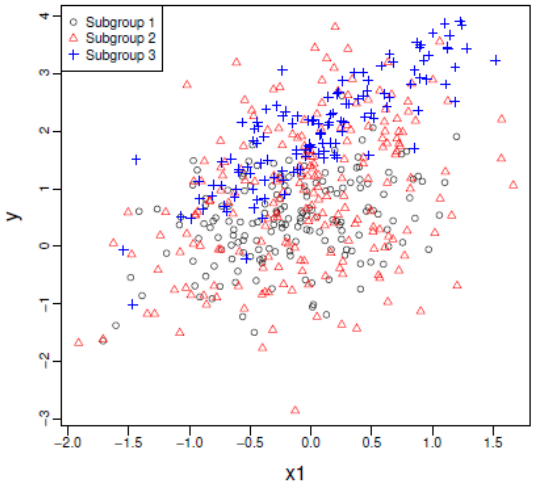
A model-based multi-threshold method for subgroup identification

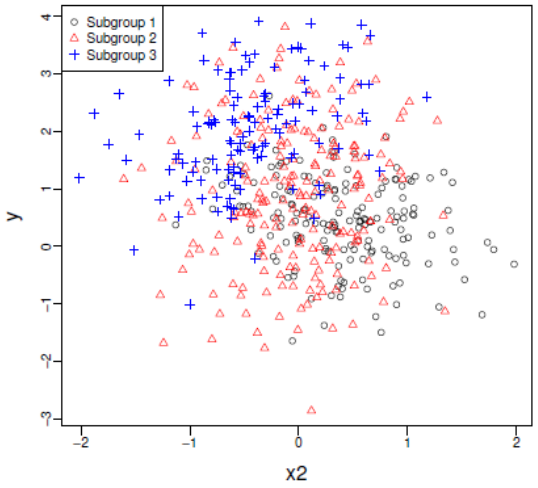
Jingli Wang | Jialiang Li | Yaguang Li | Weng Kee Wong

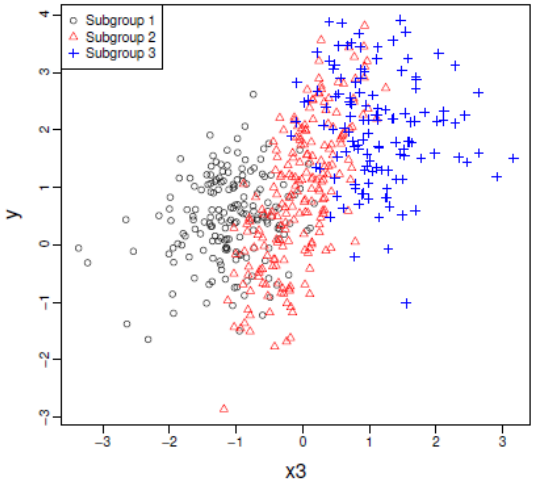
Rising of Personalized Medicine

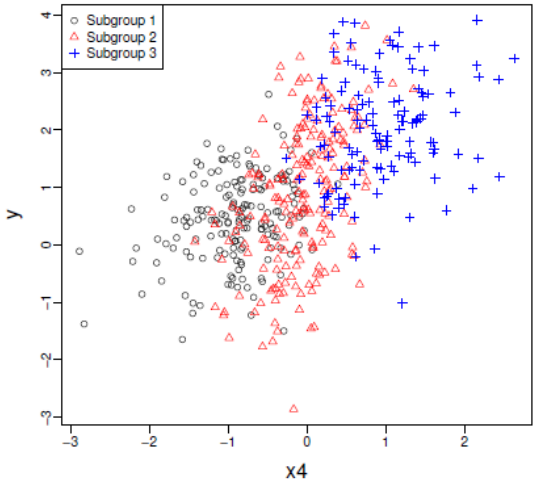
- The "one size fits all" strategy does not work for many serious diseases, such as cancer and targeted therapies based on individual traits tend to work better,
- It is much more difficult to find a treatment that works for all patients;
- Risk factors for a disease are likely going to vary among different patients groups;
- Recent advances in genomics, computational biology, medical imagining and regenerative medicine have made targeted therapies more feasible.

