Individualized Multi-directional Variable Selection

Xiwei Tang and Annie Qu*

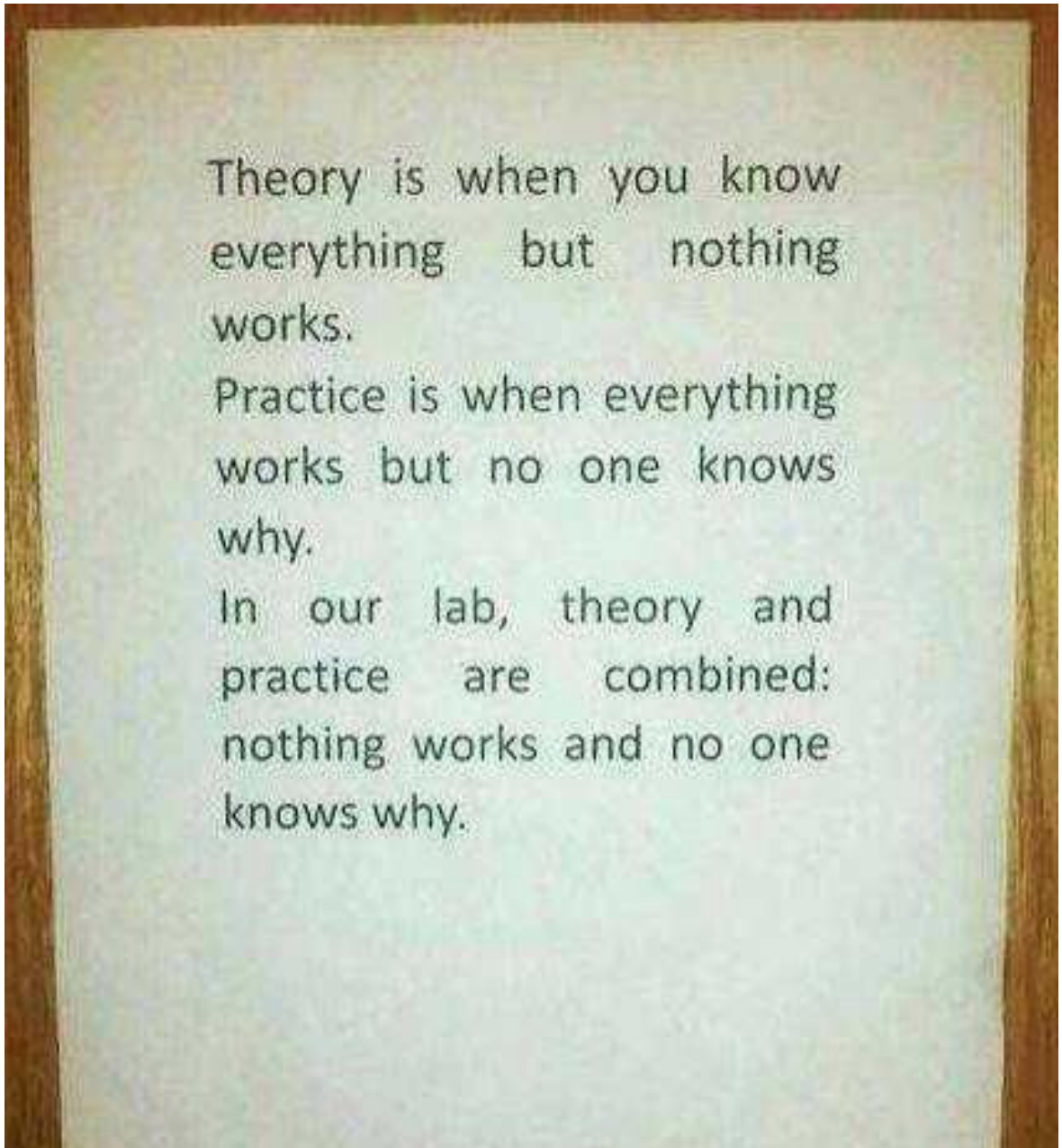
Department of Statistics, University of Illinois at Urbana-Champaign

National University of Singapore, July, 2017

SLDS/Nonparamatric Conference in 2018

- The Conference on Statistical Learning and Data Science / Nonparametric Statistics
- Date: Monday, June 4 - Wednesday, June 6, 2018
- Place: Columbia University, NYC
- Conference is chaired by Annie Qu (anniequ@illinois.edu) and Cynthia Rudin (cynthia@cs.duke.edu).
- Local Chair: Tian Zheng, Columbia U
- Still looking for industry sponsorship!

Prelude


Theory is when you know everything but nothing works.
Practice is when everything works but no one knows why.
In our lab, theory and practice are combined: nothing works and no one knows why.

Individualized Modeling V.S. Population Modeling

- Individualized modeling
  - Heterogeneity variation
  - Subject-specific effect

- Population modeling
  - Homogeneity assumption
  - Marginal average effect

Why Individualized Modeling (Variable Selection)?

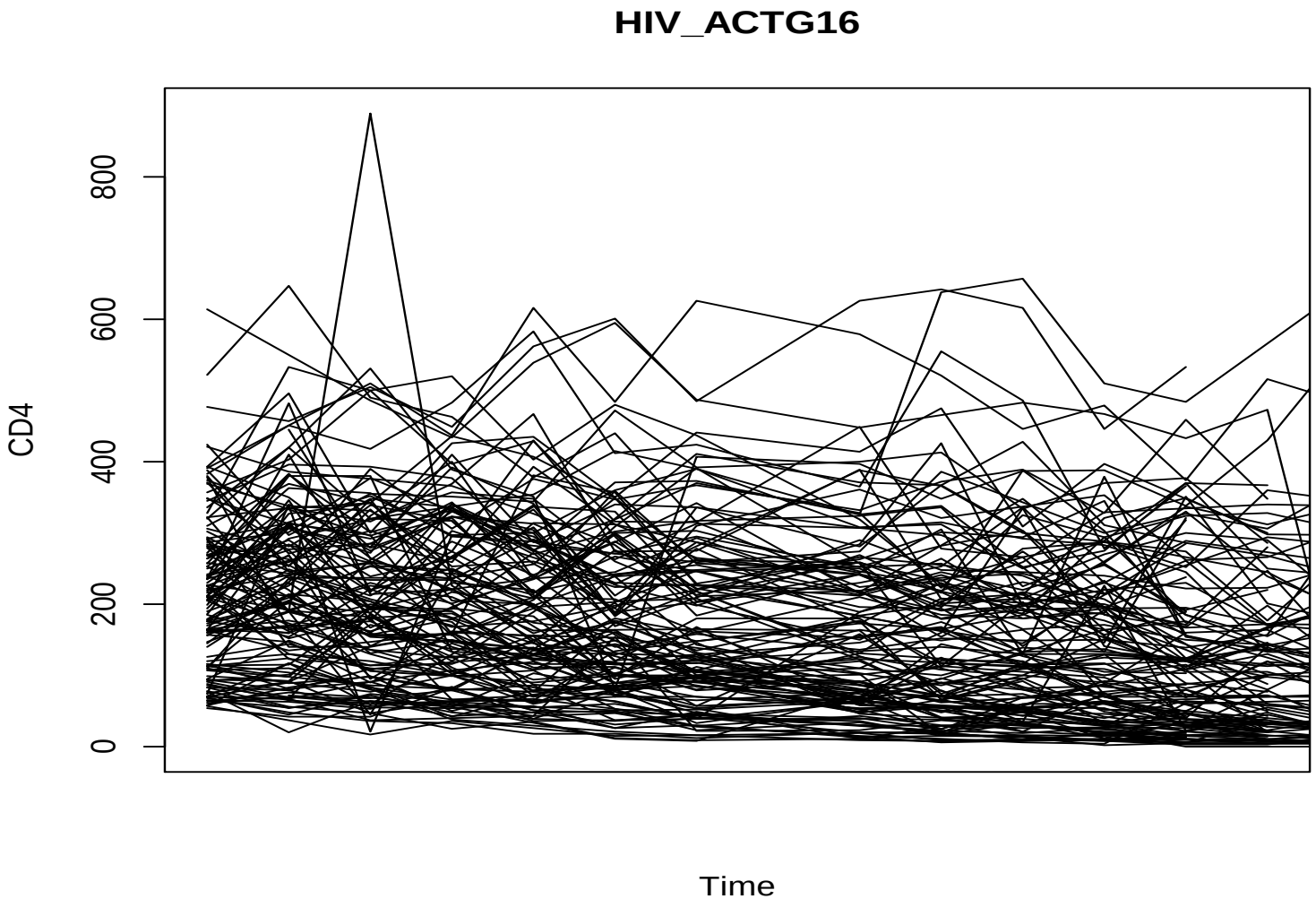
- Abundant individual information collected

- More precise prediction for individuals
  - Personalized medicine
  - Individualized recommendation

Real Data Example: HIV Data (ACTG16)

- Harvard AIDS clinical trial group: **longitudinal** data study
  - 140 patients with measurements over 14 time points
  - Response: **CD4 counts** (missing rate: 8.5%)
  - Main variable of interest: **Zidovudine (ZDV)** treatment effect
  - Control variables: Age, Gender

