A random matrix approach for discriminant analysis

Abla KAMMOUN

King's Abdullah University of Science and Technology, Saudi Arabia

November 14, 2017

Outline

Machine learning

- An increase interest in the field of machine learning
- Has Become an important tool to many fields
  - Medicine, Biology, Sociology, health care
  - Security
  - Finance, economy
- **Goal** Build programs so that computers can perform tasks in an intelligent way

Machine learning

Shai Shalev-Shwarts and Shai Ben-David: Understanding Machine Learning: From theory to Algorithms, 2014 Cambridge University Press

Why do we need machine learning ?

- Tasks that are too complex
  - Tasks performed by humans but we do not know how we do them:
    - Image understanding, speech recognition,
  - Tasks beyond human capabilities:
    - Analysis of very large and complex data sets: weather prediction, analysis of genomic data, web search engines and electronic commerce, to name a few
    - Understand meaningful information buried in large and complex data sets.
- Tasks needs to be tailored to the input data
  - Decoding handwritten text
  - Spam detection programs
  - Speech recognition programs.

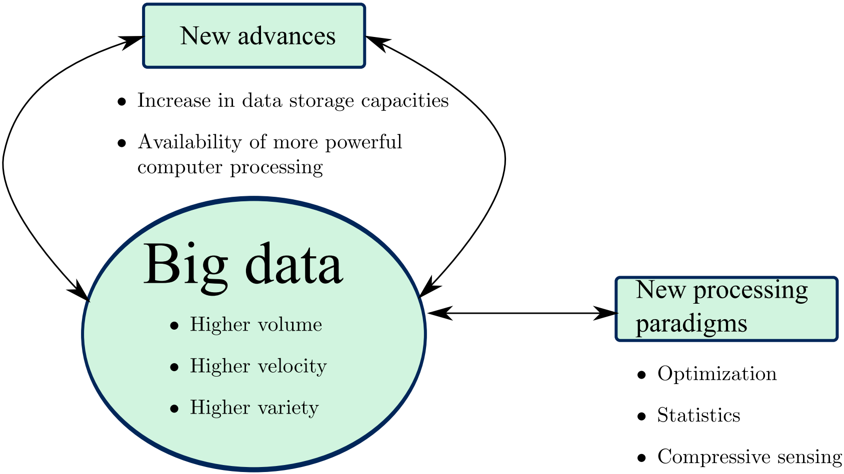
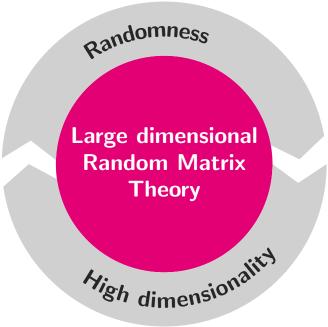


Figure: Big data: Challenges and potential tools

Random matrix theory: **Study the behavior of large random matrices**

- Allow the prediction of the behavior of random quantities depending on large random matrices
- Key of success: Randomness + High dimensionality



Random matrix theory: Example

Large random matrices

- High dimensional random matrices
  - Self-averaging effect mechanism similar to that met in the law of large numbers
  - More determinism in the system

Emblematic result from random matrix theory

- Let $\mathbf{H} \in \mathbb{C}^{n \times p}$ with i.i.d entries with zero mean and variance $\frac{1}{n}$.
- We assume that $p, n \to \infty$ with $\frac{p}{n} \to c$.

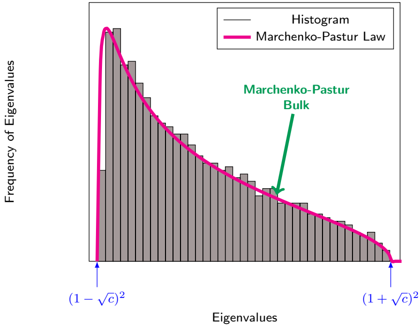


Figure: Histogram of eigenvalues of $\mathbf{H}\mathbf{H}^{\mathsf{H}}$

As n, p tends to infinity with $\frac{p}{n} \to c$, the histogram can be approximated by a "Deterministic" curve !

