High Dimensional Classification in the Presence of Correlation: A Factor Model Approach

A. PEDRO DUARTE SILVA*

Faculdade de Economia e Gestão /
Centro de Estudos em Gestão e Economia

Universidade Católica Portuguesa
Centro Regional do Porto

PARIS, 23-28 August 2010

Compstat' 2010

(*) Supported by: FEDER / POCI 2010

Ciência.Inovação 2010 Programa Operacional Ciência e Inovação 2010
MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E ENSINO SUPERIOR

FCT Fundação para a Ciência e a Tecnologia
MINISTÉRIO DA CIÊNCIA E DA TECNOLOGIA

High Dimensional
Correlation Adjusted Classification

Overview

1. A Factor-model linear classification rule for High-Dimensional correlated data

2. Asymptotic properties with $p \to \infty$

3. Variable selection for problems with “rare” and “mostly weak” group differences

4. Performance in Micro-Array problems

5. Conclusions and Perspectives

High Dimensional
Correlation Adjusted Classification

Problem Statment:

$(Y ; X)$ $Y \in \{0,1\}$ $X \in \Re^p$

We want to find a rule that predicts Y given X

Bayes rule: $\hat{Y} = \mathbf{argmax_g} \; \pi_g f_g(X)$

Assuming $X \mid Y \sim N_p(\mu_{(Y)}, \Sigma)$

$\Rightarrow$ Bayes rule:

$$\hat{Y} = 1\left\{\Delta^T \Sigma^{-1}\left(X_i - \frac{1}{2}(\mu_0 + \mu_1)\right) > \log \frac{\pi_0}{\pi_1}\right\} \qquad \Delta = \mu_{(1)} - \mu_{(0)}$$

How to estimate Σ^{-1} when p > n and the X correlations are important ?

High Dimensional
Correlation Adjusted Classification

A Factor-Model Approach

$X_i = \mu_{(Yi)} + B\ f_i + \varepsilon_i \qquad f_i \in \Re^q \qquad \varepsilon_i \in \Re^P \qquad q << p$

$f_i \sim N_q(0, I_q) \qquad \varepsilon_i \sim N_p(0, D_\varepsilon) \qquad \forall j\ D_\varepsilon(j) > k_0 \in \Re_0$

$\Rightarrow$

$$\Sigma = B\ B^T + D_\varepsilon$$

$$\Sigma^{-1} = D_\varepsilon^{-1} - D_\varepsilon^{-1} B\ [I_q + B^T D_\varepsilon^{-1} B]^{-1} B^T D_\varepsilon^{-1}$$

$$\hat{\Sigma}_{RFctq} = \hat{B}\hat{B}^T + \hat{D}_\varepsilon$$

$$\hat{B}, \hat{D}_\varepsilon = \arg\min_{\hat{B},\hat{D}_\varepsilon} \| \hat{V}^{-1/2} \hat{\Sigma}_{RFctq} \hat{V}^{-1/2} - \hat{V}^{-1/2} S\, \hat{V}^{-1/2} \|_F^2$$

High Dimensional
Correlation Adjusted Classification

Asymptotic Properties

We will compare empirical linear rules

$$\delta_L = 1\left\{\hat{\Delta}^T \hat{\Sigma}_{\delta_L}^{-1}\left(X_i - \frac{1}{2}(\overline{X}_0 + \overline{X}_1)\right) > \log\frac{n_0}{n_1}\right\}$$

For some parameter space Γ_{δ_L} and Δ estimator $\hat{\Delta}$ satisfying

$$\max_{\Gamma_{\delta_L}} E_\theta \|\hat{\Delta} - \Delta\|^2 = o(1) \qquad \textbf{(C1)}$$

based on the criterion

$$\overline{W}_{\Gamma_{\delta_L}}(\delta_L) = \max_{\Gamma_{\delta_L}} P_\theta\left(\delta_L(Y_i) = 1 \mid Y_i = 0\right) = \max_{\Gamma_{\delta_L}}\left(1 - \Phi\left(\frac{\hat{\Delta}^T \hat{\Sigma}_{\delta_L}^{-1} \hat{\Delta}}{2\sqrt{\hat{\Delta}^T \hat{\Sigma}_{\delta_L}^{-1} \Sigma \hat{\Sigma}_{\delta_L}^{-1} \hat{\Delta}}}\right)\right)$$

when $p \to \infty$; $\frac{n(p)}{p} \to d < \infty$

High Dimensional
Correlation Adjusted Classification

Asymptotic Properties

Main Result

when

$$\theta = \left(\mu_{(0)}, \mu_{(1)}\ , \Sigma\right) \in \Gamma_{F_q}(k_0, k_1, k_2, q, B, c) = \left\{ \begin{array}{l} \theta :\ \Delta^T \Sigma^{-1} \Delta \geq c^2 , \\ k_1 \leq\ \lambda_{min}(\Sigma) \leq\ \lambda_{max}(\Sigma)\ \leq k_2 \\ \Delta \in B \\ \forall\, j,a \sum_{j',l'} \left| \frac{\partial \beta(j,a)}{\partial R(j',l')} \right| < \infty \\ \forall\, j \sum_{j',l'} \left| \frac{\partial D_\varepsilon(j)}{\partial R(j',l')} \right| < \infty \end{array} \right\}$$