Model-based Subgroups









Subgrouping Methods

- Tree-like partitioning algorithms: Automatic Interaction Detection (AID), theta automatic interaction detection (THAID) and the classification and regression tree (CART), generalized unbiased interaction detection and estimation (GUIDE), model-based recursive partitioning (MOB), interaction tree (IT), subgroup identification based on difference effect search (SIDES) and Virtual twins (VT).

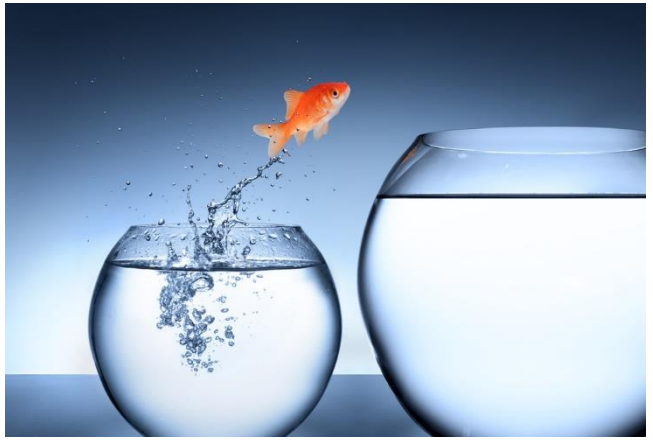
Our model involves change points

- (i) one needs to decide the number of change points and (ii) one must estimate the change point locations accurately.
- Earlier authors proposed iterative cumulative-sum methods which could be computationally intensive.
- Recent authors adopted the penalization method which accelerates the change point detection.

Thresholding variable

- One first decides on the choice of the thresholding variable for which the change point is sought.
- Typically, simple thresholding variable is used. For example, in econometric time series modeling, the thresholding variable is simply the time.
- Existing methods usually take one of the covariates as the thresholding variable, which may be inadequate for partitioning and prediction purposes.
- The same applies when we choose the splitting variable for a tree method.

True model



$$y_i = \sum_{j=1}^{s+1} \boldsymbol{x}_i' \boldsymbol{\beta}_j 1_{\{a_{j-1} < Z_i \le a_j\}} + \epsilon_i,$$

$$= \boldsymbol{x}_i' \left[\boldsymbol{\beta}_1 + \sum_{j=1}^{s} (\boldsymbol{\beta}_{j+1} - \boldsymbol{\beta}_j) 1_{\{a_j < Z_i \le a_{s+1}\}} \right] + \epsilon_i, \quad i = 1, \cdots, n,$$

Previous theoretical works

- Li and Jin (2018, AOS) proposed a two-step multi-change point detection (TSMCPD) method for the model and proved the theoretical properties for estimators.
- Estimated number of change points is strongly consistent to the true number.
- Estimated locations of change points are strongly consistent to the true locations.
- Estimators for regression parameters enjoy the oracle properties.
- Allow censored survival time response too.

(1) Splitting stage

$$\sum_{j=1}^{q_n+1} \frac{b_j}{2n} \sum_{i \in \mathcal{I}_j} w_{[r_{Y_i}, \mathcal{I}_j]} \left(Y_i - \mathbf{X}_i^\top (\beta_1 + \sum_{k=1}^{j-1} \mathbf{d}_k) \right)^2 + \sum_{k=1}^{q_n} p_{\lambda_n, \gamma_n}(\|\mathbf{d}_k\|)$$

(2) Refining stage

Algorithm 1 TSMCD

for $\ell = 1, 2, \ldots, 20$ **do**

Step 1: Splitting stage

Set $m = \lfloor 0.1\ell\sqrt{n^*} \rfloor$ and $q_n = \lfloor n^*/m \rfloor - 1$ where n^* is the number of events;

Split the data sequence into $q_n + 1$ segments $\mathcal{I}_j, j = 1, \ldots, q_n + 1$ as Section 2.1;

Estimate $\hat{\theta} = ((\hat{\theta}_1)^\top, \ldots, (\hat{\theta}_{q_n+1})^\top)^\top$ by minimizing (2.3) or (2.4);

Compute the index sets $\hat{\mathcal{A}} = \{j : \hat{\theta}_j \neq 0, j = 1, \ldots, q_n + 1\}$ and $\hat{\mathcal{A}}^* = \{j : j \in \hat{\mathcal{A}}, j - 1 \notin \hat{\mathcal{A}}, j = 2, \ldots, q_n + 1\} \equiv \{\hat{k}_1, \ldots, \hat{k}_{\hat{s}}\}$ where $\hat{k}_1 < \hat{k}_2 < \ldots < \hat{k}_{\hat{s}}$ and $\hat{s} = \sharp\hat{\mathcal{A}}^*$;

