High-Dimensional Classification Methods for Sparse Signals and Their Applications in Gene Expression Data

Dawit Tadesse, Ph.D.
Department of Mathematical Sciences
University of Cincinnati

Biostatistics Epidemiology & Research Design Monthly Seminar Series
Cincinnati Children's Hospital Medical Center

November 11, 2014

Contents

- **1. Introduction**

Contents

- **1. Introduction**
- **2. Classification with Sparse Signals**

Contents

- **1. Introduction**
- **2. Classification with Sparse Signals**
- **3. Feature Selection**

Contents

- **1. Introduction**
- **2. Classification with Sparse Signals**
- **3. Feature Selection**
- **4. Simulation Results**

Contents

- 1. **Introduction**
- 2. **Classification with Sparse Signals**
- 3. **Feature Selection**
- 4. **Simulation Results**
- 5. **Applications to Gene Expression Data**

Contents

- **1. Introduction**
- **2. Classification with Sparse Signals**
- **3. Feature Selection**
- **4. Simulation Results**
- **5. Applications to Gene Expression Data**
- **6. Conclusion**

Contents

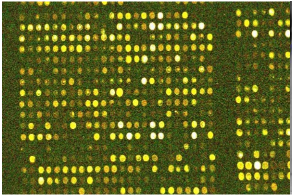
- **1. Introduction**
- **2. Classification with Sparse Signals**
- **3. Feature Selection**
- **4. Simulation Results**
- **5. Applications to Gene Expression Data**
- **6. Conclusion**
- **7. Selected Bibliography**

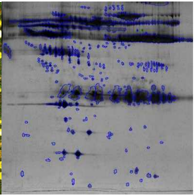
1. Introduction

- High-dimensional classification arises in many contemporary statistical problems.

1. Introduction

- High-dimensional classification arises in many contemporary statistical problems.
- • Bioinformatic: disease classification using microarray, proteomics, fMRI data.

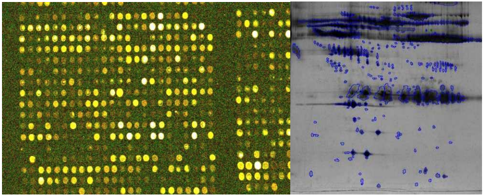




1. Introduction

- High-dimensional classification arises in many contemporary statistical problems.

- • Bioinformatic: disease classification using microarray, proteomics, fMRI data.

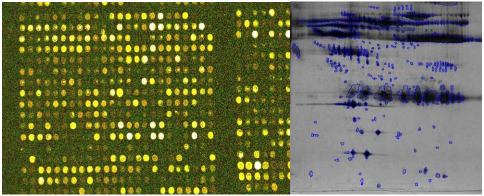


- • Document or text classification: E-mail spam.

1. Introduction

- High-dimensional classification arises in many contemporary statistical problems.

- • Bioinformatic: disease classification using microarray, proteomics, fMRI data.



- • Document or text classification: E-mail spam.

- • Voice recognition, hand written recognition, etc.

1. Introduction

Well known classification methods include:

- ♠ Logistic Regression

1. Introduction

Well known classification methods include:

- ♠ Logistic Regression
- ♠ Fisher discriminant analysis

1. Introduction

Well known classification methods include:

- ♠ Logistic Regression
- ♠ Fisher discriminant analysis
- ♠ Naive Bayes classifier

1. Introduction

Well known classification methods include:

- ♠ Logistic Regression
- ♠ Fisher discriminant analysis
- ♠ Naive Bayes classifier

For high-dimensional data (i.e. when $p >> n$), the above methods doesn't work well.

1. Introduction

Well known classification methods include:

- ♠ Logistic Regression
- ♠ Fisher discriminant analysis
- ♠ Naive Bayes classifier

For high-dimensional data (i.e. when $p >> n$), the above methods doesn't work well.

♦ **Bickel and Levina (2004)** showed that Fisher breaks down for high-dimensions and suggested Naive Bayes rule.

1. Introduction

Well known classification methods include:

- ♠ Logistic Regression
- ♠ Fisher discriminant analysis
- ♠ Naive Bayes classifier

For high-dimensional data (i.e. when $p >> n$), the above methods doesn't work well.

