QUDA: A Direct Approach for Sparse Quadratic Discriminant Analysis

Chenlei Leng

Warwick Statistics

Joint with Binyan Jiang (HKPU) and Xiangyu Wang (Duke)

OxWaSP Workshop

9 Oct, 2015

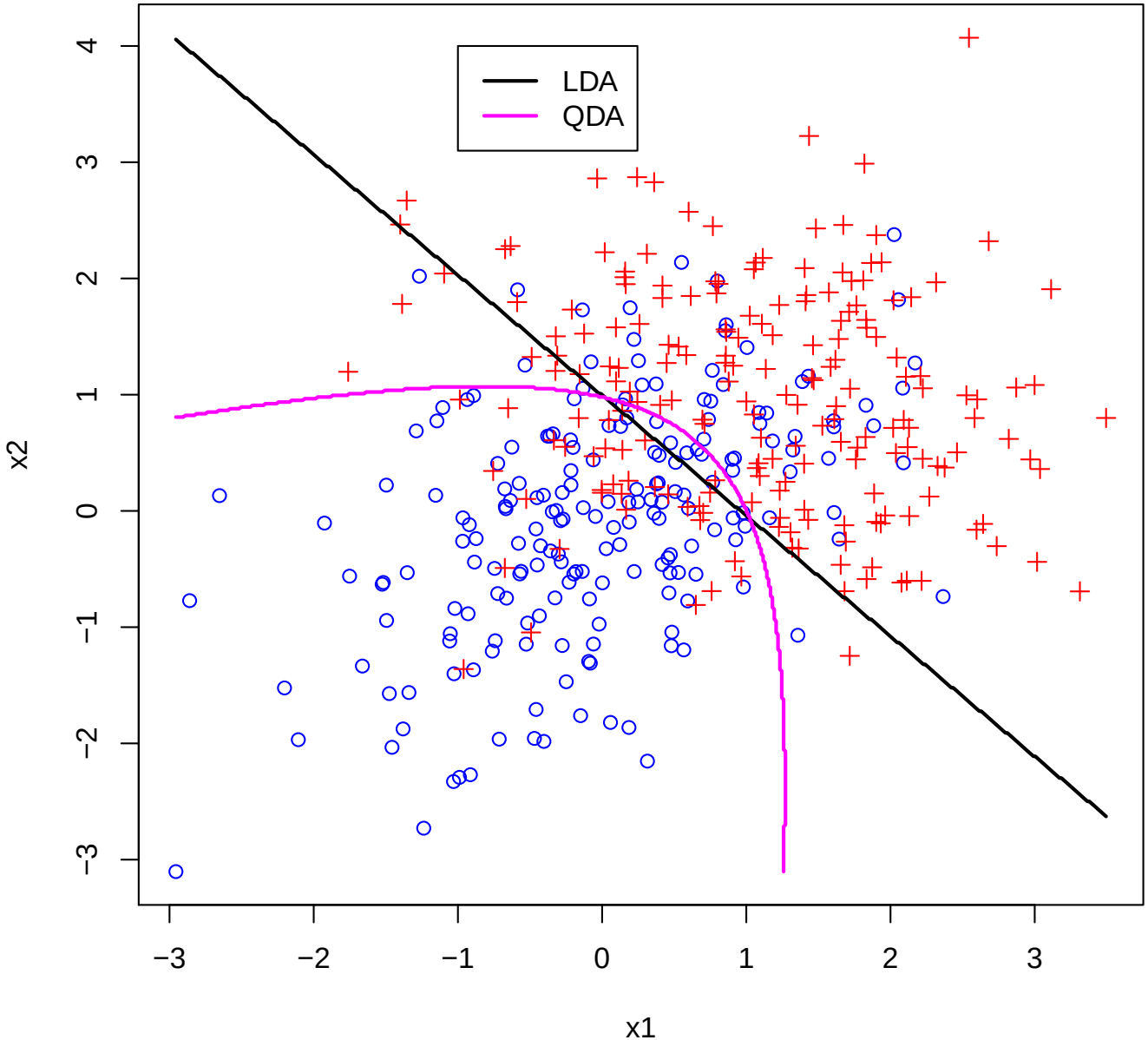
Motivation

Warwick Statistics

OxWaSP: Statistical Science for 21st Century data-intensive environments and large-scale models

- A paradigm shift from hypothesis driving to data driven scientific research: Huge datasets are collected in -omics, brain imaging, astronomy, engineering, and so on.
- The numbers of the features (variables) are huge relative to the sample size: "Fat Data".
- Features interact in interesting ways.
- The aim of this talk is to introduce a "simple model" for "big data" classification.

