Random Matrix Theory: Selected Applications from Statistical Signal Processing and Machine Learning

Ph.D. Thesis Defense

Khalil Elkhalil

Committee Chairperson: Dr. Tareq Y. Al-Naffouri
Committee Co-Chair: Dr. Mohamed-Slim Alouini
Committee Members: Dr. Xiangliang Zhang and Dr. Abla Kammoun
External Examiner: Dr. Alfred Hero

June 24, 2019

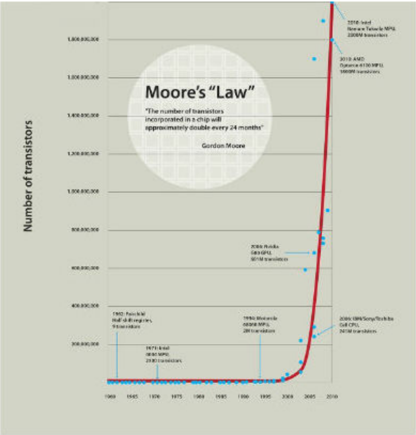
King Abdullah University of Science and Technology
Computer, Electrical and Mathematical Sciences and Engineering

KAUST
King Abdullah University of Science and Technology

Table of contents

Introduction

- The # of transistors that you can fit into a piece of silicon doubles every couple of years.



[1]

[1] C. M. Bishop, Microsoft research, Cambridge.

Big data

- Large sample size
- High dimensional data
- High variability

New processing techniques

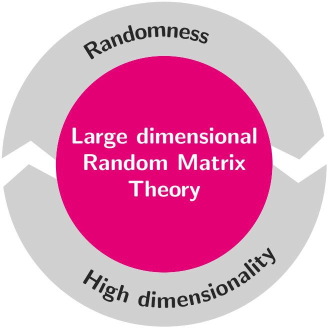
- Optimization
- Statistics

We need a tool that embraces these challenges

Random matrix theory (RMT)

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



Statistical Signal Processing

- Large number of antenna arrays vs large number of observations

→ Improved signal processing techniques

Wireless Communications

- Large # of antennas, Large # of users

→ Improved transmission and detection strategies

→ Low complexity design

Machine learning

- Supervised[2]/semi-supervised[3]/unsupervised learning[4].

→ A better fundamental understanding

→ Improved classification performance

---

[2] Z. Liao, R. Couillet, "A Large Dimensional Analysis of Least Squares Support Vector Machines", submitted

[3] X. Mai, R. Couillet, "A random matrix analysis and improvement of semi-supervised learning for large dimensional data", submitted.

[4] R. Couillet, F. Benaych-Georges, "Kernel Spectral Clustering of Large Dimensional Data", Electronic Journal of Statistics, vol. 10, no. 1, pp. 1393-1454, 2016.

RMT: Example

How does this work ?

- Self-averaging effect mechanism similar to that met in the law of large numbers

- $\mathbf{h}_1, \cdots, \mathbf{h}_n \in \mathbb{C}^p$ with i.i.d entries with zero mean and variance $\frac{1}{n}$.
- $\mathbf{H}\mathbf{H}^{\mathsf{H}}$ is an estimator of the cov. matrix with $\mathbf{H} = [\mathbf{h}_1, \cdots, \mathbf{h}_n]$.

p fixed, $n \to \infty$

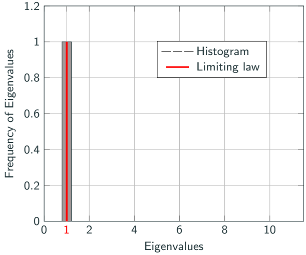


Figure 1.1: Histogram of eigenvalues of $\mathbf{H}\mathbf{H}^{\mathsf{H}}$

$p, n \to \infty$, $p/n \to c$

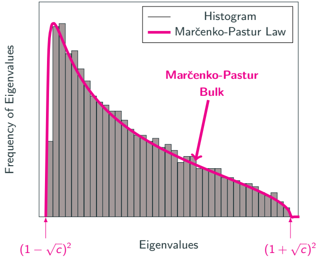


Figure 1.2: Histogram of eigenvalues of $\mathbf{H}\mathbf{H}^{\mathsf{H}}$

Why is this useful?

The same result can be extended in the correlated case[5]

Certain functionals of $\mathbf{H}\mathbf{H}^{\mathsf{H}}$ can be evaluated when $p, n \to \infty$**,** $p/n \to c$.

$$\boxed{f\left(\mathbf{H}\mathbf{H}^{\mathsf{H}}\right)}$$

- $\frac{1}{n}\operatorname{tr}\left(\mathbf{H}\mathbf{H}^{\mathsf{H}}\right)$
- $\frac{1}{n}\operatorname{tr}\left(\mathbf{H}\mathbf{H}^{\mathsf{H}}\right)^{-k}$: performance of linear est. techniques
- $\frac{1}{n}\log\det\left(\mathbf{H}\mathbf{H}^{\mathsf{H}}\right)$: MIMO systems, linear estimation (LCE)
- $\lambda_{\min}\left(\mathbf{H}\mathbf{H}^{\mathsf{H}}\right)$, $\lambda_{\max}\left(\mathbf{H}\mathbf{H}^{\mathsf{H}}\right)$,... : WEV in linear estimation

What happens for the moments in the finite regime?

$$\boxed{\mathbb{E}_{\mathbf{H}\sim\mathcal{D}} f\left(\mathbf{H}\mathbf{H}^{\mathsf{H}}\right)}$$

[5] J. W. Silverstein and Z. D. Bai, On the Empirical Distribution of Eigenvalues of a Class of Large Dimensional Random Matrices, Journal of Multivariate Analysis, vol. 54, pp. 175192, May 2002.

Moments of Correlated Gram matrices

Linear estimation

Let $m < n$ and $\mathbf{H} \in \mathbb{C}^{n\times m}$ with i.i.d zero mean unit variance Gaussian entries and $\mathbf{\Lambda}$ is positive definite matrix with distinct eigenvalues $\theta_1, \theta_2, \cdots, \theta_n$.

$$\mathbf{y}_{n\times 1} = \mathbf{H}_{n\times m}\mathbf{v}_{m\times 1} + \mathbf{\Lambda}^{-\frac{1}{2}}\mathbf{z}_{n\times 1}. \tag{1}$$

Define the correlated Gram matrix

$$\mathbf{G} = \mathbf{H}^*\mathbf{\Lambda}\mathbf{H}. \tag{2}$$

	LS	LMMSE
MSE	$\mathbb{E}\,\mathrm{tr}\,\mathbf{G}^{-1}$	$\mathbb{E}\,\mathrm{tr}\left(\mathbf{G}+\mathbf{R}_v^{-1}\right)^{-1}$

Sample covariance matrix (SCM): $\mathbf{u}(k) = \mathbf{R}^{\frac{1}{2}}\mathbf{h}(k)$

$$\hat{\mathbf{R}}(n) = (1-\lambda)\sum_{k=1}^{n}\lambda^{n-k}\mathbf{u}(k)\mathbf{u}^*(k) = \mathbf{R}^{\frac{1}{2}}\mathbf{H}\mathbf{\Lambda}(n)\mathbf{H}^*\mathbf{R}^{\frac{1}{2}}, \tag{3}$$

$$\begin{aligned} Loss(n) &\triangleq \mathbb{E}\left\|\mathbf{R}^{\frac{1}{2}}\hat{\mathbf{R}}^{-1}(n)\mathbf{R}^{\frac{1}{2}} - \mathbf{I}_m\right\|_F^2 \\ &= m + \mathbb{E}\,\mathrm{tr}\,\mathbf{G}_n^{-2} - 2\mathbb{E}\,\mathrm{tr}\,\mathbf{G}_n^{-1} \end{aligned}$$

Define the negative moments of $\mathbf{G}$ as

$$\boxed{\mu_{\mathbf{\Lambda}}(-k) \triangleq \mathbb{E}\,\mathrm{tr}\left(\mathbf{G}^{-k}\right), \quad k \in \mathbb{N}.}$$

