

# Improved Discriminant Analysis Classifiers based on Spiked Model

Houssein Sifaou, Abba 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

- $\mathbf{x}_1, \dots, \mathbf{x}_n$  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:

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

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

- R-LDA classifier:

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

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

### How to set the regularization parameter $\gamma$ ?

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

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

- ◊  $\sigma^2 > 0$
- ◊  $\lambda_1 \geq \dots \geq \lambda_r > 0$
- ◊  $\mathbf{v}_1, \dots, \mathbf{v}_r$  are orthonormal

- Several real applications

- ◊ Detection
- ◊ EEG signals (EEG: electroencephalogram)
- ◊ Financial econometrics

- Observation:

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

- Proposed estimator:

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

- ◊  $\mathbf{u}_1, \dots, \mathbf{u}_r$ : eigenvectors associated with the largest eigenvalues of  $\hat{\boldsymbol{\Sigma}}$

- ◊  $\{w_j\}_{j=1}^r$ : some design parameters to be optimized

- Discriminant score:

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

- Misclassification rate associated with class  $i$ :

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

- ◊  $G(\boldsymbol{\mu}_i, \bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \hat{\mathbf{C}}) = \left( \boldsymbol{\mu}_i - \frac{\bar{\mathbf{x}}_0 + \bar{\mathbf{x}}_1}{2} \right)^T \hat{\mathbf{C}}^{-1} (\bar{\mathbf{x}}_0 - \bar{\mathbf{x}}_1)$

- ◊  $D(\boldsymbol{\mu}_i, \bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \hat{\mathbf{C}}, \boldsymbol{\Sigma}) = (\bar{\mathbf{x}}_0 - \bar{\mathbf{x}}_1)^T \hat{\mathbf{C}}^{-1} \boldsymbol{\Sigma} \hat{\mathbf{C}}^{-1} (\bar{\mathbf{x}}_0 - \bar{\mathbf{x}}_1)$

- 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:

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

where  $\mathbf{w} = [w_1, \dots, w_r]^T$ .

- 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 :  $n, p \rightarrow \infty$  with fixed ration  $c = p/n$

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

$$\bullet \epsilon^{\text{I-LDA}} = \pi_0 \Phi \left[ \frac{-\sqrt{\alpha}(\bar{G}(\mathbf{w}) - \eta)}{2\sqrt{\bar{D}(\mathbf{w}) + \kappa}} \right] + \pi_1 \Phi \left[ \frac{-\sqrt{\alpha}(\bar{G}(\mathbf{w}) + \eta)}{2\sqrt{\bar{D}(\mathbf{w}) + \kappa}} \right]$$

- ◊  $\alpha = \frac{\|\boldsymbol{\mu}\|^2}{\sigma^2}$ ,  $\eta = \frac{1}{\sigma} [c/\pi_0 - c/\pi_1 + 2 \log \frac{\pi_1}{\pi_0}]$ ,  $\kappa = \frac{1}{\sigma} [p/n_0 + p/n_1]$
- ◊  $\bar{G}(\mathbf{w}) = 1 + \sum_{j=1}^r a_j b_j w_j$ ,
- ◊  $\bar{D}(\mathbf{w}) = 1 + \sum_{j=1}^r [\lambda_j b_j + 2a_j b_j (\lambda_j + 1) w_j] + \sum_{j=1}^r [a_j b_j (1 + \lambda_j a_j) w_j^2]$
- ◊  $\boldsymbol{\mu} = \boldsymbol{\mu}_0 - \boldsymbol{\mu}_1$ ,  $a_j = \frac{1 - c/\lambda_j^2}{1 + c/\lambda_j}$ ,  $b_j = \frac{\boldsymbol{\mu}^T \mathbf{v}_j \mathbf{v}_j^T \boldsymbol{\mu}}{\|\boldsymbol{\mu}\|^2}$ ,  $j = 1, \dots, r$

## Optimal Parameters

- Optimal parameters :  $w_j^* = \frac{u_j^* \gamma_j}{\beta \alpha_j} - \beta_j$ ,  $j = 1, \dots, r$

