Model Selection for Correlated Data with Diverging Number of Parameter

Annie Qu
University of Illinois at Urbana-Champaign
Joint Work with Peng Wang and Jianhui Zhou

Dec. 12-16, 2011
A Data Example

- Impact of air pollution on asthmatic patients, Ontario, 1992.
- Based on 39 patients, cluster size is 21.
- Response: observations of asthmatic status on 21 consecutive days, i.e. presence (1) or absence (0) of difficulties in breathing.
- Covariates: pollution levels of 7 pollutants, daily mean temperature and daily mean humidity, total 9 covariates.
- **GEE method with “unspecified” correlation structure does not converge**
Importance of Selecting the Correct Correlation Structure

- Improve efficiency of regression parameter estimation.
- Reduce the bias of parameter estimation in nonparametric modeling (Wang, 2003)
- Increase statistical power for hypothesis testing.
Our Approach

- Current literature focuses on the estimation of covariance matrix: Huang et al., 2007, 2008 (Cholesky decomposition) Bickel and Levina, 2008a; 2008b (tapering and banding, threshholding); Rothman et al., 2009 (inverse of covariance); Cai et al., 2010; Yuan, 2010 (multivariate linear regression)

- Our approach avoids the estimation of each individual entry of the correlation matrix, useful when cluster size is large.

- Reduce the dimension of the parameter involved in the estimation.

- Does not require the specification of the likelihood.

- Can be applied to non-normal response.

- Diverging cluster sizes

- Enjoys consistency and oracle property
Consider the marginal model

\[ E(y_i) = g(X_i \beta), \quad i = 1, \ldots, n \]

- \( y_i = (y_{i1}, \ldots, y_{im})' \) is the response variable
- \( t = 1, \ldots, m \) are the time points
- \( X_i \) is a known \( m \times \text{dim}(\beta) \) covariate matrix
- \( \beta \) is a parameter vector
- \( g(\cdot) \) is the link function
Groups of Basis Matrices

- In the quadratic inference function approach (Qu, Lindsay and Li, 2000), $R^{-1} \approx \sum_{j=1}^{t} a_j M_j$

- (Zhou and Qu, 2012) The basis matrices can be divided into different groups, i.e.

$$R^{-1} \approx \sum_{j=1}^{J_m} \sum_{b=1}^{B_j} \alpha_{jb} M_{jb} = \sum_{j=1}^{J_m} \alpha_j G_j$$

- $M_{jb}$ is the $b$th basis matrix in the $j$th group
- The $j$th group $G_j$ consisting of $B_j$ basis matrices $M_{j1}, \ldots, M_{jB_j}$
- The associated coefficient vector $\alpha_j = (\alpha_{j1}, \ldots, \alpha_{jB_j})$
Example 1: AR(1) Correlation Structure

If $R$ has an AR(1) structure with the parameter $\rho$, $R^{-1}$ can be represented as

$$R^{-1} = \alpha_{11} I_m + \alpha_{21} M_{2,1} + \alpha_{22} M_{2,2}$$

- $I_m$ is the identity matrix in group $G_1$
- $M_{2,1}$ and $M_{2,2}$ are two basis matrices in group $G_2$
- $M_{2,1}$ has 1 on the sub-diagonal, and 0 elsewhere
- $M_{2,2}$ has 1 on the $(1, 1)$ and $(m, m)$ components and, 0 elsewhere
- $\alpha_{11} = (1 + \rho^2)/(1 - \rho^2)$ and $\alpha_2 = (\alpha_{21}, \alpha_{22}) = (-\rho/(1 - \rho^2), -\rho^2/(1 - \rho^2))$
Example 2: Exchangeable Correlation Structure

If \( R \) is exchangeable with the correlation parameter \( \rho \), we have

\[
R^{-1} = \alpha_{11} I_m + \alpha_{31} M_{3,1}
\]

- \( I_m \) is the identity matrix in group \( G_1 \)
- The second basis matrix \( M_{3,1} \) has 0 on its main diagonal, and 1 elsewhere
- \( \alpha_{11} = -\{(m - 2)\rho + 1\}/\{(m - 1)\rho^2 - (m - 2)\rho - 1\} \) and \( \alpha_{31} = \rho/\{(m - 1)\rho^2 - (m - 2)\rho - 1\} \)
Example 3: Sub Block Structures

