Sparse Linear Discriminant Analysis With High Dimensional Data

Jun Shao

East China Normal University
University of Wisconsin

Joint work with Yazhen Wang, Xinwei Deng, Sijian Wang

Outline

- Introduction
- Linear discriminant analysis and asymptotic results
- Sparse linear discriminant analysis and asymptotic results
- Application and simulation
- Conclusion and discussion

Introduction

The classification problem

Classify a subject to class 1 or class 2 based on an observed vector $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \Sigma)$

$N_p(\boldsymbol{\mu}, \Sigma)$: the p-dimensional normal distribution with mean vector $\boldsymbol{\mu} = \boldsymbol{\mu}_k$, $k = 1, 2$, and covariance matrix Σ

The dimension of x

In traditional applications, p is small (a few variables)

Modern technologies: a large p (many variables)

- genetic and microarray data
- data from biomedical imaging
- data from signal processing
- climate data
- high-frequency financial data.

Introduction

The classification problem

Classify a subject to class 1 or class 2 based on an observed vector $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

$N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$: the p-dimensional normal distribution with mean vector $\boldsymbol{\mu} = \boldsymbol{\mu}_k$, $k = 1, 2$, and covariance matrix $\boldsymbol{\Sigma}$

The dimension of $\mathbf{x}$

In traditional applications, p is small (a few variables)

Modern technologies: a large p (many variables)

- genetic and microarray data
- data from biomedical imaging
- data from signal processing
- climate data
- high-frequency financial data.

Example: Classifying human acute leukemias into two types

- Gene expression microarray (Golub et al., 1999)
- Two types of human acute leukemias
  - acute myeloid leukemia (AML)
  - acute lymphoblastic leukemia (ALL)
- Distinguishing ALL from AML is crucial for successful treatment
- Classification based solely on gene expression monitoring
- $p = 7,129$ genes
- A training data set
  - 47 ALL
  - 25 AML
  - $n = 47 + 25 = 72$
- p is much larger than the sample size
- $p/n \approx 100$

When the distribution of $\mathbf{x}$ is known (μ and Σ are known)

- An optimal classification rule exists, which classifies $\mathbf{x}$ to class 1 if and only if

$$\delta'\Sigma^{-1}(\mathbf{x} - \overline{\mu}) \geq 0$$

$\delta = \mu_1 - \mu_2$, $\overline{\mu} = (\mu_1 + \mu_2)/2$

- It minimizes the average misclassification rate
- The optimal misclassification rate is

$$R_{\text{OPT}} = \Phi(-\Delta_p/2), \qquad \Delta_p = \sqrt{\delta'\Sigma^{-1}\delta}$$

Φ: the standard normal distribution function

- This rule is the Bayes rule with equal prior probabilities for two classes
- The dimension p: the larger, the better

$$\lim_{\Delta_p \to \infty} R_{\text{OPT}} = 0, \qquad \lim_{\Delta_p \to 0} R_{\text{OPT}} = 1/2$$

When μ and Σ are unknown

- We have a training sample $\mathbf{X} = \{\mathbf{x}_{ki}, i = 1, ..., n_k, k = 1, 2\}$
- $\mathbf{x}_{ki} \sim N_p(\boldsymbol{\mu}_k, \Sigma)$, $k = 1, 2$
- $n = n_1 + n_2$
- All $\mathbf{x}_{ki}$'s are independent and $\mathbf{X}$ is independent of $\mathbf{x}$

Statistical issue

How to use the training sample to construct a rule having a misclassification rate close to R_{OPT}

Traditional application: small-p-large-n

The well known linear discriminant analysis (LDA) replaces unknown δ, $\overline{\mu}$, and Σ by $\widehat{\delta} = \overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2$, $\widehat{\overline{\mu}} = \overline{\mathbf{x}} = (\overline{\mathbf{x}}_1 + \overline{\mathbf{x}}_2)/2$, and $\widehat{\Sigma}^{-1} = \mathbf{S}^{-1}$ where

$$\overline{\mathbf{x}}_k = \frac{1}{n_k} \sum_{i=1}^{n_k} \mathbf{x}_{ki}, \quad k = 1, 2, \qquad \mathbf{S} = \frac{1}{n} \sum_{k=1}^{2} \sum_{i=1}^{n_k} (\mathbf{x}_{ki} - \overline{\mathbf{x}}_k)(\mathbf{x}_{ki} - \overline{\mathbf{x}}_k)'$$

are the maximum likelihood estimators

When μ and Σ are unknown

- We have a training sample $\mathbf{X} = \{\mathbf{x}_{ki}, i = 1, ..., n_k, k = 1, 2\}$
- $\mathbf{x}_{ki} \sim N_p(\mu_k, \Sigma)$, $k = 1, 2$
- $n = n_1 + n_2$
- All $\mathbf{x}_{ki}$'s are independent and $\mathbf{X}$ is independent of $\mathbf{x}$

Statistical issue

How to use the training sample to construct a rule having a misclassification rate close to R_{OPT}

Traditional application: small-p-large-n

The well known linear discriminant analysis (LDA) replaces unknown δ, $\overline{\mu}$, and Σ by $\widehat{\delta} = \bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2$, $\widehat{\overline{\mu}} = \bar{\mathbf{x}} = (\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2)/2$, and $\widehat{\Sigma}^{-1} = \mathbf{S}^{-1}$ where

$$\bar{\mathbf{x}}_k = \frac{1}{n_k}\sum_{i=1}^{n_k} \mathbf{x}_{ki}, \quad k = 1, 2, \qquad \mathbf{S} = \frac{1}{n}\sum_{k=1}^{2}\sum_{i=1}^{n_k}(\mathbf{x}_{ki} - \bar{\mathbf{x}}_k)(\mathbf{x}_{ki} - \bar{\mathbf{x}}_k)'$$

are the maximum likelihood estimators

When μ and Σ are unknown

- We have a training sample $\mathbf{X} = \{\mathbf{x}_{ki}, i = 1, ..., n_k, k = 1, 2\}$
- $\mathbf{x}_{ki} \sim N_p(\mu_k, \Sigma)$, $k = 1, 2$
- $n = n_1 + n_2$
- All $\mathbf{x}_{ki}$'s are independent and $\mathbf{X}$ is independent of $\mathbf{x}$

