Borrowing information from relevant microarray studies for sample classification using weighted partial least squares

Xiaohong Huang

Eli Lilly and Company
Indianapolis, IN 46285
Outline

- Introduction
- Statistical classification methods for microarray data
  - Partial least squares (PLS)
  - Penalized partial least squares (PPLS)
  - Applications of PPLS to Minnesota data and PGA data
- Classification with combined data of multiple studies
  - PLS with a conjugate gradient path
  - Weighted PLS/PPLS
  - Experiment: Combined data
- Summary and discussion
Introduction

- Traditional medical diagnosis/classification method is very subjective
  - Based on morphological characteristics, pathological features
  - Depends on highly trained pathologists

Limitation: Hard to diagnose disease subtypes that are morphologically similar but follow different clinical courses.

- New classification method is objective
  - Based on microarray gene expression data.
  - Can be highly accurate.

Potentials: diagnose disease subtypes; predict clinical outcomes...
**Example: Two-class microarray**

**Notations:**

<table>
<thead>
<tr>
<th></th>
<th>Class 1</th>
<th>Class 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1           ...  n₁</td>
<td>n₁ + 1     ...  n₁ + n₂ = n</td>
</tr>
<tr>
<td>gene 1</td>
<td>X₁,₁     ...  X₁,n₁</td>
<td>X₁,n₁+₁  ...  X₁,n</td>
</tr>
<tr>
<td>gene 2</td>
<td>X₂,₁     ...  X₂,n₁</td>
<td>X₂,n₁+₁  ...  X₂,n</td>
</tr>
<tr>
<td>...</td>
<td>...       ...</td>
<td>...         ...</td>
</tr>
<tr>
<td>gene p</td>
<td>Xₚ,₁     ...  Xₚ,n₁</td>
<td>Xₚ,n₁+₁  ...  Xₚ,n</td>
</tr>
</tbody>
</table>

**Outcome** \( Y = (y₁, y₂, \ldots, yₙ)' \).

**Covariates** \( Xᵢ = (xᵢ₁, xᵢ₂, \ldots, xᵢₙ)' \), \( i = 1, \ldots, p \).

Covariates are often standardized \( \text{var}(xᵢ) = 1 \).
A special feature of microarray data:

Small $n$, large $p$

A simple prediction problem:

- Our goal is to predict $Y$ from $X_1, X_2, \ldots, X_p$ by a linear model.
- Especially interested in problems where $p \gg n$.

Many new methods have appeared

- Weighted voting, Compound covariate,
- Penalized regression: Shrunken centroids, LASSO,
- Machine learning: SVM, Bagging/boosting trees,
Penalized Partial Least Squares

- Partial Least Squares (PLS)
  - Particularly suited for constructing linear models when there are more variables than observations.
  - Robust to the collinearity between covariates.
  - Suited for fitting linear models with microarray data.

- Penalized Partial Least Squares (PPLS)
  - A penalized regression method built on the framework of PLS.

Application to the Minnesota data: PPLS

Minnesota data

- Oligonucleotide microarray data obtained by Hall et al. (2003) in a heart failure study conducted at the Medical School of UMN.
- Contain 30 samples: 10 ischemic, 7 ischemic with acute MI and 13 idiopathic.
- Affy HG-U133A chips: Contain 22,283 genes.
- Initially processed in MAS 5.0.

**Goal:** Distinguish between the ischemic and the idiopathic etiology classes.
Initial gene ranking

Given gene $i$

$$F_i = \frac{MS_{\text{class}}}{MS_{\text{error}}} = \frac{\sum_{c=1}^{C} n_c (\bar{x}_{ic} - \bar{x}_i)^2)/(C - 1)}{\left( \sum_{c=1}^{C} \sum_{j \in c} (x_{ij} - \bar{x}_{ic})^2 \right)/(n - C)}.$$  

$x_{ij}$: gene expression intensity of gene $i$ and sample $j$.

$n_c$: number of samples in class $c$, $n = \sum_{c=1}^{C} n_c$: total sample size.

$C$: number of classes.

$\bar{x}_{ic}$: mean gene expression of class $c$.

$\bar{x}_i$: overall mean gene expression.