Then,

	LS (Exact)	LMMSE ($\mathbf{R}_x = \sigma_x^2\mathbf{I}$, $\sigma_x^2 \gg 1$)
MSE	$\mu_{\mathbf{\Lambda}}(-1)$	$\sum_{k=0}^{l} \frac{(-1)^k}{\sigma_x^{2k}} \mu_{\mathbf{\Lambda}}(-k-1) + o\left(\sigma_x^{-2l}\right)$

$$Loss\,(n) = m + \mu_{\mathbf{\Lambda}(n)}(-2) - 2\mu_{\mathbf{\Lambda}(n)}(-1).$$

Theorem (Negative moments)[a] Let $p = \min(m, n-m)$, then for $1 \leq k \leq p$, we have

$$\mu_{\mathbf{\Lambda}}(-k) = L \sum_{j=1}^{k} \sum_{i=1}^{m} \mathcal{D}(i,j) \frac{(-1)^{k-j}}{(k-j)!} \mathbf{b}_i^t \mathbf{\Psi}^{-1} \mathbf{D}_i \mathbf{a}_{j,k}.$$

[a] K. Elkhalil, A. Kammoun, T. Al-Naffouri and M.-S. Alouini. Analytical Derivation of the Inverse Moments of One-sided Correlated Gram Matrices with Applications. IEEE Trans. Signal Processing, 2016.

$$\mathbf{\Psi} = \begin{bmatrix} 1 & \theta_1 & \cdots & \theta_1^{n-m-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \theta_{n-m} & \cdots & \theta_{n-m}^{n-m-1} \end{bmatrix}$$

Limitations

$$\mu_{\boldsymbol{\Lambda}}(-k) = L \sum_{j=1}^{k} \sum_{i=1}^{m} \mathcal{D}(i,j) \frac{(-1)^{k-j}}{(k-j)!} \mathbf{b}_i^t \boldsymbol{\Psi}^{-1} \mathbf{D}_i \mathbf{a}_{j,k}.$$

$$\boldsymbol{\Psi} = \begin{bmatrix} 1 & \theta_1 & \cdots & \theta_1^{n-m-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \theta_{n-m} & \cdots & \theta_{n-m}^{n-m-1} \end{bmatrix}$$

- So complicated formula
- Not useful if the eigenvalues of $\boldsymbol{\Lambda}$ are close to each other (We treat this issue for positive moments) [6].
- Not numerically stable if the dimensions are large.
- Not insightful!
- Not universal: the result will be different if we change the distribution from Gaussian.

[6] K. Elkhalil, A. Kammoun, T. Y. Al-Naffouri and M.-S. Alouini: Numerically Stable Evaluation of Moments of Random Gram Matrices With Applications. IEEE Signal Process. Lett. 24(9): 1353-1357 (2017)

Theorem (Silverstein and Bai [7])

Consider the Gram matrix $\mathbf{G} = \mathbf{H}^*\mathbf{\Lambda}\mathbf{H}$ *with the following assumptions*

- $m, n \to \infty$ *with* $\frac{m}{n} \to c \in (0, \infty)$
- $\|\mathbf{\Lambda}\| = O(1)$ *with* $rank(\mathbf{\Lambda}) = O(m)$.

$$\frac{1}{m} \operatorname{tr} \mathbf{G}^{-1} - \delta \to_{a.s.} 0, \;\; \delta = \left[\frac{1}{m} \operatorname{tr} \mathbf{\Lambda} \left(\mathbf{I}_n + \delta \mathbf{\Lambda}\right)^{-1}\right]^{-1}.$$

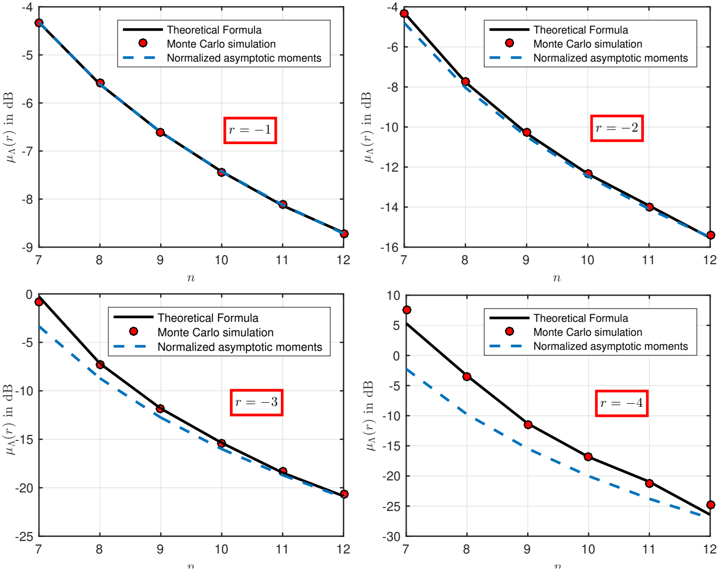
Higher inverse moments can be computed using an iterative process[a]

[a] Khalil Elkhalil, Abla Kammoun, Tareq Y. Al-Naffouri, Mohamed-Slim Alouini: Analytical Derivation of the Inverse Moments of One-Sided Correlated Gram Matrices With Applications. IEEE Trans. Signal Processing 64(10): 2624-2635 (2016)

[7] J. W. Silverstein and Z. D. Bai, On the Empirical Distribution of Eigenvalues of a Class of Large Dimensional Random Matrices, Journal of Multivariate Analysis, vol. 54, pp. 175192, May 2002.

Validation of the inverse moments

$m = 3$



Optimal λ for SCM estimation

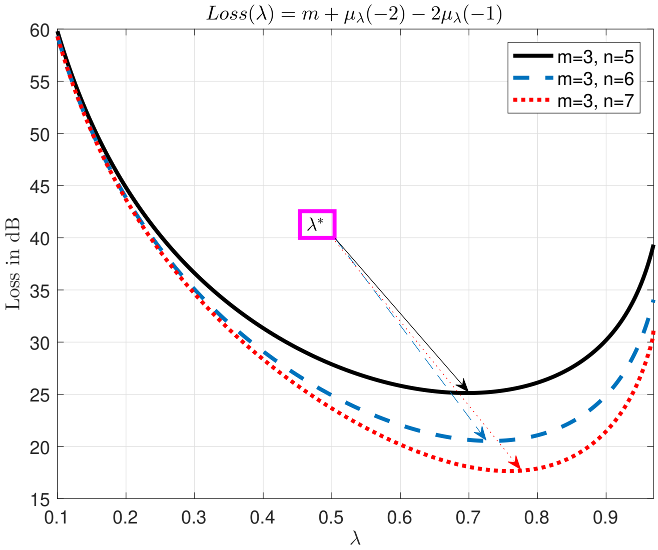
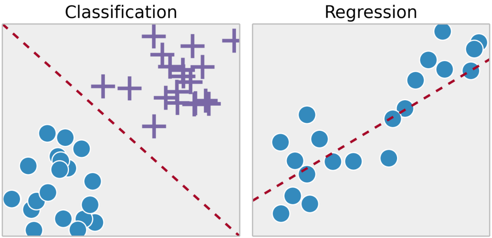


Figure 2.1: The estimation loss as a function of λ (Exact formula).

Regularized discriminant analysis with large dimensional data

Supervised learning

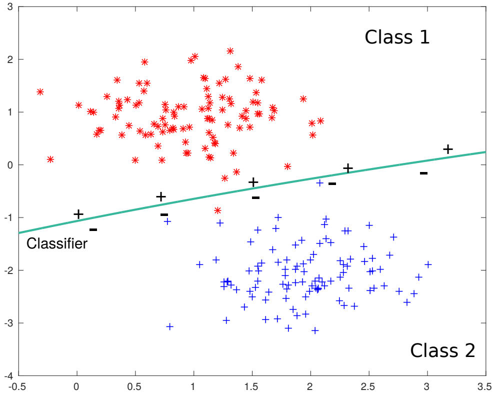
- We are provided with labeled data $\left(\text{features}_i, \text{response/label}_i\right)_{1 \leq i \leq n}$.
- Fit a model to the data.



https://towardsdatascience.com

Classification

- Principle: Build 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.

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

Model based classification

- Data is assumed to be sampled from a certain dist.
- The decision rule is constructed based on that.
- The MAP rule is considered in the design

$$\boxed{\widehat{k} = \arg \max_{k:classes} \mathbb{P}\left[\mathcal{C}_k | \mathbf{x}\right]}$$

The classifier is designed to satisfy this rule.