Statistical issue

How to use the training sample to construct a rule having a misclassification rate close to R_{OPT}

Traditional application: small-p-large-n

The well known linear discriminant analysis (LDA) replaces unknown δ, $\bar{\mu}$, and Σ by $\widehat{\delta} = \bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2$, $\widehat{\bar{\mu}} = \bar{\mathbf{x}} = (\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2)/2$, and $\widehat{\Sigma}^{-1} = \mathbf{S}^{-1}$ where

$$\bar{\mathbf{x}}_k = \frac{1}{n_k} \sum_{i=1}^{n_k} \mathbf{x}_{ki}, \quad k = 1, 2, \qquad \mathbf{S} = \frac{1}{n} \sum_{k=1}^{2} \sum_{i=1}^{n_k} (\mathbf{x}_{ki} - \bar{\mathbf{x}}_k)(\mathbf{x}_{ki} - \bar{\mathbf{x}}_k)'$$

are the maximum likelihood estimators

Modern application: large-p-small-n (large-p-not-so-large-n)

- How do we construct a rule when $p > n$?
- The LDA needs an estimator of Σ^{-1} (a generalized inverse $\mathbf{S}^{-}$?)
- The larger p, the better?
- A larger p results in more information , but produces more uncertainty when the distribution of $\mathbf{x}$ is unknown
- A greater challenge for data analysis since the training sample size n cannot increase as fast as p
- Bickel and Levina (2004) showed that the LDA is as bad as random guessing when $p/n \to \infty$
- In some studies researchers found that it is better to ignore some information (such as the correlation among the p components of $\mathbf{x}$) Domingos and Pazzani (1997), Lewis (1998), Dudoit et al. (2002).

Our task

To construct a nearly optimal rule for large dimension data

Modern application: large-p-small-n (large-p-not-so-large-n)

- How do we construct a rule when $p > n$?
- The LDA needs an estimator of Σ^{-1} (a generalized inverse $\mathbf{S}^{-}$?)
- The larger p, the better?
- A larger p results in more information , but produces more uncertainty when the distribution of $\mathbf{x}$ is unknown
- A greater challenge for data analysis since the training sample size n cannot increase as fast as p
- Bickel and Levina (2004) showed that the LDA is as bad as random guessing when $p/n \to \infty$
- In some studies researchers found that it is better to ignore some information (such as the correlation among the p components of $\mathbf{x}$) Domingos and Pazzani (1997), Lewis (1998), Dudoit et al. (2002).

Our task

To construct a nearly optimal rule for large dimension data

Linear discriminant analysis and asymptotic results

Regularity conditions

There is a constant c_0 (not depending on p or n) such that

- $c_0^{-1} \leq$ all eigenvalues of $\boldsymbol{\Sigma} \leq c_0$
- $c_0^{-1} \leq \max_{j \leq p} \delta_j^2 \leq c_0$
  δ_j is the jth component of δ

Consequences

- $\Delta_p \geq c_0^{-1}$, $\Delta_p = \sqrt{\delta' \Sigma^{-1} \delta}$
- $R_{\mathrm{OPT}} \leq \Phi(-(2c_0)^{-1}) < 1/2$
- $\Delta_p^2 = O(\|\delta\|^2)$ and $\|\delta\|^2 = O(\Delta_p^2)$

Asymptotic setting

- $n = n_1 + n_2$, $n_1/n \to c \in (0, \infty)$ as $n \to \infty$
- p is a function of n, $p/n \to b \in [0, \infty]$ as $n \to \infty$

Linear discriminant analysis and asymptotic results

Regularity conditions

There is a constant c_0 (not depending on p or n) such that

- $c_0^{-1} \leq$ all eigenvalues of $\mathbf{\Sigma} \leq c_0$
- $c_0^{-1} \leq \max_{j \leq p} \delta_j^2 \leq c_0$
  δ_j is the jth component of $\boldsymbol{\delta}$

Consequences

- $\Delta_p \geq c_0^{-1}$, $\Delta_p = \sqrt{\boldsymbol{\delta}' \mathbf{\Sigma}^{-1} \boldsymbol{\delta}}$
- $R_{\text{OPT}} \leq \Phi(-(2c_0)^{-1}) < 1/2$
- $\Delta_p^2 = O(\|\boldsymbol{\delta}\|^2)$ and $\|\boldsymbol{\delta}\|^2 = O(\Delta_p^2)$

Asymptotic setting

- $n = n_1 + n_2$, $n_1/n \to c \in (0, \infty)$ as $n \to \infty$
- p is a function of n, $p/n \to b \in [0, \infty]$ as $n \to \infty$

Linear discriminant analysis and asymptotic results

Regularity conditions

There is a constant c_0 (not depending on p or n) such that

- $c_0^{-1} \leq$ all eigenvalues of $\boldsymbol{\Sigma} \leq c_0$
- $c_0^{-1} \leq \max_{j \leq p} \delta_j^2 \leq c_0$
  δ_j is the jth component of $\boldsymbol{\delta}$

