Improved LDA Classifier based on Spiked Models

Houssem Sifaou , Abla Kammoun and Mohamed-Slim Alouini

KAUST

- Linear discriminant analysis (LDA) classifier: originally proposed by R. A. Fisher.
- Applied in several fields: taxonomic problems, detection, face and speech recognition, cancer genomics..
- Optimal under the assumption that data follow a Gaussian mixture model with a common covariance matrix.
- In the case of different covariance matrices among classes, Quadratic Discriminant Analysis (QDA) becomes the optimal classifier.
- Unknown statistics: mean vectors, population covariance matrix.
- Widely adopted approach: empirical estimates.
- Empirical estimates: inaccurate in high dimensional settings, where the sample size is of the same order as the number of samples.
- Proposed solution: improved LDA classifier based on spiked model suited for high dimensional settings.

LDA and R-LDA Classifiers

- $\mathbf{x}_1, \cdots, \mathbf{x}_n$ in $\mathbb{R}^{p\times 1}$ belonging to two classes $\mathcal{C}_0$ and $\mathcal{C}_1$

$$\mathbf{x}_i \in \mathcal{C}_i \Leftrightarrow \mathbf{x}_i \sim \mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$$

  - $\boldsymbol{\mu}_i$: mean of class i
  - $\boldsymbol{\Sigma}$: covariance matrix common to both classes.
- LDA classifier:

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

  - $\overline{\mathbf{x}}_i = \frac{1}{n_i}\sum_{\ell\in\mathcal{T}_i}\mathbf{x}_\ell$
  - $\hat{\boldsymbol{\Sigma}} = \frac{(n_0-1)\hat{\boldsymbol{\Sigma}}_0+(n_1-1)\hat{\boldsymbol{\Sigma}}_1}{n-2}$, with $\hat{\boldsymbol{\Sigma}}_i = \frac{1}{n_i-1}\sum_{\ell\in\mathcal{T}_i}(\mathbf{x}_\ell - \overline{\mathbf{x}}_i)(\mathbf{x}_\ell - \overline{\mathbf{x}}_i)^T, \quad i = 0, 1$
- R-LDA classifier:

$$W^{\text{R-LDA}}(\mathbf{x}) = \left(\mathbf{x} - \frac{\overline{\mathbf{x}}_0+\overline{\mathbf{x}}_1}{2}\right)^T \mathbf{H}(\overline{\mathbf{x}}_0 - \overline{\mathbf{x}}_1) - \log\frac{\pi_1}{\pi_0},$$

$$\mathbf{H} = \left(\mathbf{I}_p + \gamma\hat{\boldsymbol{\Sigma}}\right)^{-1}, \quad \gamma > 0.$$

How to set the regularization parameter γ?

- Classical approach: grid search
- Recent approach: set the regularization parameter to the value that minimizes the asymptotic classification error.

Tow major drawbacks of the latter approach:

- The estimation of the optimal regularization parameter is computationally expensive (grid search)
- It does not take advantage from available prior information on the structure of the covariance matrix

Improved LDA Classifier

- Assumption: spiked model

$$\boldsymbol{\Sigma} = \sigma^2\mathbf{I}_p + \sigma^2\sum_{j=1}^{r}\lambda_j\mathbf{u}_j\mathbf{u}_j^T,$$

  - $\sigma^2 > 0$
  - $\lambda_1 \geq \cdots \geq \lambda_r > 0$
  - $\mathbf{u}_1, \cdots, \mathbf{u}_r$ are orthonormal
- Several real applications
  - Detection
  - EEG signals (EEG: electroencephalogram)
  - Financial econometrics
- Observation:

$$\boldsymbol{\Sigma}^{-1} = \frac{1}{\sigma^2}\left[\mathbf{I}_p - \sum_{j=1}^{r}\frac{\lambda_j}{1+\lambda_j}\mathbf{u}_j\mathbf{u}_j^T\right]$$

- Proposed estimator:

$$\hat{\mathbf{C}}^{-1} = \frac{1}{\sigma^2}\left[\mathbf{I}_p + \sum_{j=1}^{r}w_j\mathbf{v}_j\mathbf{v}_j^T\right]$$