$$\pi_0 = \pi_1 = 1/2$$

(C1) is satisfied

$$\Sigma_{RFctq} = B\,B^T + D_\varepsilon^2 = \arg\min_{B,D_\varepsilon} \| R_{RFctq} - V^{-1/2} \Sigma V^{-1/2} \|_F^2 \qquad R_{RFctq} = V^{-1/2}\, \Sigma_{RFct_q} V^{-1/2}$$

It follows that: when $p \to \infty\ ;\ \frac{n(p)}{\log p} \to \infty$

$$\overline{W}_{\Gamma_{\delta_{Fq}}}(\delta_{Fq}) \to 1 - \Phi\left(\frac{\sqrt{K_{0Fq}}}{1 + K_{0Fq}}\, c \right) \qquad K_{0Fq} = \max_{\Gamma_{F_q}} \frac{\lambda_{max}(\Sigma_{0Fq})}{\lambda_{min}(\Sigma_{0Fq})} \qquad \Sigma_{0F_q} = \Sigma_{RFct_q}^{-\frac{1}{2}}\ \Sigma\ \Sigma_{RFct_q}^{-\frac{1}{2}}$$

High Dimensional
Correlation Adjusted Classification

Selecting Predictors

1 - Rank variables acording to two-sample t-scores

2 – Choose a selection cut-off for the score values

Higher Criticism **(Donoho e Jin 2004)**

Given p ordered p-values: $\pi_1, \ldots, \pi_p$

$$\mathrm{HC}(j;\pi_j) = \sqrt{p}\,\frac{(j/p) - \pi_j}{\sqrt{(j/p)(1-(j/p))}}$$

$$\mathrm{HC}^* = \max_{j \le \alpha_0} \mathrm{HC}(j;\pi_j)$$

High Dimensional
Correlation Adjusted Classification

Selecting Predictors

Higher Criticism

In a two-group homokedastic model, with :

- Diagonal classification rules
- p-values derived from two-group t-scores
- Independent variables
- Rare "effects" (mean group diferences)
- Weak effects

when p $\rightarrow\infty$

HC* is asymptotically equivalent to the optimal selection threshold

(Donoho e Jin 2009)

High Dimensional
Correlation Adjusted Classification

Selecting Predictors

Control of false discovery rates

Given a sequence of p <u>independent</u> tests with ordered p-values: $\pi_1, \ldots, \pi_p$

Reject the null hypothesis (H_{0j}) where $j \le k$, with

$$k = \max\left\{ j : \pi_j \le \frac{j}{p}\alpha \right\}$$

(Benjamini e Hochberg 1995)

Given a sequence of p <u>dependent</u> tests with ordered p-values: $\pi_1, \ldots, \pi_p$

Reject the null hypothesis (H_{0j}) where $j \le k$, with

$$k = \max\left\{ j : \pi_j \le \frac{j}{p\sum_{i=1}^{p}\frac{1}{i}}\alpha \right\}$$

(Benjamini e Yekutieli 2001)

High Dimensional
Correlation Adjusted Classification

Selecting Predictors

Expanded Higher Criticism

A selection scheme for problems where effects are rare and **most** (but not necessarily all) effects are weak

1 - Include all variables that satisfy Benjamini and Yekutieli's criterion

2 - Estimate an "empirical null distributiuon"

3 - Compute p-values for the effects of non-selected variables, based on the null estimated in step 2

4 - Find the HC* threshold from the p-values computed in step 3

High Dimensional
Correlation Adjusted Classification

Singh's Prostate Cancer Data – p=6033; n=50+52

Rule	Error Estimate (std error)	# Variables kept (min – median - max)
Fisher's LDA*	0.2146 (0.0101)	58 – **134.5** – 421
Naive Bayes*	**0.0670** (0.0052)	58 – **134.5** – 421
Support Vector Machines*	**0.0642** (0.0052)	58 – **134.5** – 421
Nearest Shruken Centroids	0.0838 (0.0063)	108 – **356** – 1771
Regularized DA	**0.0741** (0.0053)	82 – **390** – 1201
Shrunken DA*	**0.0650** (0.0051)	58 – **134.5** – 421
Factor-based LDA* (q=1)	**0.0641** (0.0052)	58 – **134.5** – 421
NLDA*	**0.0720** (0.0052)	58 – **134.5** – 421

* After variable selection by the maximum of FDR (False Discovery Rates) and HC (Higher Criticism), both derived from Independence based T-scores.
The p-values used in the HC computations are derived from empirical Null distributions

