Estimating subgroup specific treatment effects via concave fusion

Jian Huang
University of Iowa

April 6, 2016

Outline

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.

A simulated example

Example 1. Consider a regression model with heterogeneous treatment effects :

$$y_i = \mathbf{z}_i^\top \boldsymbol{\eta} + x_i\beta_i + \varepsilon_i, i = 1, \ldots, n, \tag{1}$$

where $\mathbf{z}_i \in \mathbb{R}^5$. We randomly assign the treatment coefficients to two groups with equal probabilities, so that

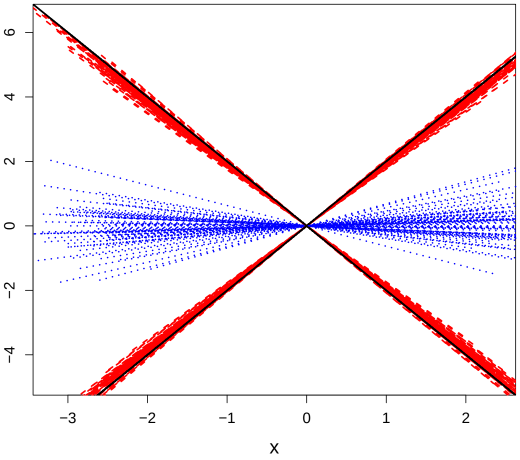
$$\beta_i = 2 \text{ for } i \in \mathcal{G}_1 \text{ and } \beta_i = -2 \text{ for } i \in \mathcal{G}_2.$$

Consider the two approaches:

- Least squares regression without taking into account heterogeneity.
- The proposed method.

Example

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.

Model and approach

We consider the model

$$y_i = \boldsymbol{z}_i^{\mathsf{T}} \boldsymbol{\eta} + \boldsymbol{x}_i^{\mathsf{T}} \boldsymbol{\beta}_i + \varepsilon_i, i = 1, \ldots, n. \tag{2}$$

Heterogeneous treatment effects: let $\mathcal{G} = (\mathcal{G}_1, \ldots, \mathcal{G}_K)$ be a partition of $\{1, \ldots, n\}$. Assume $\boldsymbol{\beta}_i = \alpha_k$ for all $i \in \mathcal{G}_k$, where α_k is the common value for the $\boldsymbol{\beta}_i$'s from group $\mathcal{G}_k$.

- Goal: estimate K and identify the subgroups; estimate $(\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_K)$ and $\boldsymbol{\eta}$.
- 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.)

Subgroup Analysis via Concave Pairwise Fusion

Consider the concave pairwise fusion penalized least squares criterion

$$Q_n(\boldsymbol{\eta}, \boldsymbol{\beta}; \lambda) = \frac{1}{2}\sum_{i=1}^{n}(y_i - \boldsymbol{z}_i^{\mathsf{T}}\boldsymbol{\eta} - \boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{\beta}_i)^2 + \sum_{1\le i<j\le n} p(\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|, \lambda), \quad (3)$$

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

$$(\widehat{\boldsymbol{\eta}}(\lambda), \widehat{\boldsymbol{\beta}}(\lambda)) = \underset{\boldsymbol{\eta}\in\mathbf{R}^q,\, \boldsymbol{\beta}\in\mathbf{R}^{np}}{\operatorname{argmin}} Q_n(\boldsymbol{\eta}, \boldsymbol{\beta}; \lambda). \quad (4)$$

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

$$\{\hat{\boldsymbol{\beta}}(\lambda) : \lambda \in [\lambda_{\min}, \lambda_{\max}]\}.$$

Concave Pairwise Fusion

The penalty shrinks some of the pairs $\beta_j - \beta_k$ to zero. Based on this, we can partition the sample into subgroups.