As n, p tends to infinity with $\frac{p}{n} \to c$, all the eigenvalues are contained in the interval $\left[(1-\sqrt{c})^2, (1+\sqrt{c})^2\right]$

Applications of random matrix theory

Signal processing

- Large number of antenna arrays, vs large number of observations
- **Outcomes** Improved signal processing techniques

Wireless Communication

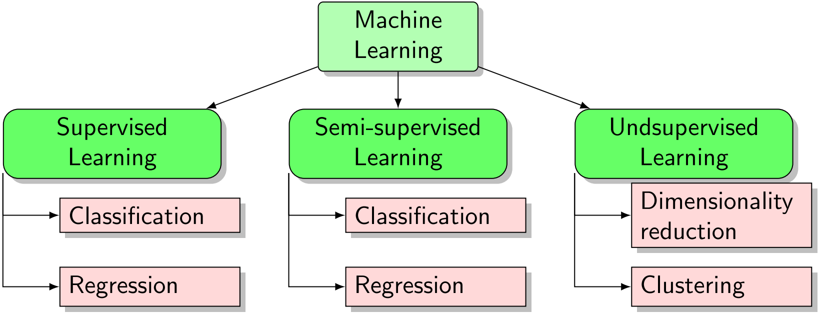
- Large-scale MIMO systems, Large number of users
- **Outcomes** Improved transmission and detection strategies

Outline

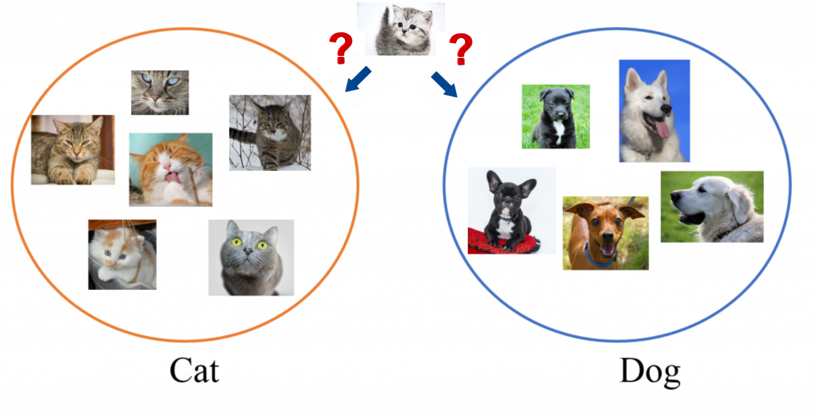
Machine Learning	Random Matrix Theory
▸ Goal: Design algorithms that allow computers to perform intelligent processing ▸ Applications: Hyperspectral imagery, Biology, Business ▸ Challenges: High dimensional data & data of different kind ▸ Data is often considered as deterministic! $\Longrightarrow$ Dimensionality is viewed as "a curse"	▸ Attribute: Efficient handling of high dimensional **data** ▸ Has proved to bring **important results** to several engineering disciplines ▸ Main ingredient: Consider Data as **random** $\Longrightarrow$ Dimensionality is viewed as a **"blessing"**

RMT tools to rise to the challenges of Machine Learning

Random matrix approach for machine learning



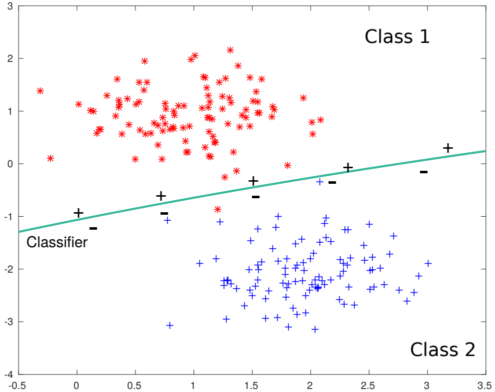
Classification: Classification is the task of selecting the best match for any input among a set of the underlying categories.



Outline

Discriminant analysis

- Widely used statistical method for supervised classification
- Principle: Builds a classification rule that allows to assign for an unseen observation its corresponding class.



Let $\mathbf{x}$ be the input data and f be the classification rule.

$$\text{Classifier} \triangleq \begin{cases} \text{Assign class 1} & \text{if} \quad f(\mathbf{x}) < 0 \\ \text{Assign class 2} & \text{if} \quad f(\mathbf{x}) > 0 \end{cases}$$

Discriminant analysis

Basic assumptions

- We assume that there are $\mathbf{x}_1, \cdots, \mathbf{x}_n$ observations with known classes.
- Observations are independent but take different distributions across classes.
- We assume the prior probability of class k is π_k.
- We assume known the class-conditional probability associated with each class

$$\text{class-conditional probability} \quad \triangleq \quad \mathbb{P}\left[\mathbf{X} = \mathbf{x} \mid \mathbf{x} \in \text{class} \quad k\right]$$