♦ **Bickel and Levina (2004)** showed that Fisher breaks down for high-dimensions and suggested Naive Bayes rule.

♦ **Fan and Fan (2008)** showed that even for Naive Bayes using all the features increases the error rate and suggested FAIR.

1. Introduction

⧫ **Fan and Fan (2008)** showed that the two-sample t-test can get important features.

1. Introduction

⧫ **Fan and Fan (2008)** showed that the two-sample t-test can get important features.

⧫ **Fan and etal.(2012)** showed that Naive Bayes increase error rates if there is correlation among the features.

1. Introduction

♦ **Fan and Fan (2008)** showed that the two-sample t-test can get important features.

♦ **Fan and etal.(2012)** showed that Naive Bayes increase error rates if there is correlation among the features.

♦ **My Works**:

1. Introduction

♦ **Fan and Fan (2008)** showed that the two-sample t-test can get important features.

♦ **Fan and etal.(2012)** showed that Naive Bayes increase error rates if there is correlation among the features.

♦ **My Works**:

• I will show that even under high-correlation Naive Bayes can perform better than Fisher.

• I propose a generalized test statistic and give the condition under which it selects important features.

2. Classification with Sparse Signals

Fisher discriminant rule

$$\delta_F(\boldsymbol{X}, \boldsymbol{\mu}_d, \boldsymbol{\mu}_a, \Sigma) = \mathbf{1}\left\{\boldsymbol{\mu}_d^T \Sigma^{-1}(\boldsymbol{X} - \boldsymbol{\mu}_a) > 0\right\}, \tag{1}$$

2. Classification with Sparse Signals

Fisher discriminant rule

$$\delta_F(\boldsymbol{X}, \boldsymbol{\mu}_d, \boldsymbol{\mu}_a, \Sigma) = \mathbf{1}\left\{\boldsymbol{\mu}_d^T \Sigma^{-1}(\boldsymbol{X} - \boldsymbol{\mu}_a) > 0\right\}, \qquad (1)$$

with corresponding misclassification error rate

$$W(\delta_F, \boldsymbol{\theta}) = \bar{\Phi}\left(\frac{(\boldsymbol{\mu}_d^T \Sigma^{-1} \boldsymbol{\mu}_d)^{1/2}}{2}\right). \qquad (2)$$

2. Classification with Sparse Signals

Naive Bayes rule

$$\delta_{NB}(\boldsymbol{X}, \boldsymbol{\mu}_d, \boldsymbol{\mu}_a, D) = \mathbf{1}\left\{\boldsymbol{\mu}_d^T D^{-1}(\boldsymbol{X} - \boldsymbol{\mu}_a) > 0\right\}, \qquad (3)$$

2. Classification with Sparse Signals

Naive Bayes rule

$$\delta_{NB}(\boldsymbol{X}, \boldsymbol{\mu}_d, \boldsymbol{\mu}_a, D) = \mathbf{1}\left\{\boldsymbol{\mu}_d^T D^{-1}(\boldsymbol{X} - \boldsymbol{\mu}_a) > 0\right\}, \tag{3}$$

whose misclassification error rate is

$$W(\delta_{NB}, \boldsymbol{\theta}) = \bar{\Phi}\left(\frac{\boldsymbol{\mu}_d^T D^{-1}\boldsymbol{\mu}_d}{2(\boldsymbol{\mu}_d^T D^{-1}\Sigma D^{-1}\boldsymbol{\mu}_d)^{1/2}}\right). \tag{4}$$

2. Classification with Sparse Signals

Definition: Suppose that $\boldsymbol{\mu}_d = (\alpha_1, \alpha_2, \ldots, \alpha_s, 0, \ldots, 0)^T$ is the $p \times 1$ mean difference vector where $\alpha_j \in \mathbb{R}\backslash\{0\}, j = 1, 2, \ldots, s$. We say that $\boldsymbol{\mu}_d$ is sparse if $s = o(p)$. Signal is defined as $C_s = \boldsymbol{\mu}_d^T D^{-1} \boldsymbol{\mu}_d = \displaystyle\sum_{j=1}^{s} \frac{\alpha_j^2}{\sigma_j^2}$ where σ_j^2 is the common variance for feature j in the two classes.

