Estimating subgroup specific treatment effects via concave fusion

Jian Huang
University of Iowa

April 6, 2016
Outline

1 Motivation and the problem

2 The proposed model and approach
   • Concave pairwise fusion
   • Computational algorithm: ADMM
   • Initial value and solution path

3 Statistical properties

4 Numerical studies
   • Simulations
   • ACTG175 data

5 Concluding remarks
Motivation: Precision medicine

- Most medical treatments have been designed for the “average patient.” As a result of this “one-size-fits-all” approach, treatments can be very successful for some patients but not for others.

- Precision medicine is an approach to disease treatment and prevention that seeks to maximize effectiveness by taking into account individual variability in genes, environment, and lifestyle.

- However, it does not mean the creation of drugs or medical devices that are unique to a patient, but rather the ability to classify individuals into subpopulations that differ in their susceptibility to a particular disease, in the genetic factors of a diseases, or in their response to a specific treatment.
Motivation: subgroup analysis

- Subgroup analysis: subgrouping (clustering) with respect to how a clinical outcome is related to individual characteristics, including possibly unobserved ones.

- Estimation of subgroup specific treatment effects: subgrouping (clustering) with respect to heterogeneous treatment effects.

- Estimation of treatment assignment rules: this may need to take into account heterogeneity in the target patient population.
Example 1. Consider a regression model with heterogeneous treatment effects:
\[ y_i = z_i^T \eta + x_i \beta_i + \varepsilon_i, \quad i = 1, \ldots, n, \]  
(1)

where \( z_i \in \mathbb{R}^5 \). We randomly assign the treatment coefficients to two groups with equal probabilities, so that
\[ \beta_i = 2 \quad \text{for } i \in G_1 \quad \text{and} \quad \beta_i = -2 \quad \text{for } i \in G_2. \]

Consider the two approaches:
- Least squares regression without taking into account heterogeneity.
- The proposed method.
Figure 1: Simulated example, the two solid black lines represent $y = 2x$ and $y = -2x$
Some existing approaches

- Mixture model analysis (Gaussian mixture model): used widely for data clustering and classification (Banfield and Raftery (1993); Hastie and Tibshirani (1996); McNicholas (2010); Wei and Kosorok (2013), Shen and He (2015)).

  This approach requires specifying the number of subgroups in the population and a parametric model assumption.

- Methods of estimating homogeneity effects of covariates (Tibshirani et al. (2005); Bondell and Reich (2008); Shen and Huang (2010); Ke, Fan and Wu (2013), among others). These works consider grouping covariates, not observations.
We consider the model

\[ y_i = z_i^T \eta + x_i^T \beta_i + \varepsilon_i, \quad i = 1, \ldots, n. \]  

(2)

**Heterogeneous treatment effects:** let \( G = (G_1, \ldots, G_K) \) be a partition of \( \{1, \ldots, n\} \). Assume \( \beta_i = \alpha_k \) for all \( i \in G_k \), where \( \alpha_k \) is the common value for the \( \beta_i \)'s from group \( G_k \).

- **Goal:** estimate \( K \) and identify the subgroups; estimate \( (\alpha_1, \ldots, \alpha_K) \) and \( \eta \).
- **Method:** a concave pairwise fusion penalized least squares approach.
- **Algorithm:** an alternating direction method of multipliers (ADMM, Boyd et al. 2011).

**Challenge:** information of subgroups are unknown (the number of subgroups, which subjects belong to which subgroups, etc.)
Consider the concave pairwise fusion penalized least squares criterion

\[
Q_n(\eta, \beta; \lambda) = \frac{1}{2} \sum_{i=1}^{n} (y_i - z_i^T \eta - x_i^T \beta_i)^2 + \sum_{1 \leq i < j \leq n} p(\|\beta_i - \beta_j\|, \lambda),
\]

where \( p(\cdot, \lambda) \) is a penalty function with a tuning parameter \( \lambda \geq 0 \).