LDA and QDA



The setup

Warwick Statistics

Consider the classification problem in which

$$\text{observations from class 1 follow } X \sim N(\mu_1, \Sigma_1),$$
$$\text{observations from class 2 folllow } Y \sim N(\mu_2, \Sigma_2),$$

where

- $\mu_1, \mu_2 \in \mathbb{R}^p$ are the mean vectors
- $\Sigma_1, \Sigma_2 \in \mathbb{R}^{p \times p}$ are the covariance matrices.

Given a dataset, the goal is to classify a future observation z to one of the two classes such that the classification error is made as small as possible.

Bayes' rule

Warwick Statistics

The optimal Bayes rule classifies an observation z to class 1 if

$$\pi_1 f(z|\mu_1, \Sigma_1) > \pi_2 f(z|\mu_2, \Sigma_2)$$

where $f(\cdot|\mu, \Sigma)$ is the multivariate normal pdf with mean μ and Σ, and π_1 and π_2 are the two prior probabilities.

- When $\Sigma_1 = \Sigma_2$, the procedure becomes linear discriminant analysis (LDA);
- When $\Sigma_1 \neq \Sigma_2$, the procedure becomes quadratic discriminant analysis (QDA). The focus of the talk.

Outline

Warwick Statistics

- The problem (with high-dimensional data)
- QUDA
- Simulation and data analysis
- Theory
- Discusssion

The oracle

The Bayes discriminant function consists of z's satisfying

$$\pi_1 f(z|\mu_1, \Sigma_1) = \pi_2 f(z|\mu_2, \Sigma_2)$$

or

$$D(z) = (z-\mu)^T \Omega (z-\mu) + \delta^T (z-\mu) + \eta,$$

where

- $\mu = (\mu_1 + \mu_2)/2$: the mean of the two centroids,
- $\Omega = \Sigma_2^{-1} - \Sigma_1^{-1}$: the difference of the two precision matrices,
- $\delta = (\Sigma_1^{-1} + \Sigma_2^{-1})(\mu_1 - \mu_2)$,
- $\eta = 2\log(\pi_1/\pi_2) - \frac{1}{4}(\mu_1-\mu_2)^T \Omega (\mu_1-\mu_2) - \log|\Sigma_2| + \log|\Sigma_1|$.

Reduces to that of LDA if $\Sigma_1 = \Sigma_2$.

Remarks

The discriminant function

$$D(z) = (z - \mu)^T \Omega (z - \mu) + \delta^T (z - \mu) + \eta.$$

- This can be seen as a two-way interaction model as in regression, hence the name QDA
- Nonzeros in Ω are the important interactions
- Nonzeros in δ are the important main effects

Sample estimates

Warwick Statistics

Given data $X_j,\ j = 1, .., n_1$ from class 1 and $Y_k,\ k = 1, ..., n_2$ from class 2, we can estimate μ_i and $\Sigma_i,\ i = 1, 2$, as

$$\hat{\mu}_1 = \frac{1}{n_1}\sum_{j=1}^{n_1} X_j,\ \hat{\mu}_2 = \frac{1}{n_2}\sum_{j=1}^{n_2} Y_j;$$

$$\hat{\Sigma}_1 = \frac{1}{n_1}\sum_{j=1}^{n_1}(X_j - \hat{\mu}_1)(X_j - \hat{\mu}_1)^T,\ \hat{\Sigma}_2 = \frac{1}{n_2}\sum_{j=1}^{n_2}(Y_j - \hat{\mu}_2)(Y_j - \hat{\mu}_2)^T;$$

$$\hat{\pi}_1 = n_1/(n_1 + n_2),\ \hat{\pi}_2 = n_2/(n_1 + n_2).$$

The plug-in estimates

- $\hat{\mu} = (\hat{\mu}_1 + \hat{\mu}_2)/2$
- $\hat{\Omega} = \hat{\Sigma}_2^{-1} - \hat{\Sigma}_1^{-1}$
- $\hat{\delta} = (\hat{\Sigma}_1^{-1} + \hat{\Sigma}_2^{-1})(\hat{\mu}_1 - \hat{\mu}_2)$
- $\hat{\eta} = 2\log(\hat{\pi}_1/\hat{\pi}_2) - \frac{1}{4}(\hat{\mu}_1 - \hat{\mu}_2)^T \hat{\Omega}(\hat{\mu}_1 - \hat{\mu}_2) - \log|\hat{\Sigma}_2| + \log|\hat{\Sigma}_1|$

Our focus

Warwick Statistics

Problem: High-dimensional $p \gg \max\{n_1, n_2\}$.

- $\hat{\mu} = (\hat{\mu}_1 + \hat{\mu}_2)/2$: ?
- $\hat{\Omega} = \hat{\Sigma}_2^{-1} - \hat{\Sigma}_1^{-1}$: ?
- $\hat{\delta} = (\hat{\Sigma}_1^{-1} + \hat{\Sigma}_2^{-1})(\hat{\mu}_1 - \hat{\mu}_2)$: ?
- $\hat{\eta} = 2\log(\hat{\pi}_1/\hat{\pi}_2) - \frac{1}{4}(\hat{\mu}_1 - \hat{\mu}_2)^T\hat{\Omega}(\hat{\mu}_1 - \hat{\mu}_2) - \log|\hat{\Sigma}_2| + \log|\hat{\Sigma}_1|$: ?