  - $\mathbf{v}_1, \cdots, \mathbf{v}_r$: eigenvectors associated with the largest eigenvalues of $\hat{\boldsymbol{\Sigma}}$
  - $\{w_j\}_{j=1}^r$: some design parameters to be optimized
- Discriminant score:

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

- Misclassification rate associated with class i:

$$\epsilon_i^{\text{Imp-LDA}} = \boldsymbol{\Phi}\left(\frac{(-1)^{i+1}G(\boldsymbol{\mu}_i, \overline{\mathbf{x}}_0, \overline{\mathbf{x}}_1, \hat{\mathbf{C}}) + (-1)^i\log\frac{\pi_1}{\pi_0}}{\sqrt{D(\boldsymbol{\mu}_i, \overline{\mathbf{x}}_0, \overline{\mathbf{x}}_1, \hat{\mathbf{C}}, \boldsymbol{\Sigma})}}\right)$$

  - $G(\boldsymbol{\mu}_i, \overline{\mathbf{x}}_0, \overline{\mathbf{x}}_1, \hat{\mathbf{C}}) = \left(\boldsymbol{\mu}_i - \frac{\overline{\mathbf{x}}_0+\overline{\mathbf{x}}_1}{2}\right)\hat{\mathbf{C}}^{-1}(\overline{\mathbf{x}}_0 - \overline{\mathbf{x}}_1)$
  - $D(\boldsymbol{\mu}_i, \overline{\mathbf{x}}_0, \overline{\mathbf{x}}_1, \hat{\mathbf{C}}, \boldsymbol{\Sigma}) = (\overline{\mathbf{x}}_0 - \overline{\mathbf{x}}_1)^T\hat{\mathbf{C}}^{-1}\boldsymbol{\Sigma}\hat{\mathbf{C}}^{-1}(\overline{\mathbf{x}}_0 - \overline{\mathbf{x}}_1)$
- Total classification error rate:

$$\epsilon^{\text{Imp-LDA}} = \pi_0\epsilon_0^{\text{Imp-LDA}} + \pi_1\epsilon_1^{\text{Imp-LDA}}.$$

- The parameters w_j are optimized so that they minimize the total classification error rate:

$$\mathbf{w}^\star = \underset{\mathbf{w}}{\text{argmin}} \quad \mathbb{E}\left[\epsilon^{\text{Imp-LDA}}\right]$$

  where $\mathbf{w} = [w_1, \cdots, w_r]^T$.
- Finding the optimal $\mathbf{w}$ could not be achieved in practice since the misclassification rate involves unknown quantities. To get around this problem, we invoke results from random matrix theory that approximate the total misclassification error rate.

Asymptotic Analysis

Asymptotic regime :$n, p \to \infty$ with fixed ration $c = p/n$

$$\epsilon^{\text{Imp-LDA}} - \bar{\epsilon}^{\text{Imp-LDA}} \xrightarrow{\text{a.s.}} 0$$

- $\bar{\epsilon}^{\text{Imp-LDA}} = \pi_0\boldsymbol{\Phi}\left(\frac{-\alpha(\overline{G}(\mathbf{w})-\eta)}{\sqrt{\overline{D}(\mathbf{w})+\omega}}\right) + \pi_1\boldsymbol{\Phi}\left(\frac{-\alpha(\overline{G}(\mathbf{w})+\eta)}{\sqrt{\overline{D}(\mathbf{w})+\omega}}\right)$
  - $\alpha = \frac{\|\boldsymbol{\mu}\|}{2\sigma}$, $\eta = \frac{\sigma^2}{\|\boldsymbol{\mu}\|^2}[\frac{c}{\pi_0} - \frac{c}{\pi_1} + 2\log\frac{\pi_1}{\pi_0}]$, $\omega = \frac{\sigma^2}{\|\boldsymbol{\mu}\|^2}[\frac{p}{n_0} + \frac{p}{n_1}]$
  - $\overline{G}(\mathbf{w}) = 1 + \sum_{j=1}^r a_jb_jw_j$
  - $\overline{D}(\mathbf{w}) = 1 + \sum_{j=1}^r[\lambda_jb_j + 2a_jb_j(\lambda_j+1)w_j] + \sum_{j=1}^r\left[a_jb_j(1+\lambda_ja_j)w_j^2\right]$
  - $\boldsymbol{\mu} = \boldsymbol{\mu}_0 - \boldsymbol{\mu}_1$, $a_j = \frac{1-c/\lambda_j^2}{1+c/\lambda_j}$, $b_j = \frac{\boldsymbol{\mu}^T\mathbf{v}_j\mathbf{v}_j^T\boldsymbol{\mu}}{\|\boldsymbol{\mu}\|^2}$, $j = 1, \cdots, r$