Consequences

- $\Delta_p \geq c_0^{-1}$, $\Delta_p = \sqrt{\boldsymbol{\delta}' \boldsymbol{\Sigma}^{-1} \boldsymbol{\delta}}$
- $R_{\mathrm{OPT}} \leq \Phi(-(2c_0)^{-1}) < 1/2$
- $\Delta_p^2 = O(\|\boldsymbol{\delta}\|^2)$ and $\|\boldsymbol{\delta}\|^2 = O(\Delta_p^2)$

Asymptotic setting

- $n = n_1 + n_2$, $n_1/n \to c \in (0, \infty)$ as $n \to \infty$
- p is a function of n, $p/n \to b \in [0, \infty]$ as $n \to \infty$

Conditional and uncoditional misclassification rate

T: a classification rule

- $R_T(\mathbf{X})$: the average of the conditional probabilities of making two types of misclassification, where the conditional probabilities are with respect to $\mathbf{x}$, given the training sample $\mathbf{X}$
- $R_T = E[R_T(\mathbf{X})]$: unconditional misclassification rate of T

Asymptotic optimality ($n \to \infty$)

- T is asymptotically optimal if $R_T(\mathbf{X})/R_{\mathrm{OPT}} \to_P 1$
- T is asymptotically sub-optimal if $R_T(\mathbf{X}) - R_{\mathrm{OPT}} \to_P 0$
- T is asymptotically worst if $R_T(\mathbf{X}) \to_P 1/2$

Note

- If $R_{\mathrm{OPT}} \not\to 0$ (i.e., $\Delta_p = \sqrt{\delta'\Sigma^{-1}\delta}$ is bounded), then the asymptotic sub-optimality is the same as the asymptotic optimality.
- If $R_{\mathrm{OPT}} \to 0$, however, we hope not only $R_T(\mathbf{X}) \to_P 0$ in probability, but also $R_T(\mathbf{X})$ and R_{OPT} have the same convergence rate.

Conditional and uncoditional misclassification rate

T: a classification rule

- $R_T(\mathbf{X})$: the average of the conditional probabilities of making two types of misclassification, where the conditional probabilities are with respect to $\mathbf{x}$, given the training sample $\mathbf{X}$
- $R_T = E[R_T(\mathbf{X})]$: unconditional misclassification rate of T

Asymptotic optimality ($n \to \infty$)

- T is asymptotically optimal if $R_T(\mathbf{X})/R_{\mathrm{OPT}} \to_P 1$
- T is asymptotically sub-optimal if $R_T(\mathbf{X}) - R_{\mathrm{OPT}} \to_P 0$
- T is asymptotically worst if $R_T(\mathbf{X}) \to_P 1/2$

Note

- If $R_{\mathrm{OPT}} \not\to 0$ (i.e., $\Delta_p = \sqrt{\delta'\Sigma^{-1}\delta}$ is bounded), then the asymptotic sub-optimality is the same as the asymptotic optimality.
- If $R_{\mathrm{OPT}} \to 0$, however, we hope not only $R_T(\mathbf{X}) \to_P 0$ in probability, but also $R_T(\mathbf{X})$ and R_{OPT} have the same convergence rate.

Conditional and uncoditional misclassification rate

T: a classification rule

- $R_T(\mathbf{X})$: the average of the conditional probabilities of making two types of misclassification, where the conditional probabilities are with respect to $\mathbf{x}$, given the training sample $\mathbf{X}$
- $R_T = E[R_T(\mathbf{X})]$: unconditional misclassification rate of T

Asymptotic optimality ($n \to \infty$)

- T is asymptotically optimal if $R_T(\mathbf{X})/R_{\rm OPT} \to_p 1$
- T is asymptotically sub-optimal if $R_T(\mathbf{X}) - R_{\rm OPT} \to_p 0$
- T is asymptotically worst if $R_T(\mathbf{X}) \to_p 1/2$

Note

- If $R_{\rm OPT} \not\to 0$ (i.e., $\Delta_p = \sqrt{\delta'\Sigma^{-1}\delta}$ is bounded), then the asymptotic sub-optimality is the same as the asymptotic optimality.
- If $R_{\rm OPT} \to 0$, however, we hope not only $R_T(\mathbf{X}) \to_p 0$ in probability, but also $R_T(\mathbf{X})$ and $R_{\rm OPT}$ have the same convergence rate.

Linear discriminant analysis ($p < n$)

For what kind of p (which may diverge to ∞), the LDA is asymptotically optimal or sub-optimal?

Theorem 1

Suppose that $s_n = p\sqrt{\log p}/\sqrt{n} \to 0$.

(i) The conditional misclassification rate of the LDA is equal to

$$R_{\text{LDA}}(\mathbf{X}) = \Phi\big(-[1 + O_P(s_n)]\Delta_p/2\big).$$

(ii) If $\Delta_p = \sqrt{\delta'\Sigma^{-1}\delta}$ is bounded, then the LDA is asymptotically optimal and

$$\frac{R_{\text{LDA}}(\mathbf{X})}{R_{\text{OPT}}} - 1 = O_P(s_n).$$

(iii) If $\Delta_p \to \infty$, then the LDA is asymptotically sub-optimal.

(iv) If $\Delta_p \to \infty$ and $s_n\Delta_p^2 = (p\sqrt{\log p}/\sqrt{n})\Delta_p^2 \to 0$, then the LDA is asymptotically optimal.

Linear discriminant analysis ($p < n$)

For what kind of p (which may diverge to ∞), the LDA is asymptotically optimal or sub-optimal?

Theorem 1

Suppose that $s_n = p\sqrt{\log p}/\sqrt{n} \to 0$.