Step 2: Refining stage

if $\hat{s} = 0$ **then** go to the step 9;

else if $\hat{s} > 0$ **then** estimate the threshold a_j in $(\tilde{Z}_{(n^*-(q_n-\hat{k}_j+3)m)}, \tilde{Z}_{(n^*-(q_n-\hat{k}_j+1)m)}]$ by (2.5);

end if

Estimate the coefficient $\theta^* = ((\beta_1^*)^\top, (\mathbf{d}_1^*)^\top, \ldots, (\mathbf{d}_{\hat{s}}^*)^\top)^\top$ by minimizing (2.6);

Return the set of estimated thresholds $\hat{\mathcal{M}}_\ell = \{\hat{a}_{1,\ell}, \ldots, \hat{a}_{\hat{s}_\ell,\ell}\}$ in step 7 and the estimator of the coefficient $\hat{\theta}_\ell^*$ in step 9, and compute $BIC_{\hat{\mathcal{M}}_\ell}$ by (2.8);

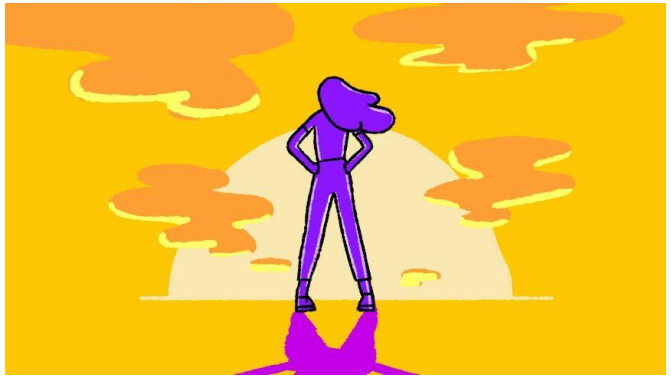
end for

Choose $\hat{\ell}$ that minimizes the $BIC_{\hat{\mathcal{M}}_\ell}$ and obtain the final estimators $\hat{\mathcal{M}}_{opt} = \hat{\mathcal{M}}_{\hat{\ell}}$ and $\hat{\theta}_{opt}^* = \hat{\theta}_{\hat{\ell}}^*$.

Thresholding variable

- When the thresholding variable Z is given, we may apply the two stage multiple change-point detection (TSMCD) method in Li and Jin (2018) to find change points and estimate the regression parameters.
- In practice, the choice of thresholding variable is crucial on the identification and interpretation of the subgroups.
- We consider 4 methods to specify the thresholding variable Z.

Method 1. Single covariate



$$Z \in \{X_j, j = 1, \cdots, p\}.$$

Method 2. Weighted combination

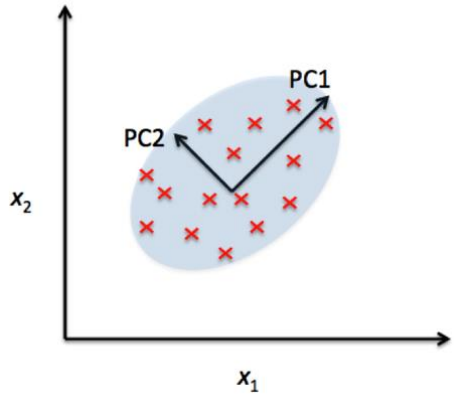
$$Z = \sum_{j=1}^{p} w_j X_j,$$

Method 3. Factor Analysis

$$\tilde{X} = \mu + \Lambda f + \varepsilon$$

$$Z \in \{\hat{f}_1, \cdots, \hat{f}_m\}$$

Method 4. Principal Component Analysis



$$g_1 = \boldsymbol{e}_1' \tilde{\boldsymbol{X}}, \text{ with } \boldsymbol{e}_1 \in \mathbb{R}^p \text{ and } \boldsymbol{e}_1' \boldsymbol{e}_1 = 1 \text{ such that}$$

$$\mathtt{var}(\boldsymbol{e}_1' \tilde{\boldsymbol{X}}) = \max_{\boldsymbol{e}} \mathtt{var}(\boldsymbol{e}' \tilde{\boldsymbol{X}}) = \max_{\boldsymbol{e}} \boldsymbol{e}' \boldsymbol{\Sigma} \boldsymbol{e}.$$