Optimal Parameters

- Optimal parameters : $w_j^\star = \frac{u^\star\gamma_j}{\beta\,\alpha_j} - \beta_j, \quad j = 1, \cdots, r$
  - $\alpha_j = \lambda_ja_j^2b_j + a_jb_j$, $\beta_j = \frac{\lambda_j+1}{\lambda_ja_j+1}$, $\gamma_j = a_jb_j$, $j = 1, ..., r,$,
  - $\beta = \sqrt{\sum_{j=1}^r\gamma_j^2/\alpha_j}$
  - $u^\star$ is the unique positive solution of the following equation: $h(u) = 0$,

$$h(u) = \pi_0(\beta b - d_0u)e^{\frac{-\alpha^2}{2}\frac{(\beta u+d_0)^2}{u^2+b}} + \pi_1(\beta b - d_1u)e^{\frac{-\alpha^2}{2}\frac{(\beta u+d_1)^2}{u^2+b}},$$

  - $b = 1 + \omega + \sum_{j=1}^r\left[\lambda_jb_j - \frac{(\lambda_ja_jb_j+a_jb_j)^2}{\lambda_ja_j^2b_j+a_jb_j}\right]$
  - $d_0 = 1 - \eta - \sum_{j=1}^r\frac{\lambda_ja_jb_j+a_jb_j}{\lambda_ja_j+1}$
  - $d_1 = 1 + \eta - \sum_{j=1}^r\frac{\lambda_ja_jb_j+a_jb_j}{\lambda_ja_j+1}$

- The determination of the design parameters w_j involve finding the unique positive zero $u^\star$ of function.
- This can be numerically obtained using, for instance, a bisection algorithm with a few number of iterations since tight upper and lower bounds of $u^\star$ can be easily obtained.
- In the case of equiprobable classes, $u^\star$ can be computed in closed form expression.

- The optimal design parameters depend on unobservable quantities λ_j and b_j
- Consistent estimators for these quantities need to be retrieved.
- Under asymptotic regime assumption ($n, p \to \infty$ with fixed ration $c = p/n$)

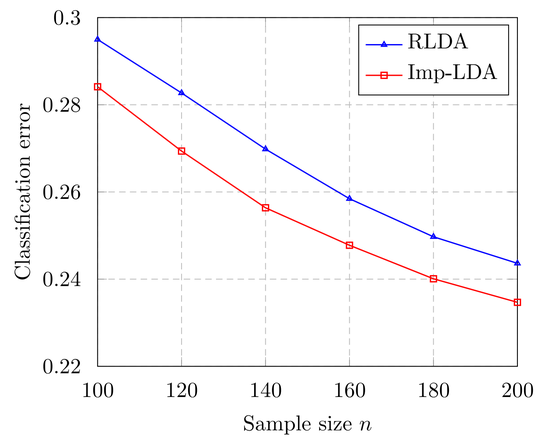
$$|\lambda_j - \hat{\lambda}_j| \xrightarrow{\text{a.s.}} 0, \quad \text{and} \quad |b_j - \hat{b}_j| \xrightarrow{\text{a.s.}} 0,$$

  - $\hat{\lambda}_j = \frac{s_j/\sigma^2+1-c+\sqrt{(s_j/\sigma^2+1-c)^2-4s_j/\sigma^2}}{2}$,
  - $\hat{b}_j = \frac{1+c/\hat{\lambda}_j}{1-c/\hat{\lambda}_j^2}\frac{\hat{\boldsymbol{\mu}}^T\mathbf{v}_j\mathbf{v}_j^T\hat{\boldsymbol{\mu}}}{\|\hat{\boldsymbol{\mu}}\|^2}$, $j = 1, \cdots, r$
  - $\hat{\boldsymbol{\mu}} = \overline{\mathbf{x}}_0 - \overline{\mathbf{x}}_1$
  - s_j: j-th largest eigenvalue of the pooled covariance matrix