Principle

- For an unseen data $\mathbf{x}$ compute using the Bayes rule the maximum posterior probability

$$\hat{\mathbf{g}}_k(\mathbf{x}) = \mathbb{P}\left[\mathbf{x} \in \text{ class } k \ \mid \ \mathbf{x}\right]$$

- Assign to $\mathbf{x}$ the class with the highest posterior probability

$$\arg\max_k \hat{\mathbf{g}}_k(\mathbf{x})$$

Gaussian discriminant analysis

Gaussian mixture model for binary classification (2 classes)

- $\mathbf{x}_1, \cdots, \mathbf{x}_n \in \mathbb{R}^p$
- Class k is formed by $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, $k = 0, 1$

Linear discriminant analysis (LDA) Different mean but equal covariances. $\boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}_1$.

$$W^{LDA} = \left(\mathbf{x} - \frac{\boldsymbol{\mu}_0 + \boldsymbol{\mu}_1}{2}\right)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1) - \log\frac{\pi_1}{\pi_0}$$

$$\begin{cases} \text{Assign} \quad \mathbf{x} \quad \text{to class} \quad 0 \quad \text{if} \quad W^{LDA} > 0 \\ \text{Assign} \quad \mathbf{x} \quad \text{to class} \quad 1 \quad \text{otherwise} \end{cases}$$

$\rightarrow$ Decision rule is linear in $\mathbf{x}$. The LDA is a linear classifier

Quadratic discriminant analysis Different mean and covariances across classes:

$$W^{QDA} = -\frac{1}{2}\log\frac{|\boldsymbol{\Sigma}_0|}{|\boldsymbol{\Sigma}_1|} - \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_0)^T \boldsymbol{\Sigma}_0^{-1}(\mathbf{x} - \boldsymbol{\mu}_0) + \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_1)^T \boldsymbol{\Sigma}_1^{-1}(\mathbf{x} - \boldsymbol{\mu}_1)$$

$$\begin{cases} \text{Assign} \quad \mathbf{x} \quad \text{to class} \quad 0 \quad \text{if} \quad W^{QDA} > 0 \\ \text{Assign} \quad \mathbf{x} \quad \text{to class} \quad 1 \quad \text{otherwise} \end{cases}$$

$\rightarrow$ Decision rule is quadratic in $\mathbf{x}$, hence the name quadratic classifier.

Outline

Linear discriminant analysis

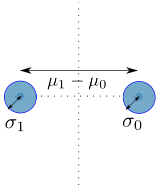
- Assume $\mathbf{\Sigma}$, $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ known.
- Equal priors : $\pi_1 = \pi_2 = 0.5$
- No asymptotic regime, p is fixed.

The total misclassification rate is equal to :

$$R = \Phi\left(-\frac{\Delta}{2}\right)$$

where $\Delta = \sqrt{(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^T \mathbf{\Sigma}^{-1} (\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)}$ and Φ the CDF of a standard normal random variables

- Assume $\boldsymbol{\Sigma}$, $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ known.
- Equal priors : $\pi_1 = \pi_2 = 0.5$
- No asymptotic regime, p is fixed.



The total misclassification rate is equal to :

$$R = \Phi\left(-\frac{\Delta}{2}\right)$$

where $\Delta = \sqrt{(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)}$ and Φ the CDF of a standard normal random variables

Takeaways:

- The higher is the difference in means, the lower is the misclassification rate,
- The variance tends to have a side effect on the classification performance.

Linear discriminant Analysis

- In practice, the means and covariance matrices on which depends the decision rule are unknown.
- Moreover, $\mathbf{\Sigma}_0 \neq \mathbf{\Sigma}_1$.
- We assume availability of **Training data**: observations for which the class label is known.

$$\underbrace{\boxed{\mathcal{N}(\mu_0, \Sigma_0)}}_{n_0 \text{ observations}} \quad \underbrace{\boxed{\mathcal{N}(\mu_1, \Sigma_1)}}_{n_1 \text{ observations}}$$

Figure: Training data

- We use empirical means and sample covariance matrices as plug-in estimators.

$$\text{Sample mean in class } i: \quad \overline{\mathbf{x}}_i = \frac{1}{n_i} \sum_{\mathbf{x}_i \in \text{class}_i} \mathbf{x}_i$$

$$\text{Covariance matrix in class } i: \quad \hat{\mathbf{\Sigma}}_i = \frac{1}{n_i} \left(\mathbf{x}_i - \overline{\mathbf{x}}_i\right) \left(\mathbf{x}_i - \overline{\mathbf{x}}_i\right)^T$$

$$\text{Pooled covariance matrix}: \quad \hat{\mathbf{\Sigma}} = \frac{n_1}{n} \hat{\mathbf{\Sigma}}_1 + \frac{n_2}{n} \hat{\mathbf{\Sigma}}_2$$