Genes with larger F-statistics -> higher rank.
Experiment: Minnesota data

LOOCV error (Isch vs Idio, $n = 23$)

<table>
<thead>
<tr>
<th># of top genes</th>
<th>PPLS</th>
<th>SC</th>
<th>LASSO</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>10</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>100</td>
<td>7</td>
<td>5</td>
<td>7</td>
</tr>
<tr>
<td>200</td>
<td>9</td>
<td>7</td>
<td>11</td>
</tr>
<tr>
<td>800</td>
<td>6</td>
<td>8</td>
<td>4</td>
</tr>
<tr>
<td>1600</td>
<td>5</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>9600</td>
<td>9</td>
<td>8</td>
<td>7</td>
</tr>
<tr>
<td>16000</td>
<td>8</td>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td>22283</td>
<td>9</td>
<td>5</td>
<td>7</td>
</tr>
</tbody>
</table>
Application to the PGA data: PPLS

PGA data

- Oligonucleotide microarray data obtained in a heart failure study conducted at the PGA Medical School.
- Contain 36 samples: 11 normal, 11 ischemic, and 14 idiopathic.
- Affy HG-U133 plus 2 chips: Contain ~ 54,000 genes.
- Initially processed in MAS 5.0.

Goal: Distinguish between the ischemic and the idiopathic etiology classes.
Experiment: PGA data

LOOCV error (Isch vs Idio, $n = 25$)

<table>
<thead>
<tr>
<th># of top genes</th>
<th>PPLS</th>
<th>SC</th>
<th>LASSO</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>3</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>100</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>200</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>800</td>
<td>1</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>1600</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>9600</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>16000</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>22277</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
A summary on the above experiments

- The LOOCV misclassification error
  - ranges from 5 to 11 for the Minnesota data (23 samples).
  - ranges from 1 to 3 for the PGA data (25 samples).

- These highlight some existing differences underlying the two datasets.

- A question: Is there any signal/predictive information in the data?
### Permutation test: Minnesota data

#### LOOCV error (Isch vs Idio, $n = 23$)

<table>
<thead>
<tr>
<th># of top genes</th>
<th>Original data</th>
<th>Permutated data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CV errors</td>
<td>P-value</td>
</tr>
<tr>
<td>50</td>
<td>5</td>
<td>.00</td>
</tr>
<tr>
<td>100</td>
<td>5</td>
<td>.00</td>
</tr>
<tr>
<td>400</td>
<td>7</td>
<td>.06</td>
</tr>
<tr>
<td>1600</td>
<td>8</td>
<td>.08</td>
</tr>
<tr>
<td>6400</td>
<td>9</td>
<td>.12</td>
</tr>
</tbody>
</table>
Borrow information from relevant studies

- To increase the statistical power, borrow information from other relevant studies.

- A key difference from meta-analysis:
  - Not assuming current study shares a common set of parameters with other studies.

- Example: Identifying genes associated with ventilator-associated lung injury (VALI) based on a human study.
  - Meta-analysis: Only interested in the genes associated with VALI which are conserved across the species over the evolutionary history (Grigoryev et al. 2004).
  - Our analysis: Interested in inference on a set of parameters specific for humans.
Combining Minnesota and PGA data

- Classification with the combined data.
  - Goal: Distinguish etiologies of heart failure for Minnesota patients while treat the PGA data as secondary.
  - Problem: Unobserved differences in patient characteristics.
  - Solution: Treat samples in different studies unequally, e.g., assign different weights.

Combining Minnesota data and PGA data.
- Technically easy: all probe sets present on U133A chip are identically replicated on U133 Plus 2 chip.
- Data mapped by probe set ID (6 could not be found).
Notations

- Given $X$, to predict $Y$ with a linear model
  
  \[ F(X, a) = a_0 + \sum_{i=1}^{p} a_i x_i, \]
  
  the goal is to minimize the expected loss (risk):
  
  \[ R(a) = E_Y L(Y, F(X, a)). \]

- $L(Y, F(X, a))$: loss criterion.

- An empirical estimate of the expected loss:
  
  \[ \hat{R}(a) = \frac{1}{n} \sum_{j=1}^{n} L(y_j, a_0 + \sum_{i=1}^{p} a_i x_{ij}). \]

- The optimal values of $a$:
  
  \[ \hat{a} = \arg \min_a \frac{1}{n} \sum_{j=1}^{n} L(y_j, a_0 + \sum_{i=1}^{p} a_i x_{ij}). \]
Partial Least Squares (PLS)

Conjugate gradient procedure under squared error loss:

\[ \hat{a}_{k+1} = \hat{a}_k + \rho_k s_k \]
\[ s_k = g_k + \frac{g_k^T g_k}{g_{k-1}^T g_{k-1}} s_{k-1} \]

- \( g_k \): negative gradient at \( \hat{a}_k \)
  \[ g_k = -\frac{\partial}{\partial a} \hat{R}(a) \big|_{a=\hat{a}_k} \]

- \( \rho_k \): step size
  \[ \rho_k = \arg\min_\rho \hat{R}(\hat{a}_k + \rho s_k) \]

- \( k \): number of PLS components.