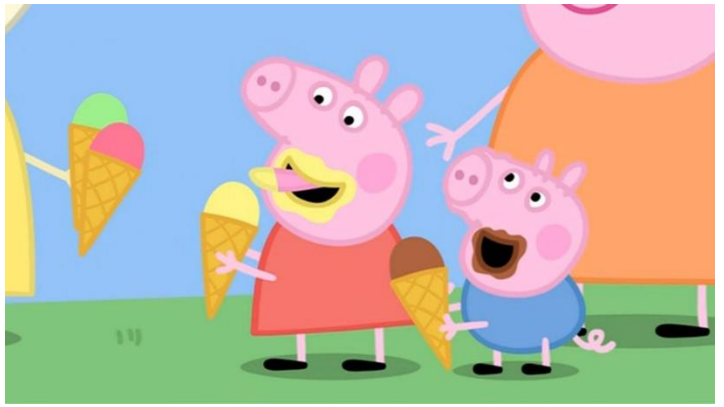
Simulations

For Cases I to IV, we apply the factor analysis method to estimate the thresholding variable and generate the covariates $\boldsymbol{x}_i = (x_{i1}, \cdots, x_{ip})'$ from the following factor model:

$$\boldsymbol{x}_i = \boldsymbol{\Lambda} \boldsymbol{f}_i + \boldsymbol{\epsilon}_i, i = 1, \cdots, n.$$

Here $\boldsymbol{\Lambda}$ is a $p \times m$ loading matrix where each row is generated from the standard multivariate normal distribution, the factors $\boldsymbol{f}_i = (f_{i1}, \cdots, f_{im})'$ are generated from the multivariate normal distribution $N(\boldsymbol{0}, \boldsymbol{I}_m)$, and the stochastic error $\boldsymbol{\epsilon}_i$ is generated from the multivariate normal distribution $N(\boldsymbol{0}, 0.1\boldsymbol{I}_p)$. There are $p = 10$ covariates and $m = 3$ factors.

Simulations



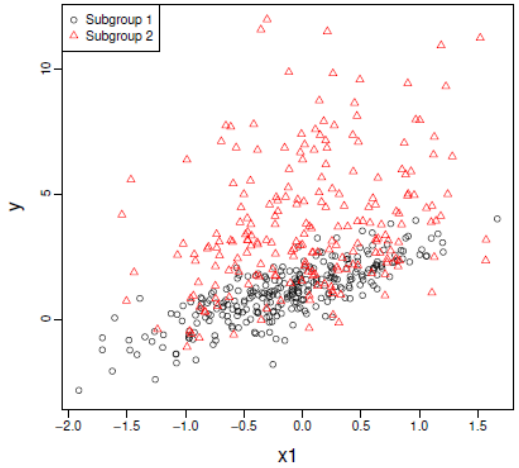
Case I: We generate data from the model without any change point (i.e. no subgroups)

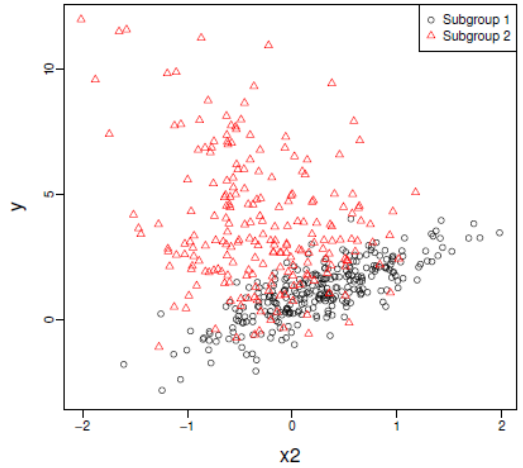
$$y_i = 1 + x_{i1} + \frac{1}{2}x_{i2} + 2(x_{i3} + x_{i4}) + \epsilon_i,$$