High Dimensional
Correlation Adjusted Classification

Golubs's Leukemia Data –- p = 7 129 ; n = 47+25

Rule	Error Estimate (std error)	# Variables kept (min – median - max)
Fisher's LDA*	0.2558 (0.0109)	326 – **478** – 712
Naive Bayes*	0.480 (0.0085)	326 – **478** – 712
Support Vector Machines*	0.0405 (0.0049)	326 – **478** – 712
Nearest Shruken Centroids	**0.0201** (0.0039)	703 – **3166** – 7129
Regularized DA	0.0491 (0.0062)	12 – **1934** – 7124
Shrunken DA*	0.0276 (0.0044)	326 – **478** – 712
Factor-based LDA* (q=1)	**0.0174** (0.0034)	326 – **478** – 712
NLDA*	0.1510 (0.0085)	326 – **478** – 712

* After variable selection by the maximum of FDR (False Discovery Rates) and HC (Higher Criticism), both derived from Independence based T-scores. The p-values used in the HC computations are derived from empirical Null distributions

High Dimensional
Correlation Adjusted Classification

Alon's Colon Data -– p = 2 000 ; n = 40+22

Rule	Error Estimate (std error)	# Variables kept (min – median - max)
Fisher's LDA*	0.3285 (0.0143)	3 – **71.5** – 200
Naive Bayes*	0.2275 (0.0133)	3 – **71.5** – 200
Support Vector Machines*	**0.1576** (0.0095)	3 – **71.5** – 200
Nearest Shruken Centroids	**0.1563** (0.0098)	7 – **39** – 527
Regularized DA	0.2174 (0.0126)	14 – **425** – 2000
Shrunken DA*	0.1865 (0.0100)	3 – **71.5** – 200
Factor-based LDA* (q=1)	**0.1746** (0.0098)	3 – **71.5** – 200
NLDA*	0.2614 (0.0114)	3 – **71.5** – 200

* After variable selection by the maximum of FDR (False Discovery Rates) and HC (Higher Criticism), both derived from Independence based T-scores. The p-values used in the HC computations are derived from empirical Null distributions

High Dimensional Correlation Adjusted Classification

Conclusions

✓ **A factor-model classification rule, designed for high-dimensional correlated data, was proposed**

❖ Asymptotic Analysis show that

As p $\rightarrow\infty$ the new rule can approach a low expected error rate

Often, much lower than

unrestricted covariance rules

independence-based rules

❖ Empirical comparisons sugest that

when combined with sensible variable selection schemes

the new rule is highly competitive in MicroArray Applications

High Dimensional
Correlation Adjusted Classification

Open Questions

❖ Should correlations **also** be incorporated the selection scheme ?

When and How ?

❖ How do factor-based rules perform in problems with more than two groups ?

❖ Do differences in misclassification costs affect the relative standing of different classification rules ?

…

High Dimensional
Correlation Adjusted Classification

References

∗ Ahdesmaki, P. and Strimmer, K. (2009). Feature selection in "omics"prediction problems using cat scores and non-discovery rate control. *rXiv,stat.AP:0903.2003v1.*

∗ Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society* B 57, 289-300.

∗ Benjamini, Y. and Yekutileli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics* 29, 1165-1188.

∗ Donoho, D. and Jin, J. (2004). Higher criticism for detecting sparse heterogeneous mixtures. *Annals of Statistics*, 32, 962-944.

∗ Donoho, D. and Jin, J. (2008). Higher criticism thresholding: Optimal feature selection when useful features are rare and weak. *Proc. Natl. Acad. Sci*, USA 105, 14790-14795.

∗ Donoho, D. and Jin, J. (2009). Feature selection by higher criticism thresholding: Optimal phase diagram. *Philosophical Transactions of the Royal Society A,* 367, 4449-4470.

∗ Duarte Silva, A.P. (2009). Linear Discriminant Analysis with more Variables than Observations. A not so Naïve Approach. In: *Classification as a Tool for Research. Proceedings of the 11th IFCS Biennial Conference and 33rd Annual Conference of the Gesellschaft für Klassifikation*. Dresden, Germany, 227-234.

∗ Efron, B. (2008). Microarrays, empirical Bayes and the two-groups model. *Statistical Science* 1, 1-22.

∗ Guo, Y., Hastie, T. and Tibshirani, T. (2007). Regularized discriminant analysis and its application in microarrays. *Biostatistics* 8, 86-100.

∗ Tibshirani, R., Hastie, B., Narismhan, B. and Chu, G. (2003). Class prediction by nearest shrunken centroids with applications to DNA microarrays. *Statistical Science*, 18, 104-117.

∗ Thomaz, C.E. and Gillies, D.F. (2005). A maximum uncertainty lda-based approach for limited sample size problems with application to face recognition. In*: 18th Brazilian Symposium on computer Graphics and Image Processimg. SIBGRAPI 2005*, 89-96.