Variable Selection for High-Dimensional Multivariate Outcomes

Tamar Sofer, Lee Dicker and Xihong Lin

Department of Biostatistics
Harvard School of Public Health

July 29, 2012
Motivation

We study the effect of covariates on the mean of a set of outcomes.

- Study the association between a SET of exposures to a SET of outcomes (e.g. gene expressions)
- Looking at one gene at a time may not be powerful
- Model the entire pathway together
  - Potentially more efficient
The multivariate regression model

\[ y_{ij} = x_i \beta_j + \epsilon_{ij}, \quad i = 1, \ldots, n \]

Where
- \( x_i \) is \( p_0 \times 1 \)
- \( \beta_j \) is a \( p_0 \times 1 \) vector modeling the effect of the covariates on the \( j \)th outcome.

Or in matrix form:

\[ y_i = X_i \beta + \epsilon_i, \quad i = 1, \ldots, n \]

- \( y_i \) is \( m \times 1 \)
- \( X_i = I_m \otimes x_i \) is \( m \times p \) with \( p = p_0 m \)
- \( \beta \) is \( p \times 1 \) (stacked \( \beta_j \)'s)
- \( \epsilon_i \) is \( m \times 1 \) and \( \epsilon_i \sim \mathcal{N}(0, \Sigma) \)
  - \( \Sigma^{-1} = \Omega \)

No structure is assumed on \( \beta \).
The likelihood

The model log-likelihood (up to a constant):

$$
\ell(y_i, x_i | \beta) = \log(|\Omega|) - (y_i - X_i\beta)^T \Omega(y_i - X_i\beta)
$$

Goal: estimate $\beta$ sparsely.

Alternative approaches:

- Set $\Omega = \Lambda$, where $\Lambda$ is a fixed working (possibly misspecified) precision matrix.
- Estimate both $\beta$ and $\Omega$ jointly.
Part 1: Regression parameter estimation with a fixed working precision matrix

- The working log-likelihood is:

\[ \ell(y_i, x_i | \beta) = - (y_i - X_i \beta)^T \Lambda (y_i - X_i \beta) \]

with \( \Lambda \) fixed, possibly misspecified.

- Estimate \( \beta \) from

\[
\min_{\beta} \left\{ \sum_{i=1}^{n} (y_i - X_i \beta)^T \Lambda (y_i - X_i \beta) + \sum_{j=1}^{p} P_\lambda(\beta_j) \right\}
\]

* \( P_\lambda(\cdot) \) could be (for instance) the Lasso (\( \ell_1 \)) penalty, or an oracle penalty - our focus!
Estimation

To estimate $\beta$:

- Transform $\tilde{X}_i \leftarrow \Lambda^{1/2}X_i$, $\tilde{Y}_i \leftarrow \Lambda^{1/2}Y_i$

- Standardize $\tilde{X}$, $\tilde{Y}$ to have variance 1 and restructure in i.i.d regression form

- We use a cyclical coordinate descent algorithm

Now select the best $\widehat{\beta}(\lambda)$!
Tuning parameter selection

- **Data validation**: minimize the prediction error.
  - Use part of the data $X_{\text{train}}, Y_{\text{train}}$ to estimate $\hat{\beta}_{\text{train}}$
  - Use the rest of the data $X_{\text{test}}, Y_{\text{test}}$ to calculate prediction error. Minimize:

$$PE(\hat{\beta}(\lambda)) = \sum_{i=1}^{n} \|y_{i,\text{test}} - X_{i,\text{test}}\hat{\beta}_{\text{train}}(\lambda)\|^2$$

- **BIC**. Minimize:

$$\text{BIC}(\hat{\beta}(\lambda)) = \sum_{i=1}^{n} (y_{i} - X_{i}\hat{\beta}(\lambda))^T \Lambda (y_{i} - X_{i}\hat{\beta}(\lambda)) + \hat{s}k_n$$

$$\hat{s} = |\{j : \hat{\beta}_j \neq 0\}|$$
Some notation

- $A = \{ j : \beta_j \neq 0 \}$ (the ‘true model’)
- $\hat{A} = \{ j : \hat{\beta}_j \neq 0 \}$ (an estimated model)
- $\beta_{\hat{A}}$ - the restriction of $\beta$ to the coordinates in $\hat{A}$
- $\lambda_{\hat{A}}$ - a tuning parameter value that yielded model $\hat{A}$ for $\hat{\beta}(\lambda_{\hat{A}})$