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): $\boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}$

$$W^{LDA}(\mathbf{x}) = \left(\mathbf{x} - \frac{\boldsymbol{\mu}_0 + \boldsymbol{\mu}_1}{2}\right)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1) - \log \frac{\pi_1}{\pi_0} \underset{\mathcal{C}_1}{\overset{\mathcal{C}_0}{\gtrless}} 0.$$

→ Decision rule is linear in $\mathbf{x}$.

Quadratic discriminant analysis: $\boldsymbol{\Sigma}_0 \neq \boldsymbol{\Sigma}_1$

$$W^{QDA}(\mathbf{x}) = -\frac{1}{2}\log\frac{|\boldsymbol{\Sigma}_0|}{|\boldsymbol{\Sigma}_1|} - \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_0)^T \boldsymbol{\Sigma}_0^{-1}(\mathbf{x} - \boldsymbol{\mu}_0) + \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_1)^T \boldsymbol{\Sigma}_1^{-1}(\mathbf{x} - \boldsymbol{\mu}_1) \underset{\mathcal{C}_1}{\overset{\mathcal{C}_0}{\gtrless}} 0.$$

→ Decision rule is quadratic in $\mathbf{x}$.

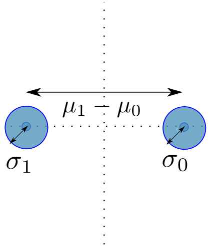
How does this perform ?

LDA

- Assume $\boldsymbol{\Sigma}$, $\boldsymbol{\mu}_0$ and $\boldsymbol{\mu}_1$ known.
- Equal priors : $\pi_0 = \pi_1 = 0.5$
- No asymptotic regime, p is fixed.

The total misclassification rate is given by [8]

$$\epsilon = \Phi\left(-\frac{\Delta}{2}\right), \quad \Delta = \|\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1\|_{\boldsymbol{\Sigma}^{-1}}$$



What happens when the statistics are not known ?

[8] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. The Elements of Statistical Learning. Springer, 2009.

LDA: Asymptotic regime (equal covariances)

Asymptotic growth regime

Let $n = n_0 + n_1$.

- $n_0, n_1, p \to \infty$ such that $\frac{p}{n} \to c < 1$.
- $\boldsymbol{\mu}_0$ and $\boldsymbol{\mu}_1$ are known.
- $\boldsymbol{\Sigma}$ is replaced by its sample estimate $\widehat{\boldsymbol{\Sigma}} = \frac{1}{n-2} \sum_{k=1}^{2} \sum_{i=1}^{n_k} (\mathbf{x}_{k,i} - \overline{\mathbf{x}}_k)(\mathbf{x}_{k,i} - \overline{\mathbf{x}}_k)^T$.

Wang et al. 2018 [a]

$$\epsilon_{LDA} - \Phi\left[-\frac{\Delta}{2}\sqrt{1-c}\right] \to_{prob.} 0$$

[a] Cheng Wang and Binyan Jiang. On the dimension effect of regularized linear discriminant analysis, arXiv:1710.03136v1

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

→ 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 $c < 1$.

What happens if $p > n$?

LDA: High dimensionality

Regularization

$$\mathbf{H} = \left(\mathbf{I}_p + \gamma\widehat{\boldsymbol{\Sigma}}\right)^{-1}.$$

Optimal γ ?

Dimensionality reduction

$$\text{data}_{(d)} = \boldsymbol{W}_{d\times p} \times \text{data}_{(p)}$$

Best d ?

LDA with random projections

Random projections

$$\mathbb{R}^p \longrightarrow \mathbb{R}^d$$
$$\boldsymbol{x} \longmapsto \boldsymbol{W}\boldsymbol{x}$$

Projection matrix

We shall assume that the projection matrix $\mathbf{W}$ writes as $\mathbf{W} = \frac{1}{\sqrt{p}}\mathbf{Z}$, where the entries $Z_{i,j}$ ($1 \leq i \leq d$, $1 \leq j \leq p$) of $\mathbf{Z}$ are centered with unit variance and independent identically distributed random variables satisfying the following moment assumption. There exists $\epsilon > 0$, such that $\mathbb{E}\left|Z_{i,j}\right|^{4+\epsilon} < \infty$.

Johnsonn-Lindenstrauss Lemma

For a given n data points $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n$ in $\mathbb{R}^p$, $\epsilon \in (0,1)$ and $d > \frac{8 \log n}{\epsilon^2}$, there exists a linear map $f : \mathbb{R}^p \to \mathbb{R}^d$ such that

$$(1-\epsilon)\left\|\boldsymbol{x}_i - \boldsymbol{x}_j\right\|^2 \leq \left\|\boldsymbol{W}\boldsymbol{x}_i - \boldsymbol{W}\boldsymbol{x}_j\right\|^2 \leq (1+\epsilon)\left\|\boldsymbol{x}_i - \boldsymbol{x}_j\right\|^2, \tag{4}$$

Conditional risk after projection

$$\epsilon_i^{\text{P-LDA}} = \Phi\left[-\frac{1}{2}\sqrt{\boldsymbol{\mu}^\top \mathbf{W}^\top \left(\mathbf{W}\boldsymbol{\Sigma}\mathbf{W}^\top\right)^{-1}\mathbf{W}\boldsymbol{\mu}} + \frac{(-1)^{i+1}\log\frac{\pi_0}{\pi_1}}{\sqrt{\boldsymbol{\mu}^\top \mathbf{W}^\top \left(\mathbf{W}\boldsymbol{\Sigma}\mathbf{W}^\top\right)^{-1}\mathbf{W}\boldsymbol{\mu}}}\right].$$

Performance of LDA with random projections

Asymptotic Performance[a]

$$\epsilon_i^{\text{P-LDA}} - \Phi\left[\frac{-\frac{1}{2}\boldsymbol{\mu}^\top\left(\boldsymbol{\Sigma}+\delta_d\mathbf{I}_p\right)^{-1}\boldsymbol{\mu}+(-1)^{i+1}\log\frac{\pi_0}{\pi_1}}{\sqrt{\boldsymbol{\mu}^\top\left(\boldsymbol{\Sigma}+\delta_d\mathbf{I}_p\right)^{-1}\boldsymbol{\mu}}}\right] \to_{prob.} 0, \quad (5)$$

$$\delta_d \operatorname{tr}\left(\boldsymbol{\Sigma}+\delta_d\mathbf{I}_p\right)^{-1} = p - d. \quad (6)$$

δ_d can be seen as a penalty on projection.

[a] K. Elkhalil, A. Kammoun, R. Calderbank, T. Al-Naffouri and M.-S. Alouini. Asymptotic Performance of Linear Discriminant Analysis with Random Projections. ICASSP 2019.

LDA

equal priors: $\Phi\left[-\frac{1}{2}\sqrt{\boldsymbol{\mu}^\top\boldsymbol{\Sigma}\boldsymbol{\mu}}\right]$

$\boldsymbol{\Sigma} = \mathbf{I}_p$: $\Phi\left[-\frac{1}{2}\|\boldsymbol{\mu}\|\right]$

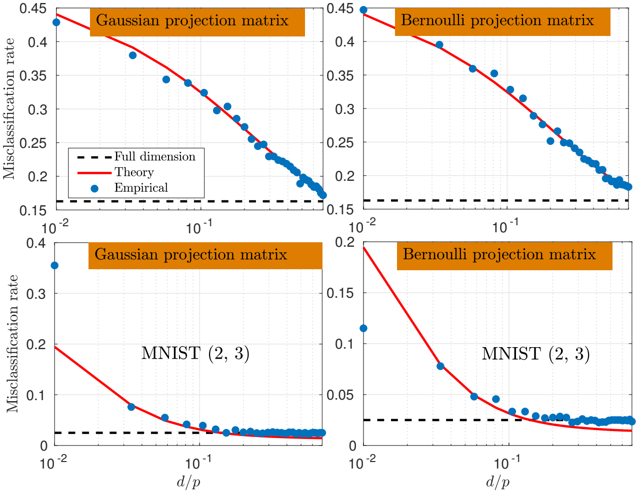
P-LDA

$$\Phi\left[-\frac{1}{2}\sqrt{\boldsymbol{\mu}^\top\left(\boldsymbol{\Sigma}+\delta_d\mathbf{I}_p\right)^{-1}\boldsymbol{\mu}}\right]$$

$$\Phi\left[-\frac{1}{2}\sqrt{d/p}\,\|\boldsymbol{\mu}\|\right]$$

P-LDA: Experiments

- $p = 800$.
- $\boldsymbol{\mu}_0 = \mathbf{0}_p$ and $\boldsymbol{\mu}_1 = \frac{3}{\sqrt{p}}\mathbf{1}_p$.
- $\boldsymbol{\Sigma} = \{0.4^{|i-j|}\}_{i,j}$.



