社会科学における LASSO の応用について

担当：金髙

Applications of LASSO in Social Sciences

- 社会科学 (e.g. 経済学、政治学、心理学など) の分野においても LASSO は広く利用されている

Applications of LASSO in Social Sciences

- 社会科学 (e.g. 経済学、政治学、心理学など) の分野においても LASSO は広く利用されている
  - 予測・分類に焦点を置いたもの
    - 政治文書の分類 (Widmann and Wich 2022)、Swing voters の予測 (Hare and Kutsuris 2022)、紛争予測 (Mueller and Rauh 2018) など

Applications of LASSO in Social Sciences

- 社会科学 (e.g. 経済学、政治学、心理学など) の分野においても LASSO は広く利用されている
  - 予測・分類に焦点を置いたもの
    - 政治文書の分類 (Widmann and Wich 2022)、Swing voters の予測 (Hare and Kutsuris 2022)、紛争予測 (Mueller and Rauh 2018) など
  - 変数選択 (+ 因果推論)
    - 民族的境界線と分極化 (Bazzi and Gudgeon 2021)、道徳的価値と投票 (Enke 2020)、映画消費におけるスピルオーバーエフェクト (Gilchrist and Sands 2016)、父親の育児休暇が子のジェンダーバイアスに及ぼす影響 (Tavits et al. 2023)、北宋における政治リーダーの血縁ネットワークの分析 (Wang 2022) など
    - Heterogeneous treatment effect の推定: Blackwell and Olson (2022), De la Cuesta, Egami, and Imai (2022), Grimmer, Messing, and Westwood (2017), and Imai and Ratkovic (2013) など

Applications of LASSO in Social Sciences

- 社会科学 (e.g. 経済学、政治学、心理学など) の分野においても LASSO は広く利用されている
  - 予測・分類に焦点を置いたもの
    - 政治文書の分類 (Widmann and Wich 2022)、Swing voters の予測 (Hare and Kutsuris 2022)、紛争予測 (Mueller and Rauh 2018) など
  - 変数選択 (+ 因果推論)
    - 民族的境界線と分極化 (Bazzi and Gudgeon 2021)、道徳的価値と投票 (Enke 2020)、映画消費におけるスピルオーバーエフェクト (Gilchrist and Sands 2016)、父親の育児休暇が子のジェンダーバイアスに及ぼす影響 (Tavits et al. 2023)、北宋における政治リーダーの血縁ネットワークの分析 (Wang 2022) など
    - Heterogeneous treatment effect の推定: Blackwell and Olson (2022), De la Cuesta, Egami, and Imai (2022), Grimmer, Messing, and Westwood (2017), and Imai and Ratkovic (2013) など

Our (main) quantity of interest

- 社会科学の主な quantity of interest は「予測・分類」よりも「効果」にある
  - 計量経済学や因果推論
  - どうすれば因果効果をバイアスなく推定できるのか
  - e.g. 最低賃金上昇と雇用増加 (Card and Krueger 1994)

左から：David Card (UC Berkeley), Joshua Angrist (MIT), Guido Imbens (Stanford). Source: https://economictimes.indiatimes.com/news/international/world-news/economists-card-angrist-and-imbens-win-2021-nobel-prize/articleshow/86936176.cms?from=mdr

Why LASSO?

- なぜ LASSO を使うのか？

Why LASSO?

- なぜ LASSO を使うのか？
  - 社会科学の観察データでは、少ないサンプルサイズに対しての共変量が多い場合がよくある

Why LASSO?

- なぜ LASSO を使うのか？
  - 社会科学の観察データでは、少ないサンプルサイズに対しての共変量が多い場合がよくある
    - 変数選択の必要性
    - 不要な変数を特定し、モデルの誤特定を防ぎたい
    - そもそも $p \gg n$ のような状況では OLS できない

Why LASSO?

- なぜ LASSO を使うのか？
  - 社会科学の観察データでは、少ないサンプルサイズに対しての共変量が多い場合がよくある
    - 変数選択の必要性
    - 不要な変数を特定し、モデルの誤特定を防ぎたい
    - そもそも $p \gg n$ のような状況では OLS できない
  - しかしながら、間違えて割と重要な変数を落としてしまう可能性も (欠落変数バイアス)

Why LASSO?

- なぜ LASSO を使うのか？
  - 社会科学の観察データでは、少ないサンプルサイズに対しての共変量が多い場合がよくある
    - 変数選択の必要性
    - 不要な変数を特定し、モデルの誤特定を防ぎたい
    - そもそも $p \gg n$ のような状況では OLS できない
  - しかしながら、間違えて割と重要な変数を落としてしまう可能性も (欠落変数バイアス)
- このような状況の中で処置効果を推定するには ?????