Let

\[
(\hat{\eta}(\lambda), \hat{\beta}(\lambda)) = \arg\min_{\eta \in \mathbb{R}^q, \beta \in \mathbb{R}^{np}} Q_n(\eta, \beta; \lambda).
\]

We compute \( (\hat{\eta}(\lambda), \hat{\beta}(\lambda)) \) for \( \lambda \in [\lambda_{\text{min}}, \lambda_{\text{max}}] \), where \( \lambda_{\text{max}} \) is the value that forces a constant \( \hat{\beta} \) solution and \( \lambda_{\text{min}} \) is a small positive number. We are particularly interested in the path

\[
\{\hat{\beta}(\lambda) : \lambda \in [\lambda_{\text{min}}, \lambda_{\text{max}}]\}.
\]
The penalty shrinks some of the pairs $\beta_j - \beta_k$ to zero. Based on this, we can partition the sample into subgroups.

Let $\{\hat{\alpha}_1, \ldots, \hat{\alpha}_{\hat{K}}\}$ be the distinct values of $\hat{\beta}$. Let

$$\hat{G}_k = \{i : \hat{\beta}_i = \hat{\alpha}_k, 1 \leq i \leq n\}, 1 \leq k \leq \hat{K}.$$ 

Then $\{\hat{G}_1, \ldots, \hat{G}_{\hat{K}}\}$ constitutes a partition of $\{1, \ldots, n\}$. 
**Penalty function**

$L_1$ penalty: $p_\gamma(t, \lambda) = \lambda t$, leads to biased estimates; in our numerical studies, the $L_1$ penalty tends to either yield a large number of subgroups or no subgroups on the solution path.

A penalty which can produce nearly unbiased estimates is more appealing.

- The SCAD penalty (Fan and Li 2001):
  
  \[
  p_\gamma(t, \lambda) = \lambda \int_0^t \min\{1, (\gamma - x/\lambda)_+/(\gamma - 1)\} dx, \gamma > 2
  \]

- The MCP (Zhang 2010):
  
  \[
  p_\gamma(t, \lambda) = \lambda \int_0^t (1 - x/(\gamma \lambda))_+ dx, \gamma > 1
  \]
Introduce a new set of parameters $\delta_{ij} = \beta_i - \beta_j$.

The minimization of (3) is equivalent to minimizing

$$L_0(\eta, \beta, \delta) = \frac{1}{2} \sum_{i=1}^{n} (y_i - z_i^T \eta - x_i^T \beta_i)^2 + \sum_{i<j} p_{\gamma}(\|\delta_{ij}\|, \lambda),$$

subject to $\beta_i - \beta_j - \delta_{ij} = 0$, \hspace{1cm} (5)

where $\delta = \{\delta_{ij}^T, i < j\}^T$. 

The augmented Lagrangian is

$$L(\eta, \beta, \delta, \nu) = \frac{1}{2} \sum_{i=1}^{n} (y_i - z_i^T \eta - x_i^T \beta_i)^2 + \sum_{j<k} p_{\gamma}(\|\delta_{jk}\|, \lambda) + \sum_{j<k} \langle \nu_{jk}, \beta_j - \beta_k - \delta_{jk} \rangle + \frac{\vartheta}{2} \sum_{j<k} \|\beta_j - \beta_k - \delta_{jk}\|^2.$$ (6)

For a given \((\delta^m, \nu^m)\) at step \(m\), the iteration goes as follows:

\[
(\eta^{m+1}, \beta^{m+1}) = \arg\min_{\eta, \beta} L(\eta, \beta, \delta^m, \nu^m), \quad \tag{7}
\]

\[
\delta^{m+1} = \arg\min_{\delta} L(\eta^{m+1}, \beta^{m+1}, \delta, \nu^m), \quad \tag{8}
\]

\[
\nu^{m+1}_{ij} = \nu^{m}_{ij} + \vartheta (\beta^{m+1}_i - \beta^{m+1}_j - \delta^{m+1}_{ij}). \quad \tag{9}
\]
Step (7) is a quadratic minimization problem.