Example:

\[
\beta = (1, 0, 0, 3, 0, 0, 0)^T, \quad A = \{1, 4\}, \quad \beta_A = (1, 3)^T
\]
\[
\hat{\beta} = (1.1, 0, 0.3, 3.4, 0, 0, 0)^T, \quad \hat{A} = \{1, 3, 4\}, \quad \hat{\beta}_A = (1.1, 3.4)^T
\]
Asymptotic results, for an arbitrary $\Lambda$, oracle penalty

<table>
<thead>
<tr>
<th>When $p/n \to 0$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Consistency:</strong> $</td>
</tr>
<tr>
<td><strong>Sparsity:</strong> $P\left(\text{For all } j \text{ s.t. } \beta_j = 0, \hat{\beta}_j = 0\right) \to 1$</td>
</tr>
<tr>
<td><strong>Asymptotic normality:</strong> $\hat{\beta}_A \overset{\mathcal{L}}{\to} N(\beta_A, \tilde{\Sigma})$, if $\Lambda = \Omega$ then $\tilde{\Sigma} = \Sigma$ ($\hat{\beta}_A$ efficient)</td>
</tr>
<tr>
<td><strong>There exists</strong> $k_n \to \infty$ <strong>such that for all</strong> $\hat{A} \neq A$</td>
</tr>
<tr>
<td>$P\left(BIC(\hat{\beta}(\lambda_{\hat{A}}) - BIC(\hat{\beta}(\lambda_A)) &gt; 0\right) \to 0$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>When $p &gt; n$, $\log(p)/n \to 0$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Consistency:</strong> $</td>
</tr>
<tr>
<td><strong>Sparsity:</strong> $P\left(\text{For all } j \text{ s.t. } \beta_j = 0, \hat{\beta}_j = 0\right) \to 1$.</td>
</tr>
</tbody>
</table>
Joint regression parameters and precision matrix estimation

- For $\Lambda$ an arbitrary precision matrix, $\hat{\beta}$ is consistent.
- Most efficient if $\Lambda = \Omega$

**Estimate $\beta$ and $\Omega$ jointly!**

The objective function for joint estimation:

$$Q(\beta, \Omega) = -n \log(|\Omega|) + \sum_{i=1}^{n} (y_i - X_i \beta)^T \Omega (y_i - X_i \beta)$$

$$+ \sum_{j=1}^{p} P_{\lambda}(|\beta_j|) + \sum_{k=1}^{m} \sum_{l=1 \neq k}^{m} P_{\gamma}(|\omega_{kl}|)$$
Joint regression parameters and precision matrix estimation

Let

\[
Q(\beta|\Omega) = \sum_{i=1}^{n} (y_i - X_i \beta)^T \Omega (y_i - X_i \beta) + \sum_{j=1}^{p} P_\lambda(|\beta_j|)
\]

(Estimate \(\beta\) given \(\Omega\))

\[
Q(\Omega|\beta) = -n \log(|\Omega|) + \sum_{i=1}^{n} (y_i - X_i \beta)^T \Omega (y_i - X_i \beta) + \sum_{k=1}^{m} \sum_{l=1}^{m} P_\gamma(|\omega_{kl}|)
\]

(Estimate \(\Omega\) given \(\beta\))
The two-stage procedure for Joint estimation

Stage 1 Consistent estimation:
- $\hat{\beta}^{(1)} = \min_{\beta} Q(\beta | I_m)$
- $\hat{\Omega}^{(1)} = \min_{\Omega} Q(\Omega | \hat{\beta}^{(1)})$