Estimating Ω

Warwick Statistics

Assume Ω sparse. Note

$$\Sigma_2\Omega\Sigma_1 = \Sigma_1\Omega\Sigma_2 = \Sigma_1 - \Sigma_2.$$

If we define a loss function as

$$Tr\left(\Omega^T\Sigma_1\Omega\Sigma_2\right) - Tr\left(\Omega(\Sigma_1 - \Sigma_2)\right),$$

which is minimised when $\Omega = \Sigma_2^{-1} - \Sigma_1^{-1}$.

Define the estimator as

$$\hat{\Omega} = \arg\min_{\Omega \in R^{p\times p}} \frac{1}{2} Tr\left(\Omega^T\hat{\Sigma}_1\Omega\hat{\Sigma}_2\right) - Tr\left(\Omega(\hat{\Sigma}_1 - \hat{\Sigma}_2)\right) + \lambda\|\Omega\|_1,$$

where $\|\Omega\|_1$ is the ℓ_1 penalty of the vectorized Ω.

Algorithm

The formulation is a convex optimisation problem. Use the alternating direction method of multipliers (ADMM) by writing

$$\min_{\Omega \in R^{p\times p}} \frac{1}{2} Tr\left(\Omega^T \hat{\Sigma}_1 \Omega \hat{\Sigma}_2\right) - Tr\left(\Omega(\hat{\Sigma}_1 - \hat{\Sigma}_2)\right) + \lambda \|\Psi\|_1, \ s.t. \ \Psi = \Omega.$$

The augmented Lagrangian as

$$L(\Omega, \Psi, \Lambda) = \frac{1}{2} Tr\left(\Omega^T \hat{\Sigma}_1 \Omega \hat{\Sigma}_2\right) - Tr\left(\Omega(\hat{\Sigma}_1 - \hat{\Sigma}_2)\right) + \lambda \|\Psi\|_1 + Tr(\Lambda, \Omega - \Psi) + \frac{\rho}{2}\|\Omega - \Psi\|_F^2.$$

Given the current estimate $\Omega^k, \Psi^k, \Lambda^k$, we update successively

$$\Omega^{k+1} = \arg\min_{\Omega \in R^{p\times p}} L(\Omega, \Psi^k, \Lambda^k),$$

$$\Psi^{k+1} = \arg\min_{\Psi \in R^{p\times p}} L(\Omega^{k+1}, \Psi, \Lambda^k),$$

$$\Lambda^{k+1} = \Lambda^k + \rho(\Omega^{k+1} - \Psi^{k+1}).$$

Algorithm

Warwick Statistics

Turns out the updates for Ω and Ψ have closed-form solutions.

1. Initialize Ω, Ψ and Λ. Fix ρ. Compute the singular value decomposition $\hat{\Sigma}_1 = U_1 D_1 U_1^T$ and $\hat{\Sigma}_2 = U_2 D_2 U_2^T$, and compute B where $B_{jk} = 1/(d_{1j} d_{2k} + \rho)$. Repeat steps 2-4 until convergence;

2. Compute $A = (\hat{\Sigma}_1 - \hat{\Sigma}_2) - \Lambda + \rho\Psi$. Then update Ω as $\Omega = U_1[B \circ (U_1^T A U_2)]U_2^T$;

3. Update Ψ by soft-thresholding $\Omega + \frac{\Lambda}{\rho}$ elementwise by $\frac{\lambda}{\rho}$;

4. Update Λ by $\Lambda \leftarrow \Lambda + \rho(\Omega - \Psi)$.

The convergence properties of ADMM are well studied. Improves the computational complexity of the algorithm in Zhao, Cai and Li (Bka, 2014) from $O(p^4)$ to $O(p^3)$.

Estimating the linear index δ

Warwick Statistics

Recall $\delta = (\Sigma_1^{-1} + \Sigma_2^{-1})(\mu_1 - \mu_2)$. Note that

$$(\Sigma_1 + \Sigma_2)\delta = 4(\mu_1 - \mu_2) + (\Sigma_1 - \Sigma_2)\Omega(\mu_1 - \mu_2).$$

A similar trick leads to the following estimator

$$\hat{\delta} = \arg\min_{\delta \in R^p} \frac{1}{2}\delta^T(\hat{\Sigma}_1 + \hat{\Sigma}_2)\delta - \hat{\gamma}^T\delta + \lambda_\delta\|\delta\|_1,$$

where $\hat{\gamma} = 4(\hat{\mu}_1 - \hat{\mu}_2) + (\hat{\Sigma}_1 - \hat{\Sigma}_2)\hat{\Omega}(\hat{\mu}_1 - \hat{\mu}_2)$ and $\|\cdot\|_1$ is the vector ℓ_1 penalty.