- ◊  $\alpha_j = \lambda_j a_j^2 b_j + a_j b_j$ ,  $\beta_j = \frac{\lambda_j + 1}{\lambda_j a_j + 1}$ ,  $\gamma_j = a_j b_j$ ,  $j = 1, \dots, r$ ,
- ◊  $\beta = \sqrt{\sum_{j=1}^r \gamma_j^2 / \alpha_j}$
- ◊  $u^*$  is the minimizer of the scalar function  $\tilde{g}(u)$  given by

$$\tilde{g}(u) = \pi_0 \Phi \left( -\frac{\sqrt{\alpha} \beta u + d_0}{2 \sqrt{u^2 + b}} \right) + \pi_1 \Phi \left( -\frac{\sqrt{\alpha} \beta u + d_1}{2 \sqrt{u^2 + b}} \right),$$

$$\diamond b = 1 + \kappa + \sum_{j=1}^r \left[ \lambda_j b_j - \frac{(\lambda_j a_j b_j + a_j b_j)^2}{\lambda_j a_j^2 b_j + a_j b_j} \right]$$

$$\diamond d_0 = 1 - \eta - \sum_{j=1}^r \frac{\lambda_j a_j b_j + a_j b_j}{\lambda_j a_j + 1}$$

$$\diamond d_1 = 1 + \eta - \sum_{j=1}^r \frac{\lambda_j a_j b_j + a_j b_j}{\lambda_j a_j + 1}$$

## Optimal Bias

$$\hat{W}^{\text{OII-LDA}} = \left( \mathbf{x} - \frac{\bar{\mathbf{x}}_0 + \bar{\mathbf{x}}_1}{2} \right)^T \hat{\mathbf{C}}^{-1} (\bar{\mathbf{x}}_0 - \bar{\mathbf{x}}_1) + \theta,$$

where  $\theta$  is a parameter that will be optimized. The corresponding misclassification rate can be expressed in this case as,

$$\epsilon^{\text{OII-LDA}} = \pi_0 \Phi \left[ \frac{-G(\boldsymbol{\mu}_0, \bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \hat{\mathbf{C}}) - \theta}{\sqrt{D(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \hat{\mathbf{C}}, \boldsymbol{\Sigma})}} \right] + \pi_1 \Phi \left[ \frac{G(\boldsymbol{\mu}_1, \bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \hat{\mathbf{C}}) + \theta}{\sqrt{D(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \hat{\mathbf{C}}, \boldsymbol{\Sigma})}} \right],$$

Following similar steps, the asymptotic equivalent of  $\epsilon^{\text{OII-LDA}}$  can be obtained as for  $\epsilon^{\text{I-LDA}}$  and the corresponding optimal  $\theta$  is given by

$$\theta^* = \frac{1}{2} \left[ \frac{p}{n_0} - \frac{p}{n_1} \right] - \frac{\bar{D}(\mathbf{w}) + \kappa}{\bar{G}(\mathbf{w})} \log \frac{\pi_1}{\pi_0}.$$

The optimal parameter vector  $\mathbf{w}$  that minimizes  $\epsilon^{\text{OII-LDA}}$  is given by

$$w_j^* = \frac{b \gamma_j}{d \alpha_j} - \beta_j, \quad j = 1, \dots, r.$$

where  $d$  is given by

$$d = 1 - \sum_{j=1}^r \frac{\lambda_j a_j b_j + a_j b_j}{\lambda_j a_j + 1}.$$

Interestingly, the optimal parameters of OII-LDA are obtained in closed-form even in the case of unbalanced classes. Thus, in practice OII-LDA should be used instead I-LDA proposed in the previous section, since it yields better performance and the optimization of its parameters are less computationally demanding.

- The optimal design parameters depend on unobservable quantities  $\lambda_j$  and  $b_j$

- Consistent estimators for these quantities need to be retrieved.

- Under asymptotic regime assumption ( $n, p \rightarrow \infty$  with fixed ration  $c = p/n$ )

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

$$\diamond \hat{\alpha} = \frac{\|\hat{\boldsymbol{\mu}}\|^2}{\hat{\sigma}^2} - c_1 - c_0,$$

$$\diamond \hat{\lambda}_j = \frac{s_j / \hat{\sigma}^2 + 1 - c + \sqrt{(s_j / \hat{\sigma}^2 + 1 - c)^2 - 4s_j / \hat{\sigma}^2}}{2} - 1,$$

$$\diamond \hat{b}_j = \frac{1 + c/\hat{\lambda}_j}{1 - c/\hat{\lambda}_j^2} \frac{\hat{\boldsymbol{\mu}}^T \mathbf{u}_j \mathbf{u}_j^T \hat{\boldsymbol{\mu}}}{\|\hat{\boldsymbol{\mu}}\|^2 - c_1 \hat{\sigma}^2 - c_0 \hat{\sigma}^2}, \quad j = 1, \dots, r.$$