Examples of Sparse situations in real life:

2. Classification with Sparse Signals

Definition: Suppose that $\boldsymbol{\mu}_d = (\alpha_1, \alpha_2, \ldots, \alpha_s, 0, \ldots, 0)^T$ is the $p \times 1$ mean difference vector where $\alpha_j \in \mathbb{R}\backslash\{0\}, j = 1, 2, \ldots, s$. We say that $\boldsymbol{\mu}_d$ is sparse if $s = o(p)$. Signal is defined as $C_s = \boldsymbol{\mu}_d^T D^{-1} \boldsymbol{\mu}_d = \sum_{j=1}^{s} \frac{\alpha_j^2}{\sigma_j^2}$ where σ_j^2 is the common variance for feature j in the two classes.

Examples of Sparse situations in real life:

- $\star$ Gene Expression data (**Eg**: p genes from Leukemia and Normal, only s of them distinguish Leukemia and Normal).

2. Classification with Sparse Signals

Definition: Suppose that $\boldsymbol{\mu}_d = (\alpha_1, \alpha_2, \ldots, \alpha_s, 0, \ldots, 0)^T$ is the $p \times 1$ mean difference vector where $\alpha_j \in \mathbb{R}\backslash\{0\}, j = 1, 2, \ldots, s$. We say that $\boldsymbol{\mu}_d$ is sparse if $s = o(p)$. Signal is defined as $C_s = \boldsymbol{\mu}_d^T D^{-1} \boldsymbol{\mu}_d = \sum_{j=1}^{s} \frac{\alpha_j^2}{\sigma_j^2}$ where σ_j^2 is the common variance for feature j in the two classes.

Examples of Sparse situations in real life:

- ⋆ Gene Expression data (**Eg**: p genes from Leukemia and Normal, only s of them distinguish Leukemia and Normal).

- ⋆ Author Identification (**Eg**: two novels from two authors and there are only s few words which distinguish them).

2. Classification with Sparse Signals

Theorem 2.1: If $m \leq s, \boldsymbol{\mu}_d^{(m)} = (\alpha, \alpha, \ldots, \alpha)^T = \alpha\mathbf{1}, \alpha \neq 0$ and $\Sigma^{(m)}$ is the truncated $m \times m$ equicorrelation matrix, then we have

$$W(\delta_F, \boldsymbol{\theta}^{(m)}) = W(\delta_{NB}, \boldsymbol{\theta}^{(m)}),$$

where $\boldsymbol{\theta}^{(m)}$ is the truncated parameter.

2. Classification with Sparse Signals

Theorem 2.1: If $m \leq s, \boldsymbol{\mu}_d^{(m)} = (\alpha, \alpha, \ldots, \alpha)^T = \alpha\mathbf{1}, \alpha \neq 0$ and $\Sigma^{(m)}$ is the truncated $m \times m$ equicorrelation matrix, then we have

$$W(\delta_F, \boldsymbol{\theta}^{(m)}) = W(\delta_{NB}, \boldsymbol{\theta}^{(m)}),$$

where $\boldsymbol{\theta}^{(m)}$ is the truncated parameter.

We **define** $\bar{\boldsymbol{\rho}}^{(m)}$ and $\boldsymbol{\rho}_{\max}^{(m)}$ are equicorrelation matrices with off diagonals the mean of the correlation coefficients and largest of the absolute values of the correlation coefficients respectively.

2. Classification with Sparse Signals

Theorem 2.2: Suppose $\rho^{(m)}$ is an $m \times m$ correlation matrix and $\boldsymbol{\mu}_d^{(m)}$ is an $m \times 1$ mean difference vector.