This is exactly lasso!

Literature review

Warwick Statistics

- Linear classifiers for high-dimensional data widely studied, Bickel and Levina (Bernoulli, 2004), Witten and Tibshirani (JRSSB, 2011), Shao et al. (AoS, 2011), Cai and Liu (JASA, 2011), Fan, Feng, and Tong (JRSSB, 2012), Mai, Zou, and Yuan (Bka, 2012).
- Quadratic classifiers attract attention more recently, Li and Shao (2015), Fan et al. (AoS, 2015).
- Li and Shao: Too many assumptions
- Fan et al: A two-step method by first screening then penalised logistic regression, work for Ω with a special structure
- Zhao, Cai and Li (Bka, 2014): Apply Dantzig selector for estimating Ω, computationally demanding

Numerical studies

Warwick Statistics

Methods to be compared

- QUDA
- LDA and QDA whenever possible
- DSDA (Mai, Zou and Yuan, 2012)
- Penalised logistic regression with ℓ_1 penalty, main effects only (PLR), main and interactions (PLR2)
- IIS-SQDA (Fan et al. AoS, 2015)

Simulation Setup

Warwick Statistics

$p = 50,\ 200$, or 500, $n_1 = n_2 = 100$.
$X \sim N(u_1, \Sigma_1)$ and $Y \sim N(u_2, \Sigma_2)$ with $u_2 = 0$ and $u_1 = \Sigma_1 \beta$ with $\beta = (0.6, 0.8, 0, \cdots, 0)^T$. Denote $\Omega_i = \Sigma_i^{-1}$.

- Model 1: Ω_1 is a band matrix with $(\Omega_1)_{ii} = 1$ and $\Omega_{ij} = 0.3$ for $|i - j| = 1$. $\Omega_2 = \Omega_1 + \Omega$, where a 3×3 submatrix of Ω is nonzero.
- Model 2: We set $(\Omega_1)_{ij} = 0.5^{|i-j|}$ and let $\Omega_2 = \Omega_1 + \Omega$, where $\Omega = I_p$.
- Model 3: Ω_1 is the same as Model 2 and $\Omega_2 = \Omega_1$.
- Model 4: Ω_1 is the same as Model 2 and Ω is a band matrix defined as $(\Omega)_{ii} = 1$ and $(\Omega)_{ij} = 0.5$ for $|i - j| = 1$. Let $\Omega_2 = \Omega_1 + \Omega$.
- Model 5: $\Omega_1 = I_p$ and $\Omega_2 = \Omega_1 + \Omega$ where Ω is a random sparse symmetric matrix with conditional number 10 and non-zero density $n_1/p^2 \times 0.7$.

Table 1: Model 1: a $3 * 3$ dense submatrix

p	Method	MR (%)	FP.main	FP.inter	FN.main	FN.inter
50	LDA	39.43 (0.15)	–	–	–	–
	QDA	43.47 (0.10)	–	–	–	–
	PLR	36.12 (0.26)	5.95 (0.93)	–	1.21 (0.04)	–
	DSDA	35.05 (0.22)	8.81 (1.06)	–	0.07 (0.03)	–
	PLR2	30.15 (0.44)	0.51 (0.14)	11.26 (2.78)	0.60 (0.05)	2.62 (0.09)
	IIS-SQDA	27.56 (0.27)	5.60 (0.82)	2.16 (0.32)	0.19 (0.04)	2.05 (0.09)
	QUDA	26.50 (0.28)	0.85 (0.18)	35.26 (4.72)	0.39 (0.07)	3.74 (0.14)
	Oracle	23.04 (0.09)	–	–	–	–
200	PLR	37.62 (0.34)	7.82 (1.87)	–	1.47 (0.05)	–
	DSDA	36.34 (0.30)	15.06 (3.37)	–	0.36 (0.05)	–
	PLR2	32.55 (0.53)	0.25 (0.06)	17.44 (3.63)	0.90 (0.05)	2.72 (0.08)
	IIS-SQDA	26.94 (0.31)	6.43 (1.24)	0.78 (0.17)	0.42 (0.05)	2.22 (0.08)
	QUDA	26.51 (0.20)	0.29 (0.07)	25.48 (2.75)	0.82 (0.08)	4.14 (0.12)
	Oracle	21.93 (0.08)	–	–	–	–
500	PLR	38.82 (0.33)	9.31 (1.99)	–	1.58 (0.05)	–
	DSDA	37.10 (0.29)	16.06 (3.02)	–	0.42 (0.05)	–
	PLR2	35.45 (0.64)	0.34 (0.09)	55.69 (12.67)	0.99 (0.05)	3.05 (0.10)
	IIS-SQDA	26.78 (0.31)	3.22 (1.09)	0.23 (0.05)	0.98 (0.02)	2.65 (0.09)
	QUDA	26.68 (0.27)	0.14 (0.06)	10.96 (1.38)	1.02 (0.08)	4.36 (0.09)
	Oracle	21.81 (0.09)	–	–	–	–