Individuals' CD4 counts (response) over time



A Traditional Model Testing ZDV Effect

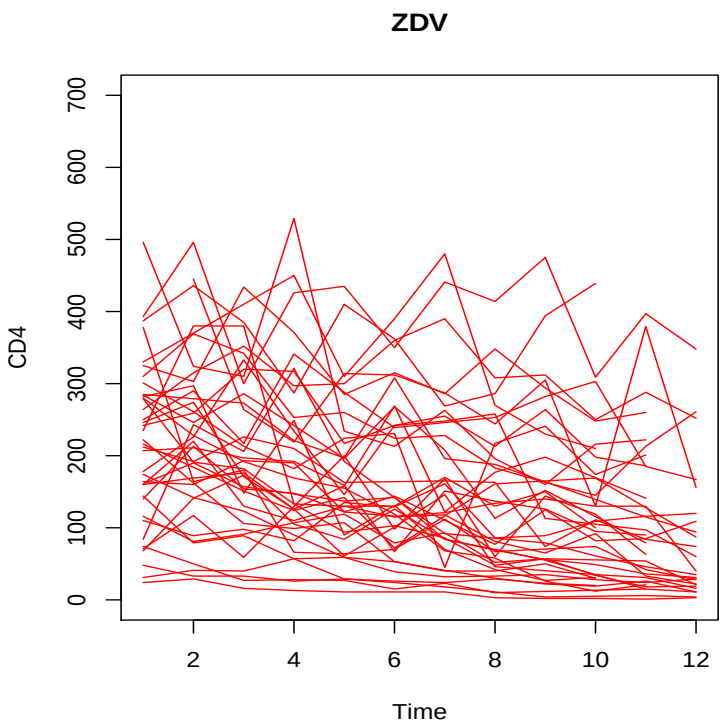
- A marginal model:

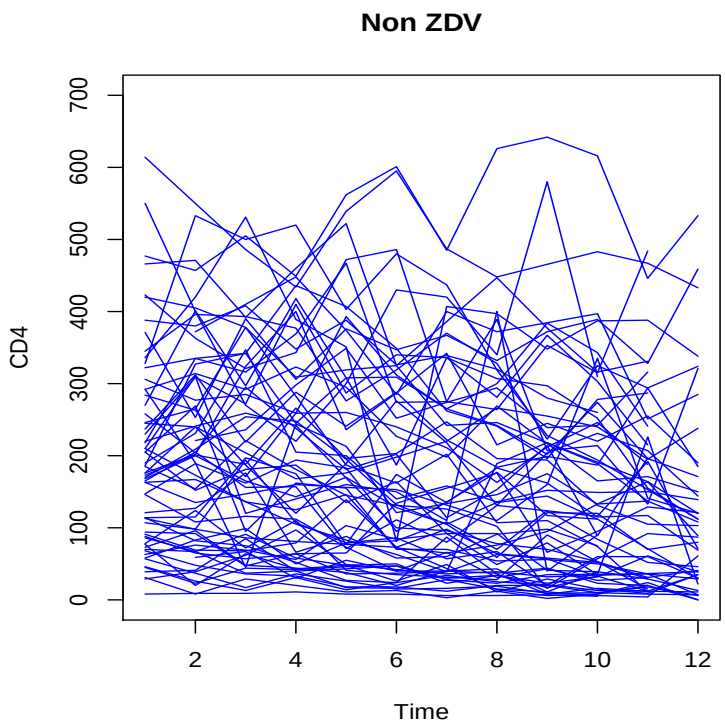
$$y_{it} = \beta_0 + \beta_t * Time + \beta_z * ZDV + \beta_{zt} * ZDV * Time + \beta_a * Age + \beta_g * Gender + \varepsilon_{it}$$

- Treatment indicator: ZDV (treatment group=1, control group=0)
- Treatment effect β_{zt}: difference in **slope** compared to the control group

ZDV Treatment Effect

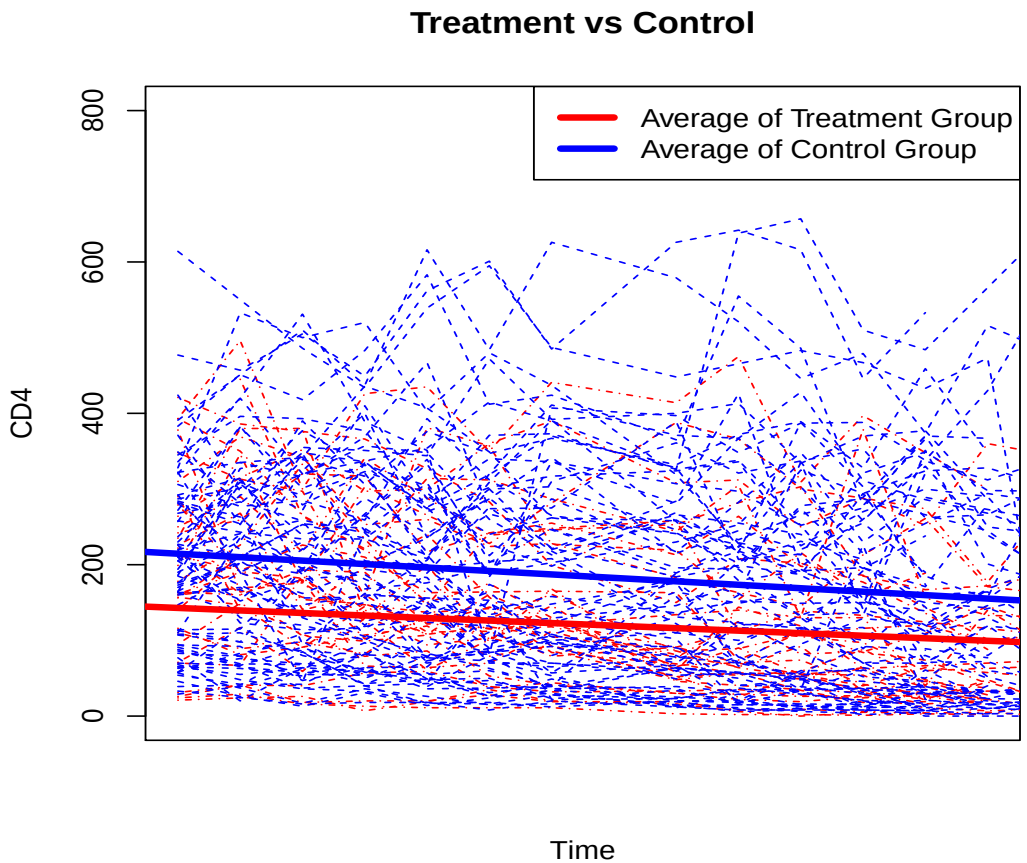
- Examine the marginal treatment effect over time: $\hat{\beta}_{zt}$
- $\hat{\beta}_{zt}$ is **Not significant!** (p-value=0.113)
  $\Longrightarrow$ **No effect** of ZDV on average



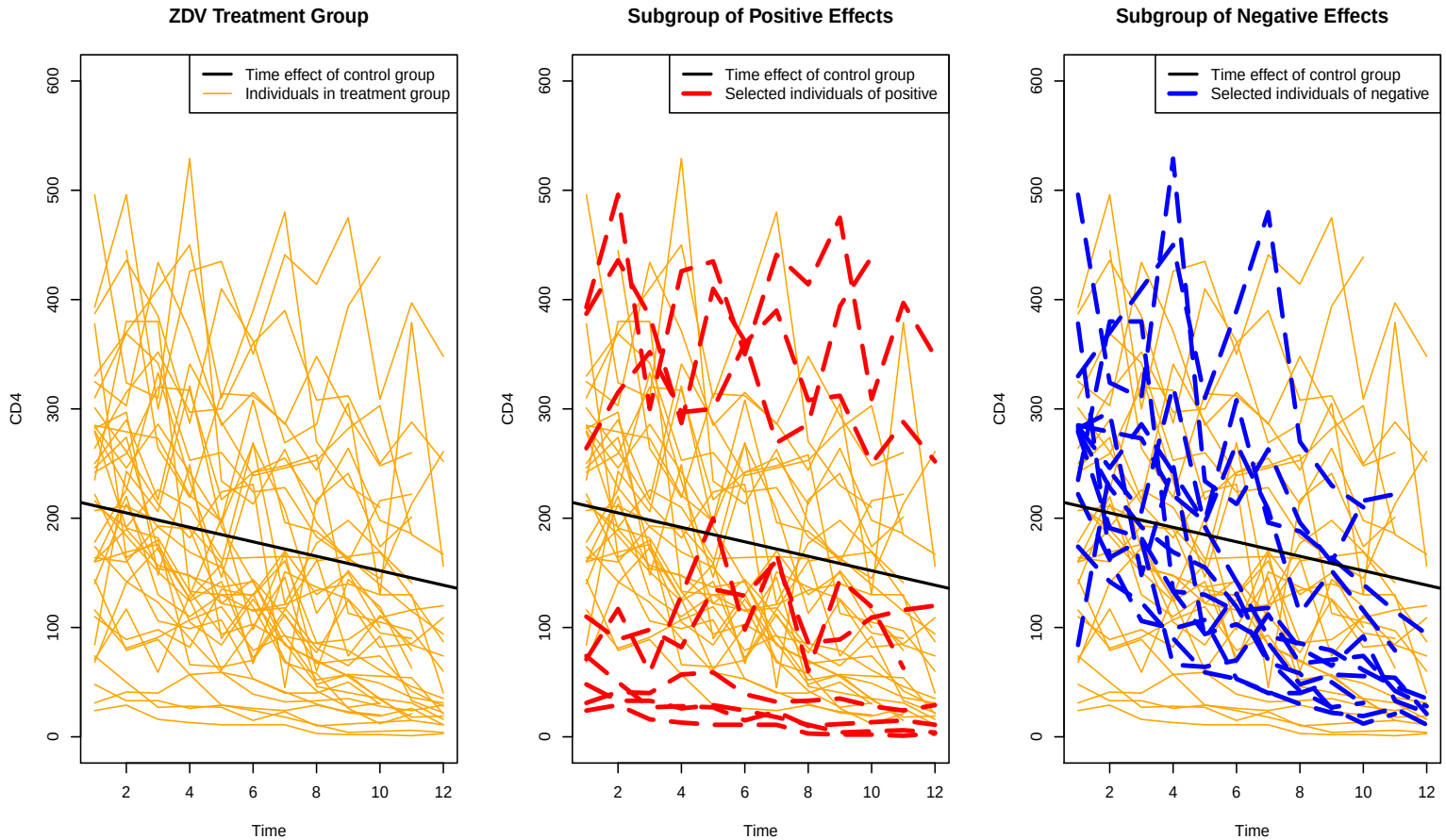


Average Treatment Effect

- Treatment effect over time



Heterogeneity within ZDV Treatment Group



* Subgroup individuals are selected based on subject-wise OLS estimators

Individualized Regression Model Framework

Individualized regression model under **clustered data** framework:

$$\boldsymbol{y_i} = \boldsymbol{X_i}\boldsymbol{\beta_i} + \boldsymbol{Z_i}\boldsymbol{\alpha} + \boldsymbol{\varepsilon_i}, \quad i = 1, ..., N$$