- The LDA discriminant function becomes:

$$\hat{W}^{LDA} = \left(\mathbf{x} - \frac{\overline{\mathbf{x}}_0 + \overline{\mathbf{x}}_1}{2}\right)^T \hat{\mathbf{\Sigma}}^{-1} \left(\overline{\mathbf{x}}_0 - \overline{\mathbf{x}}_1\right) - \log \frac{\pi_1}{\pi_0}$$

Cheng Wang and Binyan Jiang. On the dimension effect of regularized linear discriminant analysis

Asymptotic growth regime. Let $n = n_0 + n_1$.

- $n_0, n_1, p \to \infty$ such that $\frac{n_0}{n_1} \to 1$ and $\frac{p}{n} \to c < 1$
- $\boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}_1$
- $\boldsymbol{\mu} \triangleq \boldsymbol{\mu}_0 - \boldsymbol{\mu}_1$ is such that $\|\boldsymbol{\mu}\| = O(1)$.

Under these assumptions, the misclassification rate converges to:

$$R_{LDA} \to \Phi\Big[-\frac{\Delta}{2}\underbrace{\sqrt{1-c}}_{\text{Price of dimensionality}}\Big]$$

Takeaways:

- When $c \to 1$, the misclassification rate tends to 0.5.

$\rightarrow$ For the LDA to result in acceptable performance, we need c close to 0.

- Because its use of the inverse of the pooled covariance matrix, the LDA applies only when $c < 1$.

Outline

Regularized linear discriminant analysis

Regularizaed Linear Discriminant Analysis :R-LDA

- Applies for $c \in (0, \infty)$.
- Uses a regularized estimation of the inverse of the covariance matrix:

$$\hat{\mathbf{\Sigma}}(\gamma) = \left(\gamma\hat{\Sigma} + \mathbf{I}_p\right)$$

- The discriminant score for the R-LDA is:

$$\hat{W}^{R-LDA} = \left(\mathbf{x} - \frac{\overline{\mathbf{x}}_0 + \overline{\mathbf{x}}_1}{2}\right)^T \hat{\mathbf{\Sigma}}(\gamma)^{-1} \left(\overline{\mathbf{x}}_0 - \overline{\mathbf{x}}_1\right) - \log\frac{\pi_1}{\pi_0}$$

Analysis of regularized discriminant analysis

Assumptions

- $p, n_1, n_2 \to \infty$ with $\frac{p}{n} \to c\,(0, \infty)$, $\frac{n_1}{n} = \frac{n_2}{n} \to 0.5$
- The difference in means $\boldsymbol{\mu} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ satisfies $\|\boldsymbol{\mu}\| = O(1)$. The spectral norms of $\boldsymbol{\Sigma}_0$ and $\boldsymbol{\Sigma}_1$ are bounded

Results

Equal covariance matrices [Zollanvari 2015]

$$R - \Phi\left[-\frac{\xi}{\sqrt{D}}\right] \to 0.$$

where $\xi > 0$ and D depends on the classes statistics.

Different covariance matrices [ElKhalil, Kammoun, Couillet, 2017]

$$R - \frac{1}{2}\Phi\left[-\frac{\xi}{\sqrt{D_0}} + \frac{\beta}{\sqrt{D_0}}\right] - \frac{1}{2}\Phi\left[-\frac{\xi}{\sqrt{D_1}} - \frac{\beta}{\sqrt{D_1}}\right] \to 0.$$

with ξ, D_0 and D_1 are positive.

Takeaways

- Different misclassification rate across classes,
- The enhancement in the misclassification rate in one class is likely to be lost by the increase in the mis-classification rate of the other class.

$\to$ LDA does not leverage well the information about the class differences.