R-LDA: Asymptotic regime (equal covariances)

Asymptotic growth regime

- $n_0, n_1, p \to \infty$ such that $\frac{p}{n} \to c \in (0, \infty)$.
- $\boldsymbol{\mu}_k$ are replaced by $\overline{\boldsymbol{x}}_k = \frac{1}{n_k} \sum_{\boldsymbol{x}_i \in \mathcal{C}_k} \boldsymbol{x}_i$.
- $\boldsymbol{\Sigma}^{-1}$ is replaced by its ridge estimate $\mathbf{H} = \left(\mathbf{I}_p + \gamma \widehat{\boldsymbol{\Sigma}}\right)^{-1}$.

Hachem el al 2008. [a]

$$\mathbf{H} \sim \mathbf{T} = \left(\mathbf{I}_p + \rho \boldsymbol{\Sigma}\right)^{-1},$$

in the sense that $\boldsymbol{a}^T \left(\mathbf{H} - \mathbf{T}\right) \boldsymbol{b} \to_{prob.} 0$ and $\frac{1}{n} \operatorname{tr} \mathbf{A} \left(\mathbf{H} - \mathbf{T}\right) \to_{prob.} 0$.

[a] W. Hachem, O. Khorunzhiy, P. Loubaton, J. Najim, L. Pastur: A New Approach for Mutual Information Analysis of Large Dimensional Multi-Antenna Channels. IEEE Trans. Information Theory 54(9): 3987 - 4004 (2008).

Zollanvari and Dougherty 2015 [a]

$$\epsilon_{R-LDA}^{equal} - \Phi \left[\frac{-\boldsymbol{\mu}^T \left(\mathbf{I}_p + \rho \boldsymbol{\Sigma}\right)^{-1} \boldsymbol{\mu}}{\sqrt{D}} \right] \to_{prob.} 0$$

[a] Amin Zollanvari and Edward R. Dougherty: Generalized Consistent Error Estimator of Linear Discriminant Analysis. IEEE Trans. Signal Processing 63(11): 2804-2814 (2015)

What if the covariances are different ?

R-LDA: Asymptotic regime (dist. covariances)

Asymptotic growth regime

- $n_0, n_1, p \to \infty$ such that $\frac{p}{n} \to c \in (0, \infty)$.
- $\boldsymbol{\mu}_k$ are replaced by $\overline{\mathbf{x}}_k = \frac{1}{n_k} \sum_{\mathbf{x}_i \in \mathcal{C}_k} \mathbf{x}_i$.
- $\boldsymbol{\Sigma}^{-1}$ is replaced by its ridge estimate $\mathbf{H} = \left(\mathbf{I}_p + \gamma \widehat{\boldsymbol{\Sigma}}\right)^{-1}$.

Benaych and Couillet 2016 [a]

$$\mathbf{H} \sim \mathbf{T}_{0,1} \propto (\mathbf{I_p} + \rho_0 \boldsymbol{\Sigma}_0 + \rho_1 \boldsymbol{\Sigma}_1)^{-1}$$

[a] F. Benaych-Georges and R. Couillet, Spectral Analysis of the Gram Matrix of Mixture Models, ESAIM: Probability and Statistics, vol. 20, pp. 217237, 2016.

Elkhalil et al. 2018 [a]

$$\epsilon_{R-LDA}^{dist.} - \left\{ \frac{1}{2}\Phi\left[\frac{-\boldsymbol{\mu}^T \mathbf{T}_{0,1}\boldsymbol{\mu} + \beta}{\sqrt{D_0}}\right] + \frac{1}{2}\Phi\left[\frac{-\boldsymbol{\mu}^T \mathbf{T}_{0,1}\boldsymbol{\mu} - \beta}{\sqrt{D_1}}\right]\right\} \to_{prob.} 0$$

[a] K. Elkhalil, A. Kammoun, R. Couillet, T. Al-Naffouri and M.-S. Alouini. A Large Dimensional Study of Regularized Discriminant Analysis Classifiers. Under review in IEEE Trans. Information Theory.

How is this different from the case of equal covariances?

R-LDA: Asymptotic regime (dist. covariances)

$$\epsilon_{R-LDA}^{dist.} - \left\{ \frac{1}{2}\Phi\left[\frac{-\boldsymbol{\mu}^T \mathbf{T}_{0,1}\boldsymbol{\mu} + \beta}{\sqrt{D_0}}\right] + \frac{1}{2}\Phi\left[\frac{-\boldsymbol{\mu}^T \mathbf{T}_{0,1}\boldsymbol{\mu} - \beta}{\sqrt{D_1}}\right] \right\} \rightarrow_{prob.} 0$$

Some insights

- If $\|\boldsymbol{\Sigma}_0 - \boldsymbol{\Sigma}_1\| = o(1)$

$$\epsilon_{R-LDA}^{dist.} = \epsilon_{R-LDA}^{equal} + o(1).$$

  $\rightarrow$ R-LDA is robust against small perturbations.

- Different misclassification rates across classes.
- The enhancement in the misclassification rate in one class is likely to be lost by the other class.
- R-LDA does not leverage well the information about the covariance differences.

What about R-QDA ?

R-QDA: Asymptotic regime

R-LDA	R-QDA
n_0, n_1 samples	n_0, n_1 samples
$\widehat{\boldsymbol{\Sigma}} = \frac{1}{n-2}\sum_{k=1}^{2}\sum_{i=1}^{n_k}\left(\mathbf{x}_{k,i}-\overline{\mathbf{x}}_k\right)\left(\mathbf{x}_{k,i}-\overline{\mathbf{x}}_k\right)^T$	$\widehat{\boldsymbol{\Sigma}}_0 = \frac{1}{n_0-1}\sum_{i=1}^{n_0}\left(\mathbf{x}_{0,i}-\overline{\mathbf{x}}_0\right)\left(\mathbf{x}_{0,i}-\overline{\mathbf{x}}_0\right)^T$ $\widehat{\boldsymbol{\Sigma}}_1 = \frac{1}{n_1-1}\sum_{i=1}^{n_1}\left(\mathbf{x}_{1,i}-\overline{\mathbf{x}}_1\right)\left(\mathbf{x}_{1,i}-\overline{\mathbf{x}}_1\right)^T$

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

Asymptotic growth regime

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

CLT(Lyapunov)

$$\epsilon_i^{R-QDA} - \Phi\left[(-1)^i \frac{1/\sqrt{p}\xi_i - 1/\sqrt{p}\operatorname{tr}\mathbf{B}_i}{\sqrt{1/p2\operatorname{tr}\mathbf{B}_i^2 + 1/p4\mathbf{y}_i^T\mathbf{y}_i}}\right] \to_{prob.} 0.$$

R-QDA: Asymptotic regime

Elkhalil et al. 2017/2018 [a] [b]

$$\epsilon_i^{R-QDA} - \left\{ \frac{1}{2}\Phi\left[\frac{\overline{\xi}_0 - \overline{b}_0}{\sqrt{2\overline{B_0}}}\right] + \frac{1}{2}\Phi\left[\frac{-\overline{\xi}_1 + \overline{b}_1}{\sqrt{2\overline{B_1}}}\right]\right\} \rightarrow_{prob.} 0.$$

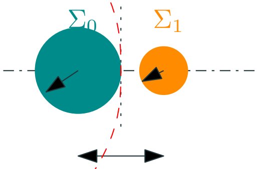
$\overline{\xi}_i$, $\overline{b}_i$ and $\overline{B}_i$ depend on the classes' statistics.