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

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

Then $\{\widehat{\mathcal{G}}_1, \ldots, \widehat{\mathcal{G}}_{\widehat{K}}\}$ 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):

$$p_\gamma(t,\lambda) = \lambda \int_0^t (1 - x/(\gamma\lambda))_+ dx, \gamma > 1$$

ADMM Algorithm

- Introduce a new set of parameters $\boldsymbol{\delta}_{ij} = \boldsymbol{\beta}_i - \boldsymbol{\beta}_j$.
- The minimization of (3) is equivalent to minimizing

$$L_0(\boldsymbol{\eta}, \boldsymbol{\beta}, \boldsymbol{\delta}) = \frac{1}{2} \sum_{i=1}^{n} (y_i - \mathbf{z}_i^{\mathsf{T}} \boldsymbol{\eta} - \mathbf{x}_i^{\mathsf{T}} \boldsymbol{\beta}_i)^2 + \sum_{i<j} p_\gamma(\|\boldsymbol{\delta}_{ij}\|, \lambda),$$
$$\text{subject to } \boldsymbol{\beta}_i - \boldsymbol{\beta}_j - \boldsymbol{\delta}_{ij} = \mathbf{0}, \qquad (5)$$

where $\boldsymbol{\delta} = \{\boldsymbol{\delta}_{ij}^{\mathsf{T}}, i < j\}^{\mathsf{T}}$.

ADMM

The augmented Lagrangian is

$$L(\boldsymbol{\eta}, \boldsymbol{\beta}, \boldsymbol{\delta}, \boldsymbol{\upsilon}) = \frac{1}{2}\sum_{i=1}^{n}(y_i - \mathbf{z}_i^{\mathsf{T}}\boldsymbol{\eta} - \mathbf{x}_i^{\mathsf{T}}\boldsymbol{\beta}_i)^2 + \sum\nolimits_{j<k} p_{\gamma}(\|\boldsymbol{\delta}_{jk}\|, \lambda) \quad (6)$$
$$+ \sum_{j<k}\langle \boldsymbol{\upsilon}_{jk}, \boldsymbol{\beta}_j - \boldsymbol{\beta}_k - \boldsymbol{\delta}_{jk}\rangle + \frac{\vartheta}{2}\sum_{j<k}\|\boldsymbol{\beta}_j - \boldsymbol{\beta}_k - \boldsymbol{\delta}_{jk}\|^2.$$

For a given $(\boldsymbol{\delta}^m, \boldsymbol{\upsilon}^m)$ at step m, the iteration goes as follows:

$$(\boldsymbol{\eta}^{m+1}, \boldsymbol{\beta}^{m+1}) = \underset{\boldsymbol{\eta}, \boldsymbol{\beta}}{\operatorname{argmin}}\, L(\boldsymbol{\eta}, \boldsymbol{\beta}, \boldsymbol{\delta}^m, \boldsymbol{\upsilon}^m), \quad (7)$$
$$\boldsymbol{\delta}^{m+1} = \underset{\boldsymbol{\delta}}{\operatorname{argmin}}\, L(\boldsymbol{\eta}^{m+1}, \boldsymbol{\beta}^{m+1}, \boldsymbol{\delta}, \boldsymbol{\upsilon}^m), \quad (8)$$
$$\boldsymbol{\upsilon}_{ij}^{m+1} = \boldsymbol{\upsilon}_{ij}^m + \vartheta(\boldsymbol{\beta}_i^{m+1} - \boldsymbol{\beta}_j^{m+1} - \boldsymbol{\delta}_{ij}^{m+1}). \quad (9)$$

ADMM

- Step (7) is a quadratic minimization problem.
- Step (8) involves minimizing

$$\frac{\vartheta}{2}\|\zeta_{jk}^m - \boldsymbol{\delta}_{jk}\|^2 + p_\gamma(\|\boldsymbol{\delta}_{jk}\|, \lambda) \tag{10}$$

  with respect to $\boldsymbol{\delta}_{jk}$, where $\zeta_{jk}^m = \boldsymbol{\beta}_j^m - \boldsymbol{\beta}_k^m + \vartheta^{-1}\boldsymbol{\upsilon}_{jk}^m$. This is a thresholding operator corresponding to p_γ.
- For the L_1 penalty,

$$\boldsymbol{\delta}_{jk}^{m+1} = S(\zeta_{jk}^m, \lambda/\vartheta), \tag{11}$$