Why LASSO?

- なぜ LASSO を使うのか？
  - 社会科学の観察データでは、少ないサンプルサイズに対しての共変量が多い場合がよくある
    - 変数選択の必要性
    - 不要な変数を特定し、モデルの誤特定を防ぎたい
    - そもそも $p \gg n$ のような状況では OLS できない
  - しかしながら、間違えて割と重要な変数を落としてしまう可能性も (欠落変数バイアス)
- このような状況の中で処置効果を推定するには ?????

⇝ **Double LASSO**

Double LASSO

- **Double LASSO** by Belloni, Chernozhukov, and Hansen (2014)
  - 以下のような誘導形を考える ($i = 1, ..., N$)

$$y_i = \tau d_i + x_i'\beta + \varepsilon_i$$
$$d_i = x_i'\gamma + v_i$$

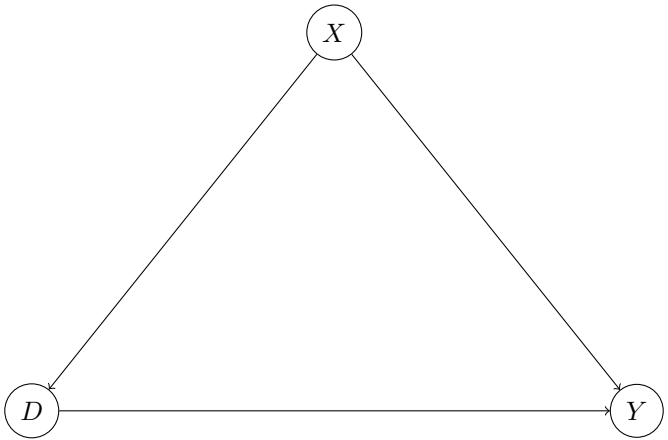
  - y_i: アウトカム
  - d_i: 処置変数
  - x_i: 共変量 (p 次元、$p \gg n$)
  - ε_i, v_i: 誤差項
- このような状況下で、どのように共変量 x_i を選べばいいのか？
  - x_i の選択によって Omitted variable/misspecification bias が発生する

Double LASSO (cont)

- **Double LASSO** (Belloni, Chernozhukov, and Hansen 2014)
  1. $d_i \sim x_i$ で LASSO して、non-zero coefficient の共変量集合 I を取得
  2. $y_i \sim x_i$ で LASSO して、non-zero coefficient の共変量集合 II を取得
  3. $I \cup II$ の変数 + d_i を用い、OLS により処置効果 τ の推定値を得る

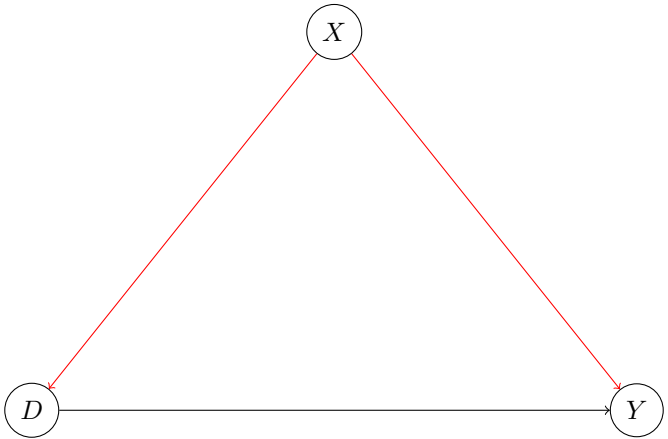
Double LASSO (cont)

- Omitted variable bias (OVB) と LASSO
  - 以下のような DAG を考える



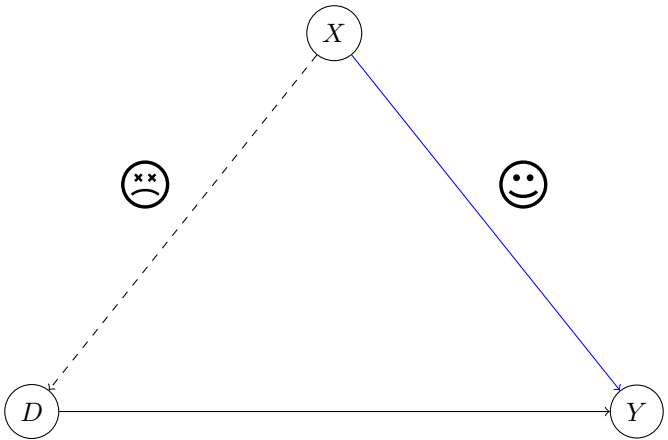
Double LASSO (cont)