- N: sample size (number of subjects)
- m_i: cluster size (set $m_i = m$, number of repeated measurements)
- $\boldsymbol{y_i}$ $(m \times 1)$: response vector; $\boldsymbol{\varepsilon_i}$ $(m \times 1)$: random error
- $\boldsymbol{X_i}$ $(m \times p)$: **individualized covariates** with $\boldsymbol{\beta_i} = (\beta_{i1}, ..., \beta_{ip})'$
- $\boldsymbol{Z_i}$ $(m \times q)$: **population-shared covariates** with $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_q)'$

A Penalized Variable Selection Approach

Minimize a penalized objective function:

$$(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}}) = \underset{\boldsymbol{\beta},\boldsymbol{\alpha}}{\operatorname{argmin}} \ \frac{1}{2} \sum_{i=1}^{N} \| \boldsymbol{y_i} - \boldsymbol{\mu_i}(\boldsymbol{\beta_i}, \boldsymbol{\alpha}) \|_2^2 + \lambda_{N,m} \sum_{i=1}^{N} \sum_{k=1}^{p} \rho(\beta_{ik}) \quad (1)$$

- $\rho(\cdot)$: Lasso (Tibshirani, 1996), SCAD (Fan and Li, 2001), elastic net (Zou, 2005), adaptive Lasso (Zou, 2006), MCP (Zhang, 2010) and TLP (Shen et al., 2012), etc.
- Subject(individual)-wise modeling, may have overfitting
- **Not** utilize **cross-subject** information

Subpopulation with Respect to Regression Coefficients

- Encourage **subgrouping** of individuals who share the similar effect
- Some existing approaches:
  - **Grouping covariates in a population model**: fused Lasso (Tibshirani et al., 2005), OSCAR (Bondell et al., 2008), grouping pursuit (Shen and Huang, 2010), CARDS (Ke et al., 2015)
  - **Mixture-of-regressions model** (Jacobs et al., 1991)
  - **Pairwise penalized clustering** (Hocking et al., 2011; Lindsten et al., 2011; Pan et al., 2013; Ma and Huang, 2016)
  - **Meta lasso** (Li et al., 2013): No subpopulation

Our Target in Two Perspectives

- **Individual feature selection**: select different relevant predictors for different individuals
- **Utilizing subgroup homogeneity**: borrow cross-subject information in both variable selection and model estimation

Subpopulation Structure Assumption

Regarding to the kth ($k = 1, \ldots, p$) individualized covariate, assume:

$$\beta_{ik} = \begin{cases} \gamma_k, & \text{if} \quad i \in \mathcal{G}_k \\ 0, & \text{if} \quad i \in \mathcal{G}_k^c \end{cases}$$

for ith subject, $i = 1, \ldots, N$

- $\mathcal{G}_k$ is the unknown index set for signal group of individuals
- γ_k is the unknown homogeneous effect shared within subgroup
- Different subgrouping with respect to different covariates
- Extensions
  - multiple subgroups: $\mathcal{G}_k^0, \mathcal{G}_k^1, \mathcal{G}_k^2, \ldots$
  - constraint homogeneous effects: $\gamma_k^+ > 0$ or $\gamma_k^- < 0$

The Proposed Approach

- Regression coefficients $\boldsymbol{\beta}_i$'s and $\boldsymbol{\alpha}$ are estimated by minimizing

$$Q_{N,m}(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma}) = \frac{1}{2}\sum_{i=1}^{N}(\boldsymbol{y_i} - \boldsymbol{\mu}_i)^T \boldsymbol{V}_i^{-1}(\boldsymbol{y_i} - \boldsymbol{\mu}_i) + \lambda_{N,m}\sum_{i=1}^{N}\sum_{k=1}^{p} s(\beta_{ik}, \gamma_k),$$

- **Multi-directional separation penalty (MDSP)**

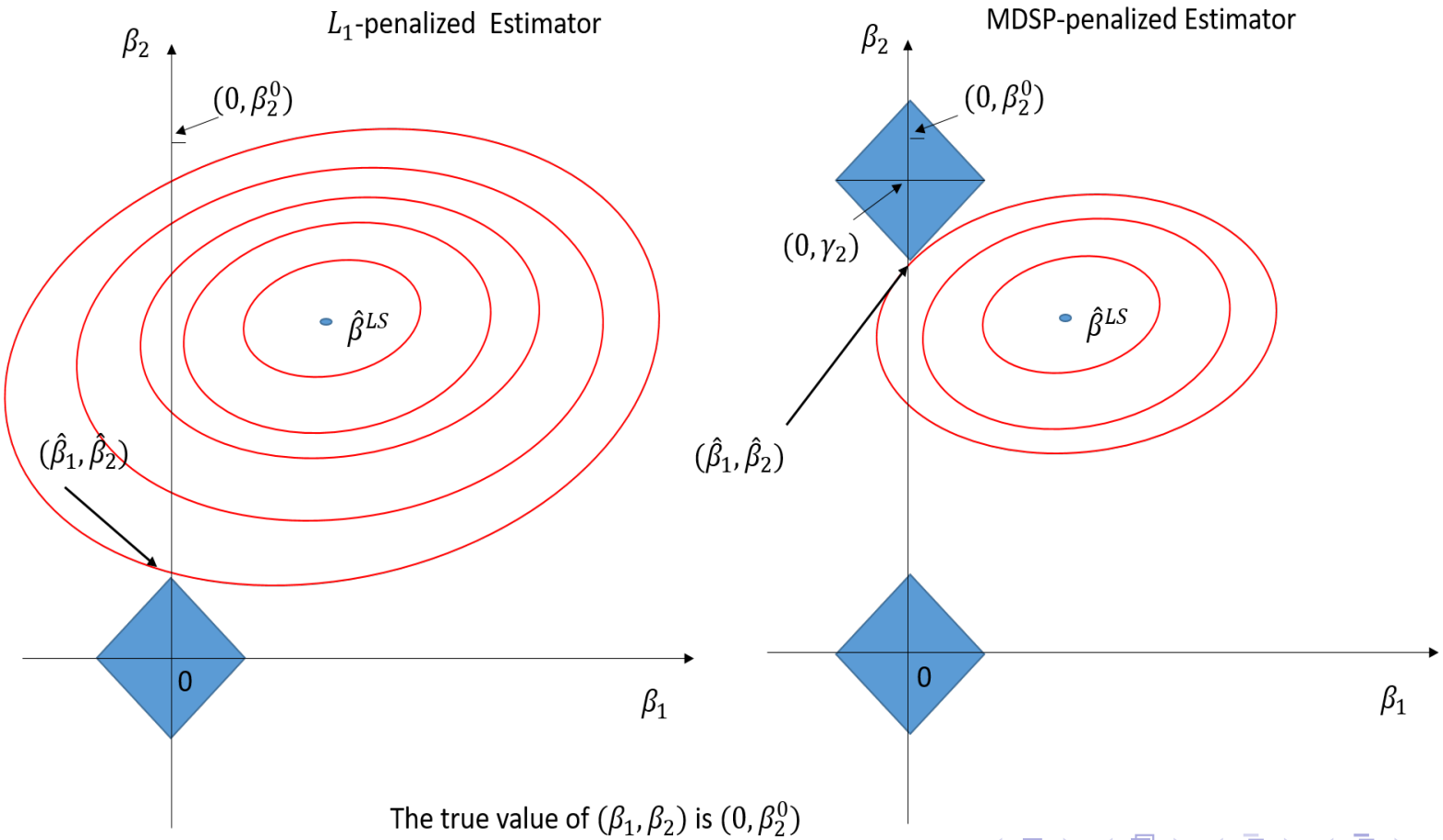
$$s(\beta_{ik}, \gamma_k) = \min(|\beta_{ik} - \gamma_k|, |\beta_{ik}|), \quad k = 1, \ldots, p$$

  - Provide multiple shrinking directions for β_{ik}: either 0 or γ_k
  - Group β_{ik}'s over different subjects

- $\boldsymbol{V}_i = \boldsymbol{A}_i^{1/2}\boldsymbol{R}_m\boldsymbol{A}_i^{1/2}$ incorporates within-subject correlation

Alternative Shrinking Direction

- **L_1-penalty**: $\lambda(|\beta_1| + |\beta_2|)$
- **MD-penalty** for β_2: $\lambda(|\beta_1| + \min(|\beta_2|, |\beta_2 - \gamma_2|))$
- **nearly unbiased estimator** of $\hat{\beta}_2^{MD}$ if $\gamma_2 \to \beta_2^0$