Step (8) involves minimizing

$$\frac{\vartheta}{2} \| \zeta_{jk}^m - \delta_{jk} \|^2 + p_\gamma(\| \delta_{jk} \|, \lambda)$$

with respect to $\delta_{jk}$, where $\zeta_{jk}^m = \beta_j^m - \beta_k^m + \vartheta^{-1} \nu_{jk}^m$. This is a thresholding operator corresponding to $p_\gamma$.

For the $L_1$ penalty,

$$\delta_{jk}^{m+1} = S(\zeta_{jk}^m, \lambda/\vartheta),$$

where $S(z, t) = (1 - t/\| z \|)_+ z$ is the groupwise soft thresholding operator. Here $(x)_+ = x$ if $x > 0$ and $= 0$, otherwise.
MCP with $\gamma > 1/\vartheta$,

$$
\delta_{ij}^{m+1} = \begin{cases} 
\frac{S(\zeta_{ij}^m, \lambda/\vartheta)}{1-1/(\gamma\vartheta)} & \text{if } \|\zeta_{ij}^m\| \leq \gamma \lambda, \\
\zeta_{ij} & \text{if } \|\zeta_{ij}^m\| > \gamma \lambda.
\end{cases}
$$

(12)

SCAD penalty with $\gamma > 1/\vartheta + 1$,

$$
\delta_{ij}^{m+1} = \begin{cases} 
\frac{S(\zeta_{ij}^m, \lambda/\vartheta)}{S(\zeta_{ij}^m, \gamma\lambda/((\gamma-1)\vartheta))} & \text{if } \|\zeta_{ij}^m\| \leq \lambda + \lambda/\vartheta, \\
\frac{\zeta_{ij}^m}{1-1/((\gamma-1)\vartheta)} & \text{if } \lambda + \lambda/\vartheta < \|\zeta_{ij}^m\| \leq \gamma \lambda, \\
\zeta_{ij}^m & \text{if } \|\zeta_{ij}^m\| > \gamma \lambda.
\end{cases}
$$

(13)
To start the ADMM algorithm, it is important to find a reasonable initial value. We consider the ridge fusion criterion given by

\[ L_R(\eta, \beta) = \frac{1}{2} \| Z\eta - X\beta - y \|^2 + \frac{\lambda^*}{2} \sum_{1 \leq j < k \leq n} \| \beta_j - \beta_k \|^2, \]

where \( \lambda^* \) is the tuning parameter having a small value. We use \( \lambda^* = 0.001 \) in our analysis.
To compute the solution path of $\eta$ and $\beta$ along the $\lambda$ values, we use the warm start and continuation strategy to update the solutions. Let $[\lambda_{\text{min}}, \lambda_{\text{max}}]$ be the interval on which we compute the solution path.

- Let $\lambda_{\text{min}} = \lambda_0 < \lambda_1 < \cdots < \lambda_K \equiv \lambda_{\text{max}}$ be a grid of $\lambda$ values in $[\lambda_{\text{min}}, \lambda_{\text{max}}]$. Compute the ridge fusion solution $(\hat{\eta}(\lambda_0), \hat{\beta}(\lambda_0))$ and use it as the initial value.

- Compute $(\hat{\eta}(\lambda_k), \hat{\beta}(\lambda_k))$ using $(\hat{\eta}(\lambda_{k-1}), \hat{\beta}(\lambda_{k-1}))$ as the initial value for $k = 1, \ldots, K$.

Note that we start from the smallest $\lambda$ value in computing the solution path.
Let $\tilde{W} = \{w_{ik}\}$ be an $n \times K$ matrix with $w_{ik} = 1$ for $i \in G_k$ and $w_{ik} = 0$ otherwise. Let $W = \tilde{W} \otimes I_p$.