2. Classification with Sparse Signals

Theorem 2.2: Suppose $\boldsymbol{\rho}^{(m)}$ is an $m \times m$ correlation matrix and $\boldsymbol{\mu}_d^{(m)}$ is an $m \times 1$ mean difference vector.
(a)

$$\bar{\Phi}\left(\frac{\sqrt{(\boldsymbol{\mu}_d^{(m)})^T(D^{(m)})^{-1}\boldsymbol{\mu}_d^{(m)}}}{2\sqrt{\lambda_{\min}(\boldsymbol{\rho}^{(m)})}}\right) \leq W(\delta_w, \boldsymbol{\theta}^{(m)}) \leq \bar{\Phi}\left(\frac{\sqrt{(\boldsymbol{\mu}_d^{(m)})^T(D^{(m)})^{-1}\boldsymbol{\mu}_d^{(m)}}}{2\sqrt{\lambda_{\max}(\boldsymbol{\rho}^{(m)})}}\right)$$

2. Classification with Sparse Signals

Theorem 2.2: Suppose $\rho^{(m)}$ is an $m \times m$ correlation matrix and $\mu_d^{(m)}$ is an $m \times 1$ mean difference vector.

(a)

$$\bar{\Phi}\left(\frac{\sqrt{(\boldsymbol{\mu}_d^{(m)})^T (D^{(m)})^{-1} \boldsymbol{\mu}_d^{(m)}}}{2\sqrt{\lambda_{\min}(\boldsymbol{\rho}^{(m)})}}\right) \leq W(\delta_w, \boldsymbol{\theta}^{(m)}) \leq \bar{\Phi}\left(\frac{\sqrt{(\boldsymbol{\mu}_d^{(m)})^T (D^{(m)})^{-1} \boldsymbol{\mu}_d^{(m)}}}{2\sqrt{\lambda_{\max}(\boldsymbol{\rho}^{(m)})}}\right)$$

(b) Suppose, further, that $\lambda_{\min}(\boldsymbol{\rho}^{(m)}) \geq \lambda_{\min}(\bar{\boldsymbol{\rho}}^{(m)}) = 1 - \bar{\rho}$. Then

$$\bar{\Phi}\left(\frac{\sqrt{(\boldsymbol{\mu}_d^{(m)})^T (D^{(m)})^{-1} \boldsymbol{\mu}_d^{(m)}}}{2\sqrt{1-\bar{\bar{\rho}}}}\right) \leq W(\delta_w, \boldsymbol{\theta}^{(m)}) \leq \bar{\Phi}\left(\frac{\sqrt{(\boldsymbol{\mu}_d^{(m)})^T (D^{(m)})^{-1} \boldsymbol{\mu}_d^{(m)}}}{2\sqrt{1+(m-1)\rho_{\max}}}\right)$$

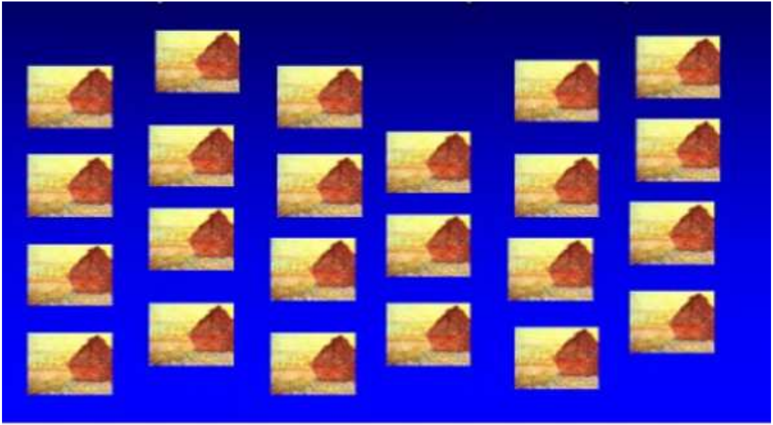
where $w = F$ or $w = NB$ for the truncated parameter $\boldsymbol{\theta}^{(m)}$.

3. Feature Selection

Goal of Feature Selection. How do i pick the best markers? Which method? Finding a needle in haystack?

3. Feature Selection

Goal of Feature Selection. How do i pick the best markers? Which method? Finding a needle in haystack?



