Regularized Quadratic Discriminant Analysis for Large Dimensional Data

Khalil Elkhalil[1], Abla Kammoun[1], Romain Couillet[2], Tareq Y. Al-Naffouri[1] and Mohamed Slim Alouini[1]

[1] CEMSE Division, King Abdullah University of Science and Technology (KAUST), KSA
[2] CNRS-CentraleSupélec-Université Paris-Sud, France

KAUST

CentraleSupélec

Introduction

Motivation

- Closed form expressions for the misclassification probability exists in the case where the training size n is larger than the number of predictors p (the signal dimension) using some properties of Wishart distribution (Exact analysis)
- Most analysis is only valid for Gaussian data limiting the scope of application.
- Most of asymptotic approaches rely on the assumption that $p \gg n$ ans assume some sparsity on the class' statistics.

Contributions

- We consider the double asymptotic regime where both $p, n \to \infty$ with $p/n \to c \in (0, \infty)$.
- We make mild assumptions on the statisitcal means and the covariance matrices just to achieve non trivial error rates.
- We derive an asymptotic limit for the misclassification probability that reveals the mathematical connection between the classification error and the statistical parameters associated with each class.
- We leverage this result to propose a more efficient design of the regularized QDA classifier by selecting the regularization parameter that minimizes the asymptotic classification error.

Regularized QDA for Binary Classification

QDA

An observation $\mathbf{x} \in \mathbb{R}^{p\times p}$ belongs to class $\mathcal{C}_i$, $i = 0, 1$, if and only if

$$\mathbf{x} = \boldsymbol{\mu}_i + \Sigma_i^{\frac{1}{2}}\omega,$$

with $\omega \sim \mathcal{N}(\mathbf{0}_{p\times 1}, \mathbf{I}_p)$. The Bayes rule classifier

$$W_i^{QDA}(\mathbf{x}) = -\frac{1}{2}\log|\Sigma_i| - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^T \Sigma_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i) + \log\pi_i, \quad i \in \{0,1\}. \tag{1}$$

The classification rule is given by

$$\begin{cases} \mathbf{x} \in \mathcal{C}_0 & \text{if } W_0^{QDA}(\mathbf{x}) > W_1^{QDA}(\mathbf{x}) \\ \mathbf{x} \in \mathcal{C}_1 & \text{otherwise.} \end{cases} \tag{2}$$

Training

Statisitcs estimation for $i \in \{0,1\}$

$$\overline{\mathbf{x}}_i = \frac{1}{n_i}\sum_{l\in\mathcal{T}_i}\mathbf{x}_l,$$

$$\widehat{\Sigma}_i = \frac{1}{n_i-1}\sum_{l\in\mathcal{T}_i}(\mathbf{x}_l-\overline{\mathbf{x}}_i)(\mathbf{x}_l-\overline{\mathbf{x}}_i)^T.$$

The empirical discriminant analysis score becomes

$$\widehat{W}_i^{QDA}(\mathbf{x}) = -\frac{1}{2}\log|\widehat{\Sigma}_i| - \frac{1}{2}(\mathbf{x}-\overline{\mathbf{x}}_i)^T\widehat{\Sigma}_i^{-1}(\mathbf{x}-\overline{\mathbf{x}}_i) + \log\pi_i. \tag{3}$$

If $n_i < p$, we use a regularized estimate for the inverse covariance matrix

$$\mathbf{H}_i = \left(\mathbf{I}_p + \gamma\widehat{\Sigma}_i\right)^{-1}, \tag{4}$$

where $\gamma > 0$ is a regularizer.

$$\widehat{W}_i^{RQDA} = \frac{1}{2}\log|\mathbf{H}_i| - \frac{1}{2}(\mathbf{x}-\overline{\mathbf{x}}_i)^T\mathbf{H}_i(\mathbf{x}-\overline{\mathbf{x}}_i) + \log\pi_i. \tag{5}$$