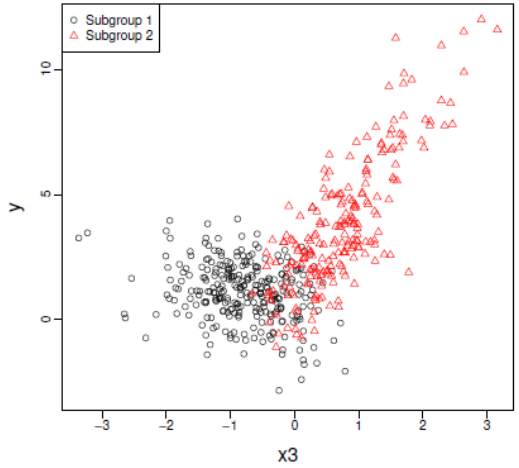
Simulations

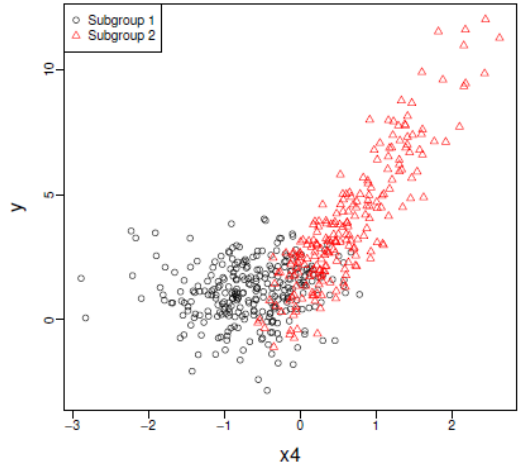
Case II: We generate data from the model

$$y_i = 1 + x_{i1} + x_{i2}1_{\{f_{i1} \le 0\}} + 2(x_{i3} + x_{i4})1_{\{f_{i1} > 0\}} + \epsilon_i,$$





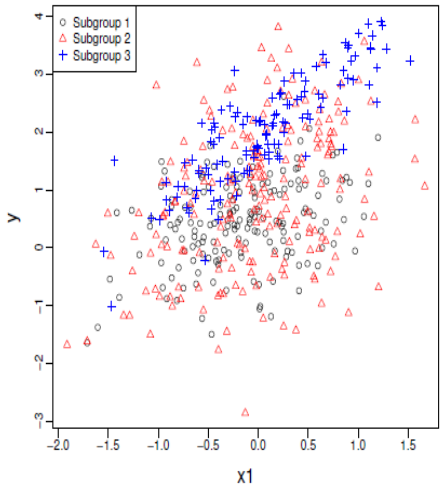


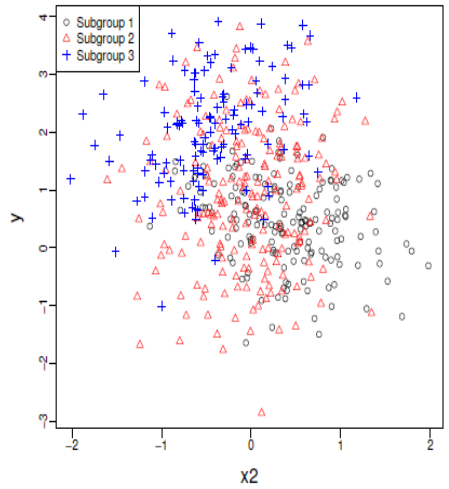


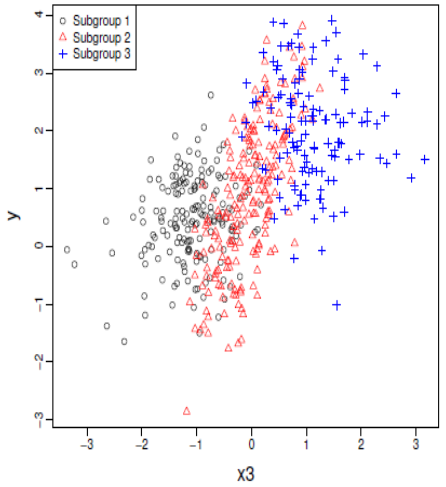
Simulations

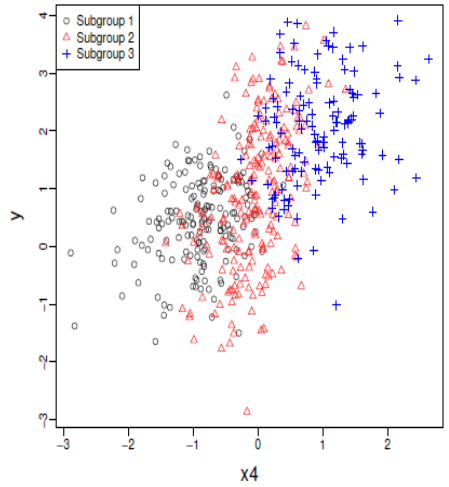
Case III: We generate data from the model