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

ADMM

- MCP with $\gamma > 1/\vartheta$,

$$\delta_{ij}^{m+1} = \begin{cases} \frac{S(\zeta_{ij}^m, \lambda/\vartheta)}{1 - 1/(\gamma\vartheta)} & \text{if } \|\zeta_{ij}^m\| \leq \gamma\lambda, \\ \zeta_{ij} & \text{if } \|\zeta_{ij}^m\| > \gamma\lambda. \end{cases} \tag{12}$$

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

$$\delta_{ij}^{m+1} = \begin{cases} S(\zeta_{ij}^m, \lambda/\vartheta) & \text{if } \|\zeta_{ij}^m\| \leq \lambda + \lambda/\vartheta, \\ \frac{S(\zeta_{ij}^m, \gamma\lambda/((\gamma-1)\vartheta))}{1 - 1/((\gamma-1)\vartheta)} & \text{if } \lambda + \lambda/\vartheta < \ \|\zeta_{ij}^m\| \leq \gamma\lambda, \\ \zeta_{ij}^m & \text{if } \|\zeta_{ij}^m\| > \gamma\lambda. \end{cases} \tag{13}$$

ADMM initial value

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

$$L_R(\boldsymbol{\eta}, \boldsymbol{\beta}) = \frac{1}{2}\|\boldsymbol{Z}\boldsymbol{\eta} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y}\|^2 + \frac{\lambda^*}{2} \sum_{1 \leq j < k \leq n} \|\boldsymbol{\beta}_j - \boldsymbol{\beta}_k\|^2,$$

where λ^* is the tuning parameter having a small value. We use $\lambda^* = 0.001$ in our analysis.

ADMM solution path

To compute the solution path of $\boldsymbol{\eta}$ and $\boldsymbol{\beta}$ along the λ values, we use the warm start and continuation strategy to update the solutions. Let $[\lambda_{\min}, \lambda_{\max}]$ be the interval on which we compute the solution path.

- Let $\lambda_{\min} = \lambda_0 < \lambda_1 < \cdots < \lambda_K \equiv \lambda_{\max}$ be a grid of λ values in $[\lambda_{\min}, \lambda_{\max}]$. Compute the ridge fusion solution $(\widehat{\boldsymbol{\eta}}(\lambda_0), \hat{\boldsymbol{\beta}}(\lambda_0))$ and use it as the initial value.
- Compute $(\widehat{\boldsymbol{\eta}}(\lambda_k), \hat{\boldsymbol{\beta}}(\lambda_k))$ using $(\widehat{\boldsymbol{\eta}}(\lambda_{k-1}), \hat{\boldsymbol{\beta}}(\lambda_{k-1}))$ as the initial value for $k = 1, \ldots, K$.

Note that we start from the smallest λ value in computing the solution path.

Statistical Properties

- Let $\widetilde{\mathbf{W}} = \{w_{ik}\}$ be an $n \times K$ matrix with $w_{ik} = 1$ for $i \in \mathcal{G}_k$ and $w_{ik} = 0$ otherwise. Let $\mathbf{W} = \widetilde{\mathbf{W}} \otimes \mathbf{I}_p$.
- Let

  $$\mathcal{M}_{\mathcal{G}} = \{\boldsymbol{\beta} \in \mathbb{R}^{np} : \boldsymbol{\beta}_i = \boldsymbol{\beta}_j, \text{ for any } i, j \in \mathcal{G}_k, 1 \le k \le K\}.$$

  For each $\boldsymbol{\beta} \in \mathcal{M}_{\mathcal{G}}$, it can be written as $\boldsymbol{\beta} = \mathbf{W}\boldsymbol{\alpha}$, where $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1^{\mathsf{T}}, \ldots, \boldsymbol{\alpha}_K^{\mathsf{T}})^{\mathsf{T}}$ and $\boldsymbol{\alpha}_k$ is a $p \times 1$ vector of the kth subgroup-specific parameter for $k = 1, \ldots, K$.
- Denote the minimum and maximum group sizes by $|\mathcal{G}_{\min}| = \min_{1 \le k \le K} |\mathcal{G}_k|$ and $|\mathcal{G}_{\max}| = \max_{1 \le k \le K} |\mathcal{G}_k|$, respectively.
- Let $\widetilde{\mathbf{X}} = \mathbf{X}\mathbf{W}$ and $\mathbf{U} = (\mathbf{Z}, \mathbf{X}\mathbf{W})$.