Outline

Regularized quadratic discriminant analysis

Regularized Quadratic Discriminant Analysis: R-QDA

- Applies for $c \in (0, \infty)$.
- Uses a regularized estimation of the inverse of the covariance matrix associated with each class

$$\hat{\mathbf{\Sigma}}_0(\gamma) = \left(\gamma \hat{\Sigma}_0 + \mathbf{I}_p\right)$$
$$\hat{\mathbf{\Sigma}}_1(\gamma) = \left(\gamma \hat{\Sigma}_1 + \mathbf{I}_p\right)$$

- The Discriminant score for the R-QDA is:

$$\hat{W}^{R-QDA} = -\frac{1}{2}\log\frac{|\hat{\mathbf{\Sigma}}_0(\gamma)|}{|\hat{\mathbf{\Sigma}}_1(\gamma)|} - \frac{1}{2}(\mathbf{x}-\mathbf{x}_0)^T\hat{\mathbf{\Sigma}}_0^{-1}(\gamma)(\mathbf{x}-\mathbf{x}_0) + \frac{1}{2}(\mathbf{x}-\mathbf{x}_1)^T\hat{\mathbf{\Sigma}}_1^{-1}(\gamma)(\mathbf{x}-\overline{\mathbf{x}}_1)$$

Assumptions

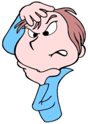
- $p, n_1, n_2 \to \infty$ with $\frac{p}{n} \to c\,(0,\infty)$, $\frac{n_1}{n} = \frac{n_2}{n} \to 0.5$
- The difference in means $\boldsymbol{\mu} = \boldsymbol{\mu}_0 - \boldsymbol{\mu}_1$ satisfies $\|\boldsymbol{\mu}\|^2 = O(\sqrt{p})$.
- The spectral norms of $\boldsymbol{\Sigma}_0$ and $\boldsymbol{\Sigma}_1$ are bounded
- Matrix $\boldsymbol{\Sigma}_0 - \boldsymbol{\Sigma}_1$ has at most $O(\sqrt{p})$ eigenvalues of order 1.

Khalil EL Khalil, Kammoun, Couillet 2017

Under these assumptions, the misclassification error rate converges to:

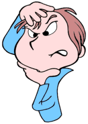
$$R_{QDA} - \frac{1}{2}\Phi\left(\frac{\overline{\xi}_0 - \overline{b}_0}{\sqrt{2\overline{B}_0}}\right) - \frac{1}{2}\Phi\left(-\frac{\overline{\xi}_1 - \overline{b}_1}{\sqrt{2\overline{B}_1}}\right) \to 0.$$

where for $i \in \{0,1\}$, $\overline{\xi}_i$, $\overline{b}_i$, $\overline{B}_i$ depends on the classes' statistics.



What happens if

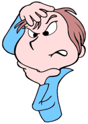
Q $\|\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1\| = O(1)$.



What happens if

Q $\|\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1\| = O(1)$.

Response The difference in means will not be asymptotically used by R-QDA. Only the information about the covariance matrices is leveraged.

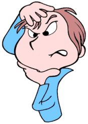


What happens if

- **Q** $\|\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1\| = O(1)$.

  Response The difference in means will not be asymptotically used by R-QDA. Only the information about the covariance matrices is leveraged.

- **Q** Matrix $\boldsymbol{\Sigma}_0 - \boldsymbol{\Sigma}_1$ has more than $O(\sqrt{p})$ eigenvalues of order 1



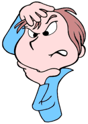
What happens if

- **Q** $\|\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1\| = O(1)$.

  Response The difference in means will not be asymptotically used by R-QDA. Only the information about the covariance matrices is leveraged.

- **Q** Matrix $\boldsymbol{\Sigma}_0 - \boldsymbol{\Sigma}_1$ has more than $O(\sqrt{p})$ eigenvalues of order 1

  Response R-QDA will perform asymptotically perfect classification.



What happens if

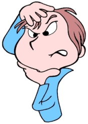
Q $\|\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1\| = O(1)$.

Response The difference in means will not be asymptotically used by R-QDA. Only the information about the covariance matrices is leveraged.

Q Matrix $\boldsymbol{\Sigma}_0 - \boldsymbol{\Sigma}_1$ has more than $O(\sqrt{p})$ eigenvalues of order 1

Response R-QDA will perform asymptotically perfect classification.

Q $\|\boldsymbol{\Sigma}_0 - \boldsymbol{\Sigma}_1\| = o(1)$ and $\|\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1\| = O(1)$



What happens if

Q $\|\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1\| = O(1)$.

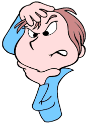
Response The difference in means will not be asymptotically used by R-QDA. Only the information about the covariance matrices is leveraged.

Q Matrix $\boldsymbol{\Sigma}_0 - \boldsymbol{\Sigma}_1$ has more than $O(\sqrt{p})$ eigenvalues of order 1

Response R-QDA will perform asymptotically perfect classification.