Let

$$M_G = \{\beta \in \mathbb{R}^{np} : \beta_i = \beta_j, \text{ for any } i, j \in G_k, 1 \leq k \leq K\}.$$ 

For each $\beta \in M_G$, it can be written as $\beta = W \alpha$, where $\alpha = (\alpha_1^T, \ldots, \alpha_K^T)^T$ and $\alpha_k$ is a $p \times 1$ vector of the $k$th subgroup-specific parameter for $k = 1, \ldots, K$.

Denote the minimum and maximum group sizes by $|G_{\min}| = \min_{1 \leq k \leq K} |G_k|$ and $|G_{\max}| = \max_{1 \leq k \leq K} |G_k|$, respectively.

Let $\tilde{X} = XW$ and $U = (Z, XW)$. 
Statistical properties

If the underlying groups $G_1, \ldots, G_K$ were known, the oracle estimator of $(\eta, \beta)$ is

$$(\hat{\eta}^{or}, \hat{\beta}^{or}) = \arg\min_{\eta\in\mathbb{R}^q, \beta\in\mathcal{M}_G} \frac{1}{2} \|y - Z\eta - X\beta\|^2, \quad (14)$$

and correspondingly, the oracle estimators for the common coefficient $\alpha$ and the coefficients $\eta$ are

$$(\hat{\eta}^{or}, \hat{\alpha}^{or}) = \arg\min_{\eta\in\mathbb{R}^q, \alpha\in\mathbb{R}^{kp}} \frac{1}{2} \|y - Z\eta - \tilde{X}\alpha\|^2$$

$$= (U^T U)^{-1} U^T y.$$

Let $\alpha^0_k$ be the true common coefficient vector for group $G_k$, $k = 1, \ldots, K$ and $\alpha^0 = ((\alpha^0_k)^T, k = 1, \ldots, K)^T$. Of course, oracle estimators are not real estimators, they are theoretical constructions useful for stating the properties of the proposed estimators.
(C1) The noise vector $\mathbf{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^T$ has sub-Gaussian tails such that $P(\lvert a^T\mathbf{\varepsilon} \rvert > \lVert a \rVert x) \leq 2 \exp(-c_1 x^2)$ for any vector $a \in \mathbb{R}^n$ and $x > 0$, where $0 < c_1 < \infty$.

(C2) Let $\rho(t) = \lambda^{-1} p_\gamma(t, \lambda)$. Suppose $\rho(t)$ is a symmetric function of $t$ and is non-decreasing and concave on $[0, \infty)$. Also, $\rho(t)$ is a constant for $t \geq a \lambda$ for some constant $a > 0$, and $\rho(0) = 0$. In addition, $\rho'(t)$ exists and is continuous except for a finite number of $t$ and $\rho'(0+) = 1$.

(C3) Assume $\sum_{i=1}^n z_{ij}^2 = n$ for $1 \leq k \leq q$, and
$\sum_{i=1}^n x_{ij}^2 1\{i \in G_k\} = \lvert G_k \rvert$ for $1 \leq j \leq p$,
$\lambda_{\min}(U^T U) \geq C_1 \lvert G_{\min} \rvert$, $\sup_i \lVert x_i \rVert \leq C_2 \sqrt{p}$ and
$\sup_i \lVert z_i \rVert \leq C_3 \sqrt{q}$ for some constants $0 < C_1 < \infty$, $0 < C_2 < \infty$ and $0 < C_3 < \infty$. 
Let
\[
\phi_n = c_1^{-1/2} C_1^{-1} \sqrt{q + Kp} |\mathcal{G}_{\text{min}}|^{-1} \sqrt{n \log n}.
\] (15)
and
\[
b_n = \min_{i \in \mathcal{G}_k, j \in \mathcal{G}_{k'}, k \neq k'} \| \beta_i^0 - \beta_j^0 \| = \min_{k \neq k'} \| \alpha_k^0 - \alpha_{k'}^0 \|
\]
be the minimal difference of the common values between two groups.