- R has a block diagonal matrix structure
- Each block is either independent, exchangeable or AR(1)
- Group $G_1$ contains the identity matrix $I_m$, and $d - 1$ matrices with block identity matrices $I_{m_i}$ ($i = 1, \ldots, d - 1$) on the first, \ldots, and $(d - 1)$th block
- For any $j$th block with AR(1) structure, the group basis matrices contain two basis matrices $M_{2,1}$ and $M_{2,2}$ as provided in Example 1
- For any block with exchangeable structure, the group basis matrices contain a basis matrix $M_{3,1}$ for the corresponding block
Selection Strategy

- Identifying which groups of basis matrices have non-zero coefficients
- Achieved by minimizing an objective function including two parts
  1. Discrepancy between the two estimating functions
     - One based on the empirical estimation
     - The other based on the approximation by basis matrices
  2. A penalty function is added to balance the complexity and sufficiency of the model
The objective functions includes two parts, the Euclidean norm of $S$ and a penalty function, i.e.

$$
\sum_{i=1}^{n} S_i^T S_i + n \dim(\beta) \sum_{j=2}^{J_m} p_\lambda(\|\alpha_j\|_2),
$$

where the discrepancy between the two estimating functions for the $i$th cluster is

$$S_i = \hat{\mu}_i^T (\hat{\beta}) A_i^{-1/2} \{ \tilde{R}^{-1} - \alpha_1 G_1 - \cdots - \alpha_{J_m} G_{J_m} \} A_i^{-1/2} (y_i - \mu_i(\hat{\beta}))$$
Objective Function (Con’t)

- $p_\lambda(\cdot)$ is the SCAD penalty function and $\lambda$ is the tuning parameter.
- $||\alpha_j||_2$ is the $L_2$-norm of $\alpha_j$.
- By imposing the $L_2$-norm, the basis matrices within the same group are selected simultaneously.
- The first group of basis matrices is not penalized.
Minimizing the Objective Function

▶ Define

\[ U_i = \hat{\mu}_i^T (\hat{\beta}) A_i^{-1/2} \tilde{R}^{-1/2} A_i^{-1/2} \{ y_i - \mu_i(\hat{\beta}) \}, \quad i = 1, \ldots, n \]

\[ V_{i,jb} = \hat{\mu}_i^T (\hat{\beta}) A_i^{-1/2} M_{jb} A_i^{-1/2} \{ y_i - \mu_i(\hat{\beta}) \} \]

\[ j = 1, \ldots, J_m, b = 1, \ldots, B_j \]

▶ Let \( V_{ij} = (V_{i,j1}, \ldots, V_{i,jB_j}) \) and \( V_i = (V_{i1}, \ldots, V_{iJ_m})^T \)

▶ Then the objective function can be written as

\[ Q(\alpha) = \sum_{i=1}^{n} \left\| U_i - \sum_{j=1}^{J_m} V_{ij} \alpha_j \right\|^2 + n \dim(\beta) \sum_{j=2}^{J_m} p_\lambda(\|\alpha_j\|) \]
Minimizing the objective function

- Transform the correlation model selection problem to be covariates model selection
- Has the same form as a penalized least square problems
- Group SCAD penalty, non-convex penalty
- Apply the one-step local approximation to SCAD penalty (Zou and Li 2008)
A New Criteria

- Choose the tuning parameter $\lambda$ using a GIC type of criteria

$$GIC_T(\lambda) = nr \log \frac{\eta_{\max} (\hat{R}^{-1} \tilde{R}^2 \hat{R}^{-1})}{\eta_{\min} (\hat{R}^{-1} \tilde{R}^2 \hat{R}^{-1})} + \log(n) k(\lambda).$$  \hspace{1cm} (1)

- $\tilde{R}$ is the empirical correlation matrix
- $\hat{R}^{-1} = \hat{\alpha}_1 G_1 + \cdots + \hat{\alpha}_J G_J$ and $\hat{\alpha}_1, \ldots, \hat{\alpha}_J$ are estimated with $\lambda$
- $k(\lambda)$ is the number of non-zero components among $\hat{\alpha}_1, \ldots, \hat{\alpha}_J$
- $\eta_{\max}(\cdot)$ is the largest eigenvalue and $\eta_{\min}(\cdot)$ is the smallest eigenvalue
- **Require an additional tuning parameter $r$**
Choice of $r$

- Analog to generalized information criteria
- Additional control over the choice of $\lambda$
- Larger $r \Rightarrow$ Smaller $\lambda \Rightarrow$ More groups of basis matrices selected
- Choose $r = m/n$, the ratio of cluster size and sample size
- Outperforms GCV, AIC and BIC
Conditions on the Penalty Function

