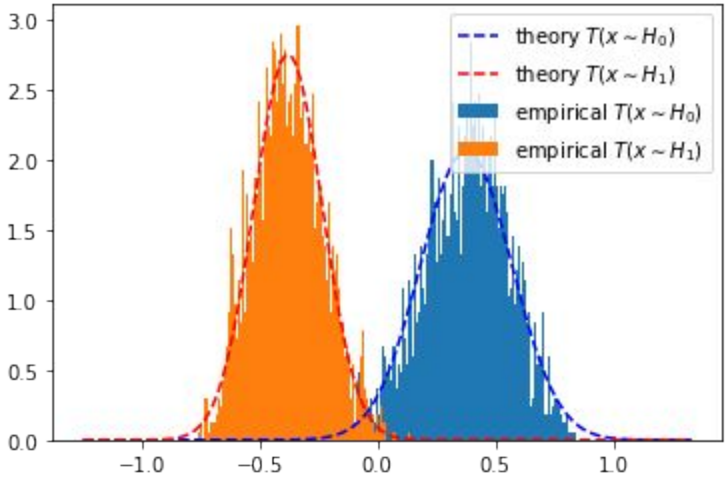
Experiments: Gaussian Mixtures

Asymmetric because mixtures have different covariances

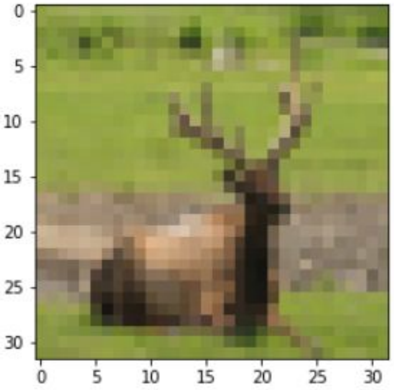


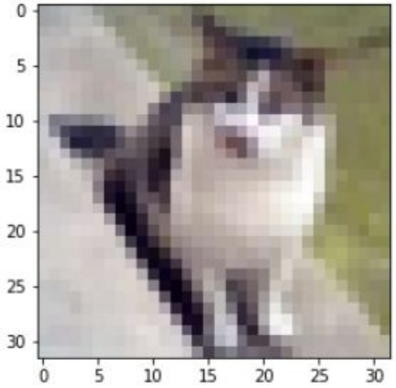
Experiments: MNIST

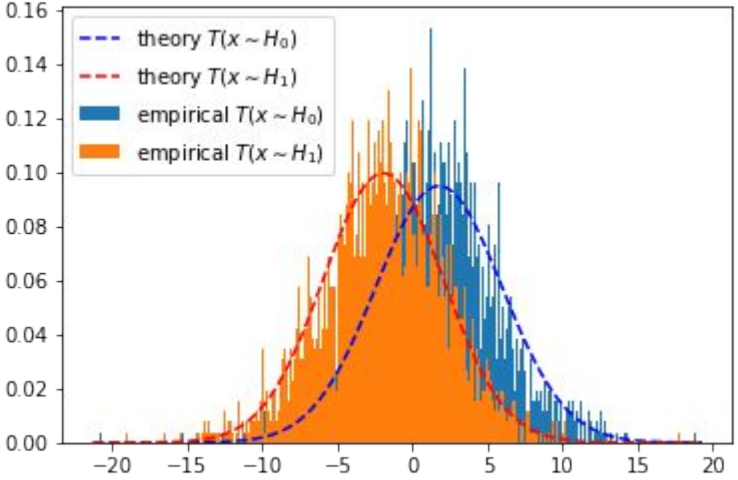




Experiments: CIFAR-10



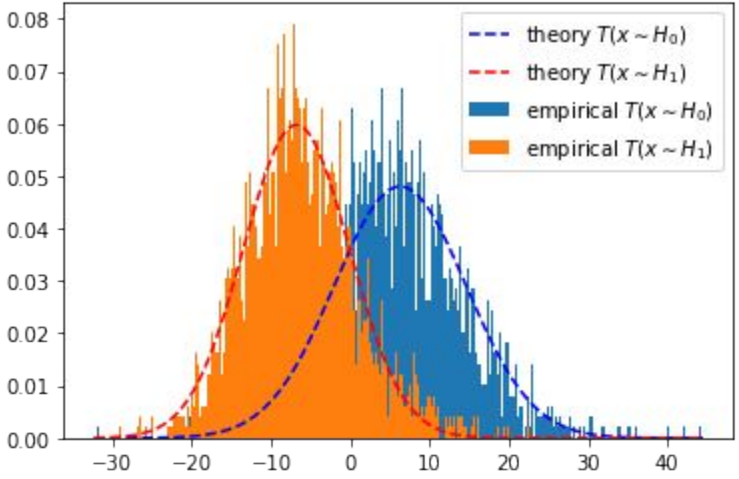




Experiments: CIFAR-10

Feature layer of simple convolutional neural network

Complicated, nonlinear transformation



2 layers of 50 channel 3x3 kernels with stride 2

CNN Accuracy: .86

LDA Accuracy: .84

Conclusion

- RMT can predict LDA statistics in the high dimensional feature limit
- Empirically robust under various conditions

Future Extensions

- Deviations from central limit
- Tool to study intermediate outputs in neural networks
- Study kernels?
- Understanding failure cases

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.