Statistical properties

If the underlying groups $\mathcal{G}_1, \ldots, \mathcal{G}_K$ were known, the oracle estimator of $(\boldsymbol{\eta}, \boldsymbol{\beta})$ is

$$(\widehat{\boldsymbol{\eta}}^{or}, \hat{\boldsymbol{\beta}}^{or}) = \underset{\boldsymbol{\eta} \in \mathbf{R}^q,\, \boldsymbol{\beta} \in \mathcal{M}_{\mathcal{G}}}{\operatorname{argmin}} \frac{1}{2} \|\mathbf{y} - \mathbf{Z}\boldsymbol{\eta} - \mathbf{X}\boldsymbol{\beta}\|^2, \tag{14}$$

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

$$\begin{aligned}(\widehat{\boldsymbol{\eta}}^{or}, \widehat{\boldsymbol{\alpha}}^{or}) &= \underset{\boldsymbol{\eta} \in \mathbf{R}^q,\, \boldsymbol{\alpha} \in \mathbf{R}^{Kp}}{\operatorname{argmin}} \frac{1}{2} \|\mathbf{y} - \mathbf{Z}\boldsymbol{\eta} - \widetilde{\mathbf{X}}\boldsymbol{\alpha}\|^2 \\ &= (\mathbf{U}^{\mathsf{T}}\mathbf{U})^{-1}\mathbf{U}^{\mathsf{T}}\mathbf{y}.\end{aligned}$$

Let $\boldsymbol{\alpha}_k^0$ be the true common coefficient vector for group $\mathcal{G}_k$, $k = 1, \ldots, K$ and $\boldsymbol{\alpha}^0 = ((\boldsymbol{\alpha}_k^0)^{\mathsf{T}}, k = 1, \ldots, K)^{\mathsf{T}}$. 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 $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^{\mathsf{T}}$ has sub-Gaussian tails such that $P(|\mathbf{a}^{\mathsf{T}}\boldsymbol{\varepsilon}| > \|\mathbf{a}\|x) \leq 2\exp(-c_1 x^2)$ for any vector $\mathbf{a} \in \mathbb{R}^n$ 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 $\sum_{i=1}^n x_{ij}^2 1\{i \in \mathcal{G}_k\} = |\mathcal{G}_k|$ for $1 \leq j \leq p$, $\lambda_{\min}(\mathbf{U}^{\mathsf{T}}\mathbf{U}) \geq C_1 \, |\mathcal{G}_{\min}|$, $\sup_i \|\mathbf{x}_i\| \leq C_2\sqrt{p}$ and $\sup_i \|\mathbf{z}_i\| \leq C_3\sqrt{q}$ for some constants $0 < C_1 < \infty$, $0 < C_2 < \infty$ and $0 < C_3 < \infty$.

Let

$$\phi_n = c_1^{-1/2} C_1^{-1} \sqrt{q + Kp}\, |\mathcal{G}_{\min}|^{-1} \sqrt{n \log n}. \qquad (15)$$

and

$$b_n = \min_{i \in \mathcal{G}_k, j \in \mathcal{G}_{k'}, k \neq k'} \|\boldsymbol{\beta}_i^0 - \boldsymbol{\beta}_j^0\| = \min_{k \neq k'} \|\boldsymbol{\alpha}_k^0 - \boldsymbol{\alpha}_{k'}^0\|$$

be the minimal difference of the common values between two groups.