Table 2: Model 2: $\Omega = I$

p	Method	MR (%)	FP.main	FP.inter	FN.main	FN.inter
50	LDA	34.53 (0.19)	–	–	–	–
	QDA	32.09 (0.25)	–	–	–	–
	PLR	31.58 (0.20)	7.51 (0.55)	–	0.07 (0.03)	–
	DSDA	29.89 (0.16)	8.52 (0.86)	–	0.16 (0.04)	–
	PLR2	5.85 (0.10)	1.14 (0.11)	45.6 (1.08)	0.14 (0.04)	14.43 (0.23)
	IIS-SQDA	11.75 (0.13)	11.41 (0.46)	11.8 (0.56)	0 (0)	37.53 (0.11)
	QUDA	1.84 (0.08)	4.12 (0.49)	110.10 (10.54)	0.28 (0.05)	1.28 (0.22)
	Oracle	0.65 (0.02)	–	–	–	–
200	PLR	33.34 (0.21)	10.79 (0.70)	–	0.16 (0.04)	–
	DSDA	30.37 (0.23)	11.91 (2.19)	–	0.29 (0.05)	–
	PLR2	1.73 (0.06)	0.01 (0.01)	12.68 (0.56)	1.08 (0.05)	119.95 (0.52)
	IIS-SQDA	11.76 (0.18)	25.39 (0.66)	4.12 (0.30)	0 (0)	186.35 (0.12)
	QUDA	0.39 (0.18)	9.03 (2.12)	724.35 (19.52)	0.21 (0.04)	6.05 (0.35)
	Oracle	0 (0)	–	–	–	–
500	PLR	34.04 (0.24)	11.17 (1.02)	–	0.30 (0.05)	–
	DSDA	30.99 (0.22)	14.61 (2.64)	–	0.44 (0.05)	–
	PLR2	1.68 (0.06)	0 (0)	5.52 (0.33)	1.19 (0.05)	401.47 (0.59)
	IIS-SQDA	12.37 (0.16)	33.56 (0.79)	1.92 (0.20)	0 (0)	485.93 (0.12)
	QUDA	0.16 (0.22)	24.33 (2.18)	4.81e3 (290.1)	0.52 (0.05)	58.09 (1.10)
	Oracle	0 (0)	–	–	–	–

Table 3: Model 3: $\Omega = 0$

p	Method	MR (%)	FP.main	FP.inter	FN.main	FN.inter
50	LDA	38.82 (0.19)	–	–	–	–
	QDA	47.57 (0.11)	–	–	–	–
	PLR	36.06 (0.23)	7.73 (0.58)	–	0.14 (0.03)	–
	DSDA	34.82 (0.24)	9.54 (1.09)	–	0.26 (0.04)	–
	PLR2	37.36 (0.34)	0.60 (0.10)	31.10 (3.21)	0.39 (0.06)	0 (0)
	IIS-SQDA	35.10 (0.22)	5.25 (0.46)	10.85 (0.96)	0.06 (0.02)	0 (0)
	QUDA	34.99 (0.58)	0.82 (0.20)	23.84 (6.69)	0.35 (0.07)	0 (0)
	Oracle	31.68 (0.10)	–	–	–	–
200	PLR	38.50 (0.31)	12.90 (1.08)	–	0.23 (0.04)	–
	DSDA	36.27 (0.28)	14.81 (2.26)	–	0.41 (0.05)	–
	PLR2	40.31 (0.45)	0.15 (0.05)	40.38 (5.05)	0.74 (0.06)	0 (0)
	IIS-SQDA	36.32 (0.25)	25.39 (0.66)	6.03 (0.50)	0 (0)	0 (0)
	QUDA	36.55 (0.74)	1.70 (1.38)	37.15 (16.39)	0.89 (0.09)	0 (0)
	Oracle	31.54 (0.10)	–	–	–	–
500	PLR	39.98 (0.32)	14.79 (1.41)	–	0.40 (0.05)	–
	DSDA	37.07 (0.29)	19.49 (3.65)	–	0.59 (0.05)	–
	PLR2	42.23 (0.53)	0.03 (0.02)	36.6 (4.32)	1.07 (0.06)	0 (0)
	IIS-SQDA	37.45 (0.26)	14.53 (1.38)	3.70 (0.32)	0.07 (0.26)	0 (0)
	QUDA	37.95 (0.76)	0.2 (0.06)	57.49 (14.74)	1.05 (0.09)	0 (0)
	Oracle	31.85 (0.12)	–	–	–	–