Define

\[ a_n = \max_{1 \leq j \leq p_m} \{ p_{\lambda_n}'(\alpha_{j0}^i), \alpha_{0}^j \neq 0 \} \]

\[ b_n = \max_{1 \leq j \leq p_m} \{ p_{\lambda_n}''(\alpha_{j0}^i), \alpha_{0}^j \neq 0 \} \]

The following conditions are associated with the penalty functions:

a. \[ a_n = O(n^{-1/2}) \]

b. \[ b_n \to 0 \text{ as } n \to \infty \]

c. \[ \liminf_{n \to \infty} \liminf_{\theta \to 0^+} p_{\lambda_n}'(\theta)/\lambda_n > 0 \]

d. There are constants \( c_1 \) and \( c_2 \), such that when \( \theta_1, \theta_2 > c_1 \lambda_n \),
\[ |p_{\lambda_n}''(\theta_1) - p_{\lambda_n}''(\theta_2)| \leq c_2 |\theta_1 - \theta_2| \]
Other Regularity Conditions

- Each element of the empirical correlation matrix is consistent
  \[ \sqrt{n} | \tilde{R}(i,j) - R(i,j) | = O_p(1), \ 1 \leq i \leq m, 1 \leq j \leq m \]

- For any \( \epsilon > 0 \), there exist constants \( l_1 \) and \( l_2 \) such that
  \[ P(0 < l_1 < \lambda_{\min}\{V_i^T V_i\} \leq \lambda_{\max}\{V_i^T V_i\} < l_2 < \infty ) > 1 - \epsilon \]

- The \( L_1 \) norm of the basis matrices is bounded, i.e., there is a constant \( K \) such that
  \[ ||M_{jb}||_1 < K, \ 1 \leq j \leq J_m, \ b = 1, \ldots, B_j \]
Theorem 1

Suppose the regularity conditions 1-4 are satisfied, if $p_m^2/n \to 0$ as $n \to \infty$, then there is a local minimizer $\hat{\alpha}$ for minimizing the objective function $Q(\alpha)$, such that

$$||\hat{\alpha} - \alpha_0|| = O_p\{\sqrt{p_m(n^{-1/2} + a_n)}\},$$

where $a_n$ is given in Condition 1 and $\alpha_0 = (\alpha_{01}, \ldots, \alpha_{0J_m})$ is the true coefficient vector associated with all the basis matrices.

- For the SCAD penalty, $a_n = 0$ when $n$ is large, therefore the SCAD estimator is consistent.
Theorem 2

Given all the regularity conditions are satisfied, if \( \lambda_n \to 0 \), \( \sqrt{n/p_m} \lambda_n \to \infty \) and \( p^2_m/n \to 0 \), then with probability tending to 1, for any given constant \( C \), and any \( \alpha_1 \) satisfying
\[
||\alpha_1 - \alpha_{01}|| = O_p(\sqrt{p_m/n}),
\]
\[
Q(\hat{\alpha}_1, 0) = \min_{||\alpha_2|| \leq C(p_m/n)^{1/2}} Q(\alpha_1, \alpha_2).
\]

- \( \hat{\alpha}_1 \) is the estimate for the non-zero coefficients
- Estimates of the zero-coefficients are shrunk to 0
Theorem 3: Oracle Property

Suppose all the regularity conditions are satisfied, if \( \lambda_n \to 0, \) \( \sqrt{n/p_m}\lambda_n \to \infty \) and \( p_m^2/n \to 0 \) as \( n \to \infty \), then with probability tending to 1, we establish the following oracle properties:

(i) (Sparsity) \( \hat{\alpha}_2 = 0 \).

(ii) (Asymptotic normality)

\[
\sqrt{n}A_mK_{m,11}^{-1/2}\left\{I_{n,11} + \frac{1}{n}\nabla^2P_{\lambda_n}(\alpha_{01})\right\}(\hat{\alpha}_{01} - \alpha_{01}) \\
+ \frac{1}{\sqrt{n}}A_mK_{m,11}^{-1/2}\nabla P_{\lambda_n}(\alpha_{01}) \xrightarrow{d} N(0, G),
\]

- \( A_m \) is any given \( q \times p_m \) matrix which satisfies \( A_m^T A_m \to G \)
- \( K_{m,11} \) is a submatrix of \( K_m \) associated with \( \alpha_1 \).
Simulation Setup

- $R$ is a block diagonal matrix, and each block with dimension $5 \times 5$ has a correlation structure either as AR(1), exchangeable or independent
- The number of blocks $d$ diverges
- $d = 5, 10, 15$ and $20 \Rightarrow m = 25, 50, 75$ and $100$
- Basis Matrices
  - $G_1$ contains the identity matrix $I_{5d}$, and $d - 1$ matrices with block identity matrices $I_5$ on the diagonal
  - Group $G_2$ contains two matrices with $M_{2,1}$ and $M_{2,2}$ on the first block
  - Group $G_3$ contains one matrix with $M_{3,1}$ for the first block
  - Other groups of basis matrices formed similarly, total $2d + 1$ groups
Normal Response