3. Feature Selection

Two-sample t-test

3. Feature Selection

Two-sample t-test

For unequal sample sizes, unequal variance, the absolute value of the two-sample t-statistic for feature j is defined as

$$T_j = \frac{|\bar{X}_{1j} - \bar{X}_{0j}|}{\sqrt{S_{1j}^2/n_1 + S_{0j}^2/n_0}}, \quad j = 1, \ldots, p. \tag{5}$$

3. Feature Selection

Two-sample t-test

For unequal sample sizes, unequal variance, the absolute value of the two-sample t-statistic for feature j is defined as

$$T_j = \frac{|\bar{X}_{1j} - \bar{X}_{0j}|}{\sqrt{S_{1j}^2/n_1 + S_{0j}^2/n_0}}, \quad j = 1, \ldots, p. \tag{5}$$

Fan and Fan (2008) gave the conditions under which the two-sample t-test can select all the important features with probability 1.

3. Feature Selection

Two-sample t-test

For unequal sample sizes, unequal variance, the absolute value of the two-sample t-statistic for feature j is defined as

$$T_j = \frac{|\bar{X}_{1j} - \bar{X}_{0j}|}{\sqrt{S_{1j}^2/n_1 + S_{0j}^2/n_0}}, \quad j = 1, \ldots, p. \tag{5}$$

Fan and Fan (2008) gave the conditions under which the two-sample t-test can select all the important features with probability 1.

In this talk we will use the two-sample t-test as feature selection method.

3. Feature Selection

They stated their theorem as follows assuming $\boldsymbol{\mu}_d$ is sparse:

Theorem 3.1: Let s be a sequence such that $\log(p-s) = o(n^{\gamma})$ and $\log s = o(n^{1/2-\gamma}\beta_n)$ for some $\beta_n \to \infty$ and $0 < \gamma < 1/3$. Suppose that $\min\limits_{1\leq j\leq s} \frac{|\mu_{d,j}|}{\sqrt{\sigma_{1j}^2 + \sigma_{0j}^2}} = n^{-\gamma}\beta_n$ where $\mu_{d,j}$ is the j^{th} feature mean difference. Then, for $x \sim cn^{\gamma/2}$ with c some positive constant, we have

$$P\left(\min_{j\leq s} T_j \geq x \text{ and } \max_{j>s} T_j < x\right) \to 1.$$

3. Feature Selection

They stated their theorem as follows assuming $\boldsymbol{\mu}_d$ is sparse:

Theorem 3.1: Let s be a sequence such that $\log(p-s) = o(n^{\gamma})$ and $\log s = o(n^{1/2-\gamma}\beta_n)$ for some $\beta_n \to \infty$ and $0 < \gamma < 1/3$. Suppose that $\min\limits_{1\le j\le s} \dfrac{|\mu_{d,j}|}{\sqrt{\sigma_{1j}^2 + \sigma_{0j}^2}} = n^{-\gamma}\beta_n$ where $\mu_{d,j}$ is the j^{th} feature mean difference. Then, for $x \sim cn^{\gamma/2}$ with c some positive constant, we have

$$P\left(\min_{j\le s} T_j \ge x \text{ and } \max_{j>s} T_j < x\right) \to 1.$$

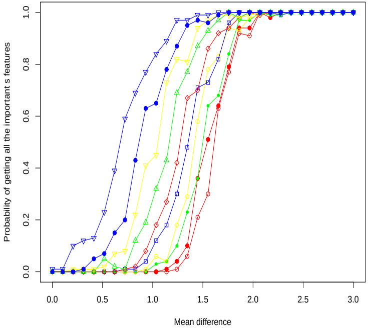
Note that asymptotically the two-sample t-test can pick up all the important features. However we are interested in the probability of selecting all the important features in the short run.

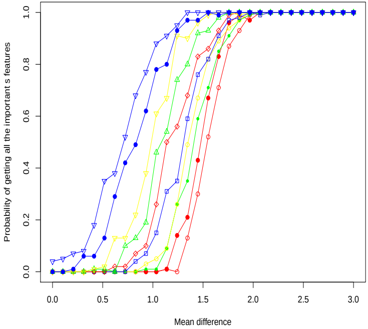
3. Feature Selection: Simulation Results

We take $p = 4500, s = 90, n_1 = n_0 = 30, \Sigma$ is equicorrelation and $\boldsymbol{\mu}_d$ equal mean difference. Simulation results for the probability of getting all the important s features in the first s and $2s$ t-statistics respectively.