**Theorem**

*Suppose (C1)-(C3) hold, $Kp = o(n)$, $q = o(n)$, and*

\[
|\mathcal{G}_{\text{min}}| \gg \sqrt{(q + Kp)n \log n}.
\]

*If $b_n > a \lambda$ and $\lambda \gg \phi_n$, for some constant $a > 0$, where $\phi_n$ is given in (15), then there exists a local minimizer $(\hat{\eta}(\lambda), \hat{\beta}(\lambda))$ of the objective function $Q_n(\eta, \beta; \lambda)$ given in (3) satisfying*

\[
P \left( (\hat{\eta}(\lambda), \hat{\beta}(\lambda)) = (\hat{\eta}^{or}, \hat{\beta}^{or}) \right) \rightarrow 1.
\]
We use the modified Bayes Information Criterion (BIC) (Schwarz, 1978; Wang, Li and Tsai, 2007) for high-dimensional data settings to select the tuning parameter by minimizing

\[
BIC(\lambda) = \log \left[ \sum_{i=1}^{n} (y_i - z_i^T \hat{\eta}(\lambda) - x_i^T \hat{\beta}_i(\lambda))^2 / n \right] + C_n \frac{\log n}{n} (\hat{K}(\lambda)p + q),
\]

where \( C_n \) is a positive number which can depend on \( n \). We use \( C_n = \log(np + q) \). We select \( \lambda \) by minimizing the modified BIC.
Example 1 (One treatment variable). Consider

\[ y_i = z_i^T \eta + x_i \beta_i + \epsilon_i, \quad i = 1, \ldots, n, \]  

(17)

where

- \( z_i = (z_{i1}, z_{i2}, \ldots, z_{i5})^T \) with \( z_{i1} = 1 \) and \( (z_{i2}, \ldots, z_{i5})^T \) generated from multivariate normal with mean 0, variance 1 and an exchangeable correlation \( \rho = 0.3 \), \( x_i \) is simulated from \( N(0, 1) \).

- \( \epsilon_i \) are i.i.d. \( N(0, 0.5^2) \).

- \( \eta = (\eta_1, \ldots, \eta_5)^T \) with \( \eta_k \) simulated from \( \text{Uniform}[1, 2] \) for \( k = 1, \ldots, 5 \).

- We randomly assign the treatment coefficients to two groups with equal probabilities, i.e., \( p(i \in G_1) = p(i \in G_2) = 1/2 \), so that \( \beta_i = \alpha_1 \) for \( i \in G_1 \) and \( \beta_i = \alpha_2 \) for \( i \in G_2 \), where \( \alpha_1 = 2 \) and \( \alpha_2 = -2 \).

- We consider \( n = 100, 200 \).
Figure 2: Fusiongram: Solution paths for $(\hat{\beta}_1(\lambda), \ldots, \hat{\beta}_n(\lambda))$ against $\lambda$ with $n = 200$ for data from Example 1.
Table 1: The sample mean, median and standard deviation (s.d.) of \( \hat{K} \) and the percentage (per) of \( \hat{K} \) equaling the true number of subgroups by MCP and SCAD based on 100 replications with \( n = 100 \) and 200 in Example 1.

<table>
<thead>
<tr>
<th></th>
<th>( n = 100 )</th>
<th></th>
<th>( n = 200 )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>mean</td>
<td>median</td>
<td>s.d.</td>
</tr>
<tr>
<td>MCP</td>
<td>2.380</td>
<td>2.000</td>
<td>0.716</td>
</tr>
<tr>
<td>SCAD</td>
<td>2.340</td>
<td>2.000</td>
<td>0.708</td>
</tr>
</tbody>
</table>
Table 2: The sample mean, median and asymptotic standard deviation (ASD) of the estimators $\hat{\alpha}_1$ and $\hat{\alpha}_2$ by MCP and SCAD and oracle estimators $\hat{\alpha}_{1 \text{or}}$ and $\hat{\alpha}_{2 \text{or}}$ based on 100 replications with $n = 100, 200$ in Example 1.