Theorem

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

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

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

$$P\left((\widehat{\boldsymbol{\eta}}(\lambda), \widehat{\boldsymbol{\beta}}(\lambda)) = (\widehat{\boldsymbol{\eta}}^{or}, \widehat{\boldsymbol{\beta}}^{or}) \right) \to 1.$$

Simulation Studies

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

$$\mathrm{BIC}(\lambda) = \log[\sum_{i=1}^{n} (y_i - \mathbf{z}_i^{\mathsf{T}} \widehat{\boldsymbol{\eta}}(\lambda) - \mathbf{x}_i^{\mathsf{T}} \widehat{\boldsymbol{\beta}}_i(\lambda))^2 / n] + C_n \frac{\log n}{n} (\widehat{K}(\lambda) p + q), \tag{16}$$

where C_n is a positive number which can depend on n. We use $C_n = \log(np + q)$. We select λ by minimizing the modified BIC.

Example 1

Example 1 (One treatment variable). Consider

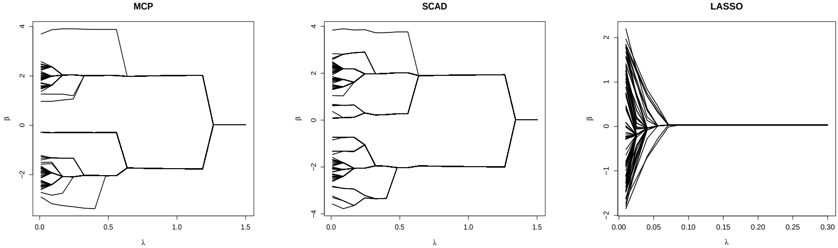
$$y_i = \mathbf{z}_i^{\mathsf{T}} \boldsymbol{\eta} + x_i \beta_i + \varepsilon_i, i = 1, \ldots, n, \qquad (17)$$

where

- $\mathbf{z}_i = (z_{i1}, z_{i2}, \ldots, z_{i5})^{\mathsf{T}}$ with $z_{i1} = 1$ and $(z_{i2}, \ldots, z_{i5})^{\mathsf{T}}$ generated from multivariate normal with mean 0, variance 1 and an exchangeable correlation $\rho = 0.3$, x_i is simulated from $N(0, 1)$.
- ε_i are i.i.d. $N(0, 0.5^2)$.
- $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_5)^{\mathsf{T}}$ with η_k simulated from Uniform$[1, 2]$ for $k = 1, \ldots, 5$.
- We randomly assign the treatment coefficients to two groups with equal probabilities, i.e., $p(i \in \mathcal{G}_1) = p(i \in \mathcal{G}_2) = 1/2$, so that $\beta_i = \alpha_1$ for $i \in \mathcal{G}_1$ and $\beta_i = \alpha_2$ for $i \in \mathcal{G}_2$, where $\alpha_1 = 2$ and $\alpha_2 = -2$.
- We consider $n = 100, 200$.

Example 1

Figure 2 : *Fusiongram: Solution paths for* $(\hat{\beta}_1(\lambda), \ldots, \hat{\beta}_n(\lambda))$ *against* λ *with* $n = 200$ *for data from Example 1.*



Example 1

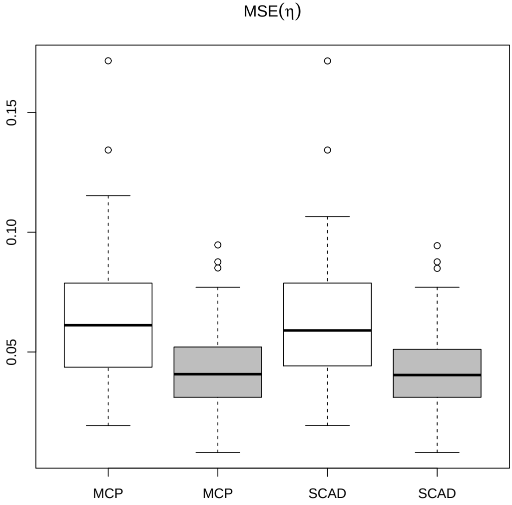
Table 1 : The sample mean, median and standard deviation (s.d.) of $\widehat{K}$ and the percentage (per) of $\widehat{K}$ equaling the true number of subgroups by MCP and SCAD based on 100 replications with $n = 100$ and 200 in Example 1.