(i) The conditional misclassification rate of the LDA is equal to

$$R_{\text{LDA}}(\mathbf{X}) = \Phi\left(-[1 + O_P(s_n)]\Delta_p/2\right).$$

(ii) If $\Delta_p = \sqrt{\delta'\Sigma^{-1}\delta}$ is bounded, then the LDA is asymptotically optimal and

$$\frac{R_{\text{LDA}}(\mathbf{X})}{R_{\text{OPT}}} - 1 = O_P(s_n).$$

(iii) If $\Delta_p \to \infty$, then the LDA is asymptotically sub-optimal.

(iv) If $\Delta_p \to \infty$ and $s_n\Delta_p^2 = (p\sqrt{\log p}/\sqrt{n})\Delta_p^2 \to 0$, then the LDA is asymptotically optimal.

Linear discriminant analysis ($p > n$)

When $p > n$, $\mathbf{S}^{-1}$ does not exist.

But the estimation of Σ^{-1} is not the only problem

Even if Σ^{-1} is known (so that the LDA can use the prefect "estimator" of Σ^{-1}), the performance of the LDA may still be bad

Theorem 2

Suppose that $p/n \to \infty$ and that Σ is known so that the LDA classifies $\mathbf{x}$ to class 1 if and only if $\widehat{\delta}'\Sigma^{-1}(\mathbf{x} - \widehat{\overline{\mu}}) \geq 0$, where $\widehat{\delta} = \bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2$, and $\widehat{\overline{\mu}} = \bar{\mathbf{x}}$.

(i) If $\Delta_p^2/\sqrt{p/n} \to 0$ (which is true when $\Delta_p = \sqrt{\delta'\Sigma^{-1}\delta}$ is bounded), then $R_{\mathrm{LDA}}(\mathbf{X}) \to_P 1/2$.

(ii) If $\Delta_p^2/\sqrt{p/n} \to c$ with $0 < c < \infty$, then $R_{\mathrm{LDA}}(\mathbf{X}) \to_P \Phi\left(-c/(2\sqrt{2})\right)$ and $R_{\mathrm{LDA}}(\mathbf{X})/R_{\mathrm{OPT}} \to_P \infty$.

(iii) If $\Delta_p^2/\sqrt{p/n} \to \infty$, then $R_{\mathrm{LDA}}(\mathbf{X}) \to_P 0$ but $R_{\mathrm{LDA}}(\mathbf{X})/R_{\mathrm{OPT}} \to_P \infty$.

Linear discriminant analysis ($p > n$)

When $p > n$, $\mathbf{S}^{-1}$ does not exist.

But the estimation of Σ^{-1} is not the only problem

Even if Σ^{-1} is known (so that the LDA can use the prefect "estimator" of Σ^{-1}), the performance of the LDA may still be bad

Theorem 2

Suppose that $p/n \to \infty$ and that Σ is known so that the LDA classifies $\mathbf{x}$ to class 1 if and only if $\widehat{\delta}'\Sigma^{-1}(\mathbf{x} - \widehat{\mu}) \geq 0$, where $\widehat{\delta} = \bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2$, and $\widehat{\mu} = \bar{\mathbf{x}}$.

(i) If $\Delta_p^2/\sqrt{p/n} \to 0$ (which is true when $\Delta_p = \sqrt{\delta'\Sigma^{-1}\delta}$ is bounded), then $R_{\mathrm{LDA}}(\mathbf{X}) \to_P 1/2$.

(ii) If $\Delta_p^2/\sqrt{p/n} \to c$ with $0 < c < \infty$, then $R_{\mathrm{LDA}}(\mathbf{X}) \to_P \Phi\left(-c/(2\sqrt{2})\right)$ and $R_{\mathrm{LDA}}(\mathbf{X})/R_{\mathrm{OPT}} \to_P \infty$.

(iii) If $\Delta_p^2/\sqrt{p/n} \to \infty$, then $R_{\mathrm{LDA}}(\mathbf{X}) \to_P 0$ but $R_{\mathrm{LDA}}(\mathbf{X})/R_{\mathrm{OPT}} \to_P \infty$.

Linear discriminant analysis ($p > n$)

Reason for bad performance of the LDA when $p > n$

- Too many parameters in δ to be estimated, even if Σ is known
- Similarly, too many parameters in Σ to be estimated, even if μ_k is known

Solutions?

A reasonable classification rule can be obtained if both δ and Σ are sparse

Sparsity

- Many elements of δ are 0 or very small
- Many off-diagonal elements of Σ are 0 or very small
- Both are true in many applications

Linear discriminant analysis ($p > n$)

Reason for bad performance of the LDA when $p > n$

- Too many parameters in δ to be estimated, even if Σ is known
- Similarly, too many parameters in Σ to be estimated, even if μ_k is known

Solutions?

A reasonable classification rule can be obtained if both δ and Σ are sparse

Sparsity

- Many elements of δ are 0 or very small
- Many off-diagonal elements of Σ are 0 or very small
- Both are true in many applications

Linear discriminant analysis ($p > n$)

Reason for bad performance of the LDA when $p > n$

- Too many parameters in δ to be estimated, even if Σ is known
- Similarly, too many parameters in Σ to be estimated, even if μ_k is known

Solutions?