<table>
<thead>
<tr>
<th></th>
<th>$n = 100$</th>
<th></th>
<th>$n = 200$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>mean</td>
<td>median</td>
<td>ASD</td>
</tr>
<tr>
<td>$\hat{\alpha}_1$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MCP</td>
<td>1.884</td>
<td>1.928</td>
<td>0.077</td>
</tr>
<tr>
<td>SCAD</td>
<td>1.874</td>
<td>1.964</td>
<td>0.078</td>
</tr>
<tr>
<td>$\hat{\alpha}_{1 \text{or}}$</td>
<td>1.993</td>
<td>1.998</td>
<td>0.072</td>
</tr>
<tr>
<td>$\hat{\alpha}_2$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MCP</td>
<td>−1.783</td>
<td>−1.929</td>
<td>0.078</td>
</tr>
<tr>
<td>SCAD</td>
<td>−1.770</td>
<td>−1.954</td>
<td>0.078</td>
</tr>
<tr>
<td>$\hat{\alpha}_{2 \text{or}}$</td>
<td>−1.993</td>
<td>−1.988</td>
<td>0.073</td>
</tr>
</tbody>
</table>
Figure 3: The boxplots of the MSEs of $\hat{\eta}$ using MCP and SCAD, respectively, with $n = 100$ (white) and $n = 200$ (grey) in Example 1.
Example 2 (Multiple treatment variables). We simulated data from the heterogeneous model with multiple treatment variables:

\[ y_i = z_i^T \eta + x_i^T \beta_i + \epsilon_i, \quad i = 1, \ldots, n, \]  

(18)

where

- \( z_i, \epsilon_i \) and \( \eta \) are simulated in the same way as in Example 1.
- Let \( x_i = (x_{i1}, x_{i2}, x_{i3})^T \) in which \( x_{i1} \) is simulated from standard normal and \( (x_{i2}, x_{i3})^T \) are from centered and standardized binomial with probability 0.7 for one outcome.
- We randomly assign the responses to two groups with equal probabilities, i.e., we let \( p(i \in G_1) = p(i \in G_2) = 1/2 \), so that \( \beta_i = \alpha_1 \) for \( i \in G_1 \) and \( \beta_i = \alpha_2 \) for \( i \in G_2 \), where \( \alpha_1 = (\alpha_{11}, \alpha_{12}, \alpha_{13}) \) and \( \alpha_2 = (\alpha_{21}, \alpha_{22}, \alpha_{23}) \). Let \( \alpha_{1j} = \alpha \) and \( \alpha_{2j} = -\alpha \) for \( j = 1, 2, 3 \). We let \( \alpha = 1, 2 \) for different signal-noise ratios. Let \( n = 200 \).
Figure 4: Fusiongram for $(\beta_{11}, \ldots, \beta_{1n})$, the first component in $\beta_i$’s in Example 2.
Table 3: The sample mean, median and standard deviation (s.d.) of $\hat{K}$ and the percentage (per) that $\hat{K}$ equals to the true number of subgroups by MCP and SCAD based on 100 replications with $\alpha = 1, 2$ in Example 2.

<table>
<thead>
<tr>
<th></th>
<th>$\alpha = 1$</th>
<th></th>
<th>$\alpha = 2$</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>mean</td>
<td>median</td>
<td>s.d.</td>
<td>per</td>
</tr>
<tr>
<td>MCP</td>
<td>2.700</td>
<td>3.000</td>
<td>0.717</td>
<td>0.440</td>
</tr>
<tr>
<td>SCAD</td>
<td>2.690</td>
<td>3.000</td>
<td>0.706</td>
<td>0.440</td>
</tr>
</tbody>
</table>
Example 3 (No treatment heterogeneity). We generate data from a model with homogeneous treatment effects

\[ y_i = z_i^T \eta + x_i \beta + \epsilon_i, \ i = 1, \ldots, n, \]

where \( z_i, x_i, \epsilon_i \) and \( \eta \) are simulated in the same way as in Example 1. Set \( \beta = 2 \) and \( n = 200 \).