For the normal response, we generate the data from the following longitudinal model,

\[ Y_i = \beta_0 + X_{1i}\beta_1 + X_{2i}\beta_2 + X_{3i}\beta_3 + \epsilon_i \]

- \( X_{ti}, t = 1, 2, 3 \) are the covariates generated from \( N(0, 1) \)
- \( \epsilon_i \sim N(0, R) \)
- First two blocks are AR(1), the third block is exchangeable, the remaining blocks are independent
- The covariates \( \beta = (\beta_0, \beta_1, \beta_2, \beta_3)^T = (2, 1, 1, 1)^T \)
- Sample size \( n = 200 \)
Results for Normal Response: $\rho = 0.7$

Table: Percentages of correctly identified signals and non-signals using GIC criteria with correlation $\rho = 0.7$, sample size $n = 200$, results are from 100 simulations.

<table>
<thead>
<tr>
<th>Cluster size</th>
<th>$r$</th>
<th>Signals</th>
<th>Non-signals</th>
<th>% of fits</th>
</tr>
</thead>
<tbody>
<tr>
<td>$m = 25$</td>
<td>0.125</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td></td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>$m = 50$</td>
<td>0.250</td>
<td>99</td>
<td>100</td>
<td>99.9</td>
</tr>
<tr>
<td></td>
<td></td>
<td>100</td>
<td>99.9</td>
<td>0.98</td>
</tr>
<tr>
<td>$m = 75$</td>
<td>0.375</td>
<td>96</td>
<td>97</td>
<td>99.9</td>
</tr>
<tr>
<td></td>
<td></td>
<td>97</td>
<td>99.9</td>
<td>0.92</td>
</tr>
<tr>
<td>$m = 100$</td>
<td>0.500</td>
<td>97</td>
<td>98</td>
<td>98</td>
</tr>
<tr>
<td></td>
<td></td>
<td>96</td>
<td>98</td>
<td>0.72</td>
</tr>
</tbody>
</table>

- % of correct-fitting decreases as the number of block increases
- % of identifying the AR(1) and exchangeable correlation structures are high even when $m = 100$
For the binary response, the responses are generated from the logistic regression model

\[
\text{logit}\{E(Y_i)\} = \beta_0 + X_1i\beta_1 + X_2i\beta_2 + X_3i\beta_3,
\]

- \(X_{ti}(t = 1, 2, 3)\) are the covariates, generated from a normal distribution \(N(0, 0.01)\)
- First two blocks are exchangeable, and the third block is AR(1)
- The covariates \(\beta = (\beta_0, \beta_1, \beta_2, \beta_3)^T = (0.2, 1, -1, -1)^T\)
- Correlation Parameter \(\rho = 0.6\)
Results for Binary Response: $n = 300$

Table: Percentages of Correctly Identified Signals and Non-signals using GIC criteria with correlation $\rho = 0.6$, Binary response, sample size $n = 300$

<table>
<thead>
<tr>
<th>Cluster size $m$</th>
<th>$r$</th>
<th>Signals</th>
<th>Non-signals</th>
<th>% of fits</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Correct</td>
<td>Under</td>
<td>Over</td>
</tr>
<tr>
<td>$m = 25$</td>
<td>0.833</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>$m = 50$</td>
<td>0.167</td>
<td>99</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>$m = 75$</td>
<td>0.250</td>
<td>94</td>
<td>98</td>
<td>97</td>
</tr>
<tr>
<td>$m = 100$</td>
<td>0.333</td>
<td>89</td>
<td>94</td>
<td>91</td>
</tr>
</tbody>
</table>

- Results similar to normal response with $\rho = 0.7$
- $r = m/n$ is a reasonable choice
Air Pollution Data Set

- Impact of air pollution on asthmatic patients
- Based on 39 patients, cluster size is 21
- Response: observations of asthmatic status on 21 consecutive days, i.e. presence (1) or absence (0) of difficulties in breathing
- Covariates: pollution levels of 7 pollutants, daily mean temperature and daily mean humidity, total 9 covariates
Basis Matrices