The conditional classification error can easily be shown to write as

$$\epsilon_i = \mathbb{P}\left[\omega^T\mathbf{B}_i\omega + 2\omega^T\mathbf{y}_i < \xi_i\right], \textit{where } \omega \sim \mathcal{N}(\mathbf{0}_p, \mathbf{I}_p), \tag{6}$$

$$\mathbf{B}_i = \Sigma_i^{1/2}(\mathbf{H}_1 - \mathbf{H}_0)\Sigma_i^{1/2},$$

$$\mathbf{y}_i = \Sigma_i^{1/2}\left[\mathbf{H}_1(\boldsymbol{\mu}_i - \overline{\mathbf{x}}_1) - \mathbf{H}_0(\boldsymbol{\mu}_i - \overline{\mathbf{x}}_0)\right],$$

$$\xi_i = -\log\left(\frac{|\mathbf{H}_0|}{|\mathbf{H}_1|}\right) + (\boldsymbol{\mu}_i - \overline{\mathbf{x}}_0)^T\mathbf{H}_0(\boldsymbol{\mu}_i - \overline{\mathbf{x}}_0) - (\boldsymbol{\mu}_i - \overline{\mathbf{x}}_1)^T\mathbf{H}_1(\boldsymbol{\mu}_i - \overline{\mathbf{x}}_1) + 2\log\frac{\pi_1}{\pi_0}.$$

Important Results from RMT

Technical assumptions

1. Data scaling: $\frac{n_i}{p} \to c \in (0,\infty)$, with $|n_0 - n_1| = o(p)$.
2. Mean scaling: $\|\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1\|^2 = O(\sqrt{p})$.
3. Covariance scaling: $\|\Sigma_i\| = O(1)$.
4. $\Sigma_0 - \Sigma_1$ has exactly $O(\sqrt{p})$ eigenvalues of $O(1)$.

Recent RMT tools (*Hachem et al, 2008*)

For any $\|\mathbf{A}\| = O(1)$ and under assumptions 1-4:

$$\frac{1}{p}\operatorname{tr}\mathbf{A}\mathbf{H}_i - \frac{1}{p}\operatorname{tr}\mathbf{A}\left(\mathbf{I}_p + \frac{\gamma}{1+\gamma\delta_i}\Sigma_i\right)^{-1} \xrightarrow[a.s.]{p\to\infty} 0, \textit{ where } \delta_i = \frac{1}{n_i}\operatorname{tr}\Sigma_i\left(\mathbf{I}_p + \frac{\gamma}{1+\gamma\delta_i}\Sigma_i\right)^{-1}.$$

Asymptotic analysis using RMT tools

Define

$\mathbf{T}_i = \left(\mathbf{I}_p + \frac{\gamma}{1+\gamma\delta_i}\Sigma_i\right)^{-1}$ and the scalars ϕ_i and $\tilde{\phi}_i$ as $\phi_i = \frac{1}{n_i}\operatorname{tr}\Sigma_i^2\mathbf{T}_i^2$, $\tilde{\phi}_i = \frac{1}{(1+\gamma\delta_i)^2}$.

Let $\boldsymbol{\mu} = \boldsymbol{\mu}_0 - \boldsymbol{\mu}_1$, and set $\overline{\xi}_i$, $\overline{b}_i$ and $\overline{B}_i$ to

$$\overline{\xi}_i \triangleq \frac{1}{\sqrt{p}}\left[-\log\frac{|\mathbf{T}_0|}{|\mathbf{T}_1|} + \log\frac{(1+\gamma\delta_0)^{n_0}}{(1+\gamma\delta_1)^{n_1}} + \gamma\left(\frac{n_1\delta_1}{1+\gamma\delta_1} - \frac{n_0\delta_0}{1+\gamma\delta_0}\right) + (-1)^{i+1}\boldsymbol{\mu}^T\mathbf{T}_{1-i}\boldsymbol{\mu}\right].$$

$$\overline{b}_i \triangleq \frac{1}{\sqrt{p}}\operatorname{tr}\Sigma_i(\mathbf{T}_1 - \mathbf{T}_0).$$

$$\overline{B}_i \triangleq c\left[\frac{\phi_0}{1-\gamma^2\phi_0\tilde{\phi}_0} + \frac{\phi_1}{1-\gamma^2\phi_1\tilde{\phi}_1}\right] - \frac{2}{p}\operatorname{tr}\Sigma_i\mathbf{T}_1\Sigma_i\mathbf{T}_0.$$

Asymptotic Conditional Probability of Misclassification

Under assumptions 1-4, the following convergence holds for $i \in \{0,1\}$

$$\epsilon_i - \Phi\left((-1)^i\frac{\overline{\xi}_i - \overline{b}_i}{\sqrt{2\overline{B}_i}}\right) \xrightarrow{p} 0.$$

Special cases

- $\|\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1\| = O(1)$. In this case, the classification error rate would still converge to a non trivial limit but would not asymptotically depend on the difference $\|\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1\|$.
- $\|\Sigma_0 - \Sigma_1\| = O(p^{-\frac{1}{2}-\alpha})$, $\alpha > 0$.

$$\epsilon - \Phi\left(-\frac{\boldsymbol{\mu}^T\mathbf{T}\boldsymbol{\mu}}{2\sqrt{p}}\sqrt{\frac{1-\gamma^2\phi\tilde{\phi}}{c\gamma^2\phi^2\tilde{\phi}}}\right) \xrightarrow{p} 0, \tag{7}$$

which means that RLDA is better than RQDA (Zollanvari and Dougherty, 2015).