- We use our proposed penalized estimation approach to fit the model assuming the possible existence of treatment heterogeneity.
- The sample mean of the estimated number of groups \( \hat{K} \) is 1.49 and 1.48 based on 100 replications, respectively, for the MCP and SCAD methods.
- The sample median is 1 for both methods.
Table 4: The empirical bias (Bias) of the estimates of $\beta$ and $\eta$, and the average asymptotic standard deviation (ASD) and the empirical standard deviation (ESD) of MCP and SCAD, as well as the oracle estimator (ORAC) in Example 3.

<table>
<thead>
<tr>
<th></th>
<th>$\beta$</th>
<th>$\eta_1$</th>
<th>$\eta_2$</th>
<th>$\eta_3$</th>
<th>$\eta_4$</th>
<th>$\eta_5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>MCP</td>
<td>Bias</td>
<td>$-0.005$</td>
<td>$-0.002$</td>
<td>$0.007$</td>
<td>$0.003$</td>
<td>$0.002$</td>
</tr>
<tr>
<td></td>
<td>ASE</td>
<td>$0.035$</td>
<td>$0.034$</td>
<td>$0.037$</td>
<td>$0.037$</td>
<td>$0.038$</td>
</tr>
<tr>
<td></td>
<td>ESE</td>
<td>$0.034$</td>
<td>$0.041$</td>
<td>$0.038$</td>
<td>$0.041$</td>
<td>$0.042$</td>
</tr>
<tr>
<td>SCAD</td>
<td>Bias</td>
<td>$-0.004$</td>
<td>$-0.001$</td>
<td>$0.007$</td>
<td>$0.003$</td>
<td>$0.002$</td>
</tr>
<tr>
<td></td>
<td>ASE</td>
<td>$0.035$</td>
<td>$0.034$</td>
<td>$0.037$</td>
<td>$0.037$</td>
<td>$0.037$</td>
</tr>
<tr>
<td></td>
<td>ESE</td>
<td>$0.034$</td>
<td>$0.040$</td>
<td>$0.037$</td>
<td>$0.041$</td>
<td>$0.042$</td>
</tr>
<tr>
<td>ORAC</td>
<td>Bias</td>
<td>$-0.004$</td>
<td>$-0.001$</td>
<td>$0.006$</td>
<td>$0.004$</td>
<td>$0.002$</td>
</tr>
<tr>
<td></td>
<td>ASE</td>
<td>$0.036$</td>
<td>$0.035$</td>
<td>$0.038$</td>
<td>$0.038$</td>
<td>$0.039$</td>
</tr>
<tr>
<td></td>
<td>ESE</td>
<td>$0.036$</td>
<td>$0.039$</td>
<td>$0.034$</td>
<td>$0.039$</td>
<td>$0.041$</td>
</tr>
</tbody>
</table>
We apply our method to the AIDS Clinical Trials Group Study 175 (ACTG175) (Tsiatis et al., 2007), ACTG175 was a randomized clinical trial to compare the 4 treatments:

- zidovudine with other three therapies including
  - zidovudine and didanosine,
  - zidovudine and zalcitabine,
  - didanosine

in adults infected with the human immunodeficiency virus type 1. We randomly select 300 patients from the study to consist of our dataset.
The response variable is the log-transformed value of the CD4 counts at 20±5 weeks.

We use binary variables for the treatments \( x_i = (x_{i1}, x_{i2}, x_{i3})^T \).