Squared error loss:
\[ L(Y, F(X, a)) = (Y - F(X, a))^2 / 2. \]
Propose: Weighted PLS

- Expected loss estimated by a weighted average loss:

\[
\hat{R}_w(a) = \sum_{j=1}^{n} w_j L(y_j, a_0 + \sum_{i=1}^{p} a_i x_{ij}).
\]

- Conjugate gradient procedure under squared error loss:

\[
g_k = -\frac{\partial}{\partial a} \hat{R}_w(a) \bigg|_{a=\hat{a}_k} = Z^T W (Y - Z\hat{a}_k)
\]

\[
\rho_k = \arg\min_{\rho} \hat{R}_w(\hat{a}_k + \rho s_k) = \begin{cases} 
1 & \text{if } Zs_k = 0 \\
\frac{(Zs_k)^T W (Y - Z\hat{a}_k)}{(Zs_k)^T W (Zs_k)} & \text{if } Zs_k \neq 0
\end{cases}
\]

- \(W = diag(w_1, \cdots, w_n)\): diagonal matrix with weights, \(\sum_{j=1}^{n} w_j = 1\).

- \(Z = \begin{pmatrix} 1 & X_1 & X_2 & \cdots & X_p \end{pmatrix}_{n \times (p+1)}\): covariate matrix.
Propose: Weighted PPLS

- Weighted PPLS
  - A weighted PLS model with a conjugate gradient path.
  - Penalized regression in the framework of weighted PLS.
  - Similar to the PPLS construction.

- Weighted PLS model: 
  \[ Y = b_0 + \sum_{i=1}^{p} b_i (X_i - \bar{x}_i 1) \]

- Penalize \( b_i \) by soft-thresholding:
  \[ b'_i = \text{arg} \min_{\beta_i} (\beta_i - b_i)^2 + \lambda |\beta_i| \]
  \[ b'_i = \text{sign}(b_i)(|b_i| - \lambda)_+ \]
  \[ f_+ = \max(f, 0) \]
  \( \lambda \): shrinkage parameter.
Experiment: Combined data

- LOOCV error of predicting Minnesota samples
- Weighted PPLS classifiers.
- Top 200 genes

<table>
<thead>
<tr>
<th>w = PGA</th>
<th>Shrink 0%</th>
<th>Shrink 40%</th>
<th>Shrink 80%</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PLS components k 1 2 3 4 5</td>
<td>PLS components k 1 2 3 4 5</td>
<td>PLS components k 1 2 3 4 5</td>
</tr>
<tr>
<td>0</td>
<td>5 9 6 7 7</td>
<td>5 8 8 6 5</td>
<td>5 8 7 6 7</td>
</tr>
<tr>
<td>1/4</td>
<td>6 8 6 6 5</td>
<td>5 7 4 3 5</td>
<td>8 7 2 4 3</td>
</tr>
<tr>
<td>1/2</td>
<td>5 8 6 4 6</td>
<td>6 7 3 4 5</td>
<td>7 7 2 3 2</td>
</tr>
<tr>
<td>3/4</td>
<td>4 7 7 4 6</td>
<td>6 7 4 4 6</td>
<td>8 6 2 2 2</td>
</tr>
<tr>
<td>1</td>
<td>5 7 7 4 7</td>
<td>6 7 4 4 6</td>
<td>8 5 3 2 2</td>
</tr>
</tbody>
</table>
Summary

- **Weighted PPLS:**
  - Penalized regression in a framework of weighted PLS.
  - Penalization/shrinking can improve over weighted PLS.

- **Weighted PPLS methods with combined data:**
  - Account for possible different relevances of the other studies by weighting.
  - Improve the performance of the classifier using data from a single study.
General application

- Broad scope of the weighting scheme:
  - Applicable when the PGA data only contain ischemic and normal groups.

- Further extendable:
  - The primary and secondary experiments were conducted under different (but relevant) conditions, or on different organisms.
  - Microarray data with a survival end point.
  - Other loss functions.