	$n = 100$				$n = 200$			
	mean	median	s.d.	per	mean	median	s.d.	per
MCP	2.380	2.000	0.716	0.710	2.210	2.000	0.520	0.790
SCAD	2.340	2.000	0.708	0.710	2.210	2.000	0.541	0.800

Table 2 : The sample mean, median and asymptotic standard deviation (ASD) of the estimators $\widehat{\alpha}_1$ and $\widehat{\alpha}_2$ by MCP and SCAD and oracle estimators $\widehat{\alpha}_1^{or}$ and $\widehat{\alpha}_2^{or}$ based on 100 replications with $n = 100, 200$ in Example 1.

		$n = 100$			$n = 200$		
		mean	median	ASD	mean	median	ASD
$\widehat{\alpha}_1$	MCP	1.884	1.928	0.077	1.907	1.963	0.055
	SCAD	1.874	1.964	0.078	1.899	1.928	0.057
$\widehat{\alpha}_1^{or}$		1.993	1.998	0.072	1.998	1.999	0.051
$\widehat{\alpha}_2$	MCP	-1.783	-1.929	0.078	-1.823	-1.959	0.071
	SCAD	-1.770	-1.954	0.078	-1.778	-1.921	0.071
$\widehat{\alpha}_2^{or}$		-1.993	-1.988	0.073	-2.001	-2.005	0.052

Figure 3 : *The boxplots of the MSEs of $\widehat{\eta}$ 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:

$$y_i = \mathbf{z}_i^{\mathsf{T}}\boldsymbol{\eta} + \mathbf{x}_i^{\mathsf{T}}\boldsymbol{\beta}_i + \varepsilon_i, i = 1, \ldots, n, \tag{18}$$

where

- $\mathbf{z}_i$, ε_i and $\boldsymbol{\eta}$ are simulated in the same way as in Example 1.
- Let $\mathbf{x}_i = (x_{i1}, x_{i2}, x_{i3})^{\mathsf{T}}$ in which x_{i1} is simulated from standard normal and $(x_{i2}, x_{i3})^{\mathsf{T}}$ 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 $p(i \in \mathcal{G}_1) = p(i \in \mathcal{G}_2) = 1/2$, so that $\boldsymbol{\beta}_i = \boldsymbol{\alpha}_1$ for $i \in \mathcal{G}_1$ and $\boldsymbol{\beta}_i = \boldsymbol{\alpha}_2$ for $i \in \mathcal{G}_2$, where $\boldsymbol{\alpha}_1 = (\alpha_{11}, \alpha_{12}, \alpha_{13})$ and $\boldsymbol{\alpha}_2 = (\alpha_{21}, \alpha_{22}, \alpha_{23})$. 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 $\boldsymbol{\beta}_i$'s in Example 2.*

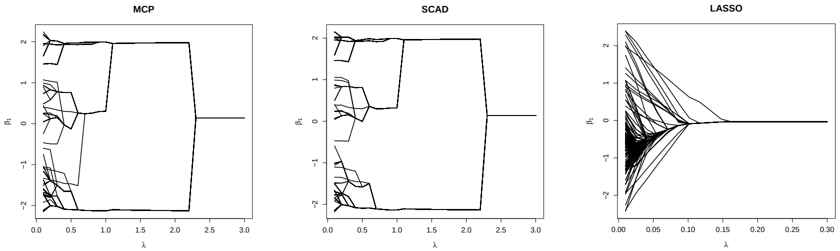


Table 3 : The sample mean, median and standard deviation (s.d.) of $\widehat{K}$ and the percentage (per) that $\widehat{K}$ equals to the true number of subgroups by MCP and SCAD based on 100 replications with $\alpha = 1, 2$ in Example 2.

	$\alpha = 1$				$\alpha = 2$			
	mean	median	s.d.	per	mean	median	s.d.	per
MCP	2.700	3.000	0.717	0.440	2.180	2.000	0.411	0.830
SCAD	2.690	3.000	0.706	0.440	2.190	2.000	0.419	0.820

Example 3 (No treatment heterogeneity). We generate data from a model with homogeneous treatment effects

$$y_i = \mathbf{z}_i^{\mathsf{T}} \boldsymbol{\eta} + x_i \beta + \varepsilon_i, i = 1, \ldots, n,$$

where $\mathbf{z}_i$, x_i, ε_i and $\boldsymbol{\eta}$ 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 $\widehat{K}$ 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 β and $\boldsymbol{\eta}$, 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.