There are 12 baseline covariates in the model,

1. age (years),
2. weight (kg),
3. Karnofsky score,
4. CD4 counts at baseline,
5. CD8 counts at baseline,
6. hemophilia (0 = no, 1 = yes),
7. homosexual activity (0 = no, 1 = yes),
8. history of intravenous drug use (0 = no, 1 = yes),
9. race (0 = white, 1 = not white),
10. gender (0 = female, 1 = male),
11. antiretroviral history (0 = naive, 1 = experienced) and
12. symptomatic status (0 = asymptomatic, 1 = symptomatic).
We fit the heterogeneous model

\[ y_i = z_i^T \eta + x_i^T \beta_i + \epsilon_i, \quad i = 1, \ldots, 300, \]

where \( z_i = (1, z_{i2}, \ldots, z_{i13})^T \) with the first component for intercept and other components being the 12 covariates described above. All the predictors are centered and standardized.

We identified two subgroups.
ACTG175 data

Figure 5: Fusiongrams for $\beta_1 = (\beta_{11}, \ldots, \beta_{1n})$, $\beta_2 = (\beta_{21}, \ldots, \beta_{2n})$, and $\beta_3 = (\beta_{31}, \ldots, \beta_{3n})$. 
Table 5: The estimates (Est.), standard deviations (s.d.) and p-values (P-value) of $\alpha_1$ and $\alpha_2$ by the MCP and SCAD methods, and those values of $\beta = \alpha_1$ by the OLS method.

<table>
<thead>
<tr>
<th></th>
<th>$\alpha_{11}$</th>
<th>$\alpha_{12}$</th>
<th>$\alpha_{13}$</th>
<th>$\alpha_{21}$</th>
<th>$\alpha_{22}$</th>
<th>$\alpha_{23}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>MCP</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Est.</td>
<td>0.141</td>
<td>-0.011</td>
<td>-0.039</td>
<td>0.835</td>
<td>0.666</td>
<td>0.687</td>
</tr>
<tr>
<td>s.d.</td>
<td>0.055</td>
<td>0.055</td>
<td>0.055</td>
<td>0.394</td>
<td>0.268</td>
<td>0.251</td>
</tr>
<tr>
<td>p-value</td>
<td>0.010</td>
<td>0.841</td>
<td>0.478</td>
<td>0.034</td>
<td>0.013</td>
<td>0.006</td>
</tr>
<tr>
<td>SCAD</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Est.</td>
<td>0.142</td>
<td>-0.010</td>
<td>-0.037</td>
<td>0.805</td>
<td>0.614</td>
<td>0.636</td>
</tr>
<tr>
<td>s.d.</td>
<td>0.055</td>
<td>0.055</td>
<td>0.055</td>
<td>0.395</td>
<td>0.268</td>
<td>0.251</td>
</tr>
<tr>
<td>p-value</td>
<td>0.010</td>
<td>0.855</td>
<td>0.501</td>
<td>0.041</td>
<td>0.022</td>
<td>0.011</td>
</tr>
<tr>
<td>OLS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Est.</td>
<td>0.212</td>
<td>0.035</td>
<td>0.036</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>s.d.</td>
<td>0.060</td>
<td>0.058</td>
<td>0.058</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>p-value</td>
<td>&lt; 0.001</td>
<td>0.550</td>
<td>0.532</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
</tbody>
</table>
Concluding remarks

- Extension to other important models (e.g., logistic regression, Cox regression) is conceptually straightforward, but theoretical analysis and computation are more difficult.

- Extension to $p \gg n$ models is also possible, but requires further sparsity assumption to ensure model identifiability, and theoretical analysis is more difficult.

- It is of interest to speed up the ADMM so that it can handle large $n$ problems.

- It is possible to weaken the conditions for the theoretical results, but this will not change the basic story.

- The theoretical results are derived for fixed $\lambda$ values. It is much more difficult to derive the results for $\lambda$ values selected based on a data-driven procedure.
Thank you!