---

[a] K. Elkhalil, A. Kammoun, R. Couillet, T. Y. Al-Naffouri, and M.-S. Alouini. Asymptotic Performance of Regularized Quadratic Discriminat Analysis-Based Classifiers. IEEE MLSP, Roppongi, Japan, Sept 2017.

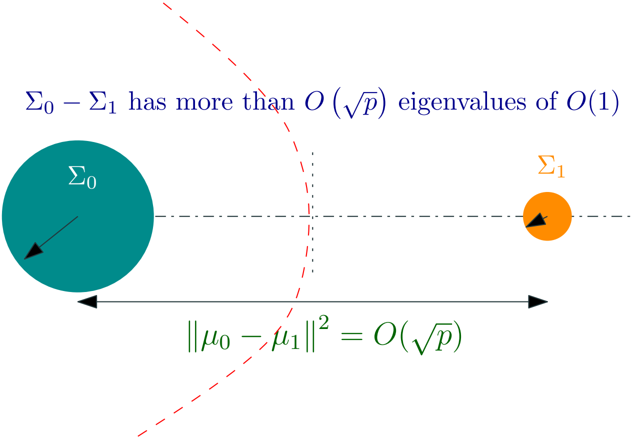
[b] K. Elkhalil, A. Kammoun, R. Couillet, T. Al-Naffouri and M.-S. Alouini. A Large Dimensional Study of Regularized Discriminant Analysis Classifiers. Under review in IEEE Trans. Information Theory.

Recall that R-QDA needs $\|\mu_0 - \mu_1\|^2 = O\left(\sqrt{p}\right)$

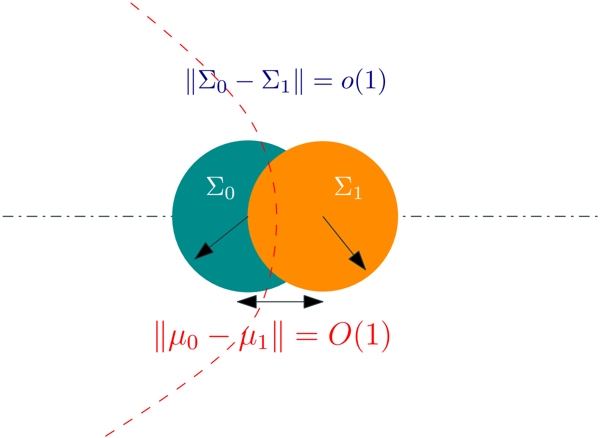


$$\|\mu_0 - \mu_1\| = O(1)$$

> The information on the distance between the means is asymptotically useless!



R-QDA achieves asymptotic perfect classification.



Classification is asymptotically impossible.

Discussion

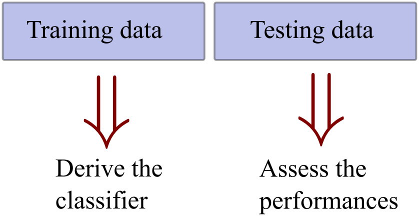
- Unbalanced training: $n_0 - n_1 = O(p)$.

R-QDA is equivalent to the naive classifier.

$$\epsilon \to \pi_0 \Phi(\infty) + \pi_1 \Phi(-\infty)$$

Parameter tuning

- Prone to estimation errors due to insufficiency in the number of observations.
- The tuning of the regularization parameter is very important



Model selection Given a set of candidate regularization factors

- Evaluate the performance using the test data for each regularization value [9]
- Select the value that presents the lowest mis-classification rate

[9] J. Friedman. Regularized discriminant analysis. Journal of the American Statistical Association, 84:165175, 1989

Consistent estimator of the classification error

Exploiting the asymptotic equivalent

- If $p = O(1)$ and $n \to \infty$, then $\left\| \widehat{\mathbf{\Sigma}}_i - \mathbf{\Sigma}_i \right\| = o_p(1)$.
- When $p, n \to \infty$, then $\left\| \widehat{\mathbf{\Sigma}}_i - \mathbf{\Sigma}_i \right\| \neq o_p(1)$.

R-LDA ($\mathbb{GE}$)

$$\widehat{\epsilon}_i^{R-LDA} - \Phi\left[\frac{(-1)^i G(\hat{\mu}_i, \hat{\mu}_0, \hat{\mu}_1, \mathbf{H})}{\sqrt{D\left(\hat{\mu}_0, \hat{\mu}_1, \mathbf{H}, \widehat{\mathbf{\Sigma}}_i\right)}}\right] \to_{p.} 0$$

R-QDA ($\mathbb{GE}$)

$$\widehat{\epsilon}_i^{R-QDA} - \Phi\left[(-1)^i \frac{\widehat{\xi}_i - \widehat{b}_i}{\sqrt{2\widehat{B}_i}}\right] \to_{p.} 0$$

Optimal regularizer

$$\widehat{\gamma}^{\star} = \arg\min_{\gamma > 0} \widehat{\epsilon}(\gamma).$$

- These results provides a glimpse on the region where the optimal γ is likely to belong.
- Perform a cross validation or testing in that region.

How well does this perform?

Performance

Benchmark estimation techniques:

- 5-fold cross-validation with 5 repetitions (5-CV).
- 0.632 bootstrap (B632).
- 0.632+ bootstrap (B632+)
- Plugin estimator consisting of replacing the stats. in the DEs by their sample estimates.

Synthetic data

- $[\boldsymbol{\Sigma}_0]_{i,j} = 0.6^{|i-j|}$.
- $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_0 + 3 \begin{bmatrix} \boldsymbol{I}_{\lceil \sqrt{p} \rceil} & \mathbf{0}_{k \times (p-k)} \\ \mathbf{0}_{(p-k) \times k} & \mathbf{0}_{(p-k) \times (p-k)} \end{bmatrix}$.
- $\boldsymbol{\mu}_0 = \left[1, \mathbf{0}_{1 \times (p-1)}\right]^T$.
- $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_0 + \frac{0.8}{\sqrt{p}} \mathbf{1}_{p \times 1}$.

Real data

- USPS dataset.
- $p = 256$ features (16×16) grayscale images.
- $n = 7291$ training examples.
- $n_{test} = 2007$ testing examples.

Performance: Synthetic data

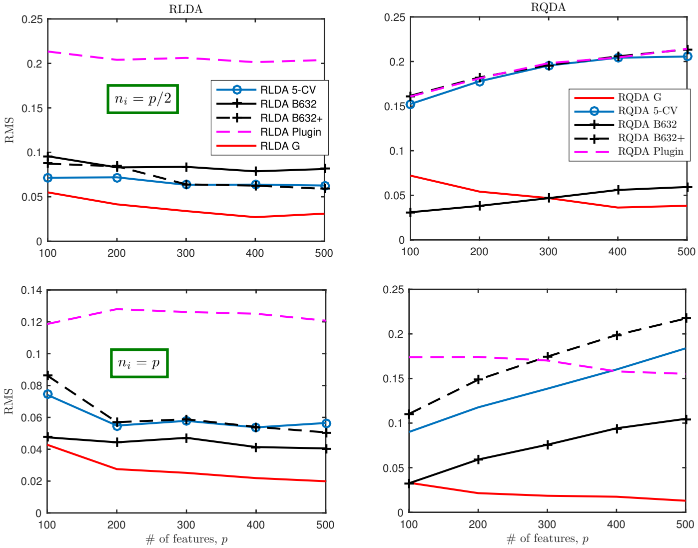


Figure 3.1: $n_0 = n_1$ and $\gamma = 1$.

Performance: Synthetic data

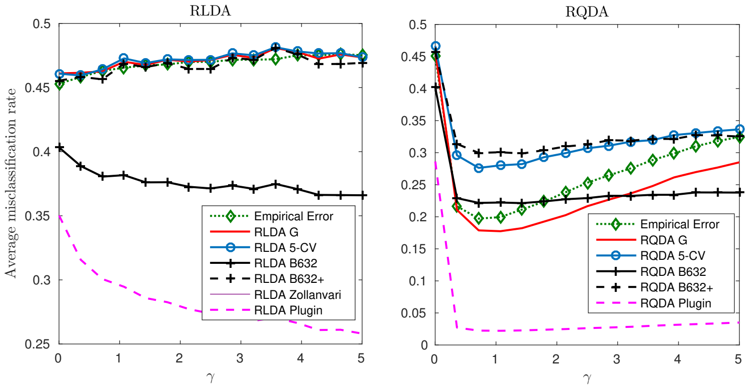


Figure 3.2: $p = 100$ features with equal training size ($n_0 = n_1 = p$).