Stage 2 Efficient estimation:
- $\hat{\beta}^{(2)} = \min_{\beta} Q(\beta | \hat{\Omega}^{(1)})$
- $\hat{\Omega}^{(2)} = \min_{\Omega} Q(\Omega | \hat{\beta}^{(2)})$

BIC for joint estimation

$$\text{BIC}(\hat{\beta}^{(2)}(\lambda)) = -n \log(|\hat{\Omega}^{(2)}|) - \sum_{i=1}^{n} (y_i - X_i\hat{\beta}^{(2)})^T \hat{\Omega}^{(2)} (y_i - X_i\hat{\beta}^{(2)}) + \hat{sk}_n$$

- 2 log likelihood
The two-stage procedure for Joint estimation

- To solve $\min_{\beta} Q(\beta|\tilde{\Omega})$, we use the procedure described before.
  - i.e. treat $I_m$ or $\tilde{\Omega}^{(1)}$ as a working precision matrix.
- To solve $\min_{\Omega} Q(\Omega|\tilde{\beta})$, we use the Graphical Lasso (GLASSO).
  - The GLASSO takes $\hat{\Sigma}(\hat{\beta})$ and estimates $\hat{\Sigma}(\hat{\beta})^{-1} \equiv \hat{\Omega}$.
  - Tuning parameter for penalized $\hat{\Omega}$ is selected via data validation or BIC.
Asymptotic results, the two stage algorithm, large $p$

- Oracle penalty is used
- Regularity conditions...

When $pm/n \to 0$, $m^2/n \to 0$

- $\Omega$ estimation:
  - Consistency: $\|\Omega - \hat{\Omega}\|^2 = O_p((m + p)m/n)$
  - Sparsity: $P\left(\text{For all } k, l \text{ s.t. } \omega_{kl} = 0, \hat{\omega}_j = 0\right) \to 1$

- $\beta$ estimation:
  - Consistency: $\|\beta - \hat{\beta}\|^2 = O_p(p/n)$
  - Sparsity: $P\left(\text{For all } j \text{ s.t. } \beta_j = 0, \hat{\beta}_j = 0\right) \to 1$
  - Asymptotic normality: $\hat{\beta}_A \overset{d}{\to} N(\beta_A, \Sigma)$
Asymptotic results, the two stage algorithm, very large $p$

- Oracle penalty is used
- Regularity conditions...

When $p > n$, $m \log(p)/n \to 0$, $m^2/n \to 0$

**$\Omega$ estimation:**
- **Consistency:** $||\Omega - \hat{\Omega}||^2 = O_p((m + \log(p))m/n)$
- **Sparsity:** $P\left(\text{For all } k, l \text{ s.t. } \omega_{kl} = 0, \hat{\omega}_j = 0\right) \to 1$

**$\beta$ estimation:**
- **Consistency:** $||\beta - \hat{\beta}||^2 = O_p(\log(p)/n)$
- **Sparsity:** $P\left(\text{For all } j \text{ s.t. } \beta_j = 0, \hat{\beta}_j = 0\right) \to 1$. 
Simulation results

Compare between the following scenarios:

\[ p = 20 \quad p = 20 \]
\[ m = 5 \quad m = 20 \]

- All simulations had \( n = 50 \) (additional \( n = 50 \) for data validation)
- Outcomes covariance matrix: \( \text{EX. } \sigma^2 = 3 \)
- 3 non-zero coefficients in the \( \beta \) matrix, \( (3, 1.5, 2) \).
- Compare between the penalties Lasso, Adaptive Lasso, SCAD and SELO.
Simulations results

Performance measures:

- **T** (‘true model’): $\mathbf{1}$ (The true model was selected)
- **FP** (‘false positives’): the number of $\beta_j = 0$ but $\hat{\beta}_j \neq 0$
- **FN** (‘false negatives’): the number of $\beta_j \neq 0$ but $\hat{\beta}_j = 0$
- **sq ME** (‘squared model error’): $(\hat{\beta} - \beta)^T \Sigma_X (\hat{\beta} - \beta)$
Simulation results: $p = 20, m = 5$