3. Feature Selection: Simulation Results

We take $p = 4500, s = 90, n_1 = n_0 = 30, \Sigma$ is equicorrelation and $\boldsymbol{\mu}_d$ equal mean difference. Simulation results for the probability of getting all the important s features in the first s and $2s$ t-statistics respectively.





3. Feature Selection

Generalized Feature Selection

Two-sample t-test depends on (approximately) normal distribution.

3. Feature Selection

Generalized Feature Selection

Two-sample t-test depends on (approximately) normal distribution.

Our test statistic T_j for feature j is defined as follows:

3. Feature Selection

Generalized Feature Selection

Two-sample t-test depends on (approximately) normal distribution.

Our test statistic T_j for feature j is defined as follows:

$$T_j = \frac{\sum_{k=1}^{n_1} w_{1kj} - \sum_{k=1}^{n_0} w_{0kj}}{SE(\sum_{k=1}^{n_1} w_{1kj} - \sum_{k=1}^{n_0} w_{0kj})} \tag{6}$$

where $w_{ikj}, i = 0, 1,$ is the statistic for feature j in class i for sample k.

3. Feature Selection

Our test statistic is a special case of Two-sample t-test, Wilcoxon Mann-Whitney, and Two-sample Proportion test.

3. Feature Selection

Our test statistic is a special case of Two-sample t-test, Wilcoxon Mann-Whitney, and Two-sample Proportion test.

Theorem 3.2: Assume that the vector $\boldsymbol{\mu}_d = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_0$ is sparse and without loss of generality only first s entries are nonzero. Let s be a sequence such that $\log(p-s) = o(n^{\gamma})$ and $log\ s = o(n^{\gamma})$ for some $0 < \gamma < 1/3$. Suppose $\min_{1 \le j \le s} |\eta_j| = n^{-\gamma} C_n$ such that $C_n / n^{\frac{3\gamma}{2}} \to c^*$. For $t \sim cn^{\frac{\gamma}{2}}$ with some constant $0 < c < c^*/2$ we have

3. Feature Selection

Our test statistic is a special case of Two-sample t-test, Wilcoxon Mann-Whitney, and Two-sample Proportion test.

Theorem 3.2: Assume that the vector $\boldsymbol{\mu}_d = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_0$ is sparse and without loss of generality only first s entries are nonzero. Let s be a sequence such that $\log(p-s) = o(n^{\gamma})$ and $log\ s = o(n^{\gamma})$ for some $0 < \gamma < 1/3$. Suppose $\min_{1 \le j \le s} |\eta_j| = n^{-\gamma} C_n$ such that $C_n / n^{\frac{3\gamma}{2}} \to c^*$. For $t \sim cn^{\frac{\gamma}{2}}$ with some constant $0 < c < c^*/2$ we have

$$P(\min_{j \le s} |T_j| \ge t, \text{ and } \max_{j > s} |T_j| < t) \to 1.$$

4. Simulation Results

We use validation data to determine the optimal number of features.

We take:

♢ $p = 4500, s = 90$

♢ Training: $n_1 = n_0 = 30$

♢ Validation: $n_1 = n_0 = 30$

♢ Testing: $n_1 = n_0 = 50$

4. Simulation Results

NB dominates Fisher

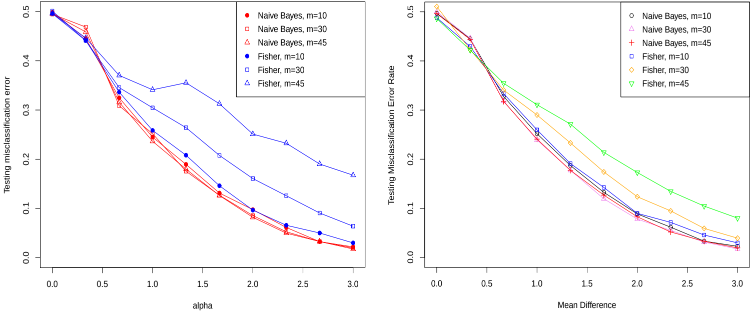
ρ	$\alpha = 1$	m NB	m F	Emp. Err.NB	Emp. Err.F
0.1	Q1	31.75	9.00	0.0375	0.1200
	Median	63.00	13.00	0.0700	0.1400
	Mean	79.98	16.42	0.0693	0.1448
	Q3	122.20	23.00	0.1000	0.1700
0.5	Q1	9.00	4.00	0.0475	0.240
	Median	53.50	10.00	0.2100	0.280
	Mean	78.96	15.72	0.1852	0.267
	Q3	155.50	25.00	0.2800	0.300
Ran. Corr.	Q1	15.00	18.75	0.0100	0.0200
	Median	20.00	21.00	0.0200	0.0300
	Mean	22.46	24.55	0.0221	0.0394
	Q3	26.25	29.25	0.0300	0.0500