Performance: USPS dataset

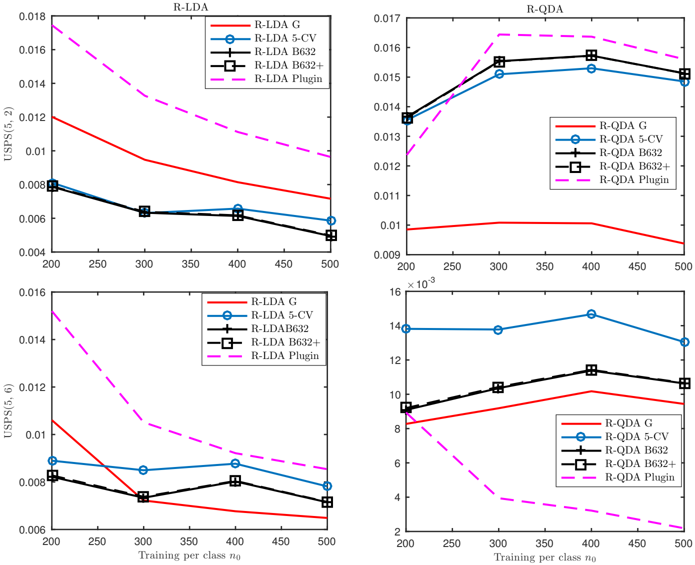
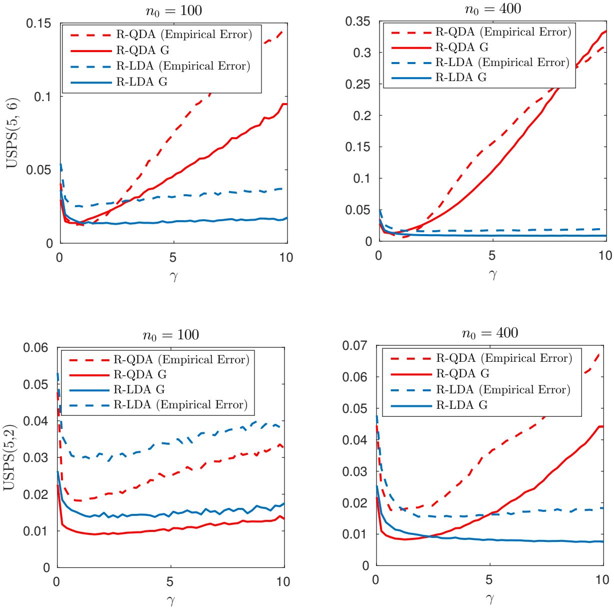


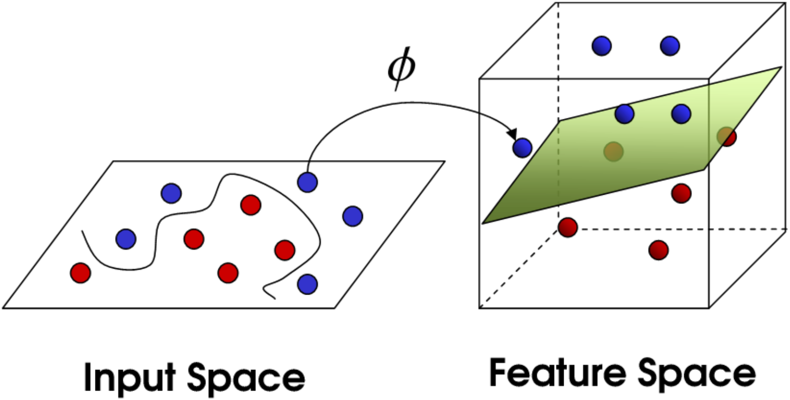
Figure 3.3: $n_0 = n_1$ and $\gamma = 1$. The first row gives the performance for the USPS data with digits (5, 2) whereas the second row considers the digits (5, 6).

Performance: USPS dataset



Centered Kernel Ridge Regression (CKRR)

KRR: Kernel trick



$y = \mathbf{w}^T \mathbf{x}$ $\qquad$ $y = \mathbf{w}^T \phi(\mathbf{x})$

KRR: Kernel trick

- $\{\boldsymbol{x}_i, y_i\}_{i=1}^{n}$ in $\mathcal{X} \times \mathcal{Y}$ s.t. $y_i = f(\boldsymbol{x}_i) + \sigma \epsilon_i$ with $\epsilon_i \sim_{\text{i.i.d}} \mathcal{N}(0,1)$.
- Feature map: $\phi : \mathcal{X} \to \mathcal{H}$, with $\mathcal{H}$ is a **RKHS**.
- Learning problem

$$\min_{\boldsymbol{\alpha}} \frac{1}{2} \|\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{\alpha}\|^2 + \frac{\lambda}{2} \|\boldsymbol{\alpha}\|^2$$

$$\boldsymbol{\Phi} = [\phi(\boldsymbol{x}_1), \cdots, \phi(\boldsymbol{x}_n)]^T \in \mathbb{R}^{|\mathcal{H}| \times n}$$

$$\boldsymbol{\alpha}^\star = \left(\boldsymbol{\Phi}^T \boldsymbol{\Phi} + \lambda \mathbf{I}_{|\mathcal{H}|}\right)^{-1} \boldsymbol{\Phi}^T \boldsymbol{y} \in \mathbb{R}^{|\mathcal{H}|}$$

$$f^*(\mathbf{x}) = \phi(\mathbf{x})^T \left(\boldsymbol{\Phi}^{\mathbf{T}} \boldsymbol{\Phi} + \lambda \mathbf{I}_{|\mathcal{H}|}\right)^{-1} \boldsymbol{\Phi}^{\mathbf{T}} \mathbf{y}$$

Woodbury

$$f^*(\mathbf{x}) = \phi(\mathbf{x})^T \boldsymbol{\Phi}^T \left(\boldsymbol{\Phi}\boldsymbol{\Phi}^T + \lambda \mathbf{I}_n\right)^{-1} \mathbf{y}$$

$$\{\boldsymbol{\Phi}\boldsymbol{\Phi}^T\}_{i,j} = \phi(\mathbf{x}_i)^T \phi(\mathbf{x}_j)$$

$$f^\star(\boldsymbol{x}) = \boldsymbol{\kappa}(\boldsymbol{x})^T (\boldsymbol{K} + \lambda \mathbf{I}_n)^{-1} \boldsymbol{y}$$

with $\boldsymbol{\kappa}(\boldsymbol{x})_i = \phi(\boldsymbol{x})^T \phi(\boldsymbol{x}_i)$ and $\boldsymbol{K}_{i,j} = \phi(\boldsymbol{x}_i)^T \phi(\boldsymbol{x}_j)$.

Centered KRR: Motivation

Inner-product kernels

$$k\left(\mathbf{x},\mathbf{x}'\right)=\phi\left(\mathbf{x}\right)^{T}\phi\left(\mathbf{x}'\right)=g\left(\mathbf{x}^{T}\mathbf{x}/p\right),\ \mathbf{x}\text{ and }\mathbf{x}'\in\mathcal{X}.$$

Asymptotic growth regime

Assumption 1.

- $p/n\to c\,(0,\infty)$.
- $\mathbb{E}\mathbf{x}_i=\mathbf{0}$ and $\mathrm{cov}\mathbf{x}_i=\mathbf{\Sigma}$ unif. bounded in p (e.g. $\mathbf{x}_i\sim\mathcal{N}\left(\mathbf{0},\mathbf{\Sigma}\right)$).

El Karoui 2010[a]

$$\left\|\mathbf{K}-\mathbf{K}^{\infty}\right\|\to_{a.s.}0$$

with $$\mathbf{K}^{\infty}=\underbrace{g\left(0\right)\mathbf{1}\mathbf{1}^{T}}_{\|.\|=O(p)}+\underbrace{g'\left(0\right)\frac{\mathbf{X}\mathbf{X}^{T}}{p}+constant(g,\mathbf{\Sigma})}_{\|.\|=O(1)}.$$

[a] N. El-Karoui, The Spectrum of Kernel Random Matrices, The Annals of Statistics, vol. 38, no. 1, pp. 150, 2010.