**BIC**

<table>
<thead>
<tr>
<th>method</th>
<th>mean T</th>
<th>mean FP</th>
<th>mean FN</th>
<th>mean sq</th>
<th>ME</th>
</tr>
</thead>
<tbody>
<tr>
<td>lasso</td>
<td>0.10</td>
<td>3.91</td>
<td>0.00</td>
<td>0.87</td>
<td></td>
</tr>
<tr>
<td>alasso</td>
<td>0.00</td>
<td>9.73</td>
<td>0.55</td>
<td>2.95</td>
<td></td>
</tr>
<tr>
<td>scad</td>
<td>0.47</td>
<td>1.86</td>
<td>0.00</td>
<td>0.30</td>
<td></td>
</tr>
<tr>
<td>selo</td>
<td>0.80</td>
<td>0.26</td>
<td>0.00</td>
<td>0.22</td>
<td></td>
</tr>
</tbody>
</table>

**Data validation**

<table>
<thead>
<tr>
<th>method</th>
<th>mean T</th>
<th>mean FP</th>
<th>mean FN</th>
<th>mean sq</th>
<th>ME</th>
</tr>
</thead>
<tbody>
<tr>
<td>lasso</td>
<td>0.00</td>
<td>12.48</td>
<td>0.00</td>
<td>0.78</td>
<td></td>
</tr>
<tr>
<td>alasso</td>
<td>0.00</td>
<td>25.00</td>
<td>0.19</td>
<td>2.38</td>
<td></td>
</tr>
<tr>
<td>scad</td>
<td>0.21</td>
<td>5.54</td>
<td>0.00</td>
<td>0.27</td>
<td></td>
</tr>
<tr>
<td>selo</td>
<td>0.79</td>
<td>0.46</td>
<td>0.00</td>
<td>0.21</td>
<td></td>
</tr>
</tbody>
</table>
Simulation results: \( p = 20, \, m = 20 \)

Data validation

<table>
<thead>
<tr>
<th></th>
<th>mean T</th>
<th>mean FP</th>
<th>mean FN</th>
<th>mean sq</th>
<th>ME</th>
</tr>
</thead>
<tbody>
<tr>
<td>lasso</td>
<td>0.00</td>
<td>24.7</td>
<td>0.00</td>
<td>1.02</td>
<td></td>
</tr>
<tr>
<td>alasso</td>
<td>0.00</td>
<td>19.94</td>
<td>0.65</td>
<td>4.01</td>
<td></td>
</tr>
<tr>
<td>scad</td>
<td>0.04</td>
<td>11.22</td>
<td>0.00</td>
<td>0.35</td>
<td></td>
</tr>
<tr>
<td>selo</td>
<td>0.72</td>
<td>1.19</td>
<td>0.00</td>
<td>0.25</td>
<td></td>
</tr>
</tbody>
</table>
Data analysis

The diabetes data set:

- Gene expression profiles of 43 males
- Similar ages
- Three Glucose Tolerance levels:
  - Normal (NGT) - 17 subjects
  - Impaired (IGT) - 9 subjects
  - Type 2 Diabetes Mellitus (DM2) - 17 subjects
- The porphyrin and chlorophyll metabolism pathway:
  - 35 probes measuring expression of genes

We studied the effect of Glucose intolerance levels on gene expression

- The 18 expressed probes were used
- We estimated the change in expression from baseline (NGT)
### Data analysis results

<table>
<thead>
<tr>
<th>Gene (probe)</th>
<th>Lasso</th>
<th>Adaptive Lasso</th>
<th>Scad</th>
<th>SELO</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>IGT</td>
<td>DM</td>
<td>NGT</td>
<td>DM</td>
</tr>
<tr>
<td>EPRS (1)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>EPRS (2)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>EPRS (3)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>BLVRB</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>GUSB</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>UROS</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>HMBS</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>FECH</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>HMOX1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>HCCS (1)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>HCCS (2)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>BLVRA (1)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>CP</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>UROD (1)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>UROD (2)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>BLVRA (2)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ADH6</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>HMOX2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Purple: estimated as non-zero. Dark: significant at the 0.05 level.
Thank you!
The two-stage algorithm