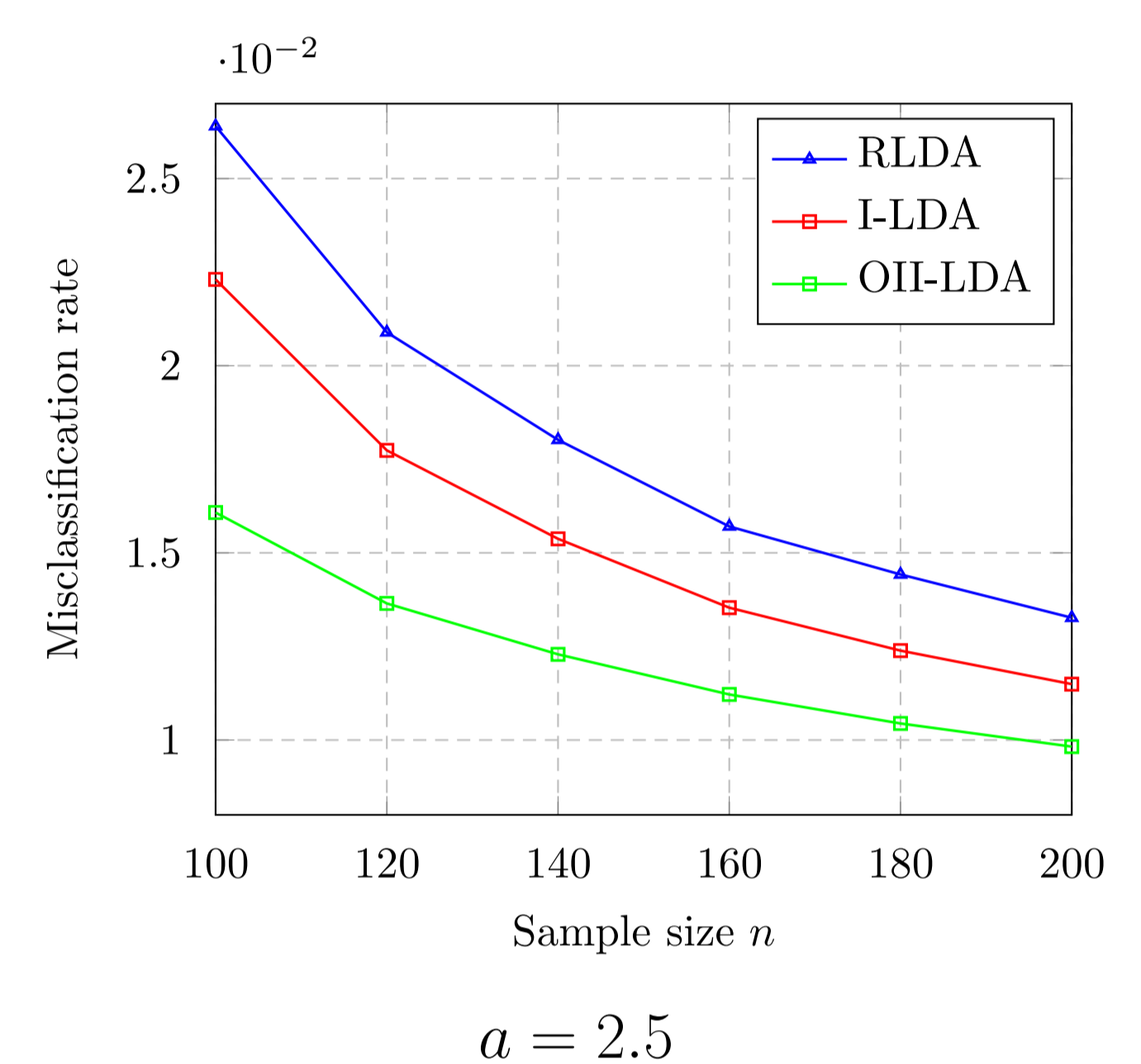
$$\diamond \hat{\boldsymbol{\mu}} = \bar{\mathbf{x}}_0 - \bar{\mathbf{x}}_1$$