$$y_i = 1 + x_{i1} - x_{i2}1_{\{f_{i1} \le a_1\}} + (2x_{i3} + \frac{1}{2}x_{i4})1_{\{a_1 < f_{i1} \le a_2\}} + 1_{\{f_{i1} > a_2\}} + \epsilon_i,$$







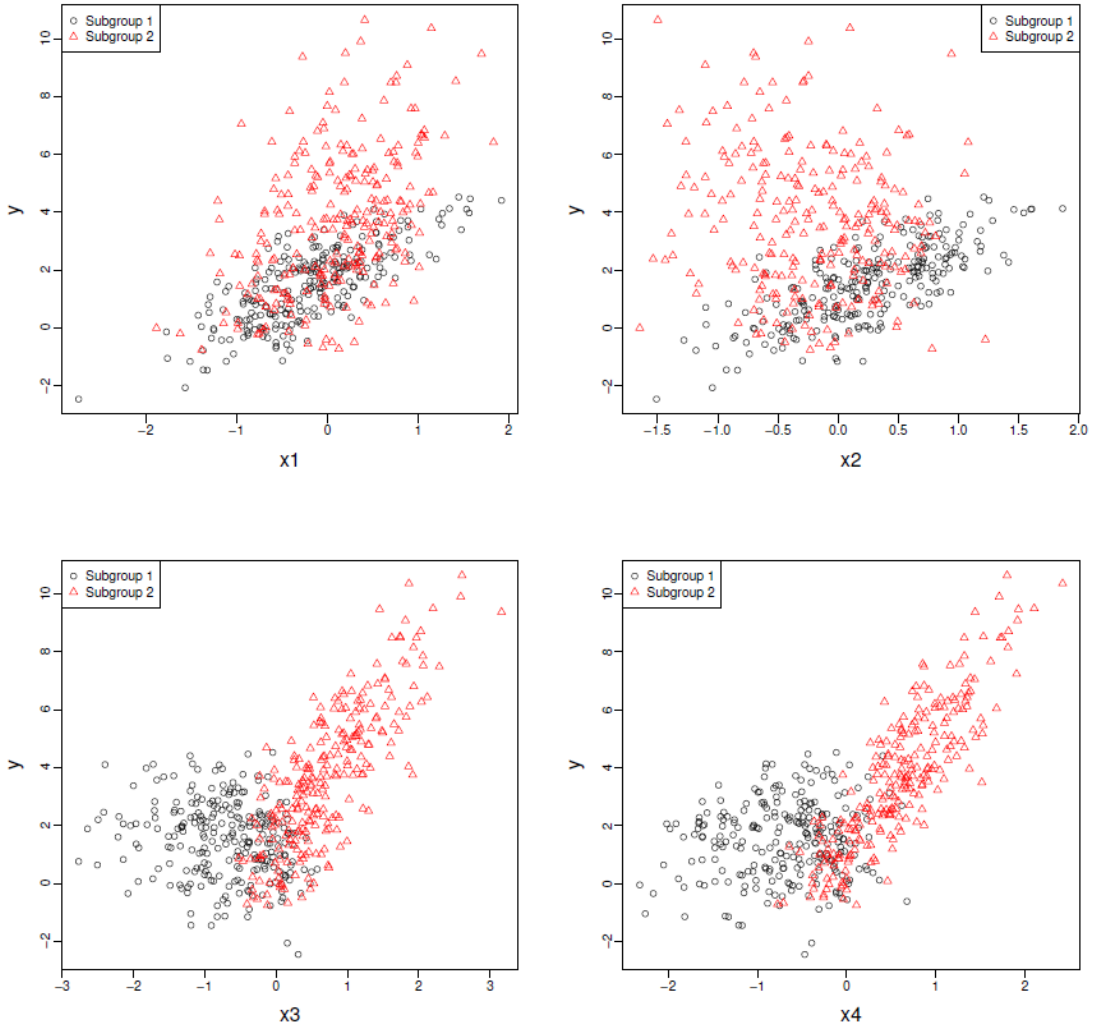


Simulations

Case IV: We generate data from the following model that contains a treatment variable

$$y_i = 1 + x_{i1} + (u_i + x_{i2})1_{\{f_{i1} \le 0\}} + 2(x_{i3} + x_{i4})1_{\{f_{i1} > 0\}} + \epsilon_i,$$

where $u_i \sim Bernoulli(0.5)$ is a binary treatment indicator. This model involves only one change point and hence two subgroups.