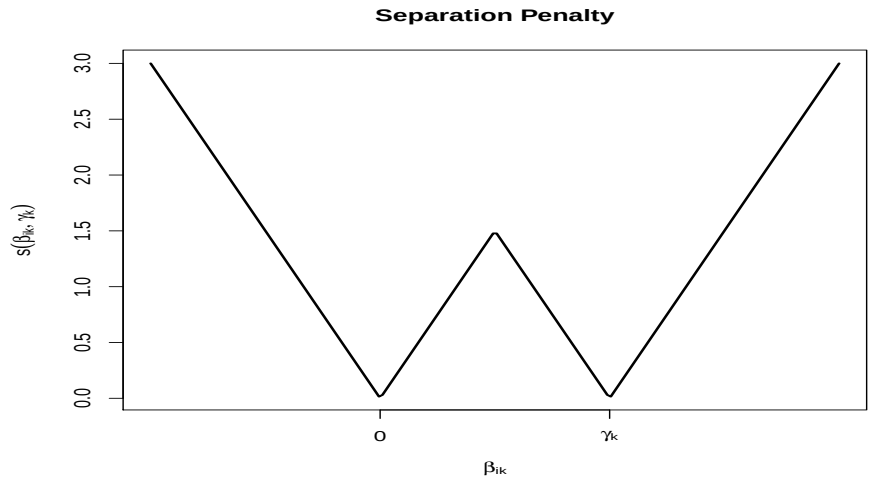


Comparison to Traditional Penalty Functions

- To overcome estimation bias due to L_1-penalty
  - **SCAD, MCP, TLP**: Non-convex penalty, control threshold by tuning parameter
  - **Adaptive Lasso**: control magnitude of penalization through initial weights
  - **Proposed penalty**: provide alternative shrinking direction

Piecewise Convex Separation Penalty and Grouping

- Different shrinking directions to separate individuals
  - Weak signals $\Rightarrow 0$
  - Strong signals $\Rightarrow \gamma_k$
- Center-based clustering
- Estimate $\hat{\gamma}_k$ as the average effect of strong signals
- Non-convex penalty



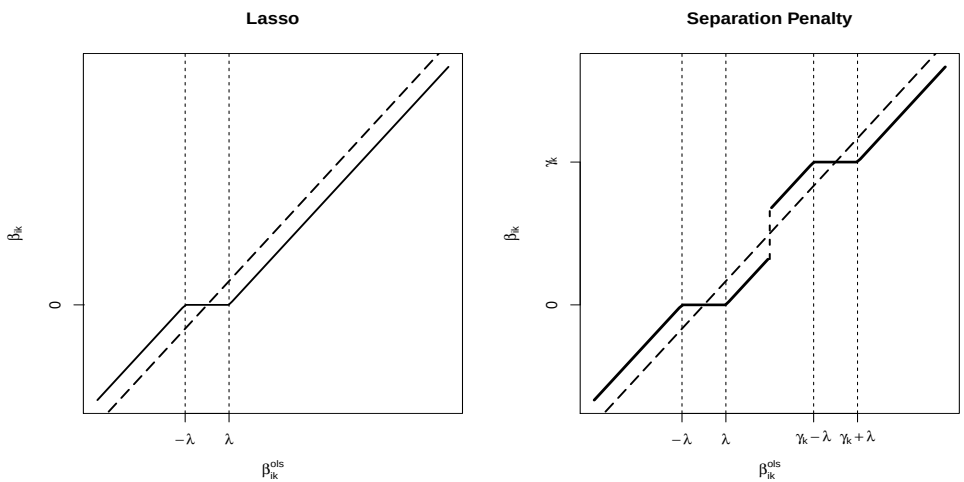


Figure 1: Thresholding functions for Lasso and the proposed separation penalty.

Algorithm for Optimization

1. (Initialization) Provide initials: $\hat{\boldsymbol{\beta}}^{(0)}, \hat{\boldsymbol{\alpha}}^{(0)}$, e.g., OLS or Lasso estimators
2. Calculate $\hat{\boldsymbol{\gamma}}^{(0)} = \text{argmin}_{\boldsymbol{\gamma}} \sum_{k=1}^{p} \sum_{i=1}^{N} \min(|\hat{\beta}_{ik}^{(0)}|, |\hat{\beta}_{ik}^{(0)} - \gamma_k|)$
3. (Regression) At the mth iteration, update $\hat{\boldsymbol{\beta}}^{(m)}$, $\hat{\boldsymbol{\alpha}}^{(m)}$ via minimizing

$$L(\boldsymbol{\beta}, \boldsymbol{\alpha}) + \lambda_{N,m} \sum_{i=1}^{N} \sum_{k=1}^{p} \left\{ (1 - \hat{\xi}_{ik}) |\beta_{ik}| + \hat{\xi}_{ik} |\beta_{ik} - \hat{\gamma}_k^{(m-1)}| \right\},$$

   where $L(\cdot)$ is the quadratic loss function, and $\hat{\xi}_{ik} = \mathbf{1}(|\hat{\beta}_{ik}^{(m-1)}| > |\hat{\beta}_{ik}^{(m-1)} - \hat{\gamma}_k^{(m-1)}|)$
4. (Grouping) Update $\hat{\boldsymbol{\gamma}}^{(m)}$ by minimizing $\sum_{k=1}^{p} \sum_{i=1}^{N} \min(|\hat{\beta}_{ik}^{(m)}|, |\hat{\beta}_{ik}^{(m)} - \gamma_k|)$
5. Iterate Step 3 and Step 4 until $\| \hat{\boldsymbol{\beta}}^{(m)} - \hat{\boldsymbol{\beta}}^{(m-1)} \|_2 + \| \hat{\boldsymbol{\alpha}}^{(m)} - \hat{\boldsymbol{\alpha}}^{(m-1)} \|_2 < \epsilon$, ϵ is a pre-specified small value

Properties of Proposed Algorithm

- **Algorithm Convergence**
  - Due to non-convex optimization, the iterative estimators **converge** to a local minimizer
- Step 3 (regression) is a Lasso-type convex optimization problem
- Step 4 (grouping): K-means algorithm (one group center is 0)
- Tuning parameter $\lambda_{N,m}$: generalized cross-validation (GCV)

Theory of Double-divergence Correlated Model

- **Two sample sizes**
  - Number of individuals N: population information
  - Number of repeated measurements m: individual information
- **Theoretical Challenges**
  - Multi-directional separation penalty
  - Both N and m could go to infinity
  - Diverging number of parameters $p_\theta = Np + q$
  - Within-subject dependence

Population-wise Oracle Estimators

- Subgroup membership $\mathcal{G}_k = \{1 \le i \le N : \beta_{ik} = \gamma_k \neq 0\}$ is known
- True signal set $(\mathcal{A}_i)$ of any individual is known, e.g.,

$$\begin{aligned}\boldsymbol{\beta}_1^{or} &= (\gamma_1, 0, \gamma_3, \gamma_4, 0, \ldots, 0)', \\ \boldsymbol{\beta}_2^{or} &= (0, \gamma_2, \gamma_3, 0, \gamma_5, \ldots, 0)', \\ &\ldots\end{aligned}$$

- The total sample size is $\sum_{i=1}^{N} m_i = mN$
- $|\mathcal{G}_k|$ denotes the subgroup size of non-zero-effect individuals, $k = 1, \ldots, p$
- $N_k = \sum_{i \in \mathcal{G}_k} m_i = m|\mathcal{G}_k|$: total information contributing to $\hat{\gamma}_k$

Subgroup Effects on Oracle Estimators' Convergence

Theorem 1

Suppose $\eta_m = \lambda_{max}(\boldsymbol{R}_m^{-1}\boldsymbol{R}_m^0) \leq C_1$ *uniformly holds for some constant* C_1*, under regularity conditions, if either (i)* $m \to \infty$ *or (ii)* $\min(|\mathcal{G}_k|) \to \infty$*, we have*

$$\eta_m^{-1/2}\| M_{Nm}^{1/2}\left(\{(\hat{\gamma}^{or})^T, (\hat{\boldsymbol{\alpha}}^{or})^T\}^T - \{(\gamma^0)^T, (\boldsymbol{\alpha}^0)^T\}^T\right)\| \leq O_p(1)$$

where $M_{Nm} = diag(\underbrace{N_1, \ldots, N_p}_{p}, \underbrace{mN, \ldots, mN}_{q})$.

- $\boldsymbol{R}_m$ is the working correlation matrix, $\boldsymbol{R}_m^0$ is the true correlation matrix
- Convergence rate of $\hat{\gamma}_k$ is $\eta_m^{-1/2}\sqrt{N_k} = \eta_m^{-1/2}\sqrt{m|\mathcal{G}_k|}$
- Both subgroup size $|\mathcal{G}_k|$ and repeated measurement size m contribute
- **Faster convergence rate** than any subject-wise estimator ($\eta_m^{-1/2}\sqrt{m}$)

Oracle property of the Diverging Proposed Estimator

Theorem 2

Let $\tau_m = \eta_m^{-1}\lambda_{min}(\boldsymbol{D}_{N,m})$, *where* $\eta_m = \lambda_{max}(\boldsymbol{R}_m^{-1}\boldsymbol{R}_m^0)$, $\boldsymbol{\theta}_{(Np+q)\times 1} = (\boldsymbol{\beta}^T, \boldsymbol{\alpha}^T)^T$. *Under regularity conditions, if* $\frac{\lambda_{N,m}}{\tau_m} \to 0$, $\frac{\lambda_{N,m}}{\sqrt{\tau_m}} \to \infty$, *and* $log(N) = o(\tau_m)$, *as* $\tau_m \to \infty$, $N(\tau_m) \to \infty$, *we have*

Oracle property: $P(\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}^{or}) \to 1$.