Centered KRR: Motivation

Inner-product kernels

$$k\left(\boldsymbol{x}, \boldsymbol{x}'\right) = \phi\left(\boldsymbol{x}\right)^T \phi\left(\boldsymbol{x}'\right) = g\left(\boldsymbol{x}^T \boldsymbol{x}/p\right), \ \boldsymbol{x} \text{ and } \boldsymbol{x}' \in \mathcal{X}.$$

Asymptotic growth regime

Assumption 1.

- $p/n \to c\,(0, \infty)$.
- $\mathbb{E}\boldsymbol{x}_i = \mathbf{0}$ and $\mathrm{cov}\boldsymbol{x}_i = \boldsymbol{\Sigma}$ unif. bounded in p (e.g. $\boldsymbol{x}_i \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Sigma}\right)$).

Centering with $\boldsymbol{P} = \mathbf{I}_n - \frac{\mathbf{1}\mathbf{1}^T}{n}$

$\boldsymbol{K}_c = \boldsymbol{P}\boldsymbol{K}\boldsymbol{P}$

El Karoui 2010[a]

$$\left\|\boldsymbol{K} - \boldsymbol{K}^{\infty}\right\| \to_{a.s.} 0$$

with $\boldsymbol{K}^{\infty} = \cancel{\underbrace{g\left(0\right)\mathbf{1}\mathbf{1}^T}_{\|.\|=O(p)}} + \underbrace{g'\left(0\right)\frac{\boldsymbol{X}\boldsymbol{X}^T}{p} + constant(g, \boldsymbol{\Sigma})}_{\|.\|=O(1)}.$

[a] N. El-Karoui, The Spectrum of Kernel Random Matrices, The Annals of Statistics, vol. 38, no. 1, pp. 150, 2010.

Centered KRR

$$\boldsymbol{P} = \boldsymbol{I}_n - \frac{\mathbf{1}\mathbf{1}^T}{n}.$$

Learning problem

$$\min_{\alpha_0, \boldsymbol{\alpha}} \frac{1}{2} \|\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{\alpha} - \alpha_0 \mathbf{1}_n\|^2 + \frac{\lambda}{2} \|\boldsymbol{\alpha}\|^2 \quad \Leftrightarrow \quad \min_{\boldsymbol{\alpha}} \frac{1}{2} \|\boldsymbol{P}(\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{\alpha})\|^2 + \frac{\lambda}{2} \|\boldsymbol{\alpha}\|^2$$

$$\boldsymbol{\alpha}^\star = \boldsymbol{\Phi}^T \boldsymbol{P} \left(\underbrace{\boldsymbol{P}\boldsymbol{K}\boldsymbol{P}}_{\boldsymbol{K}_c} + \lambda \boldsymbol{I}_n \right)^{-1} (\boldsymbol{y} - \bar{y}\mathbf{1}_n)$$

$$f_c^\star(\boldsymbol{x}) = \boldsymbol{\kappa}_c(\boldsymbol{x})^T (\boldsymbol{K}_c + \lambda \boldsymbol{I}_n)^{-1} \boldsymbol{P}\boldsymbol{y} + \bar{y}.$$

$$\boldsymbol{\kappa}_c(\boldsymbol{x}) = \boldsymbol{P}\boldsymbol{\kappa}(\boldsymbol{x}) - \frac{1}{n}\boldsymbol{P}\boldsymbol{K}\mathbf{1}_n, \quad \phi_c(\boldsymbol{x}) = \phi(\boldsymbol{x}) - \frac{1}{n}\sum_{i=1}^{n} \phi(\boldsymbol{x}_i).$$

Centered KRR $\sim$ KRR with centered kernels

What about the performance ?

CKRR: Performance

Performance metrics

$$\mathcal{R}_{train} = \frac{1}{n} \mathbb{E}_{\epsilon} \left\| \widehat{f}_c(X) - f(X) \right\|_2^2$$

$$\mathcal{R}_{test} = \mathbb{E}_{\boldsymbol{s} \sim \mathcal{D}, \epsilon} \left| \widehat{f}_c(\boldsymbol{s}) - f(\boldsymbol{s}) \right|^2$$

Assumption1. (Growth rate) As $p, n \to \infty$ we assume the following

- **Data scaling**: $p/n \to c \in (0, \infty)$.
- **Covariance scaling**: $\limsup_p \|\boldsymbol{\Sigma}\| < \infty$.

Assumptions 2. (kernel function)

$$\mathbb{E} \left| g^{(3)} \left(\frac{1}{p} \boldsymbol{x}_i^T \boldsymbol{x}_j \right) \right|^k < \infty.$$

Assumption 3. (Data generating function)

- $$\mathbb{E}_{\boldsymbol{x} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})} |f(\boldsymbol{x})|^k < \infty,$$
- $$\mathbb{E}_{\boldsymbol{x} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})} \|\boldsymbol{\nabla}_f(\boldsymbol{x})\|_2^k < \infty, \text{where } \boldsymbol{\nabla}_f(\boldsymbol{x}) = \left\{ \frac{\partial f(\boldsymbol{x})}{\partial x_l} \right\}_{l=1}^p.$$

CKRR: Limiting risk

Limiting risk[a] Let $z = -\frac{\lambda + g(\tau) - g(0) - \tau g'(0)}{g'(0)}$ with $\tau = \frac{1}{p} \operatorname{tr} \boldsymbol{\Sigma}$.

$$\mathcal{R}_{train} - \mathcal{R}_{train}^{\infty} \to_{prob.} 0,$$

$$\mathcal{R}_{test} - \mathcal{R}_{test}^{\infty} \to_{prob.} 0,$$

$$\mathcal{R}_{train}^{\infty} = \left(\frac{c\lambda m_z}{g'(0)}\right)^2 \frac{n(1+m_z)^2\left(\sigma^2 + \mathbf{var}_f\right) - n m_z (2+m_z) \|\mathbb{E}[\boldsymbol{\nabla}_f]\|^2}{n(1+m_z)^2 - p m_z^2} + \sigma^2 - 2\sigma^2 \frac{c\lambda m_z}{g'(0)}$$

$$\mathcal{R}_{test}^{\infty} = \frac{n(1+m_z)^2\left(\sigma^2 + \mathbf{var}_f\right) - n m_z (2+m_z) \|\mathbb{E}[\boldsymbol{\nabla}_f]\|^2}{n(1+m_z)^2 - p m_z^2} - \sigma^2.$$

[a] K. Elkhalil, A. Kammoun, X. Zhang, M.-S. Alouini and T. Al-Naffouri. Risk Convergence of Centered Kernel Ridge Regression with Large Dimensional Data. Submitted to IEEE Trans. Signal Processing.

Bad news ☹ Minimim prediction risk is achieved by all kernels!! $\sim$ Linear kernel

Good news ☺ kernel/regularizer can be jointly optimized!

Interesting relation between $\mathcal{R}_{train}^{\infty}$ and $\mathcal{R}_{test}^{\infty}$

$$\mathcal{R}_{test}^{\infty} = \left(\frac{c\lambda m_z}{g'(0)} \right)^{-2} \mathcal{R}_{train}^{\infty} - \sigma^2 \left(\frac{g'(0)}{c\lambda m_z} - 1 \right)^2.$$

Consistent estimator of $\widehat{\mathcal{R}}_{test}$

$$\widehat{\mathcal{R}}_{test} = \left(\frac{c\lambda \widehat{m}_z}{g'(0)} \right)^{-2} \widehat{\mathcal{R}}_{train} - \sigma^2 \left(\frac{g'(0)}{c\lambda \widehat{m}_z} - 1 \right)^2,$$

$$\widehat{m}_z = \frac{1}{p} \operatorname{tr} \left(\frac{XX^T}{p} - z\boldsymbol{I}_n \right)^{-1}.$$