Experiments

Synthetic data

$\{\Sigma_0\}_{i,j} = 0.6^{|i-j|}$, $\Sigma_1 = \Sigma_0 + 2\begin{bmatrix}\mathbf{I}_k & \mathbf{0}_{k\times(p-k)} \\ \mathbf{0}_{(p-k)\times k} & \mathbf{0}_{(p-k)\times(p-k)}\end{bmatrix}$, $k = \lfloor\sqrt{p}\rfloor$,

$\boldsymbol{\mu}_0 = \left[1, \mathbf{0}_{(p-1)\times 1}\right]$, $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_0 + p^{-\frac{1}{4}}\mathbf{1}_{p\times 1}$.

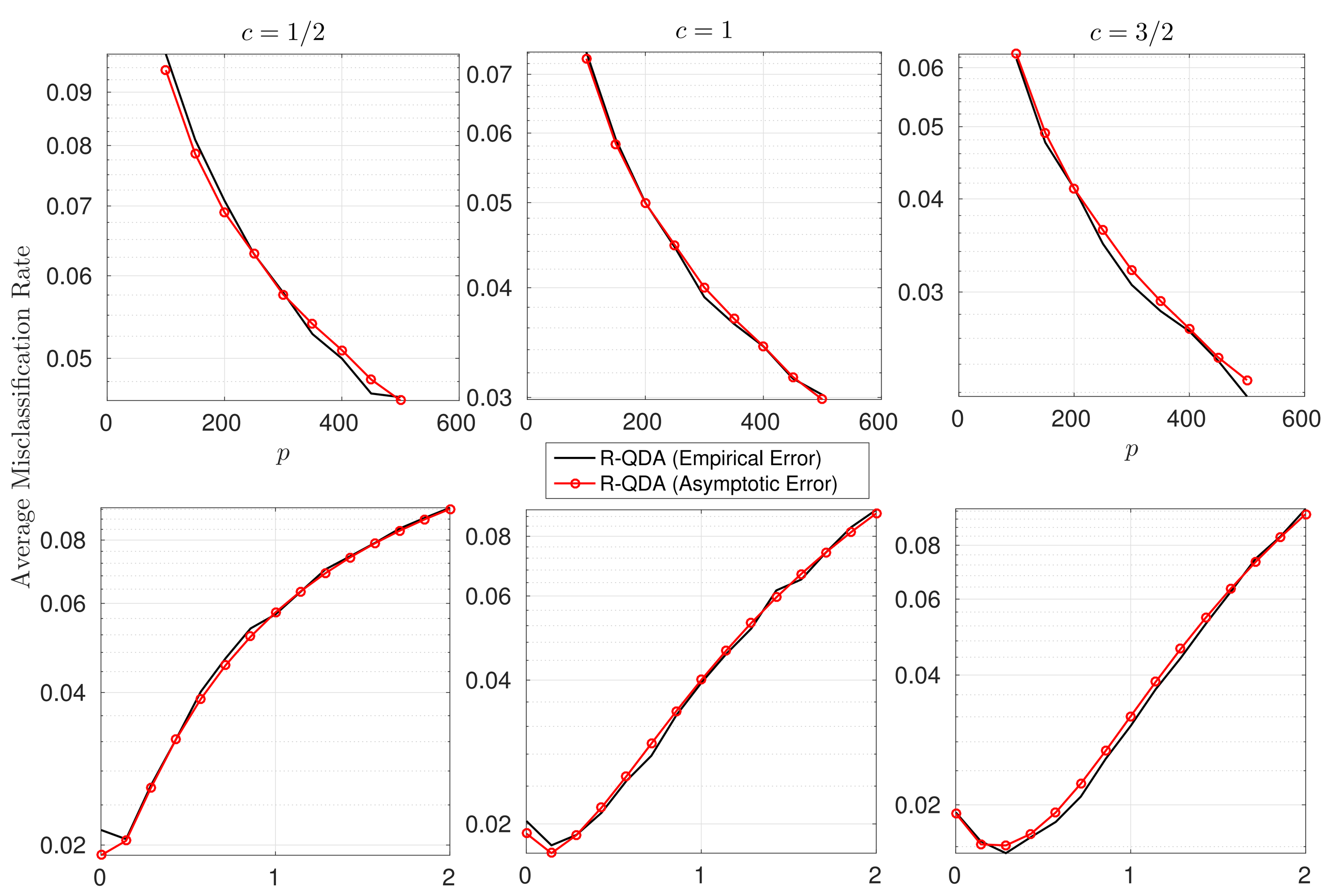

Figure: $n_0 = n_1$ for both experiments with $\gamma = 1$ in the first experiment and $p = 300$ in the second experiment.

Breast cancer data

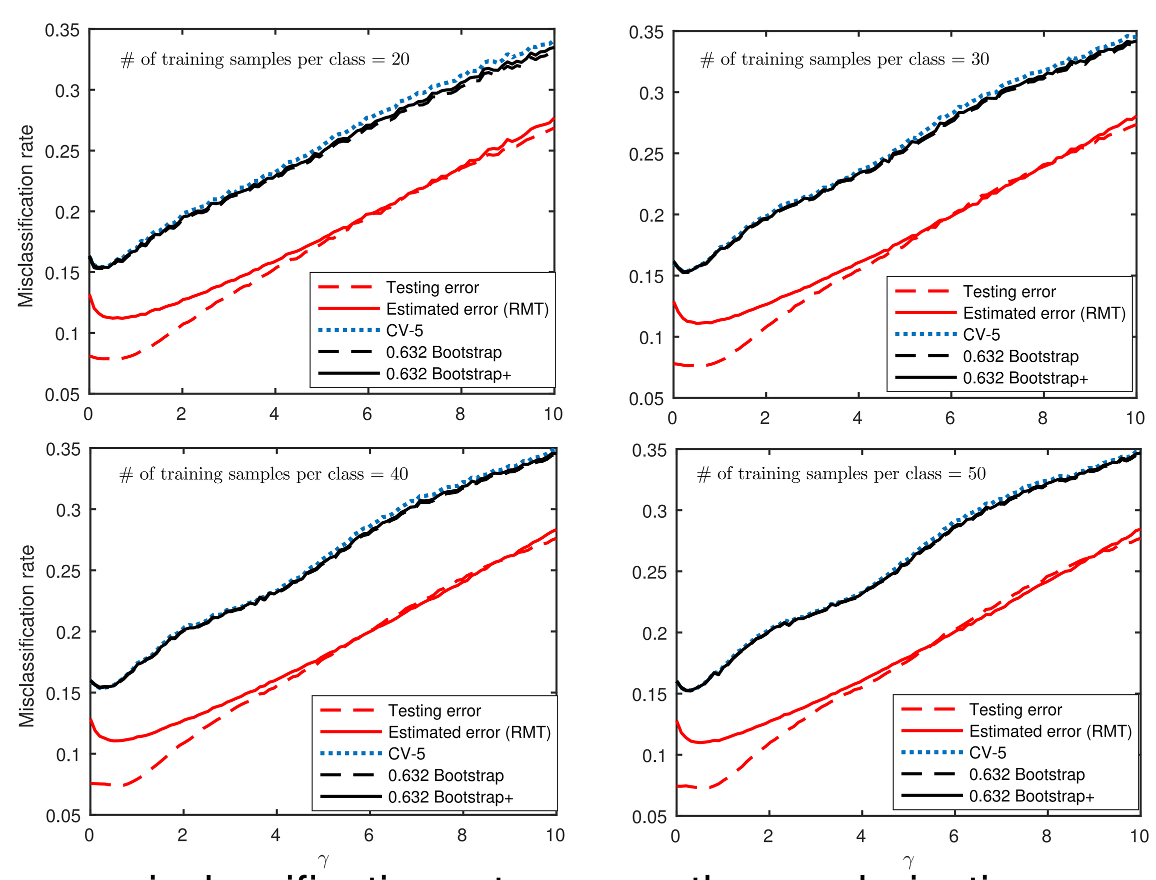

Figure: Average misclassification rate versus the regularization parameter γ with equal training size ($n_0 = n_1$).

KAUST Research Conference on Big Data Analyses in Evolutionary Biology, December 2017