Table 4: Model 4: Ω is tridiagonal

p	Method	MR (%)	FP.main	FP.inter	FN.main	FN.inter
50	LDA	35.58 (0.20)	–	–	–	–
	QDA	35.40 (0.20)	–	–	–	–
	PLR	32.42 (0.23)	8.03 (0.57)	–	0.03 (0.01)	–
	DSDA	31.39 (0.21)	11.02 (1.13)	–	0.09 (0.03)	–
	PLR2	22.42 (0.21)	1.49 (0.14)	76.22 (2.26)	0.06 (0.03)	123.72 (0.36)
	IIS-SQDA	24.67 (0.17)	9.74 (0.45)	17.44 (0.86)	0 (0)	136.56 (0.20)
	QUDA	16.91 (0.27)	0.55 (0.14)	194.98 (11.31)	0.61 (0.08)	106.51 (0.83)
	Oracle	3.22 (0.04)	–	–	–	–
200	PLR	34.93 (0.28)	12.71 (0.88)	–	0.10 (0.03)	–
	DSDA	32.64 (0.26)	15.63 (2.14)	–	0.21 (0.04)	–
	PLR2	21.82 (0.20)	0.30 (0.05)	107.80 (2.32)	0.40 (0.05)	559.23 (0.63)
	IIS-SQDA	25.25 (0.20)	21.15 (0.89)	8.56 (0.61)	0 (0)	586.60 (0.13)
	QUDA	9.59 (0.19)	0.31 (0.08)	297.38 (25.33)	0.82 (0.09)	498.61 (1.49)
	Oracle	0.28 (0.02)	–	–	–	–
500	PLR	37.19 (0.32)	15.68 (1.27)	–	0.32 (0.04)	–
	DSDA	33.83 (0.30)	22.90 (3.54)	–	0.45 (0.05)	–
	PLR2	23.06 (0.23)	0.05 (0.02)	114.94 (2.34)	0.79 (0.05)	1455 (0.65)
	IIS-SQDA	26.64 (0.21)	32.74 (1.24)	4.86 (0.36)	0 (0)	1486 (0.13)
	QUDA	4.18 (0.13)	0.20 (0.04)	298.24 (20.8)	0.42 (0.07)	1315 (2.41)
	Oracle	0 (0)	–	–	–	–

Table 5: Model 5: Ω is a random sparse matrix with 70 nonzeros

p	Method	MR (%)	FP.main	FP.inter	FN.main	FN.inter
50	LDA	39.21 (0.20)	–	–	–	–
	QDA	46.41 (0.17)	–	–	–	–
	PLR	35.76 (0.26)	6.08 (0.43)	–	0.01 (0.01)	–
	DSDA	33.73 (0.25)	8.08 (0.99)	–	0.14 (0.04)	–
	PLR2	36.62 (0.39)	1.04 (0.13)	45.83 (3.99)	0.05 (0.02)	63.69 (0.39)
	IIS-SQDA	35.56 (0.29)	8.77 (0.50)	14.85 (0.83)	0 (0)	61.18 (0.26)
	QUDA	34.32 (0.53)	0.52 (0.12)	39.76 (6.47)	0.58 (0.08)	59.76 (0.64)
	Oracle	32.36 (0.25)	–	–	–	–
200	PLR	37.73 (0.34)	9.68 (0.89)	–	0.40 (0.03)	–
	DSDA	34.58 (0.35)	10.87 (2.44)	–	0.11 (0.03)	–
	PLR2	37.40 (0.44)	0.32 (0.06)	66.44 (5.47)	0.31 (0.06)	194.46 (0.35)
	IIS-SQDA	33.22 (0.28)	19.87 (0.93)	6.16 (0.41)	0 (0)	191.37 (0.10)
	QUDA	29.35 (0.41)	0.10 (0.05)	164.24 (73.3)	1.27 (0.07)	175.8 (0.96)
	Oracle	20.09 (0.27)	–	–	–	–
500	PLR	39.13 (0.33)	14.39 (1.29)	–	0.08 (0.03)	–
	DSDA	34.76 (0.25)	9.44 (1.77)	–	0.16 (0.04)	–
	PLR2	37.44 (0.52)	0.16 (0.05)	90.78 (6.06)	0.43 (0.06)	493.48 (0.41)
	IIS-SQDA	30.50 (0.26)	33.55 (1.28)	3.18 (0.28)	0 (0)	491.32 (0.09)
	QUDA	23.75 (0.49)	4.03 (2.91)	507.92 (225.36)	1.59 (0.06)	459.96 (1.96)
	Oracle	4.16 (0.08)	–	–	–	–

Data analysis: A "small data"

Warwick Statistics

Quora answer classifier.