Issues with λ small ☹

CKRR: Consistent estimator of the prediction risk

Consistent estimator of $\widehat{\mathcal{R}}_{test}$

$$\widehat{\mathcal{R}}_{test} = \frac{1}{(cz\widehat{m}_z)^2}\left[\frac{1}{np}\mathbf{y}^T\mathbf{P}\mathbf{X}\left(z\widetilde{\mathbf{Q}}_z^2 - \widetilde{\mathbf{Q}}_z\right)\mathbf{X}^T\mathbf{P}\mathbf{y} + \mathbf{var}\left(\mathbf{y}\right)\right] - \sigma^2.$$

$$\widetilde{\mathbf{Q}}_z = \left(\frac{\mathbf{X}^T\mathbf{P}\mathbf{X}}{p} - z\mathbf{I}_p\right)^{-1}.$$

More stable with respect to λ ☺

$$\widehat{\mathcal{R}}_{test}^{\star} = \min_{z \notin \mathrm{Supp}\{XX^T/p\}} \widehat{\mathcal{R}}_{test}(z)\,,\ z^{\star} = -\frac{\lambda^{\star} + g^{\star}(\tau) - g^{\star}(0) - \tau g'^{\star}(0)}{g'^{\star}(0)}.$$

Kernels

- Linear kernels: $k\left(\mathbf{x}, \mathbf{x}'\right) = \alpha \mathbf{x}^T \mathbf{x}'/p + \beta$.
- Polynomial kernels: $k\left(\mathbf{x}, \mathbf{x}'\right) = \left(\alpha \mathbf{x}^T \mathbf{x}'/p + \beta\right)^d$.
- Sigmoid kernels: $k\left(\mathbf{x}, \mathbf{x}'\right) = \tanh\left(\alpha \mathbf{x}^T \mathbf{x}'/p + \beta\right)$.
- Exponential kernels: $k\left(\mathbf{x}, \mathbf{x}'\right) = \exp\left(\alpha \mathbf{x}^T \mathbf{x}'/p + \beta\right)$.

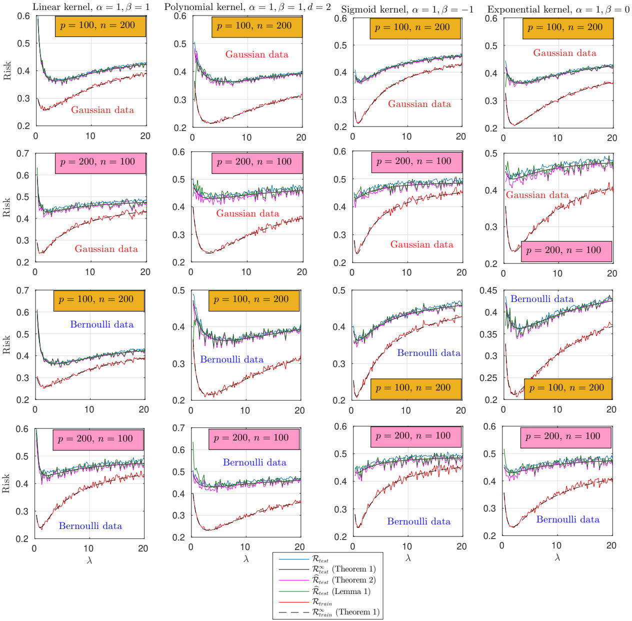
Synthetic data

- $\boldsymbol{x} \sim \boldsymbol{\Sigma}^{\frac{1}{2}} \boldsymbol{z}$ with $\boldsymbol{z}\, \{z_i\}_{i=1}^p$, $\mathbb{E}\, z_i = 0$, $\mathrm{var} z_i = 1$ and $\mathbb{E}\, z_i^k = O(1)$.
- Generating function: $f\left(\boldsymbol{x}\right) = \sin\left(\frac{\mathbf{1}^T \boldsymbol{x}}{\sqrt{p}}\right)$.

Real data

- Communities and Crime dataset.
- $p = 122$, $n_{train} = 73$ and $n_{test} = 50$.
- Prediction risk is computed by averaging over 500 data shuffling.

CKRR: Synthetic data



CKRR: Synthetic data

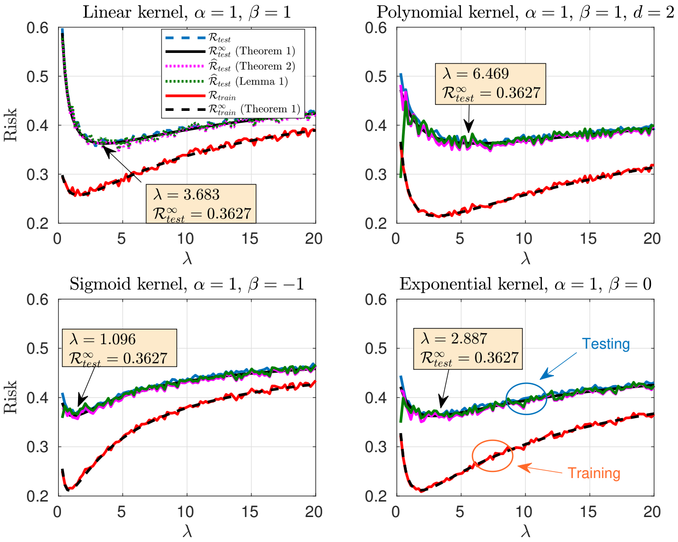


Figure 4.1: CKRR risk with respect to the regularization parameter λ on Gaussian data $(\boldsymbol{x} \sim \mathcal{N}(0_p, \{0.4^{|i-j|}\}_{i,j})$, $n = 200$ training samples and $p = 100$ predictors.

CKRR: Real data

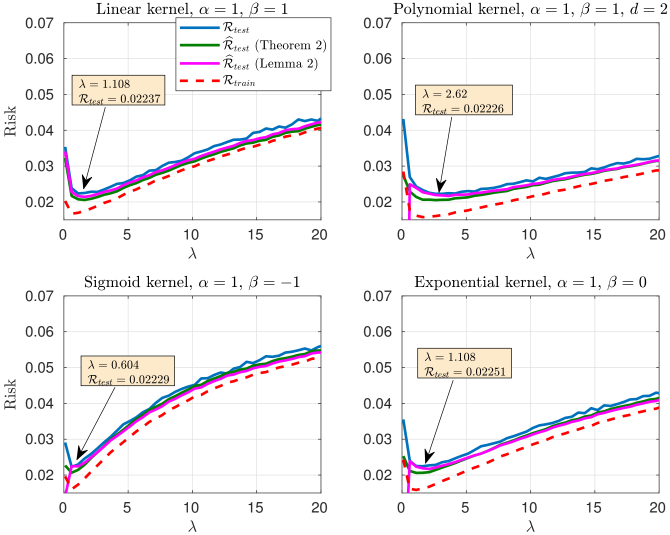


Figure 4.2: CKRR risk with respect to λ where independent zero mean Gaussian noise samples with variance $\sigma^2 = 0.05$ are added to the true response.

Conclusion

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
- Successful application of this tool has been demonstrated in the context of RDA.
- Fundamental limits of Centered kernel ridge regression.

Future research directions

Future research directions

- We can also consider the performance analysis of kernel LDA/QDA.
- Extend the analysis to *Homogenous* kernels.

- $\phi(\mathbf{x})$ is a fixed non linear feature space mapping. The kernel function is given by

$$k(\mathbf{x}, \mathbf{x}') = \phi(\mathbf{x})^T \phi(\mathbf{x}') .$$

- Homogeneous kernels

$$k(\mathbf{x}, \mathbf{x}') = f\left(\frac{\|\mathbf{x} - \mathbf{x}'\|^2}{p}\right) .$$

- $\{\mathbf{K}\}_{i,j} = k(\mathbf{x}_i, \mathbf{x}_j)$.

Theorem (Spectrum of kernel random matrices, El-Karoui 2010)

[10] *[Informal statement]*

$$\widehat{K} = f(\tau)\,\mathbf{1}\mathbf{1}^T + f'(\tau)\,\mathbf{W} + f''(\tau)\,\mathbf{Q}, \quad \frac{\|\mathbf{x} - \mathbf{x}'\|^2}{p} \to_{a.s.} \tau.$$

$$\boxed{\|\mathbf{K} - \widehat{\mathbf{K}}\| \xrightarrow{p} 0.} \tag{7}$$

This might help to analyze the performance of some kernel methods in regression or classification.

[10] N. El Karoui, The spectrum of kernel random matrices, The annals of statistics, Volume 38, Number 1 (2010), 1-50.

That’s it

Thank you for your time and attention!