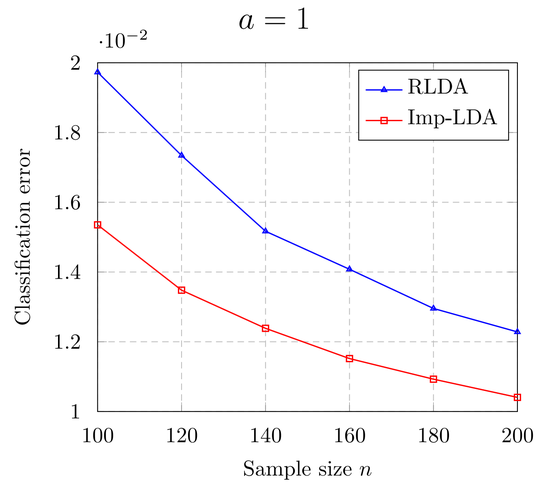
Numerical Results

Synthetic data

- $\sigma^2 = 1$ and $r = 3$
- $\mathbf{u}_1 = [1, 0, \cdots, 0]^T$, $\mathbf{u}_2 = [0, 1, 0, \cdots, 0]^T$, $\mathbf{u}_3 = [0, 0, 1, 0\cdots, 0]^T$
- $\lambda_1 = 4$, $\lambda_2 = 3$, $\lambda_3 = 2$
- $\boldsymbol{\mu}_0 = \frac{1}{\sqrt{p}}[a, a, \cdots, a]^T$ and $\boldsymbol{\mu}_1 = -\boldsymbol{\mu}_0$ where a is a finite constant ($a = 1$ and $a = 2.5$ in our simulation)



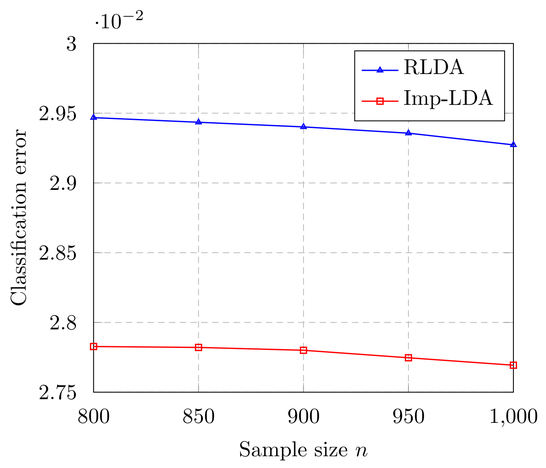
$a = 1$



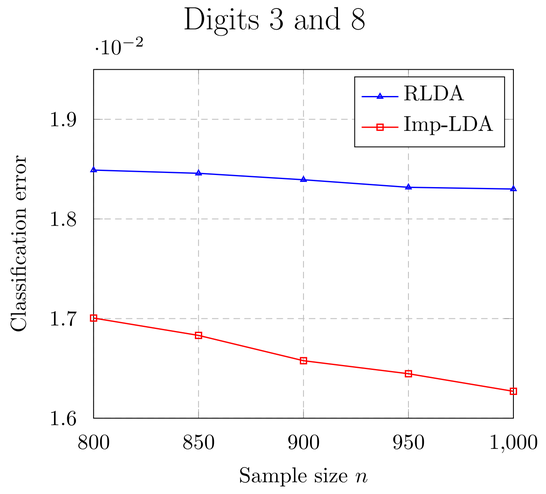
$a = 2.5$

Real data

- "USPS" dataset (standard dataset for handwritten digit recognition)



Digits 3 and 8



Digits 4 and 9

- Improved LDA classifier for spiked models.
- Lower complexity than R-LDA classifier.
- Better classification performance for both synthetic and real data.
- Apply the same approach to the quadratic discriminant analysis (QDA) classifier.
- More involved spiked models may be considered in which the population covariance matrix is low-rank perturbation of a diagonal matrix.