- This is a data challenge available at http://www.quora.com/challenges#answer_classifier.
- The training data set contains 4,500 answers from QUORA which have been annotated with either "good" or "bad".
- For each answer, 20 features were extracted from the original sentences.
- The goal of this challenge is to automatically classify a new answer based on the 20 features.

Results

Table 6: Misclassification rate (%) using 5-fold cross-validation

Method	mean	standard error
LDA	18.84	0.50
QDA	30.33	0.72
PLR	17.89	0.60
DSDA	19.11	0.56
PLR2	17.56	0.71
IIS-SQDA	17.76	0.54
QUDA	**16.44**	0.45

Data analysis: A "big data"

Warwick Statistics

Prostate cancer data, taken from Singh, et al. (Cancer Cell, 2002)

- The data set contains genetic expression levels for $N = 6033$ genes
- The sample size is 102 men with 50 normal control subjects and 52 prostate cancer patients.
- The goal is to identify genes that are linked with prostate cancer and predict potential patients.
- LDA and QDA not applicable as $n < p$.
- The difficulty lies in the interactions among genes.

Interaction

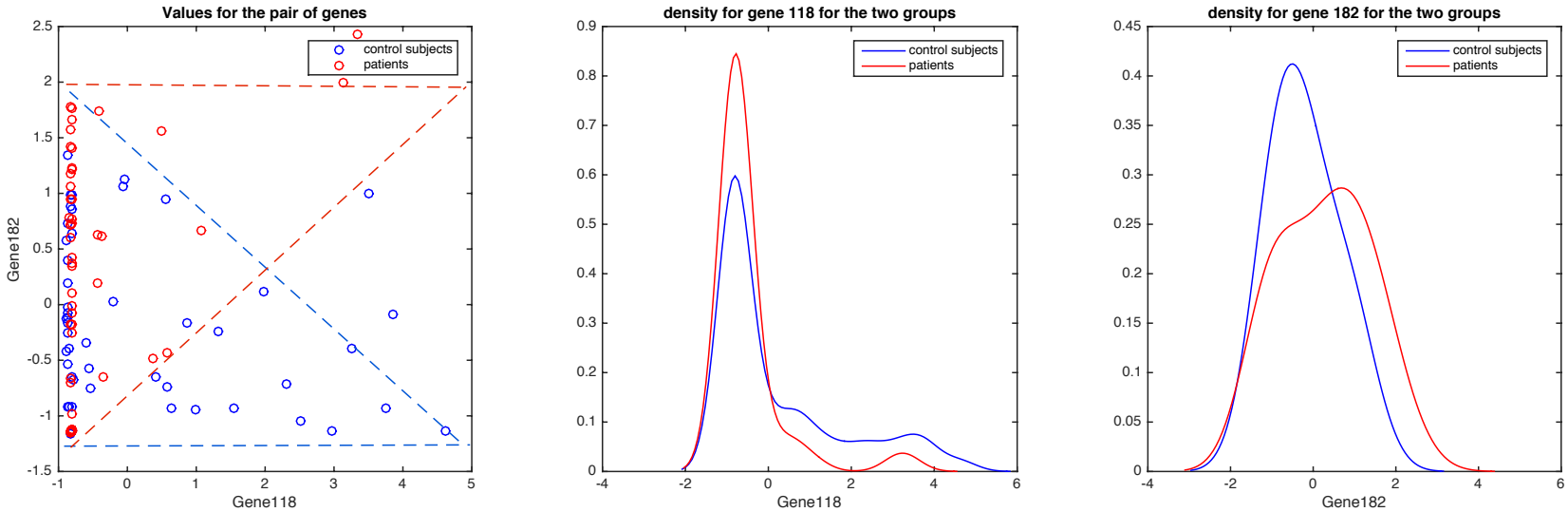


Figure 1: The plot for the gene 118 and gene 182. Left: joint scatter plot; Middle: marginal density of gene 118; Right: marginal density of gene 182.

Result

Look at the top 200 or 500 genes with the largest absolute values of the two sample t statistics.

Table 7: Misclassification rate (%) for the prostate cancer data under 5-fold cross-validation

	$p = 200$		$p = 500$	
Method	mean	std error	mean	std error
PLR	11.00	2.45	13.00	4.06
DSDA	5.00	3.32	11.00	2.92
PLR2	13.00	4.47	23.00	3.39
IIS-SQDA	12.00	2.92	15.00	2.74
QUDA	**1.00**	1.00	**2.00**	1.22