4. Simulation Results

Simulations for equicorrelation and equal mean difference with $p = 4500, s = 90, \rho = 0.5$. Balanced ($n_1 = n_0 = 30$) and unbalanced ($n_1 = 30, n_0 = 60$) respectively. The testing sample sizes are $n_1 = n_0 = 50$ for both.

4. Simulation Results

Simulations for equicorrelation and equal mean difference with $p = 4500, s = 90, \rho = 0.5$. Balanced $(n_1 = n_0 = 30)$ and unbalanced $(n_1 = 30, n_0 = 60)$ respectively. The testing sample sizes are $n_1 = n_0 = 50$ for both.

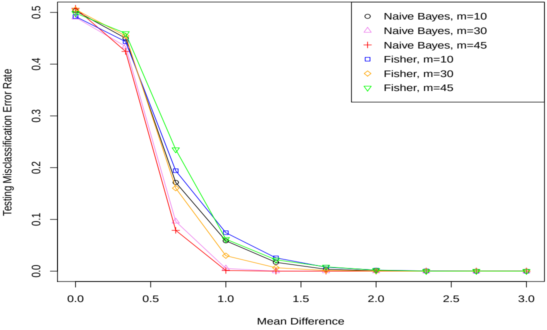


4. Simulation Results

Similar simulation as the balanced except we use random correlation. We randomly generate the eigenvalues of Σ in the interval $[0.5, 45.5]$.

4. Simulation Results

Similar simulation as the balanced except we use random correlation. We randomly generate the eigenvalues of Σ in the interval $[0.5, 45.5]$.



5. Applications to Gene Expression Data

Leukemia Data $(p = 7129, n = 72)$.

Training: $n_1 = 24$ from class ALL and $n_0 = 13$ from class AML.

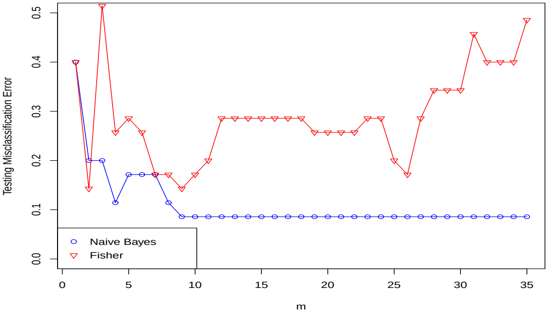
Validation: $n_1 = 23$ from class ALL and $n_0 = 12$ from class AML.

5. Applications to Gene Expression Data

Leukemia Data $(p = 7129, n = 72)$.

Training: $n_1 = 24$ from class ALL and $n_0 = 13$ from class AML.

Validation: $n_1 = 23$ from class ALL and $n_0 = 12$ from class AML.



For NB the optimal number of genes is 43 with min. error 2/35.

5. Applications to Gene Expression Data

Atopic Dermatitis (AD) Data ($p = 54675, n = 72$).

Training: $n_1 = 24$ from class AD and $n_0 = 15$ from class non-AD.

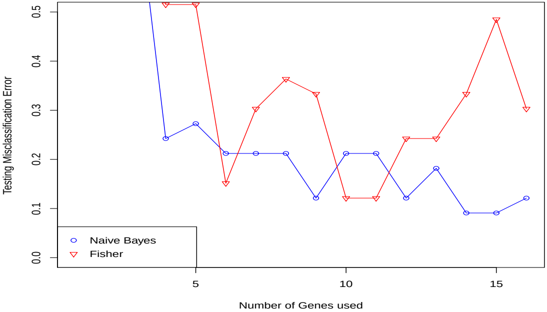
Validation: $n_1 = 25$ from class AD and $n_0 = 8$ from class non-AD.