A reasonable classification rule can be obtained if both δ and Σ are sparse

Sparsity

- Many elements of δ are 0 or very small
- Many off-diagonal elements of Σ are 0 or very small
- Both are true in many applications

Sparse linear discriminant analysis and asymptotic results

Sparsity measure for Σ

Bickel and Levina (2008) considered the following sparsity measure for Σ

$$C_{h,p} = \max_{j \le p} \sum_{l=1}^{p} |\sigma_{jl}|^h$$

σ_{jl} is the (j,l)th element of Σ
h is a constant not depending on p, $0 \le h < 1$

Special case of $h = 0$

$C_{0,p}$ is the maximum of the numbers of nonzero elements of rows of Σ

Sparsity on Σ

- Not sparse: $C_{h,p} = O(p)$
- Sparse: $C_{h,p} = O(\log p)$ or $C_{h,p} = O(n^\beta)$, $0 \le \beta < 1$

Sparse linear discriminant analysis and asymptotic results

Sparsity measure for Σ

Bickel and Levina (2008) considered the following sparsity measure for Σ

$$C_{h,p} = \max_{j \le p} \sum_{l=1}^{p} |\sigma_{jl}|^h$$

σ_{jl} is the (j,l)th element of Σ
h is a constant not depending on p, $0 \le h < 1$

Special case of $h = 0$

$C_{0,p}$ is the maximum of the numbers of nonzero elements of rows of Σ

Sparsity on Σ

- Not sparse: $C_{h,p} = O(p)$
- Sparse: $C_{h,p} = O(\log p)$ or $C_{h,p} = O(n^{\beta})$, $0 \le \beta < 1$

Sparse linear discriminant analysis and asymptotic results

Sparsity measure for Σ

Bickel and Levina (2008) considered the following sparsity measure for Σ

$$C_{h,p} = \max_{j \le p} \sum_{l=1}^{p} |\sigma_{jl}|^h$$

σ_{jl} is the (j,l)th element of Σ
h is a constant not depending on p, $0 \le h < 1$

Special case of $h = 0$

$C_{0,p}$ is the maximum of the numbers of nonzero elements of rows of Σ

Sparsity on Σ

- Not sparse: $C_{h,p} = O(p)$
- Sparse: $C_{h,p} = O(\log p)$ or $C_{h,p} = O(n^{\beta})$, $0 \le \beta < 1$

Bickel and Levina's thresholding estimator of Σ

S: sample covariance matrix

$\widetilde{\Sigma}$ is **S** thresholded at $t_n = M_1\sqrt{\log p}/\sqrt{n}$ (M_1 is a constant)

i.e., the (j,l)th element of $\widetilde{\Sigma}$ is $\widehat{\sigma}_{jl} I(|\widehat{\sigma}_{jl}| > t_n)$

$\widehat{\sigma}_{jl}$ is the (j,l)th element of **S**, and $I(A)$ is the indicator function of the set A

Consistency of $\widetilde{\Sigma}$

If

$$\frac{\log p}{n} \to 0 \qquad \text{and} \qquad d_n = C_{h,p}\left(\frac{\log p}{n}\right)^{(1-h)/2} \to 0$$

then

$$\|\widetilde{\Sigma} - \Sigma\| = O_P(d_n) \quad \text{and} \quad \|\widetilde{\Sigma}^{-1} - \Sigma^{-1}\| = O_P(d_n)$$

$\|\mathbf{A}\|$: the maximum of all eigenvalues of **A**

Bickel and Levina's thresholding estimator of Σ

S: sample covariance matrix

$\widetilde{\Sigma}$ is **S** thresholded at $t_n = M_1 \sqrt{\log p}/\sqrt{n}$ (M_1 is a constant)

i.e., the (j,l)th element of $\widetilde{\Sigma}$ is $\widehat{\sigma}_{jl} I(|\widehat{\sigma}_{jl}| > t_n)$

$\widehat{\sigma}_{jl}$ is the (j,l)th element of **S**, and $I(A)$ is the indicator function of the set A

Consistency of $\widetilde{\Sigma}$

If

$$\frac{\log p}{n} \to 0 \qquad \text{and} \qquad d_n = C_{h,p} \left(\frac{\log p}{n} \right)^{(1-h)/2} \to 0$$

then

$$\|\widetilde{\Sigma} - \Sigma\| = O_P(d_n) \quad \text{and} \quad \|\widetilde{\Sigma}^{-1} - \Sigma^{-1}\| = O_P(d_n)$$

$\|\mathbf{A}\|$: the maximum of all eigenvalues of **A**

Sparsity on δ

A large $\|\delta\|$ results in a large difference between $N_p(\mu_1, \Sigma)$ and $N_p(\mu_2, \Sigma)$

But it also results in a more difficult task of constructing a good classification rule, since δ has to be estimated based on the training sample $\mathbf{X}$ of a size that is much smaller than p.

Sparsity measure for δ

We consider the following sparsity measure for δ:

$$D_{g,p} = \sum_{j=1}^{p} \delta_j^{2g}$$

δ_j is the jth component of δ
g is a constant not depending on p, $0 \leq g < 1$

δ is sparse if $D_{g,p}$ is much smaller than p

Sparsity on δ

A large $\|\delta\|$ results in a large difference between $N_p(\mu_1, \Sigma)$ and $N_p(\mu_2, \Sigma)$

But it also results in a more difficult task of constructing a good classification rule, since δ has to be estimated based on the training sample $\mathbf{X}$ of a size that is much smaller than p.