- $\boldsymbol{D}_{N,m}(\boldsymbol{\theta}) = \sum_{i=1}^{N} \left(\frac{\partial \boldsymbol{\mu}_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T}\right)^T \boldsymbol{V}_i^{-1} \left(\frac{\partial \boldsymbol{\mu}_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T}\right)$
- If $\lambda_{max}(\boldsymbol{R}_m^0)$ is bounded (e.g., independent model), then $\tau_m = O(m)$

Simulation I: Simple Subject-wise Model

Heterogeneous model with one individualized covariate:

$$y_{ij} = \alpha_0 + \alpha_1 z_{ij1} + \alpha_2 z_{ij2} + \beta_i x_{ij} + \varepsilon_{ij}, \quad i = 1, \ldots, N, \quad j = 1, \ldots, m.$$

- Set $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_N)' = (\underbrace{\gamma, \ldots, \gamma}_{N/2}, \underbrace{0, \ldots, 0}_{N/2})'$, where $\gamma = 1$ or 2
- population parameters:$\boldsymbol{\alpha}' = (\alpha_0, \alpha_1, \alpha_2) = (1, 1, 1)$
- z_{ij1}, z_{ij2}, $x_{ij} \sim N(0, 1)$; independent $\varepsilon_{ij} \sim N(0, 1)$
- Sample size $N = 40$ or 100, cluster size $m = 10, 20$

Evaluation of Performance

- **Model estimation efficiency**
  - Root mean squared error (RMSE): $N^{-\frac{1}{2}} \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\|_2$
  - RMSE is equivalent to standard prediction error
- **Variable selection accuracy** (over individuals)
  - Correct variable selection rate (CVSR)
  - Sensitivity: true positive rate $P(\hat{\beta}_i \neq 0 | \beta_i \neq 0)$
  - Specificity: true negative rate $P(\hat{\beta}_i = 0 | \beta_i = 0)$

Comparing Methods

- Subject-wise least squares estimator
- Homogeneous least squares estimator
- Subject-wise penalized estimator
  - Lasso, adaptive Lasso, SCAD and MCP
- Fused Lasso estimator

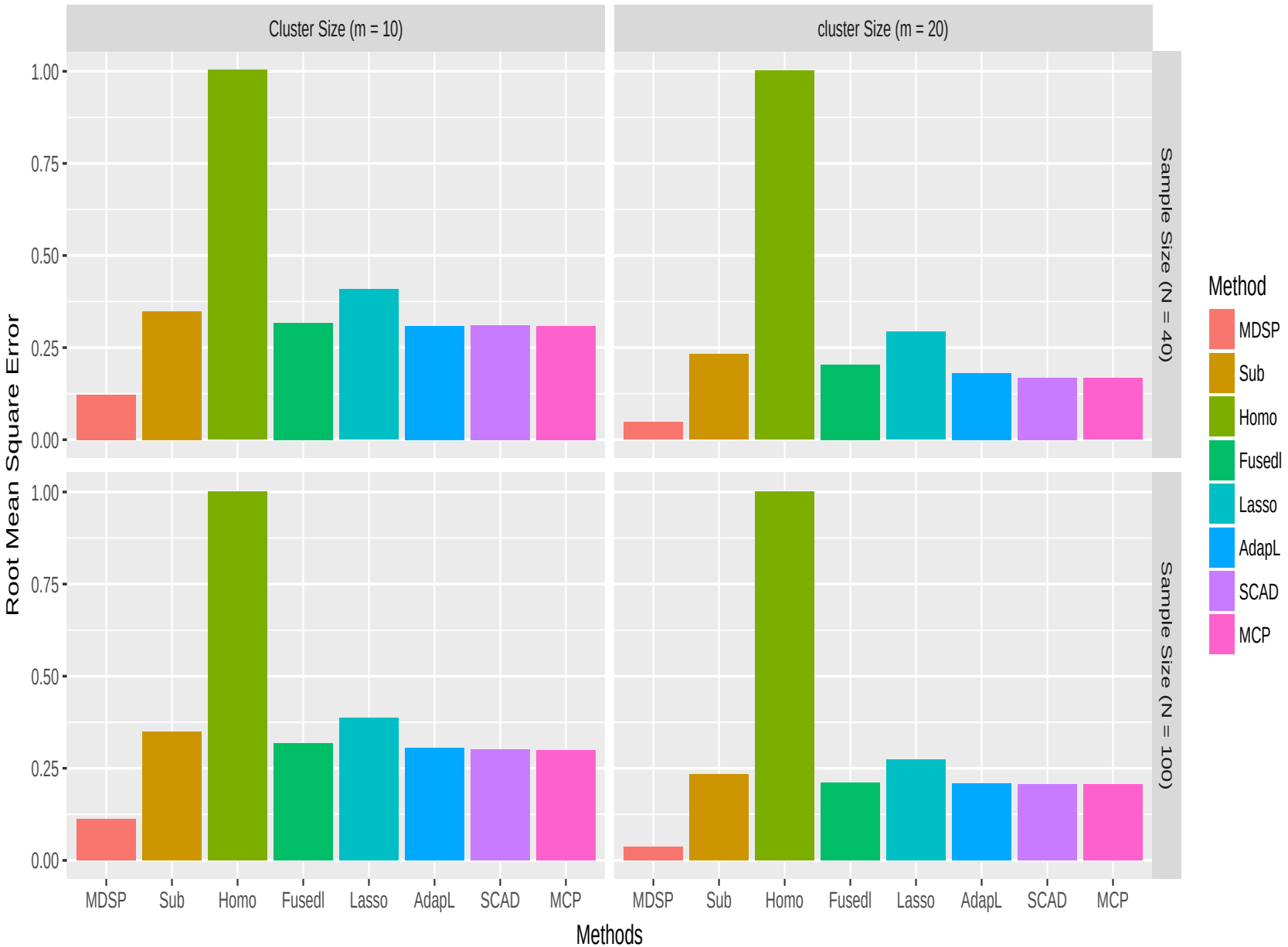
Select Number of Subgroups

- S_k: number of non-zero-effect subgroups for kth individualized covariate
- Modified Bayesian Information Criterion (Wang et al., 2007)

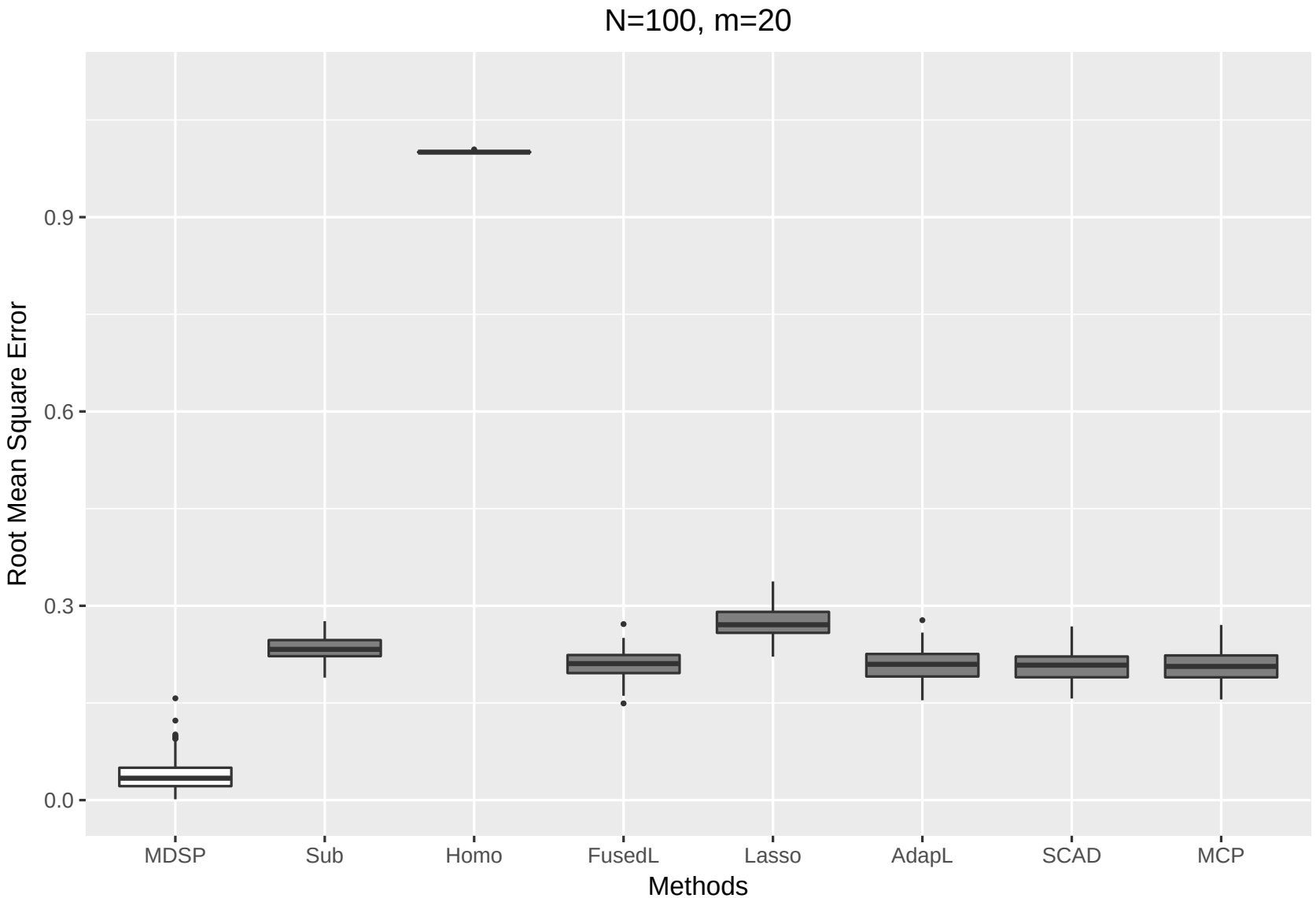
$$\text{BIC}(S_k) = \log\left(\sum_{i=1}^{N}\sum_{j=1}^{m}(y_{ij} - \hat{\mu}_{ij}(S_k))^2/mN\right) + b_{N,m}\frac{\log(mN)}{mN}(S_k + q)$$