		β	η_1	η_2	η_3	η_4	η_5
MCP	Bias	-0.005	-0.002	0.007	0.003	0.002	0.001
	ASE	0.035	0.034	0.037	0.037	0.038	0.037
	ESE	0.034	0.041	0.038	0.041	0.042	0.038
SCAD	Bias	-0.004	-0.001	0.007	0.003	0.002	0.001
	ASE	0.035	0.034	0.037	0.037	0.037	0.037
	ESE	0.034	0.040	0.037	0.041	0.042	0.038
ORAC	Bias	-0.004	-0.001	0.006	0.004	0.002	-0.001
	ASE	0.036	0.035	0.038	0.038	0.039	0.038
	ESE	0.036	0.039	0.034	0.039	0.041	0.037

ACTG175 data

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 I. We randomly select 300 patients from the study to consist of our dataset.

ACTG175 data

- The response variable is the log-transformed value of the CD4 counts at 20±5 weeks.
- We use binary variables for the treatments $\mathbf{x}_i = (x_{i1}, x_{i2}, x_{i3})^\mathsf{T}$.
- 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).

ACTG175 data

- We fit the heterogeneous model

$$y_i = \mathbf{z}_i^{\mathsf{T}} \boldsymbol{\eta} + \mathbf{x}_i^{\mathsf{T}} \boldsymbol{\beta}_i + \varepsilon_i, i = 1, \ldots, 300,$$

  where $\mathbf{z}_i = (1, z_{i2} \ldots, z_{i13})^{\mathsf{T}}$ 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* $\boldsymbol{\beta}_1 = (\beta_{11}, \ldots, \beta_{1n})$, $\boldsymbol{\beta}_2 = (\beta_{21}, \ldots, \beta_{2n})$, *and* $\boldsymbol{\beta}_3 = (\beta_{31}, \ldots, \beta_{3n})$.

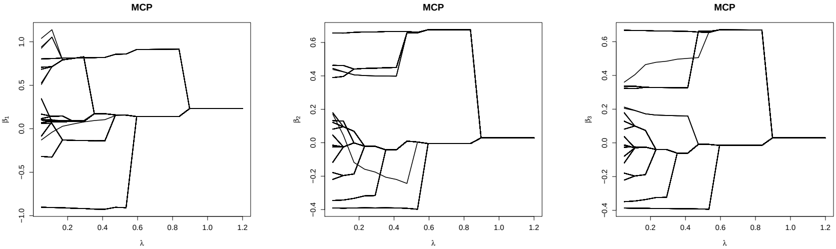


Table 5 : The estimates (Est.), standard deviations (s.d.) and p-values (P-value) of $\boldsymbol{\alpha}_1$ and $\boldsymbol{\alpha}_2$ by the MCP and SCAD methods, and those values of $\boldsymbol{\beta} = \boldsymbol{\alpha}_1$ by the OLS method.

		α_{11}	α_{12}	α_{13}	α_{21}	α_{22}	α_{23}
MCP	Est.	0.141	-0.011	-0.039	0.835	0.666	0.687
	s.d.	0.055	0.055	0.055	0.394	0.268	0.251
	p-value	0.010	0.841	0.478	0.034	0.013	0.006
SCAD	Est.	0.142	-0.010	-0.037	0.805	0.614	0.636
	s.d.	0.055	0.055	0.055	0.395	0.268	0.251
	p-value	0.010	0.855	0.501	0.041	0.022	0.011
OLS	Est.	0.212	0.035	0.036	—	—	—
	s.d.	0.060	0.058	0.058	—	—	—
	p-value	< 0.001	0.550	0.532	—	—	—

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 λ values. It is much more difficult to derive the results for λ values selected based on a data-driven procedure.

Thank you!