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

(*) Supported by: FEDER / POCI 2010
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
Problem Statement:

\[(Y ; X ) \quad Y \in \{0, 1\} \quad X \in \mathbb{R}^p\]

We want to find a rule that predicts \(Y\) given \(X\)

Bayes rule:

\[\hat{Y} = \arg \max_g \pi_g f_g(X)\]

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

\[\Rightarrow \quad \text{Bayes rule:}\]

\[\hat{Y} = \mathbf{1}\{\Delta^T \Sigma^{-1} (X_i - \frac{1}{2}(\mu_0 + \mu_1)) > \log \frac{\pi_0}{\pi_1}\} \quad \Delta = \mu_{(1)} - \mu_{(0)}\]

**How to estimate \(\Sigma^{-1}\) when \(p > n\) and the \(X\) correlations are important?**
High Dimensional
Correlation Adjusted Classification

A Factor-Model Approach

\[ X_i = \mu(Y_i) + B f_i + \varepsilon_i \quad f_i \in \mathbb{R}^q \quad \varepsilon_i \in \mathbb{R}^p \quad q << p \]

\[ f_i \sim N_q(0, I_q) \quad \varepsilon_i \sim N_p(0, D_\varepsilon) \quad \forall j \quad D_\varepsilon(j) > k_0 \in \mathbb{R}_0 \]

\[ \Rightarrow \quad \Sigma = B B^T + D_\varepsilon \]

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

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

\[ \hat{B}, \hat{D}_\varepsilon = \arg\min_{\hat{B}, \hat{D}_\varepsilon} \| \hat{V}^{-1/2} \hat{\Sigma}_{RFctq} \hat{V}^{-1/2} - \hat{V}^{-1/2} S \hat{V}^{-1/2} \|_F \]
High Dimensional Correlation Adjusted Classification

Asymptotic Properties

We will compare empirical linear rules

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

For some parameter space $\Gamma_{\delta_L}$ and $\Delta$ estimator $\hat{\Lambda}$ satisfying

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

based on the criterion

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

when $p \to \infty ; \frac{n(p)}{p} \to d < \infty$
High Dimensional Correlation Adjusted Classification

Asymptotic Properties

Main Result

when

\[ \theta = (\mu_{(0)}, \mu_{(1)}, \Sigma) \in \Gamma_{\delta_{Fq}} (k_0, k_1, k_2, q, B, c) = \begin{cases} \Lambda \in \mathbb{B} \\ \forall j, a \sum_{j', l'} \left| \frac{\partial \beta(j, a)}{\partial R(j', l')} \right| < \infty \\ \forall j \sum_{j', l'} \left| \frac{\partial D_{q}(j)}{\partial R(j', l')} \right| < \infty \end{cases} \]

\( \pi_0 = \pi_1 = 1/2 \)

\((C1)\) is satisfied

\[ \Sigma_{RFctq} = BB^T + D^2 = \arg \min_{B, D} \| R_{RFctq} - V^{-1/2} \Sigma V^{-1/2} \|_F \]

\[ R_{RFctq} = V^{-1/2} \Sigma_{RFctq} V^{-1/2} \]

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

\[ \overline{W}_{\Gamma_{\delta_{Fq}}} (\delta_{Fq}) \to 1 - \Phi \left( \sqrt{\frac{K_{0Fq}}{1 + K_{0Fq}}} \right) \]

\[ K_{0Fq} = \max_{\Gamma_{\delta_{Fq}}} \frac{\lambda_{\max}(\Sigma_{0Fq})}{\lambda_{\min}(\Sigma_{0Fq})} \]

\[ \Sigma_{0Fq} = \Sigma_{RFctq}^{-\frac{1}{2}} \sum_{RFctq} \sum_{RFctq}^{-\frac{1}{2}} \]

Compstat' 2010

PARIS, 23-28 August 2010
Selecting Predictors

1 - Rank variables according 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: $\pi_1, \ldots, \pi_p$

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

$$HC^* = \max_{j \leq a_0} HC(j; \pi_j)$$
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 differences)
- Weak effects

when $p \to \infty$

$HC^*$ 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 \( p \) independent tests with ordered \( p \)-values: \( \pi_1, \ldots, \pi_p \)

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

\[
k = \max \left\{ j : \pi_j \leq \frac{j}{p} \alpha \right\} \quad \text{(Benjamini e Hochberg 1995)}
\]

Given a sequence of \( p \) dependent tests with ordered \( p \)-values: \( \pi_1, \ldots, \pi_p \)

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

\[
k = \max \left\{ j : \pi_j \leq \frac{j}{p} \frac{1}{p} \alpha \right\} \quad \text{(Benjamini e Yekutieli 2001)}
\]
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 Yekutiel’s criterion
2. Estimate an “empirical null distribution”
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

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

* 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.

Compstat’ 2010

PARIS, 23-28 August 2010
## High Dimensional Correlation Adjusted Classification

Golub’s Leukemia Data -- p = 7129 ; n = 47+25

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

* 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

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

* 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

**Compstat’ 2010**  
PARIS, 23-28 August 2010
High Dimensional Correlation Adjusted Classification

Conclusions

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

❖ Asymptotic Analysis show that

   As \( p \to \infty \) the new rule can approach a low expected error rate

   Often, much lower than

   unrestricted covariance rules

   independence-based rules

❖ Empirical comparisons suggest 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 in 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