- $b_{N,m} = \log(\log(Np + q))$

Simulation I: Average RMSE of 200 replications (Plot)



Simulation I: RMSE of 200 replications (Boxplot)

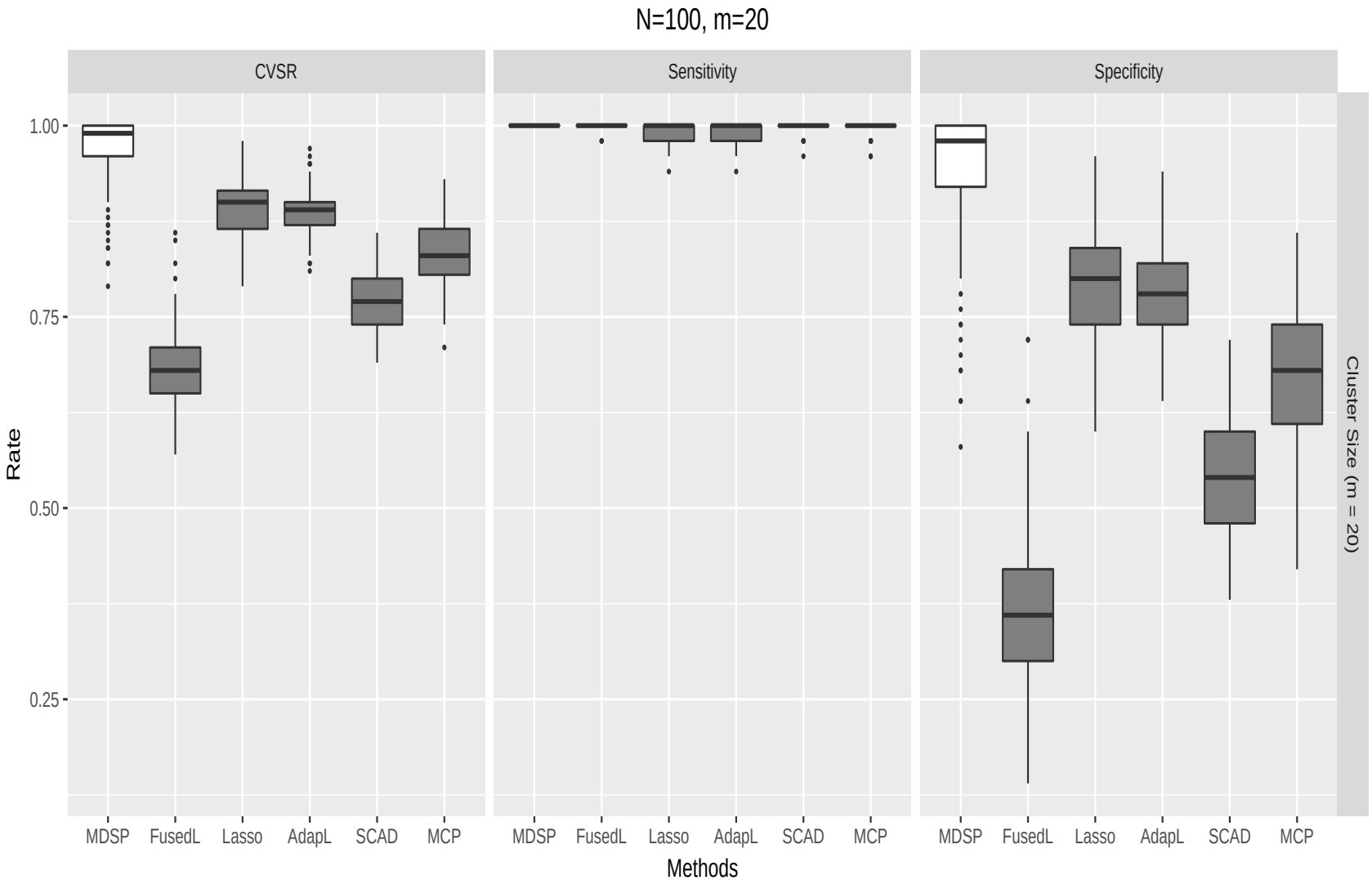


Simulation I: Average RMSE of 200 replications (Table)

Table 1: The average root mean square error (RMSE) of the proposed MDSP model compared with other approaches based on 100 simulations, with sample size $N = 40, 100$, cluster size $m = 10, 20$, and subgroup homogeneous effect $\gamma = 2$.

Sample Size (N)	Cluster Size(m)	Methods							
		MDSP	Sub	Homo	FusedL	Lasso	AdapL	SCAD	MCP
40	10	**0.122**	0.349	1.004	0.317	0.408	0.309	0.311	0.309
	20	**0.048**	0.232	1.002	0.204	0.293	0.181	0.168	0.167
100	10	**0.113**	0.350	1.001	0.318	0.387	0.305	0.300	0.299
	20	**0.037**	0.233	1.001	0.210	0.274	0.208	0.206	0.206

Simulation I: Variable Selection (Boxplot)



Simulation I: Variable Selection (Table)

Table 2: The average correct variable selection rate (CVSR), sensitivity and specificity of the proposed MDSP model compared with other approaches based on 100 simulations, with sample size $N = 40, 100$, cluster size $m = 10, 20$, and subgroup homogeneous effect $\gamma = 2$.

Variable Selection	Sample Size (N)	Cluster Size(m)	Methods **MDSP**	FusedL	Lasso	AdapL	SCAD	MCP
CVSR	40	10	0.959	0.639	0.886	0.884	0.800	0.852
		20	0.972	0.670	0.928	0.940	0.908	0.953
	100	10	0.940	0.648	0.868	0.898	0.809	0.871
		20	0.965	0.682	0.890	0.888	0.773	0.832
Sensitivity	40	10	0.997	0.996	0.997	0.998	1.000	0.998
		20	1.000	1.000	1.000	1.000	1.000	1.000
	100	10	0.998	0.997	0.998	0.998	0.999	0.999
		20	1.000	0.999	0.993	0.994	0.999	0.997
Specificity	40	10	0.922	0.282	0.774	0.771	0.602	0.705
		20	0.945	0.340	0.856	0.880	0.816	0.906
	100	10	0.882	0.299	0.738	0.797	0.620	0.744
		20	0.930	0.365	0.787	0.782	0.546	0.668

Heterogeneous model setting:

$$y_{ij} = \alpha_0 + \alpha_1 z_{ij1} + \alpha_2 z_{ij2} + \beta_{i1} x_{i1} + \beta_{i2} x_{ij2} + \varepsilon_{ij}, \quad i = 1, ..., N, \quad j = 1, ..., m$$

- Set $(\alpha_0, \alpha_1, \alpha_2) = (1, 1, 1)$; generate $\boldsymbol{z_1}$ and $\boldsymbol{z_2}$ from $N(0, 1)$
- **Within-subject invariant** covariates: $x_{i1} = -1$ or 1, coefficients β_{i1} to be either -2, 2 or 0, balanced subgroup size
- **Within-subject varying** covariates: $x_{ij2} \sim N(0, 1)$, coefficients β_{i2} to be either -1, 1 or 0, balanced subgroup size
- Sample size $N = 60$, repeated measurement size $m = 2, 5, 10$

Simulation II: Modeling Performance

Table 3: The correct variable selection rates for the proposed separation-penalty approach (β_{i1}^{MDSP}, β_{i2}^{MDSP}) and the L_1-penalized model ($\beta_{i1}^{L_1}$, $\beta_{i2}^{L_1}$), and the RMSE for two approaches.

Repeated size	Correct variable selection rate				RMSE	
	$\boldsymbol{\beta_{i1}^{MDSP}}$	$\beta_{i1}^{L_1}$	$\boldsymbol{\beta_{i2}^{MDSP}}$	$\beta_{i2}^{L_1}$	**MDSP**	L_1
$m = 2$	0.93	0.60	0.80	0.53	**0.56**	1.01
$m = 5$	0.94	0.66	0.83	0.60	**0.37**	0.64
$m = 10$	0.95	0.71	0.84	0.64	**0.18**	0.45

- performance of other subject-wise penalized models are similar to the Lasso model