Q $\|\boldsymbol{\Sigma}_0 - \boldsymbol{\Sigma}_1\| = o(1)$ and $\|\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1\| = O(1)$

ResponseThe misclassification rate of R-QDA will converge to 0.5



What happens if

- **Q** $\|\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1\| = O(1)$.

  Response The difference in means will not be asymptotically used by R-QDA. Only the information about the covariance matrices is leveraged.

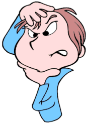
- **Q** Matrix $\boldsymbol{\Sigma}_0 - \boldsymbol{\Sigma}_1$ has more than $O(\sqrt{p})$ eigenvalues of order 1

  Response R-QDA will perform asymptotically perfect classification.

- **Q** $\|\boldsymbol{\Sigma}_0 - \boldsymbol{\Sigma}_1\| = o(1)$ and $\|\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1\| = O(1)$

  ResponseThe misclassification rate of R-QDA will converge to 0.5

- **Q** $\|\boldsymbol{\Sigma}_0 - \boldsymbol{\Sigma}_1\|_F = O(1)$ and $\|\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1\| = O(1)$



What happens if

- **Q** $\|\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1\| = O(1)$.

  Response The difference in means will not be asymptotically used by R-QDA. Only the information about the covariance matrices is leveraged.

- **Q** Matrix $\boldsymbol{\Sigma}_0 - \boldsymbol{\Sigma}_1$ has more than $O(\sqrt{p})$ eigenvalues of order 1

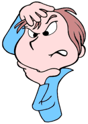
  Response R-QDA will perform asymptotically perfect classification.

- **Q** $\|\boldsymbol{\Sigma}_0 - \boldsymbol{\Sigma}_1\| = o(1)$ and $\|\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1\| = O(1)$

  ResponseThe misclassification rate of R-QDA will converge to 0.5

- **Q** $\|\boldsymbol{\Sigma}_0 - \boldsymbol{\Sigma}_1\|_F = O(1)$ and $\|\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1\| = O(1)$

  Response The misclassification rate of R-QDA will converge to 0.5



What happens if

- **Q** $\|\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1\| = O(1)$.

  Response The difference in means will not be asymptotically used by R-QDA. Only the information about the covariance matrices is leveraged.

- **Q** Matrix $\boldsymbol{\Sigma}_0 - \boldsymbol{\Sigma}_1$ has more than $O(\sqrt{p})$ eigenvalues of order 1

  Response R-QDA will perform asymptotically perfect classification.

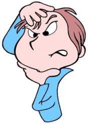
- **Q** $\|\boldsymbol{\Sigma}_0 - \boldsymbol{\Sigma}_1\| = o(1)$ and $\|\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1\| = O(1)$

  ResponseThe misclassification rate of R-QDA will converge to 0.5

- **Q** $\|\boldsymbol{\Sigma}_0 - \boldsymbol{\Sigma}_1\|_F = O(1)$ and $\|\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1\| = O(1)$

  Response The misclassification rate of R-QDA will converge to 0.5

- **Q** Unbalanced training: $\frac{n_0}{n_1}$ does not converge to 1



What happens if

- **Q** $\|\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1\| = O(1)$.

  Response The difference in means will not be asymptotically used by R-QDA. Only the information about the covariance matrices is leveraged.

- **Q** Matrix $\boldsymbol{\Sigma}_0 - \boldsymbol{\Sigma}_1$ has more than $O(\sqrt{p})$ eigenvalues of order 1

  Response R-QDA will perform asymptotically perfect classification.

- **Q** $\|\boldsymbol{\Sigma}_0 - \boldsymbol{\Sigma}_1\| = o(1)$ and $\|\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1\| = O(1)$

  ResponseThe misclassification rate of R-QDA will converge to 0.5

- **Q** $\|\boldsymbol{\Sigma}_0 - \boldsymbol{\Sigma}_1\|_F = O(1)$ and $\|\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1\| = O(1)$

  Response The misclassification rate of R-QDA will converge to 0.5

- **Q** Unbalanced training: $\frac{n_0}{n_1}$ does not converge to 1

  Response R-QDA will be equivalent to the classifier that assigns all observations to the class with the highest number of training samples.