Sparsity measure for δ

We consider the following sparsity measure for δ:

$$D_{g,p} = \sum_{j=1}^{p} \delta_j^{2g}$$

δ_j is the jth component of δ
g is a constant not depending on p, $0 \leq g < 1$

δ is sparse if $D_{g,p}$ is much smaller than p

Sparse estimator of δ

$\widetilde{\delta}$: $\widehat{\delta}$ thresholded at

$$a_n = M_2 \left(\frac{\log p}{n} \right)^{\alpha} \quad \text{with constants } M_2 > 0 \text{ and } \alpha \in (0, 1/2)$$

i.e., the jth component of $\widetilde{\delta}$ is $\widehat{\delta}_j I(|\widehat{\delta}_j| > a_n)$
$\widehat{\delta}_j$ is the jth component of $\widehat{\delta}$

A useful result

If

$$\frac{\log p}{n} \to 0,$$

then

$$P\left(|\widehat{\delta}_j| \le a_n,\ j = 1, ..., p \text{ with } |\delta_j| \le a_n/r \right) \to 1$$

and

$$P\left(|\widehat{\delta}_j| > a_n,\ j = 1, ..., p \text{ with } |\delta_j| > r a_n \right) \to 1$$

Sparse estimator of δ

$\widetilde{\delta}$: $\widehat{\delta}$ thresholded at

$$a_n = M_2 \left(\frac{\log p}{n} \right)^{\alpha} \quad \text{with constants } M_2 > 0 \text{ and } \alpha \in (0, 1/2)$$

i.e., the jth component of $\widetilde{\delta}$ is $\widehat{\delta}_j I(|\widehat{\delta}_j| > a_n)$
$\widehat{\delta}_j$ is the jth component of $\widehat{\delta}$

A useful result

If

$$\frac{\log p}{n} \to 0,$$

then

$$P\left(|\widehat{\delta}_j| \leq a_n,\ j = 1, ..., p \text{ with } |\delta_j| \leq a_n / r \right) \to 1$$

and

$$P\left(|\widehat{\delta}_j| > a_n,\ j = 1, ..., p \text{ with } |\delta_j| > r a_n \right) \to 1$$

Sparse linear discriminant analysis (SLDA) for high dimension data

Classify $\mathbf{x}$ to class 1 if and only if $\widetilde{\delta}'\widetilde{\Sigma}^{-1}(\mathbf{x}-\bar{\mathbf{x}}) \geq 0$

Theorem 3

Assume $(\log p)/n \to 0$ and

$$b_n = \max\left\{ d_n, \frac{a_n^{1-g}\sqrt{D_{g,p}}}{\Delta_p}, \frac{\sqrt{C_{h,p}q_n}}{\Delta_p\sqrt{n}} \right\} \to 0$$

$$\Delta_p = \sqrt{\delta'\Sigma^{-1}\delta}, \quad a_n = \left(\frac{\log p}{n}\right)^{\alpha}, \quad d_n = C_{h,p}\left(\frac{\log p}{n}\right)^{(1-h)/2}$$

$$C_{h,p} = \max_{j \leq p} \sum_{l=1}^{p} |\sigma_{jl}|^h, \quad D_{g,p} = \sum_{j=1}^{p} \delta_j^{2g},$$

$$q_n = \#\{j : |\delta_j| > a_n/r\}$$

Sparse linear discriminant analysis (SLDA) for high dimension data

Classify $\mathbf{x}$ to class 1 if and only if $\tilde{\delta}'\tilde{\Sigma}^{-1}(\mathbf{x}-\bar{\mathbf{x}}) \geq 0$

Theorem 3

Assume $(\log p)/n \to 0$ and

$$b_n = \max\left\{ d_n,\ \frac{a_n^{1-g}\sqrt{D_{g,p}}}{\Delta_p},\ \frac{\sqrt{C_{h,p}q_n}}{\Delta_p\sqrt{n}} \right\} \to 0$$

$$\Delta_p = \sqrt{\delta'\Sigma^{-1}\delta}, \quad a_n = \left(\frac{\log p}{n}\right)^{\alpha}, \quad d_n = C_{h,p}\left(\frac{\log p}{n}\right)^{(1-h)/2}$$

$$C_{h,p} = \max_{j\leq p}\sum_{l=1}^{p}|\sigma_{jl}|^h, \quad D_{g,p} = \sum_{j=1}^{p}\delta_j^{2g},$$

$$q_n = \#\{j : |\delta_j| > a_n/r\}$$

Theorem 3 (continued)

(i) The conditional misclassification rate of the SLDA is equal to

$$R_{\text{SLDA}}(\mathbf{X}) = \Phi\left(-[1 + O_P(b_n)]\Delta_p/2\right).$$

(ii) If Δ_p is bounded, then the SLDA is asymptotically optimal and

$$\frac{R_{\text{SLDA}}(\mathbf{X})}{R_{\text{OPT}}} - 1 = O_P(b_n).$$

(iii) If $\Delta_p \to \infty$, then the SLDA is asymptotically sub-optimal.

(iv) If $\Delta_p \to \infty$ and $b_n \Delta_p^2 \to 0$, then the SLDA is asymptotically optimal.

Situations where the SLDA is asymptotically optimal

There are two constants c_1 and c_2 such that $0 < c_1 \leq |\delta_j| \leq c_2$ for any nonzero δ_j

q_n is exactly the number of nonzero δ_j's

Δ_p^2 and $D_{0,p}$ have exactly the order q_n.