Simulation II: Estimates of Individualized Coefficients

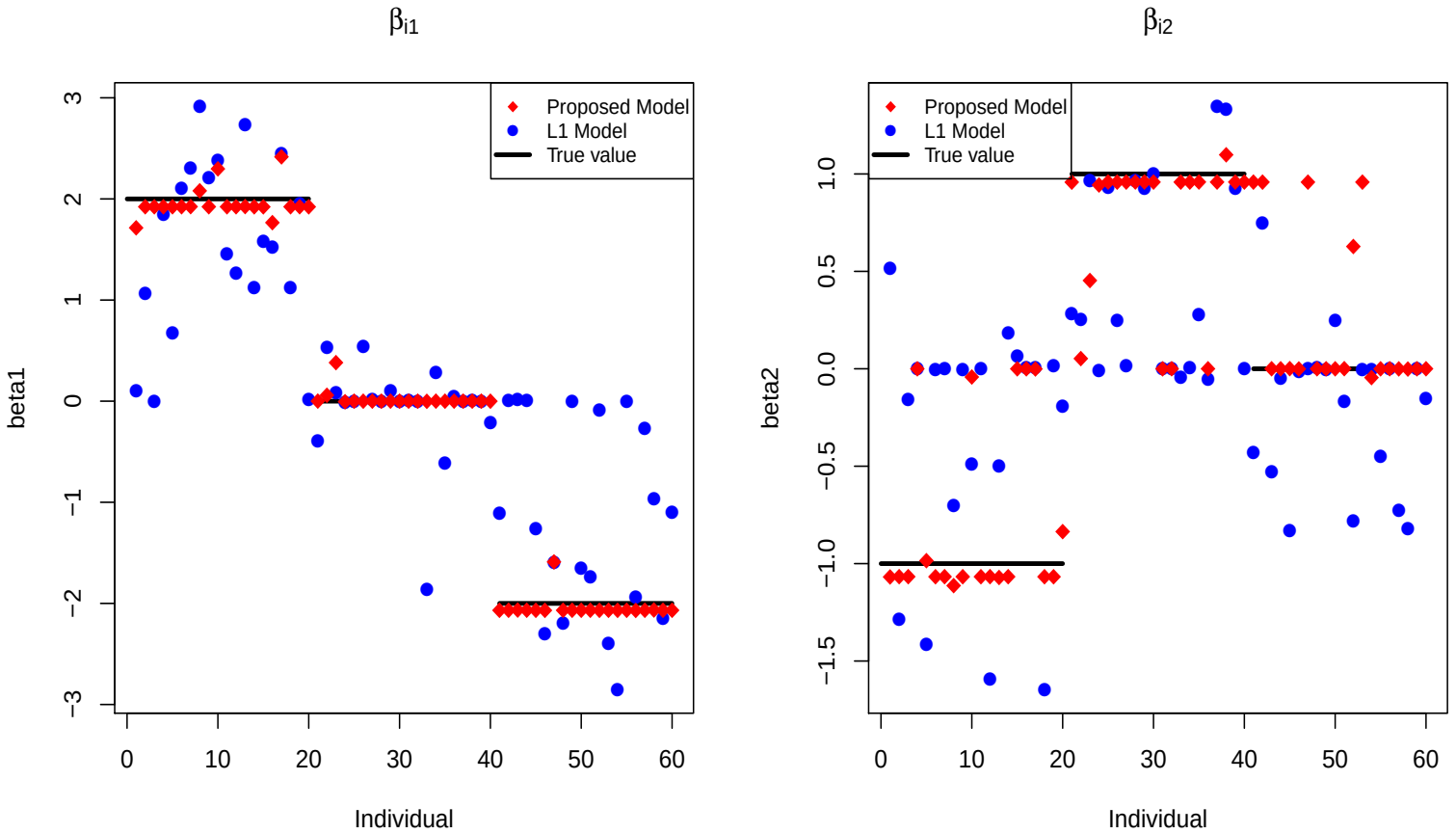


Figure 2: The estimated personalized coefficients of β_{i1}, β_{i2} for individuals when the repeated measurement size $m = 10$.

HIV Data: An Individualized Model

- Individualized effect for treatment group:

$$y_{it} = \beta_0 + \beta_t * Time + \beta_z * ZDV + \boldsymbol{\beta_{izt}} * ZDV * Time + \beta_a * Age + \beta_g * Gender + \varepsilon_{it}$$

- In control group (ZDV=0), $\boldsymbol{\beta_{izt}}$ is set to be zero
- Assuming there are three subgroups: 0, $\gamma^+ > 0$ or $\gamma^- < 0$
- Training set: time $t = 1, \ldots, 12$; Testing set: time $t = 13, 14$

Compared Methods

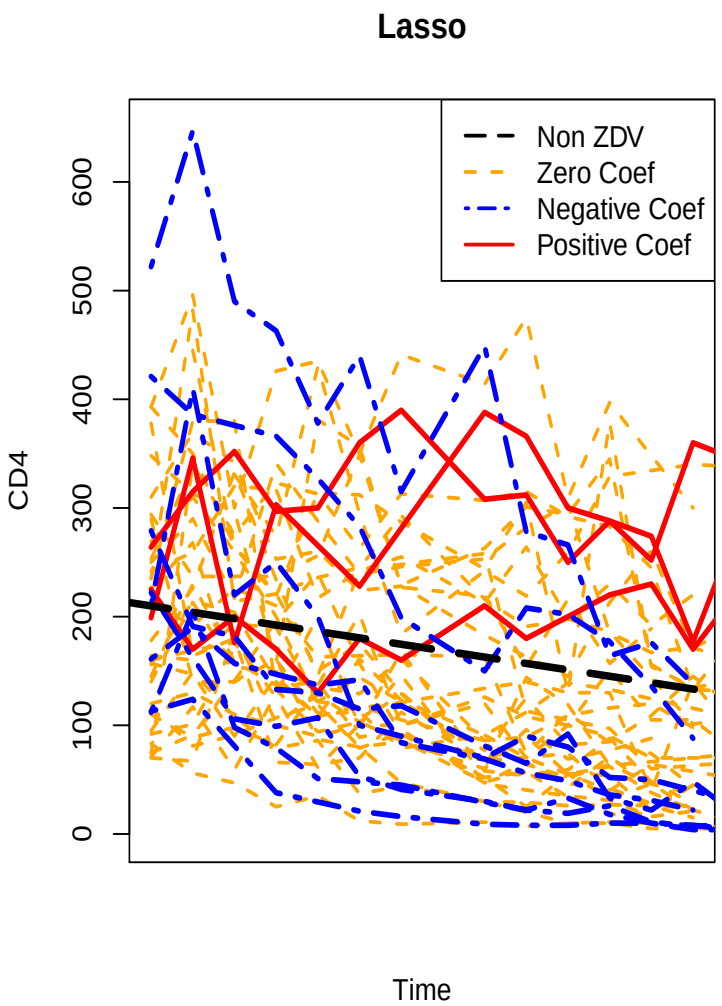
- Homogeneous Model: assume $\beta_{izt} = \beta_{zt}$
- Random-effects model: random effects on slope β_{zt}
- Subject-wise penalized (Lasso) model: $\sum_i^N \lambda |\beta_{izt}|$
- Fused Lasso model: $\lambda \sum_{i \neq j} |\beta_{izt} - \beta_{jzt}|$
- **Evaluation**: median prediction errors (MPE) on testing set

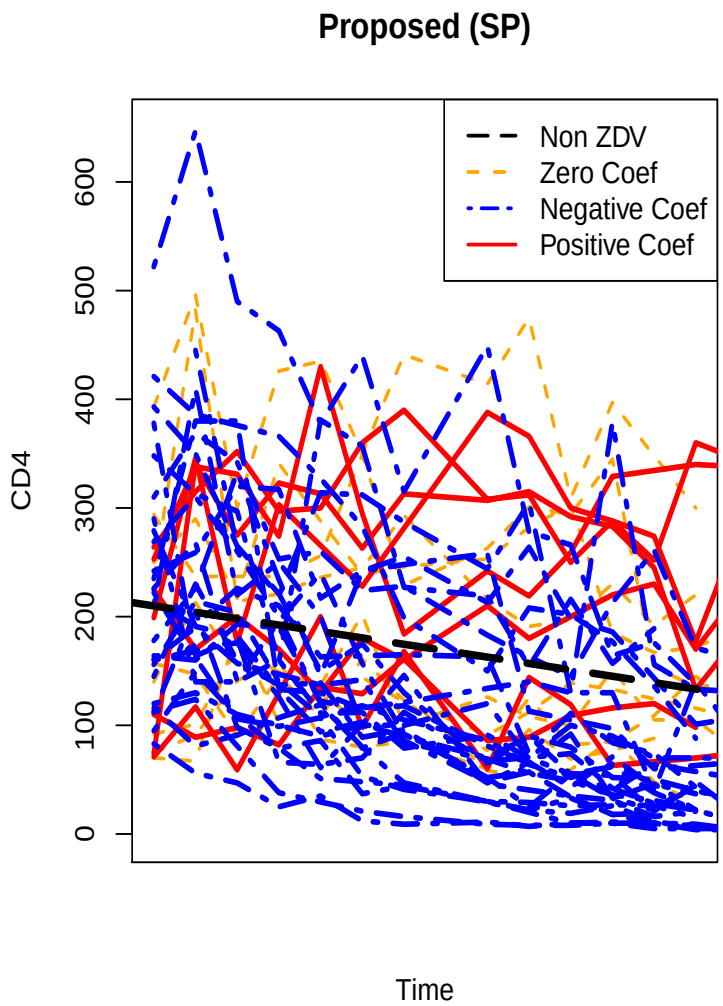
Model Estimation and Prediction

Table 4: The estimated coefficients of the population model, the random-effects model, the L_1-penalty model and the proposed model with corresponding median prediction errors (MPE) for the ACTG data. The individualized coefficient estimators $\hat{\beta}_{izt}$'s in the Lasso model, the fused Lasso (fusedL) model and the proposed (MDSP) model are not listed.

Model	$\hat{\beta}_0$	$\hat{\beta}_t$	$\hat{\beta}_z$	$\hat{\beta}_a$	$\hat{\beta}_g$	$\hat{\beta}_{zt}$	MPE
Population	3.09	−0.68	−0.54	0.01	−0.01	−0.24	1.67
Random-effects	2.56	−0.68	−0.57	0.02	−0.01	−0.29	1.70
Lasso	3.09	−0.76	−0.54	0.01	−0.01	-	1.64
fusedL	3.05	−0.72	−0.52	0.01	−0.01	-	1.62
MDSP	3.10	−0.68	−0.56	0.01	−0.01	-	**1.44**

Subpopulations for ZDV Treatment Group β_{izt}





Post-Subgrouping Marginal Model

Table 5: The treatment effect estimators within each subgroup model (zero-effect group: β_{zt}^0, negative-effect group: β_{zt}^- and positive-effect group β_{zt}^+) as well as the standard errors (s.e.) and the p-values. Each subgroup consists of the corresponding individuals in the treatment group identified by the Lasso model or the proposed model (MDSP) as well as all the individuals in the control group. The proportion of individuals with the treatment classified into each subgroup is provided.