- Group 1: Identity matrix: $I_{21}$
- Group 2: $M_{2,1}$ and $M_{2,2}$ to represent the AR(1) structure as in Example 1
- Group 3: $M_{3,1}$ to represent the exchangeable working correlation as in Example 2
- Group 4: Four additional matrices needed to represent the mixture of AR(1) and CS
- Group 5-11: Groups of basis matrices to represent the sub block structures as in Example 3 (3 sub blocks, each week is a sub block)
Results of Correlation Structure Selection

- AIC, BIC, and GCV selects all the basis matrices, except exchangeable for the third block
- GIC with $r = 21/39$ identifies the correlation structure as a simple exchangeable structure
## Comparison of GEE Estimators with Different Working Structures

<table>
<thead>
<tr>
<th>Effects</th>
<th>Independent</th>
<th>GIC</th>
<th>GCV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Meantemp</td>
<td>-0.2494</td>
<td>-0.1009</td>
<td>0.0660</td>
</tr>
<tr>
<td>s.e.</td>
<td>0.2563</td>
<td>0.0908</td>
<td>0.0892</td>
</tr>
<tr>
<td>z-value</td>
<td>-0.9733</td>
<td>-1.1112</td>
<td>0.7403</td>
</tr>
<tr>
<td>NO</td>
<td>0.2860</td>
<td>0.0553</td>
<td>-0.1362</td>
</tr>
<tr>
<td>s.e.</td>
<td>0.3419</td>
<td>0.1170</td>
<td>0.1178</td>
</tr>
<tr>
<td>z-value</td>
<td>0.8365</td>
<td>0.4724</td>
<td>-1.1555</td>
</tr>
<tr>
<td>NO2</td>
<td>-0.0105</td>
<td>-0.0335</td>
<td>0.0133</td>
</tr>
<tr>
<td>s.e.</td>
<td>0.0728</td>
<td>0.0235</td>
<td>0.0179</td>
</tr>
<tr>
<td>z-value</td>
<td>-0.1447</td>
<td>-1.4218</td>
<td>0.7425</td>
</tr>
<tr>
<td>NOX</td>
<td>-0.2717</td>
<td>-0.0728</td>
<td>0.0700</td>
</tr>
<tr>
<td>s.e.</td>
<td>0.1904</td>
<td>0.0676</td>
<td>0.0679</td>
</tr>
<tr>
<td>z-value</td>
<td>-1.4268</td>
<td>-1.0778</td>
<td>1.0298</td>
</tr>
<tr>
<td>TRS</td>
<td>-0.1784</td>
<td>-0.0037</td>
<td>-0.0063</td>
</tr>
<tr>
<td>s.e.</td>
<td>0.0947</td>
<td>0.0413</td>
<td>0.0340</td>
</tr>
<tr>
<td>z-value</td>
<td>-1.8836</td>
<td>-0.0892</td>
<td>-0.1856</td>
</tr>
<tr>
<td>OZ</td>
<td>0.1266</td>
<td>0.1190</td>
<td>0.1082</td>
</tr>
<tr>
<td>s.e.</td>
<td>0.1023</td>
<td>0.0341</td>
<td>0.0290</td>
</tr>
<tr>
<td>z-value</td>
<td>1.2384</td>
<td>3.4897</td>
<td>3.7244</td>
</tr>
<tr>
<td>CO</td>
<td>-0.0122</td>
<td>0.0200</td>
<td>-0.0504</td>
</tr>
<tr>
<td>s.e.</td>
<td>0.1504</td>
<td>0.0547</td>
<td>0.0487</td>
</tr>
<tr>
<td>z-value</td>
<td>-0.0810</td>
<td>0.3661</td>
<td>-1.0347</td>
</tr>
<tr>
<td>COH</td>
<td>0.1191</td>
<td>-0.0223</td>
<td>-0.1092</td>
</tr>
<tr>
<td>s.e.</td>
<td>0.0853</td>
<td>0.0289</td>
<td>0.0251</td>
</tr>
<tr>
<td>z-value</td>
<td>1.3967</td>
<td>-0.7740</td>
<td>-4.3530</td>
</tr>
</tbody>
</table>

- S.E.’s from working structures selected by either GIC or GCV are much smaller than that from Independent structure
- GEE with “unspecified” working structure does not converge
Discussion

- A new approach to identify the correlation structure
- Approximate the inverse of the correlation matrix with groups of basis matrices
- Objective function measures the adequacy of an approximated model
Discussion (Cont’d)

- Allow the cluster size to diverge
- Does not require likelihood function
- The estimates of the coefficients of the basis matrices have consistency and oracle property
- Simulation studies show that the proposed procedure works well for both the normal and the binary responses, even when the cluster size is large
- Handling with the unbalanced data case
- Concerns for positive definitiveness of the correlation matrix
The End

Thank you for your attention!