- If q_n is bounded (e.g., there are only finitely many nonzero δ_j's), then Δ_p is bounded and the result in Theorem 3 holds if $d_n = C_{h,p}(n^{-1}\log p)^{(1-h)/2} \to 0$
- When $q_n \to \infty$ ($\Delta_p \to \infty$), we assume that $q_n = O(n^\eta)$ and $C_{h,p} = O(n^\gamma)$ with $\eta \in (0,1)$ and $\gamma \in [0,1)$.
  Choose $\alpha = (1-h)/4$
  - If $p = O(n^\kappa)$ for a $\kappa \geq 1$, then the result in Theorem 3 holds when $\eta + \gamma < (1-h)/2$ and $\eta < (1+h)/2$
  - If $p = O(e^{n^\beta})$ for a $\beta \in (0,1)$, then the result in Theorem 3 holds if $\eta + \gamma < (1-h)(1-\beta)/2$ and $\eta < 1-(1-h)(1-\beta)/2$

Situations where the SLDA is asymptotically optimal

- Consider the case where $C_{h,p} = O(\log p)$, $D_{g,p} = O(\log p)$, and $p = O(e^{n^{\beta}})$ for a $\beta \in (0,1)$
  - If Δ_p is bounded, $d_n = O(n^{\beta+(\beta-1)(1-h)/2}) \to 0$, i.e., the SLDA is asymptotically optimal, if $\beta < (1-h)/(3-h)$
  - If $\Delta_p \to \infty$, then the largest divergence rate of Δ_p^2 is $O(\log p) = O(n^{\beta})$ and $\Delta_p^2 d_n \to 0$, i.e., the SLDA is asymptotically optimal, when $\beta < (1-h)/(5-h)$.

    When $h = 0$, this means $\beta < 1/5$.
- If $p = O(n^{\kappa})$ for a $\kappa \geq 1$ and $\max\{C_{h,p}, D_{g,p}\} = cn^{\gamma}$ for a $\gamma \in (0,1)$ and a positive constant c, then $\log p = O(\log n)$ diverges to ∞ at a rate slower than n^{γ}.

  Assume that $\Delta_p^2 = O(n^{\rho\gamma})$ with a $\rho \in [0,1]$ ($\rho = 0$ corresponds to a bounded Δ_p).

  The SLDA is asymptotically optimal if $(1+\rho)\gamma \leq (1-h)/2$ and $(1+\rho)\gamma/[2(1-g)] < \alpha \leq [1-(1+\rho)\gamma]/[2(1-g)]$

Choosing constants in thresholding: A cross-validation procedure

$\mathbf{X}_{ki}$: the data set with $\mathbf{x}_{ki}$ deleted
T_{ki}: the SLDA rule based on $\mathbf{X}_{ki}$, $i = 1, ..., n_k$, $k = 1, 2$.
The cross-validation estimator of R_{SLDA} is

$$\widehat{R}_{\mathrm{SLDA}} = \frac{1}{n} \sum_{k=1}^{2} \sum_{i=1}^{n_k} r_{ki}$$

r_{ki} is the indicator function of whether T_{ki} classifies $\mathbf{x}_{ki}$ incorrectly
If $R_{\mathrm{SLDA}} = R(n_1, n_2)$,

$$E(\widehat{R}_{\mathrm{SLDA}}) = \sum_{k=1}^{2} \sum_{i=1}^{n_k} \frac{E(r_{ki})}{n} = \frac{n_1 R(n_1 - 1, n_2) + n_2 R(n_1, n_2 - 1)}{n} \approx R_{\mathrm{SLDA}}$$

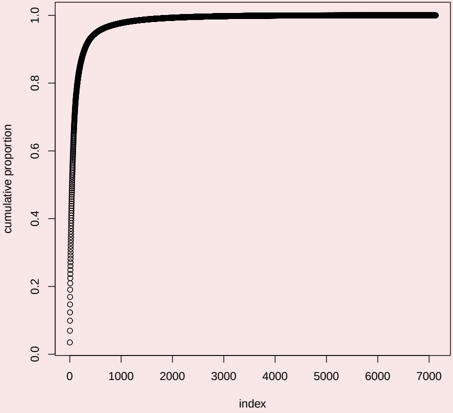
$\widehat{R}_{\mathrm{SLDA}}(M_1, M_2)$: the cross-validation estimator when (M_1, M_2) is used
Minimize $\widehat{R}_{\mathrm{SLDA}}(M_1, M_2)$ over a suitable range of (M_1, M_2)
The resulting $\widehat{R}_{\mathrm{SLDA}}$ can also be used as an estimate of R_{SLDA}

Applying the SLDA to human acute leukemias classification

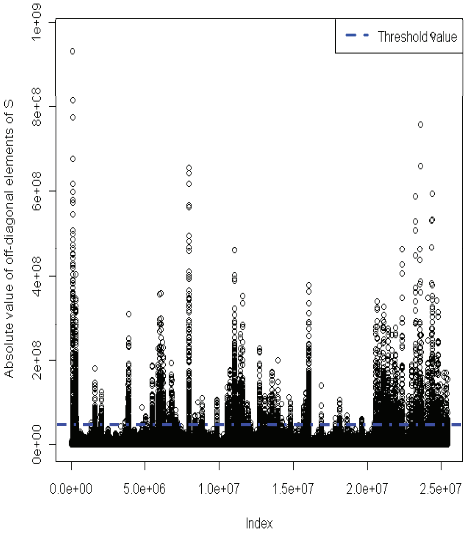
$p = 7,129$ genes

$n_1 = 47$, $n_2 = 25$, $n = 72$

Plot of the cumulative proportions of $\widehat{\delta}_j^2$

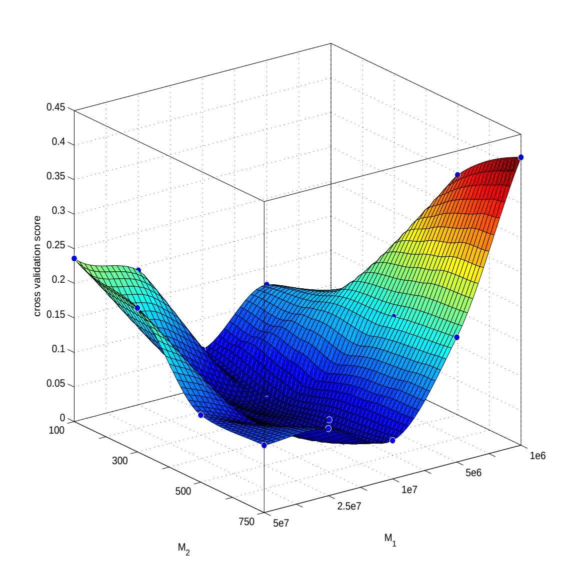


Plot of off-diagonal elements of **S** (0.45% values are above the blue line)