5. Applications to Gene Expression Data

Atopic Dermatitis (AD) Data ($p = 54675, n = 72$).

Training: $n_1 = 24$ from class AD and $n_0 = 15$ from class non-AD.

Validation: $n_1 = 25$ from class AD and $n_0 = 8$ from class non-AD.



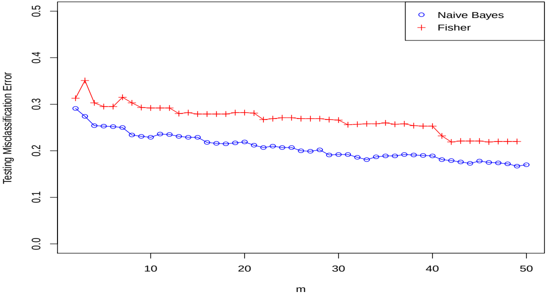
For NB the optimal number of genes is 34 with min. error 0.03.

5. Applications to Text Data

NASA flight data set ($p = 26694, n = 4567$).
Training: $n_1 = 1081, n_0 = 1486$, Validation: $n_1 = n_0 = 500$ and Testing: $n_1 = n_0 = 500$

5. Applications to Text Data

NASA flight data set ($p = 26694, n = 4567$).
Training: $n_1 = 1081, n_0 = 1486$, Validation: $n_1 = n_0 = 500$ and Testing: $n_1 = n_0 = 500$



For NB classifier the optimal number of features selected using the validation data set is 148 with corresponding testing error rate 0.116. For Fisher using 48 with corresponding testing error ≥ 0.20.

6. Conclusion

In this talk we considered a binary classification problem when the feature dimension p is much larger than the sample size n. The following are the main results:

6. Conclusion

In this talk we considered a binary classification problem when the feature dimension p is much larger than the sample size n. The following are the main results:

■ We have given conditions under which Naive Bayes is optimal for the population model.

6. Conclusion

In this talk we considered a binary classification problem when the feature dimension p is much larger than the sample size n. The following are the main results:

- We have given conditions under which Naive Bayes is optimal for the population model.

- Through theory, simulation and data analysis we have shown that Naive Bayes is practical method to use than Fisher for high-dimensional data.

6. Conclusion

In this talk we considered a binary classification problem when the feature dimension p is much larger than the sample size n. The following are the main results:

■ We have given conditions under which Naive Bayes is optimal for the population model.

■ Through theory, simulation and data analysis we have shown that Naive Bayes is practical method to use than Fisher for high-dimensional data.

■ In designing binary classification experiments, Fisher requires full correlation structure but using equicorelation structure we can design our experiment using Naive Bayes.

6. Conclusion

In this talk we considered a binary classification problem when the feature dimension p is much larger than the sample size n. The following are the main results:

- We have given conditions under which Naive Bayes is optimal for the population model.

- Through theory, simulation and data analysis we have shown that Naive Bayes is practical method to use than Fisher for high-dimensional data.

- In designing binary classification experiments, Fisher requires full correlation structure but using equicorelation structure we can design our experiment using Naive Bayes.

- Through simulation we characterized that the two-sample t-test can pick up all the important features as far the signal is not too low.

8. Selected Bibliography

- Bickel, P. J. and Levina, E. (2004). Some theory for Fisher’s linear discriminant function, ”naive Bayes”, and some alternatives when there are many more variables than observations. Bernoulli **10**, 989-1010.

- Cao, Hongyuan (2007). Moderate Deviations For Two Sample T-Statistics. ESAIM: Probability and Statistics, Vol. **11**, 264-271.

- Fan, J. and Fan, Y. (2008). High dimensional classification using features annealed independence rules. Ann. Statist., **36**, 2605-2637.

- Fan, J., Feng, Y., and Tong, X. (2012). A road to classification in high dimensional space: the regularized optimal affine discriminant. J. R. Statist. Soc. B. **74**, 745-771.

- Richard A. Johnson and Dean W. Wichern (6th edition). Applied Multivariate Statistical Analysis. Pearson Prentice Hall, 2007.

Thank You For Listening!