$$\diamond s_j: j\text{-th largest eigenvalue of the pooled covariance matrix}$$

## Numerical Results

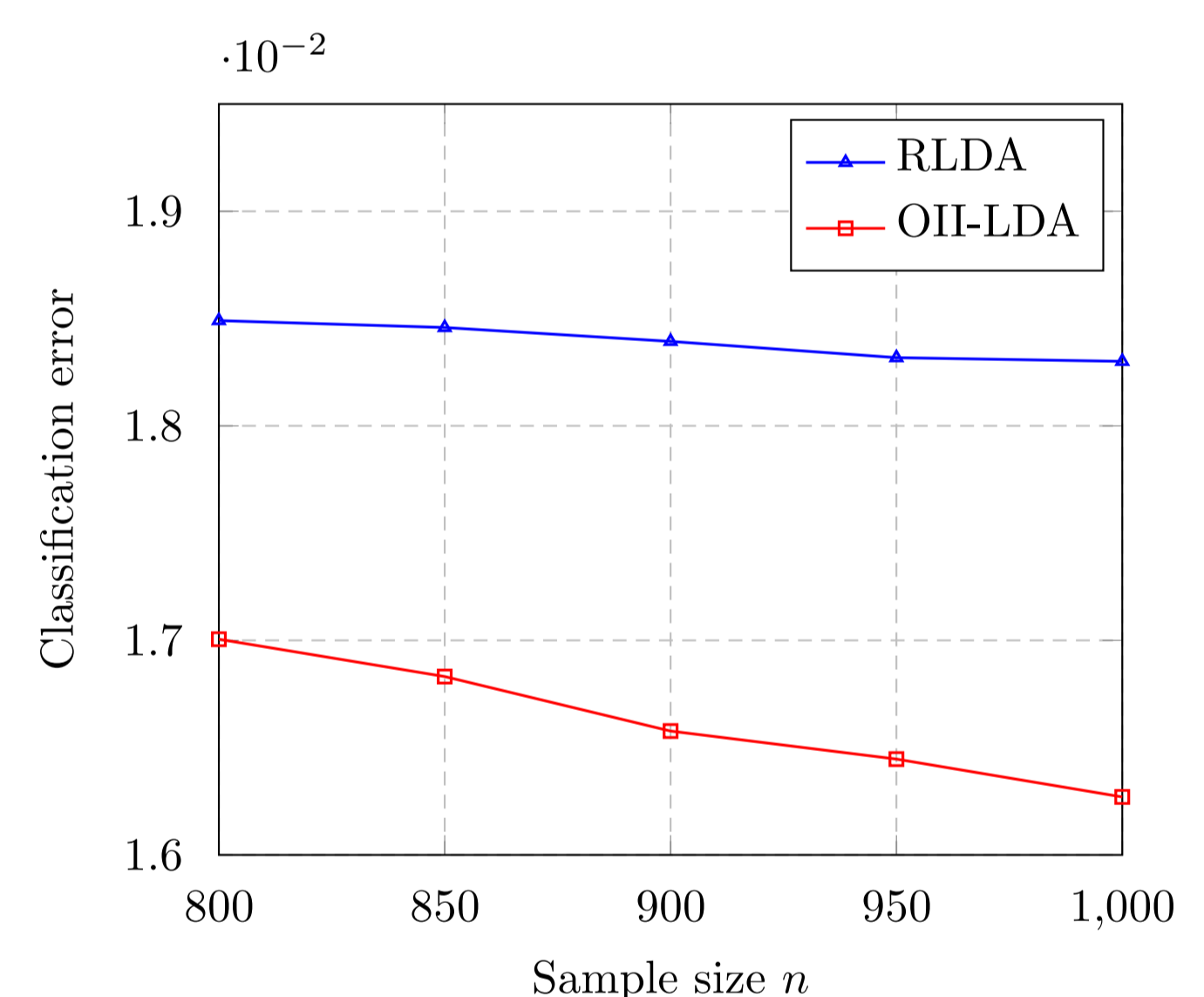
### Synthetic data

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



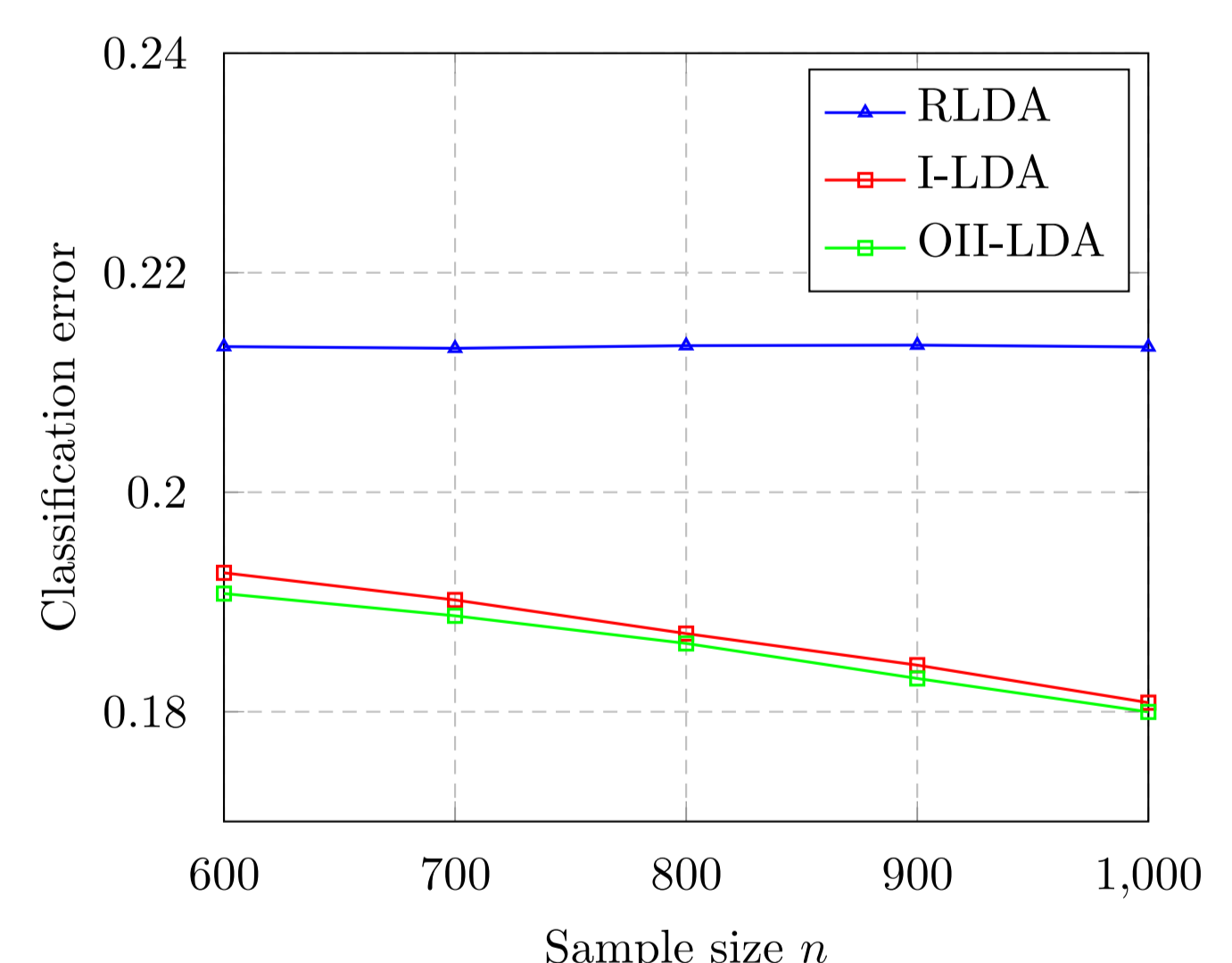
### Real data

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



### Digits 4 and 9

- Phoneme dataset (standard dataset for handwritten digit recognition)



### Phonemes 'aa' and 'ao'

Comparison between the proposed classifier, R-LDA, SVM, and KNN, using Phoneme dataset.

	$n = 800$	$n = 900$	$n = 1000$
OII-LDA	0.185	0.182	0.180
R-LDA	0.213	0.213	0.212
SVM	0.206	0.204	0.201
KNN	0.226	0.225	0.224

- Improved LDA classifier for spiked models.
- Lower complexity than R-LDA classifier.
- Better classification performance for both synthetic and real data.
- Same approach was applied 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.