Cross-validation selection of M_1 and M_2

$\alpha = 0.3$
$M_1 = 10^7$, $M_2 = 300$
2,492 nonzero $\widetilde{\delta}_j$
(35% of 7,129)
227,083 nonzero $\widetilde{\sigma}_{jk}$
(0.45% of 25,407,756)



Cross validation estimates

- Cross validation for SLDA
  - misclassification rate is 0.0278
  - 1 of 47 cases in class 1 are misclassified
  - 1 of 25 cases in class 2 are misclassified
- Cross validation for LDA
  - misclassification rate is 0.0972
  - 2 of 47 cases in class 1 are misclassified
  - 5 of 25 cases in class 2 are misclassified

Simulation

Data are generated from $N(\widehat{\mu}_1, \widetilde{\Sigma})$ and $N(\widehat{\mu}_2, \widetilde{\Sigma})$

$n_1 = 47$, $n_2 = 25$, $p = 1,714$

Misclassification rates of

- LDA = 0.152 (0.006)
- SLDA = 0.069 (0.005)
- optimal rule = 0.03

Cross validation estimates

- Cross validation for SLDA
  - misclassification rate is 0.0278
  - 1 of 47 cases in class 1 are misclassified
  - 1 of 25 cases in class 2 are misclassified
- Cross validation for LDA
  - misclassification rate is 0.0972
  - 2 of 47 cases in class 1 are misclassified
  - 5 of 25 cases in class 2 are misclassified

Simulation

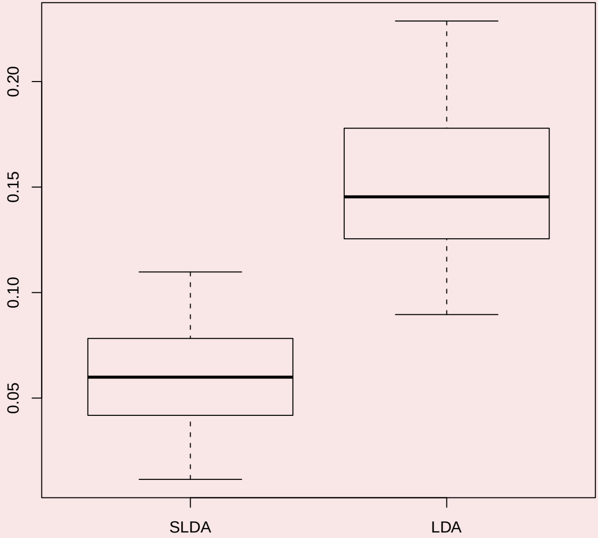
Data are generated from $N(\widehat{\boldsymbol{\mu}}_1, \widetilde{\boldsymbol{\Sigma}})$ and $N(\widehat{\boldsymbol{\mu}}_2, \widetilde{\boldsymbol{\Sigma}})$

$n_1 = 47$, $n_2 = 25$, $p = 1,714$

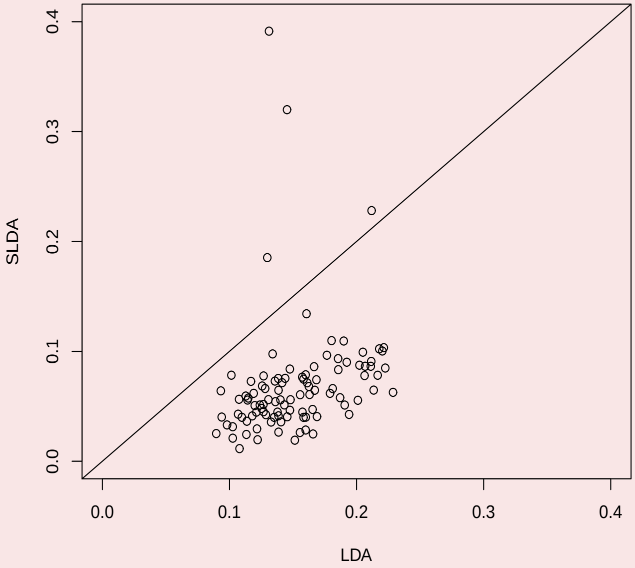
Misclassification rates of

- LDA = 0.152 (0.006)
- SLDA = 0.069 (0.005)
- optimal rule = 0.03

Boxplots of conditional misclassification rates of LDA and SLDA



Two-way plot of conditional misclassification rates: LDA vs SLDA



Conclusion and Discussion

- The ordinary linear discriminant analysis is OK if $p = o(\sqrt{n})$
- When $p/n \to \infty$, the linear discriminant analysis may be asymptotically as bad as random guessing
- When p is much larger than n, asymptotically optimal classification can be made if both the mean signal $\boldsymbol{\delta} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ and covariance matrix $\boldsymbol{\Sigma}$ are sparse
- A sparse linear discriminant analysis (SLDA) is proposed, and it is asymptotically optimal under some conditions
- SLDA is different from variable selection for $\boldsymbol{\delta}+$ LDA
  - Correlation among variables have to be considered
  - SLDA does not require the number of nonzero $\widetilde{\delta}_j$'s to be smaller than n
- Extension to non-normal data
- Extension to unequal covariance matrices: quadratic discriminant analysis