For estimating Ω, assume the irrepresentable condition on $\Gamma = \Sigma_2 \otimes \Sigma_1$ as

$$\alpha = 1 - \max_{e \in S^c} |\Gamma_{e,S} \Gamma_{S,S}^{-1}|_1 > 0.$$

For estimating δ, assume the irrepresentable condition on $\Sigma = (\Sigma_1 + \Sigma_2)/2$ as

$$\alpha_\delta = 1 - \max_{e \in S^c} |\Sigma_{e,D} \Sigma_{D,D}^{-1}|_1 > 0.$$

Results on estimating Ω

Warwick Statistics

Theorem 1

By choosing

$$\lambda = A_1 \sqrt{\frac{\kappa \log p + \log C_1}{C_2 n}},$$

for some $\kappa > 2$ *and*

$$n > A_2(\kappa \log p + \log C_1),$$

we have, with probability greater than $1 - p^{2-\kappa}$,
(i) $\hat{\Omega}_{S^c} = 0$;
(ii)

$$||\hat{\Omega} - \Omega||_\infty < A_3 \sqrt{\frac{\kappa \log p + \log C_1}{C_2 n}}$$

where A_1, A_2 *and* A_3 *are quantities depending on the sparsity* Σ_1, Σ_2 *and their Kronecker product, and* C_1, C_2 *are constant.*

Corollary 2

Under appropriate conditions, for any constant $\kappa > 0$*, choosing* $\lambda = Cd^2\sqrt{\frac{\log p}{n}}$ *for some constant* $C > 0$*, if* $d^2\sqrt{\frac{\log p}{n}} \to 0$*, we have with probability greater than* $1 - p^{2-\kappa}$*,* $\hat{\Omega}_{S^c} = 0$ *and*

$$||\hat{\Omega} - \Omega||_\infty = O\left(d^2\sqrt{\frac{\log p}{n}}\right).$$

Here d is the sparsity index of Ω.

Theorem 3

Under the same assumptions in Theorem 1 and assuming that

$$\lambda_\delta = B_1 \sqrt{\frac{\kappa \log p + \log C_1}{C_{2\delta} n}},$$

and

$$n > B_2 \cdot (\kappa \log p + \log C_1),$$

we have with probability greater than $1 - p^{2-\kappa}$,
(i) $\hat{\delta}_{D^c} = 0$;
(ii)

$$||\hat{\delta} - \delta||_\infty < B_3 \sqrt{\frac{\kappa \log p + \log C_1}{C_{2\delta} n}},$$

where B_1, B_2 *and* B_3 *depend on various quantities of the true parameters.*

Corollary 4

Let $d_0 = \max\{d, d_\delta\}$. *Under additional assumptions, for any constant* $\kappa > 0$, *by choosing* $\lambda = Cd_0^2\sqrt{\frac{\log p}{n}}$ *for some constant* $C > 0$, *and assume that* $d_0^3\sqrt{\frac{\log p}{n}} \to 0$, *we have with probability greater than* $1 - p^{2-\kappa}$,

$$||\hat{\delta} - \delta||_\infty = O\left(d_0^3\sqrt{\frac{\log p}{n}}\right).$$

Risk

- Let $R(i|j)$ be the probabilities that a new observation from class i is misclassified to class j by Bayes' rule. The Bayes risk

$$R = \pi_1 R(2|1) + \pi_2 R(1|2).$$

- $R_n(i|j)$ be the probabilities that a new observation from class i is misclassified to class j by QUDA. The misclassification rate of the QUDA rule is

$$R_n = \pi_1 R_n(2|1) + \pi_2 R_n(1|2).$$

Theorem 5

Under apprpriate assumptions, we have:

(i) if $d^2 A_3 \sqrt{\frac{\kappa \log p + \log C_1}{C_2 n}} + d_\delta B_3 \sqrt{\frac{\kappa \log p + \log C_1}{C_{2\delta} n}} \to 0$, *then*

$$R_n - R = O_p\left(d^2 A_3 \sqrt{\frac{\kappa \log p + \log C_1}{C_2 n}} + d_\delta B_3 \sqrt{\frac{\kappa \log p + \log C_1}{C_{2\delta} n}} \right);$$

(ii) with probability greater than $1 - 3p^{2-\kappa}$ *for some constant* $\kappa > 2$,

$$R_n - R = O\left(\left(d^2 A_3 \log p \sqrt{\frac{\kappa \log p + \log C_1}{C_2 n}} + d_\delta B_3 \sqrt{\log p} \sqrt{\frac{\kappa \log p + \log C_1}{C_{2\delta} n}} \right) \right).$$

Remarks

Warwick Statistics

- Results for estimating Ω similar to those in Zhao, Cai and Li (2014).
- When $\Sigma_1 = \Sigma_2$, results are similar to sparse LDA in Mai, Zou and Yuan (2011) and Cai and Liu (2011).
- The error rate of $\hat{\delta}$ is a factor times of that of $\hat{\Omega}$. It doesn't affect the misclassification error nevertheless.

Discussion

Warwick Statistics

- Lots of room to develop "old simple models" for 21st century data;
- My research tends to blend methodology, computation, theory and application;
- To students: Talk to me if you are looking for projects.

Reference: http://arxiv.org/abs/1510.00084.


KEEP
CALM
AND
USE
QUDA