of the study

R-LDA	When
	• $\|\mathbf{\Sigma}_0 - \mathbf{\Sigma}_1\| = o(1)$ and $\|\mu_0 - \mu_1\| = O(1)$ • $\|\mu_0 - \mu_1\| = O(p^\alpha)$
	• $\|\mathbf{\Sigma}_0 - \mathbf{\Sigma}_1\| = O(1)$ and $\|\mu_0 - \mu_1\| = o(1)$

of the study

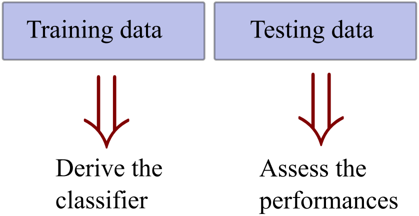
R-LDA	When
	• $\|\mathbf{\Sigma}_0 - \mathbf{\Sigma}_1\| = o(1)$ and $\|\mu_0 - \mu_1\| = O(1)$ • $\|\mu_0 - \mu_1\| = O(p^\alpha)$
	• $\|\mathbf{\Sigma}_0 - \mathbf{\Sigma}_1\| = O(1)$ and $\|\mu_0 - \mu_1\| = o(1)$

R-QDA	When
	• $\mathbf{\Sigma}_0 - \mathbf{\Sigma}_1$ has "rank" scaling at least with rate $O(\sqrt{p})$
	• $\|\mathbf{\Sigma}_0 - \mathbf{\Sigma}_1\| = o(1)$ • $\|\mathbf{\Sigma}_0 - \mathbf{\Sigma}_1\|_F = O(1)$ • Unbalanced training

R-QDA

- is prone to estimation errors due to insufficiency in the number of observations,
- The setting of the regularization parameter is very important

Evaluation of the performances



Model selection Given a set of candidate regularization factors

- Evaluate the performance using the test data for each regularization value
- Select the value that presents the lowest mis-classification rate

R-QDA Proposed method

- Provide a consistent estimator for the misclassification error rate.
- Select the regularization factor that minimizes the estimated misclassification error rate.

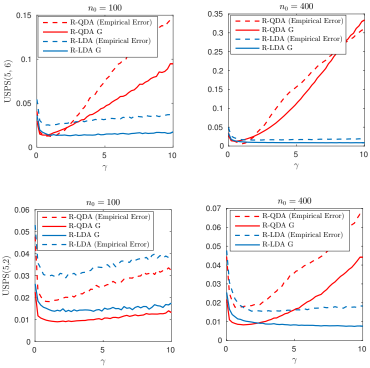


Figure: Misclassification error rate for R-QDA with respect to the regularization factor γ. The data are drawn from USPS data sets.

Outline

Rising to the high dimensionality challenge

Drawbacks of the insufficiency in the number of observations

- Instability due to the ill-conditioning of the precision matrix (LDA-QDA)
- High noise in the estimation of the covariance matrix

Solutions

- Use a regularization parameter that shrinks the covariance matrix towards identity
- Employ a dimensionality reduction method prior to classification
  - Random projection
  - PCA

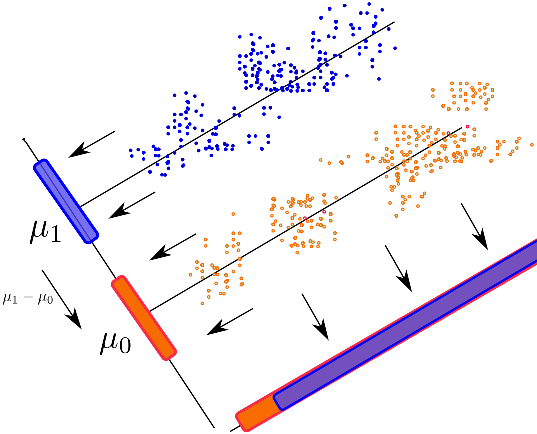


Figure: Illustration of the choice of the discriminative direction

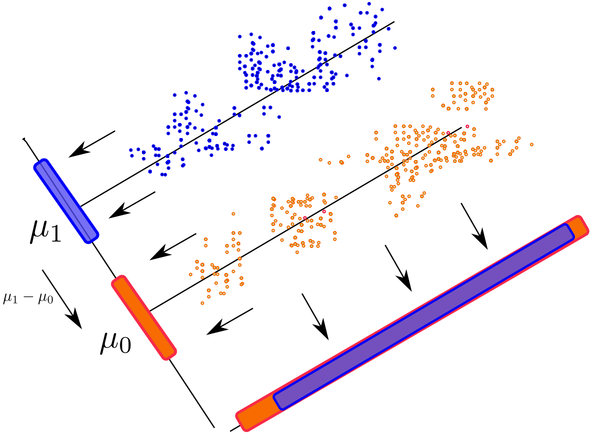


Figure: Illustration of the choice of the discriminative direction

Drawbacks

- The direction that contains the most variance is not always optimal from the classification point of view.