Model		Estimates	s.e.	p-value	Proportion
Lasso	$\hat{\beta}_{zt}^0$	−0.24	0.17	0.14	0.75
	$\hat{\beta}_{zt}^-$	−0.73	0.31	0.02	0.18
	$\hat{\beta}_{zt}^+$	0.82	0.48	0.10	0.07
MDSP	$\hat{\beta}_{zt}^0$	−0.04	0.30	0.89	0.20
	$\hat{\beta}_{zt}^-$	−0.68	0.08	**0.00**	0.64
	$\hat{\beta}_{zt}^+$	0.72	0.33	**0.02**	0.16

- Combine the identified subgroup with the control group and fit a marginal regression model

Concluding Remarks

- **Individualized** regression model with **subpopulation** structure
- **Multi-directional Separation penalty**: Center-based integration providing **multiple shrinking directions**
- Theoretical properties: **population-wise oracle property**
- Incorporate **cross-subjects** information to improve individual model's estimation and prediction

The End

Thank You!

References

- Tang, X. and Qu, A. (2017). Individualized Multi-directional Variable Selection. Submitted.
- Bondell, H. D. and Reich, B. J. (2008) Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with oscar. *Biometrics 64*, 115-123.
- Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association 96*, 1348-1360.
- Hocking, T., Joulin, A., Bach, F. and Vert, J.-P. (2011). Clusterpath: An Algorithm for Clustering using Convex Fusion Penalties. *In L. Getoor and T. Scheffer (Eds.), Proceedings of the 28th International Conference on Machine Learning (ICML'11)*, 745-752.
- Jacobs, R. A., Jordan, M. I., Nowlan, S. J., and Hinton, G. E. (1991). Adaptive mixtures of local experts. *Neural Comp. 3*, 79-87.
- Ke, T., Fan, J. and Wu, Y. (2010). Homogeneity in regression. *Journal of the American Statistical Association 110*, 175-194.
- Lindsten, F., Ohlsson, H. and Ljung, L. (2011). Clustering using sum-of-norms regularization: With application to particle filter output computation. *2011 IEEE Statistical Signal Processing Workshop (SSP)*, 201-204.
- Li, G., Wang, S., Huang, Ch., Yu, M. and Shao, J. (2014) Meta-Analysis Based Variable Selection for Gene Expression Data. *Biometrics 70*, 872-880.
- Pan, W., Shen, X. and Liu, B. (2013). Cluster analysis: unsupervised learning via supervised learning with a non-convex penalty. *Journal of Machine Learning Research 14*, 1865–1889.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Ser. B 58*, 267-288.
- Tibshirani, S., Saunders, M., Rosset, S., Zhu, J. and Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of Royal Statistical Society: Ser. B 67*, 91-108.
- Zhang, C. H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics 38*, 894-942.
- Zou, H. (2006). The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association*, 101, 1418-1429.

Regularity Conditions

(A1) The unknown parameter $\boldsymbol{\theta} = (\boldsymbol{\alpha}', \boldsymbol{\beta}')'$ belongs to a compact subset $\mathcal{B} \subseteq \mathbf{R}^{p_{\boldsymbol{\theta}}}$ and its true value $\boldsymbol{\theta}^0 = ((\boldsymbol{\alpha}^0)', (\boldsymbol{\beta}^0)')'$ lies in the interior of $\mathcal{B}$;

(A2) Random error ε_{ij} has mean 0 and variance $\sigma^2 < \infty$, $\boldsymbol{\varepsilon_i}$ is a sub-Gaussian vector

(A3) There exist $\nu_l > 0, \nu_l' > 0$, such that $\lambda_{min}(\boldsymbol{R}_i^0) > \nu_l$ and $\lambda_{min}(\boldsymbol{R}_i) > \nu_l'$ for all i and m.

(A4) $\tilde{\boldsymbol{X}}_{ij} = (\boldsymbol{Z}_{ij}', \boldsymbol{X}_{ij}')'_{(q+p)\times 1}$ belongs to a compact set $\mathcal{X} \subset \mathbf{R}^{q+p}$ for $1 \leq i \leq N$ and $1 \leq j \leq m$;

(A5) Let $\tilde{\boldsymbol{X}}_{i,k}$ denote the kth column of $\tilde{\boldsymbol{X}}_i$, assume $\|\tilde{\boldsymbol{X}}_{i\cdot,k}\|_2^2 = O_p(m)$ and $\sum_{i=1}^N m^{-1}\|\tilde{\boldsymbol{X}}_{i\cdot,k}\|_2^2 = O_p(N)$, for $1 \leq k \leq q+p$;

(A6) $m^{-1}\lambda_{min}(\boldsymbol{X}_i^T \boldsymbol{X}_i) > c_3$ for any i and $(mN)^{-1}\lambda_{min}\left(\sum_{i=1}^N \boldsymbol{Z}_i^T(\boldsymbol{I}_m - \boldsymbol{H}_{\boldsymbol{X_i}})\boldsymbol{Z}_i\right) > c_4$, where $\boldsymbol{H}_{\boldsymbol{X_i}} = \boldsymbol{X}_i(\boldsymbol{X}_i^T\boldsymbol{X}_i)^{-1}\boldsymbol{X}_i^T$, for some constants $0 < c_3 < \infty$, $0 < c_4 < \infty$.

Some Corollaries of Oracle Property

- **Uniform model selection consistency**: $P(\cap_{i=1}^{N}\{\hat{\mathcal{A}}_i = \mathcal{A}_i\}) \to 1$
- **Group identification consistency**: $P(\cap_{k=1}^{p}\{\hat{\mathcal{G}}_k = \mathcal{G}_k\}) \to 1$
- **Population-wise optimal efficiency**: $\eta_m^{-1/2}\sqrt{N_k} = \eta_m^{-1/2}\sqrt{m|\mathcal{G}_k|}$

Simulation I: Select Number of Subgroups

Table 6: The mean of identified subgroup numbers of the proposed model compared with the two-stage OLSK method based on 100 simulations, with sample size $N = 60, 120$, cluster size $m = 5, 10, 20$. The first three scenarios contain one individualized predictor ($p = 1$) of one, two and three groups, respectively. The last scenario contain two individualized predictors ($p = 2$), one with two groups and the other with three groups. The subgroup sizes are equal in each scenario. The subgroup homogeneous effects are listed as possible values for β_i in the table.

Number of individualized variables		$p = 1$						$p = 2$			
Sample Size (N)	Cluster Size(m)	$\beta_i = 0$ MDSP	$\beta_i = 0$ OLSK	$\beta_i = 0, 1$ MDSP	$\beta_i = 0, 1$ OLSK	$\beta_i = 0, 2, 5$ MDSP	$\beta_i = 0, 2, 5$ OLSK	$\beta_{1i} = 0, 2$ MDSP	$\beta_{1i} = 0, 2$ OLSK	$\beta_{2i} = -2, 0, 1$ MDSP	$\beta_{2i} = -2, 0, 1$ OLSK
60	5	1.0(100)	1.0(100)	2.0(95)	1.0(2)	2.9(88)	2.5(68)	2.0(100)	1.5(52)	3.2(85)	1.2(0)
	10	1.0(100)	1.0(100)	2.0(100)	1.3(26)	3.1(90)	2.7(74)	2.0(100)	2.0(100)	3.1(90)	2.4(44)
	20	1.0(100)	1.0(100)	2.0(100)	2.0(100)	3.1(92)	2.8(78)	2.0(100)	2.0(100)	3.0(100)	2.8(80)
120	5	1.0(100)	1.0(100)	2.0(96)	1.0(2)	3.2(86)	2.8(82)	2.0(100)	1.7(72)	3.1(90)	1.4(0)
	10	1.0(100)	1.0(100)	2.0(100)	1.2(24)	3.1(92)	2.9(86)	2.0(100)	2.0(100)	3.1(90)	2.6(64)
	20	1.0(100)	1.0(100)	2.0(100)	2.0(100)	3.0(98)	2.9(96)	2.0(100)	2.0(100)	3.1(92)	2.78(78)

Simulation II: Robustness

Table 7: The average RMSE and CVSR of the proposed MDSP model compared to the subject-wise model (Sub), the fused Lasso (FusedL), the Lasso, the adaptive Lasso (AdapI), the SCAD and the MCP penalization models, based on 100 simulations with sample size $N = 60$ and cluster size $m = 10$. The first case contains a population homogeneous effect ($K_{true} = 1$) and the second case contains an individualized predictor of three subgroups ($K_{true} = 3$) with equal subgroup size. In both cases the proposed model assumes two subgroups. The estimated subgroup homogeneous effects from the proposed model are $\hat{\gamma} = 2.01(0.06)$ and $\hat{\gamma} = -2.99(0.06)$ in these two cases (with empirical standard errors in parenthesis), respectively.

Case		**MDSP**	Sub	FusedL	Lasso	AdapL	SCAD	MCP
$K_{true} = 1$	RMSE	0.115	0.346	0.319	0.414	0.373	0.346	0.345
$(\beta_i = 2)$	CVSR	0.996	-	0.993	0.994	0.992	0.995	0.996
$K_{true} = 3$	RMSE	0.277	0.349	0.315	0.410	0.335	0.337	0.338
$(\beta_i = -3, 0, 1)$	CVSR	0.901	-	0.748	0.877	0.902	0.816	0.817

Simulation II: Robustness