Simulations

		n=300		
	Method	MSE.tr(sd)	MSE.te(sd)	$\hat{s}+1$ (sd)
Case I	AIM-rule	0.2228(0.0212)	0.2824(0.0440)	2(0)
s+1=1	seq-BT	0.2231(0.0216)	0.2818(0.0447)	2(0)
	PRIM	0.2228(0.0214)	0.2814(0.0440)	2(0)
	MOB	0.2354(0.0230)	0.2640(0.0403)	1.0180(0.1331)
	F1-TSMCD	0.2422(0.0228)	0.2591(0.0392)	1(0)
	F2-TSMCD	0.2422(0.0228)	0.2591(0.0392)	1(0)
	LM	0.2361(0.0223)	0.2660(0.0405)	1(0)
Case II	AIM-rule	0.7000(0.1394)	0.8751(0.2215)	2(0)
s+1=2	seq-BT	0.6340(0.0975)	0.7770(0.1655)	2(0)
	PRIM	0.4910(0.1625)	0.7273(0.3739)	2(0)
	MOB	0.3283(0.0534)	0.5366(0.8617)	2.0160(0.1256)
	F1-TSMCD	0.2966(0.0424)	0.3672(0.0972)	2(0)
	F2-TSMCD	0.2850(0.0392)	0.3664(0.1101)	2(0)
	TSMCD	0.2394(0.0244)	0.2714(0.0520)	2(0)
Case III	AIM-rule	0.4198(0.0455)	0.5075(0.0746)	2(0)
s+1=3	seq-BT	0.4496(0.0485)	0.5308(0.0783)	2(0)
	PRIM	0.4228(0.0425)	0.5128(0.0732)	2(0)
	MOB	0.3609(0.0536)	0.5949(0.2978)	2.5824(0.5501)
	F1-TSMCD	0.3659(0.0667)	0.4577(0.0913)	2.7118(0.5256)
	F2-TSMCD	0.3406(0.0553)	0.4309(0.0817)	2.7588(0.4373)
	TSMCD	0.2394(0.0244)	0.2858(0.0448)	3(0)
Case IV	AIM-rule	1.4016(0.3054)	1.8241(0.5061)	2(0)
s+1=2	seq-BT	1.2945(0.4429)	1.6795(0.7902)	2(0)
	PRIM	1.1633(0.3905)	1.5037(0.5387)	2(0)
	MOB	0.3584(0.0614)	0.5658(0.7032)	2.0151(0.1220)
	F1-TSMCD	0.3176(0.0479)	0.4200(0.1245)	2.0019(0.0434)
	F2-TSMCD	0.3095(0.0477)	0.4205(0.1343)	2(0)
	TSMCD	0.2375(0.0239)	0.2736(0.0530)	2(0)

Bovine Collagen Clinical Trial

- We apply our subgrouping methods to a NIH-sponsored randomized Bovine Collagen Trial for Scleroderma patients at 12 centers in the USA.
- Patients with diffuse Scleroderma were enrolled in this multicenter phase II double-blind placebo controlled trial and a total of 831 observations were collected. Patients were randomized to receive oral native collagen at a dose of 500 g/day or a similar appearing placebo.
- The Modified Rodnan Skin Score (MRSS) was the primary outcome variable.
- To implement the proposed methods to predict MRSS, we consider 6 predictor: 'over' (disease progression), 'pain' (index of pain), 'haq' (health assessment questionnaire), 'pga' (patient self assessment of disease progression), 'dlcop' (lung performance measurement) and 'age'.
- After removing missing values, we have a sample of 295 observations in the downstream analysis. All