Gaussian mixture problem

- $\mathbf{x}_1, \cdots, \mathbf{x}_n \in \mathbb{R}^p$, i.i.d.,
- 2 classes with the same number of observations in each class
- $\mathbf{x}_i$ in class j take the form:

$$\mathbf{x}_i = \left(\sigma^2 \mathbf{I}_p + \mathbf{P}\right)^{\frac{1}{2}} \mathbf{z} + \boldsymbol{\mu}_j$$

where $\mathbf{x} \sim N(\mathbf{0}, \mathbf{I}_p)$, $\mathbf{P}$ has finite rank r with distinct eigenvalues $\omega_1, \cdots, \omega_r$.

- Let $\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_n]$ Form the covariance matrix

$$\hat{\mathbf{C}} = \frac{1}{n} \mathbf{X} \left(\mathbf{I} - \frac{\mathbf{1}\mathbf{1}^T}{n} \right) \mathbf{X}^T$$

Subspace LDA

- Select the k principal eigenvectors of $\hat{\mathbf{C}}$ with $k \ll n$, $\hat{\mathbf{u}}_1, \cdots, \hat{\mathbf{u}}_k$
- Let $\hat{\mathbf{U}} = [\mathbf{u}_1, \cdots, \mathbf{u}_k]$. Project observations on the subspace spanned by $\hat{\mathbf{U}}$

$$\hat{\mathbf{U}}^T \mathbf{x}_1, \cdots, \mathbf{U}^T \hat{\mathbf{x}}_n$$

- Let $\mathbf{x}$ a test observation. Perform LDA on the projection of $\mathbf{x}$ onto the subspace of $\hat{\mathbf{U}}$.

Questions

- How to choose the directions in $\hat{\mathbf{U}}$?
- What is the optimal number k?
- How PCA-LDA compare with R-LDA?

Methodology

- Accurate eigenvalue analysis of the covariance matrix $\hat{\mathbf{C}}$.
- Two level of perturbations:
  - Additive perturbation caused by the shift in the mean vector
  - Multiplicative perturbation carried by matrix $\mathbf{P}$.

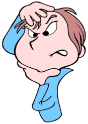
Initial results

- Worst case: Assume $(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^T \mathbf{P} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0) \to 0$.
- At most $r+1$ spikes, only one of them will have a non-vanishing alignment with $(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)$

Takeaways

- In some situations, only one dimension is relevant

$$\boxed{\text{The corresponding eigenvalue does not always lie at the edge of the spectrum}}$$



What happens if

Q If $(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^T \mathbf{P} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)$ does not go to zero

Response All the spikes matter as they will present a non vanishing alignment with $(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)$

Q If only some of the eigenvectors of $\mathbf{P}$ are aligned to $(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)$

Response Only their corresponding spikes matter + an additional spike caused by the mean perturbation

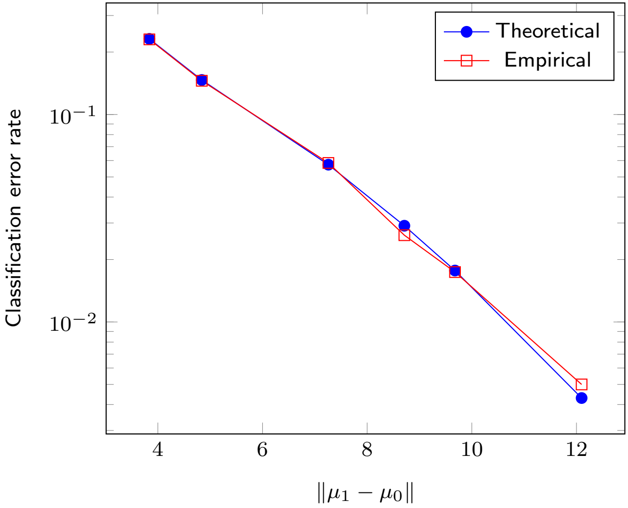


Figure: Classification error rate with respect to $\|\mu_1 - \mu_0\|$, with rank $\mathbf{P} = 4$; It has 4 non-zero eigenvalues equal to 5, 4, 2.6 and 4.7, Moreover $\mu_1 - \mu_0$ it not orthogonal to any of its eigenvectors; $p = 5000$ and $n = 12000$

Conclusion

- Random matrix theory is a powerful tool that has been applied with success to the fields wireless communications and signal processing, providing solutions to very challengning problems
- High dimensionality along with stochasticity are the sole prerequisite of this tool
- Encounter between random matrix theory and machine learning will bring about many new theoretical problems