Results

Method	MSE.tr(sd)	MSE.te(sd)	$\hat{s}+1$ (sd)
AIM-rule	0.7794(0.0093)	0.8347(0.0170)	2(0)
seq-BT	0.7725(0.0116)	0.8342(0.0169)	2(0)
PRIM	0.7793(0.0102)	0.8312(0.0151)	2(0)
MOB	0.6977(0.0096)	0.8535(0.0577)	1.9020(0.1463)
1-TSMCD	0.7657(0.0060)	0.8299(0.0199)	1.3587(0.0769)
A-TSMCD	0.7753(0.0119)	0.8396(0.0176)	1.2140(0.1891)
F1-TSMCD	0.7438(0.0154)	0.8236(0.0201)	1.5840(0.2196)
F2-TSMCD	0.7447(0.0140)	0.8228(0.0220)	1.5520(0.1888)
PC-TSMCD	0.7592(0.0123)	0.8330(0.0133)	1.4000(0.2010)
LM	0.7610(0.0013)	0.8102(0.0121)	1(0)

Breast Cancer Study

- 97 lymph nodenegative breast cancer patients who were 55 years old or younger in this study (Veer et al. (2002) Nature).
- Clinical risk factors (confounders) were age, tumor size, histological grade, angioinvasion, lymphocytic infiltration, estrogen receptor (ER) and progesterone receptor (PR) status.
- Expression levels for 24,481 gene probes were collected. After removing genes with severe missingness, we analyzed 24,188 genes.

Random selection of 20 genes: bootstrap 500 times

Method	MSE(sd)	$\hat{s}+1$(sd)	time(second)
AIM-rule	0.6763(0.0622)	2(0)	62.50
seq-BT	0.6916(0.0636)	2(0)	61.11
PRIM	0.6816(0.0732)	2(0)	18.79
1-TSMCD	0.7618(0.1042)	2.1314(0.5868)	249.21
A-TSMCD	0.7847(0.1536)	2.243(1.3715)	12.24
F1-TSMCD	0.6565(0.1190)	3.6700(1.2977)	75.82
F2-TSMCD	0.6762(0.1468)	3.0460(1.2518)	76.91
PC-TSMCD	0.6370(0.1523)	3.2900(1.3067)	57.87
LM	0.7384(0.0608)	1(0)	-

Use 20 genes screened using ROC analysis (Yu et al. 2011)

Method	MSE.tr(sd)	MSE.te(sd)	$\hat{s}+1$ (sd)
AIM-rule	0.4989(0.1201)	1.4622(0.3930)	2(0)
seq-BT	0.5224(0.1317)	1.5061(0.4940)	2(0)
PRIM	0.5259(0.1256)	1.4649(0.4564)	2(0)
1-TSMCD	0.3817(0.0422)	1.9609(0.3380)	2.5194(0.1370)
A-TSMCD	0.4138(0.1271)	2.0159(0.6202)	2.4280(0.5302)
F1-TSMCD	0.4374(0.0813)	1.3328(0.2089)	2.4440(0.4174)
F2-TSMCD	0.4401(0.0859)	1.3682(0.2392)	2.4520(0.4685)
PC-TSMCD	0.4424(0.0927)	1.3659(0.2768)	2.4920(0.4622)
LM	0.4481(0.0469)	2.5049(0.8409)	1(0)

Use 4 genes screened using nonparametric regression (Cheng et al. 2016)

Method	MSE.tr(sd)	MSE.te(sd)	$\hat{s}+1$ (sd)
AIM-rule	0.5988(0.0511)	1.0308(0.1677)	2(0)
seq-BT	0.6114(0.0551)	1.0290(0.1498)	2(0)
PRIM	0.6050(0.0591)	1.0363(0.1579)	2(0)
MOB	0.6733(0.0302)	0.9109(0.1302)	1.0050(0.05)
1-TSMCD	0.6352(0.0457)	1.1055(0.1928)	1.3350(0.1904)
A-TSMCD	0.6553(0.0868)	1.3230(1.1757)	1.2450(0.3138)
F1-TSMCD	0.6236(0.0898)	1.1850(0.9451)	1.3750(0.3718)
F2-TSMCD	0.6236(0.0898)	1.1850(0.9451)	1.3750(0.3718)
PC-TSMCD	0.6220(0.065)	0.9859 (0.1434)	1.3950(0.3356)
LM	0.6742(0.0285)	0.9105(0.1301)	1(0)

Future extensions

- Nonparametric models;
- Causal inference;
- Software development.

thank you!