- X は confounder
  - これを考慮しないモデルでは深刻な OVB が発生する



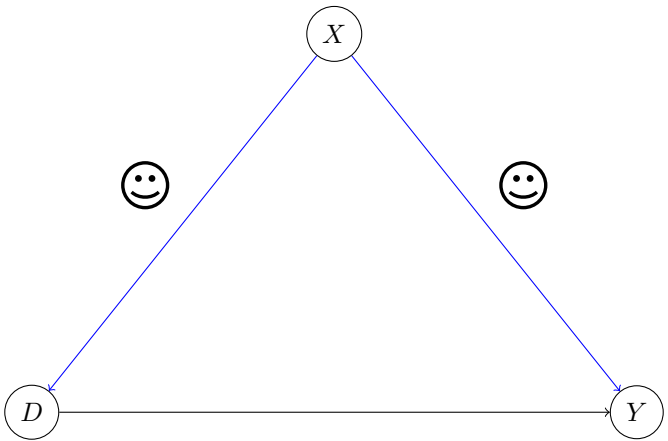
Double LASSO (cont)

- Single LASSO ($Y \sim X$ のみ) では 青い矢印 の変数選択しかできない
  - D と associated で、Y には微小に associated (e.g. $|\beta| \propto 1/\sqrt{n}$) な X を drop する可能性 $\rightarrow$ OVB



Double LASSO (cont)

- Double LASSO は両方の変数選択を行える
  - より欠落変数に対処しやすい



Double LASSO (cont)

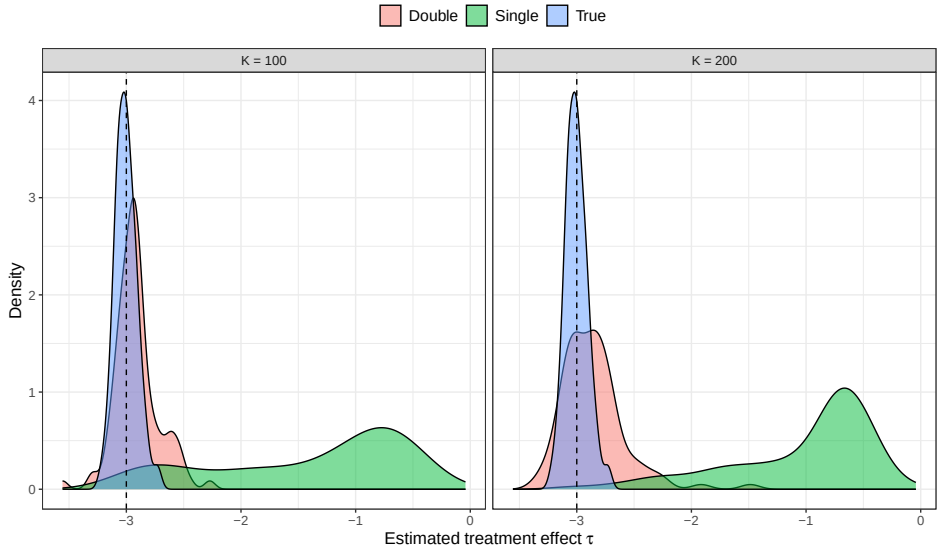
- Double LASSO の手順再掲
  1. $d_i \sim x_i$ で LASSO して、non-zero coefficient の共変量集合 I を取得
  2. $y_i \sim x_i$ で LASSO して、non-zero coefficient の共変量集合 II を取得
  3. $I \cup II$ の変数 + d_i を用い、OLS により処置効果 τ の推定値を得る

Monte Carlo Simulation of Double LASSO

- サンプル: $N = 100$; $K = \{100, 200\}$
- 処置効果: $\tau = -3$
- 共変量: $X \sim N(0, \Sigma)$
  - $\Sigma_{ij} = 0.5^{|i-j|}$
- 処置変数: $D = X\gamma + v$
  - $\gamma = [1, 1, 1, 0, \cdots 0]'$, $v \sim N(0, 1)$
- アウトカム: $Y = \tau D + X\beta + \varepsilon$
  - $\beta = [4, 2, 2, 4, 2, 0, \cdots, 0]'$, $\varepsilon \sim N(0, 1)$
- R で 100 回シミュレーションを行った

Simulation results of DL

- モデル間の比較 (シミュレーション回数 = 100)



Simulation results of DL (cont)

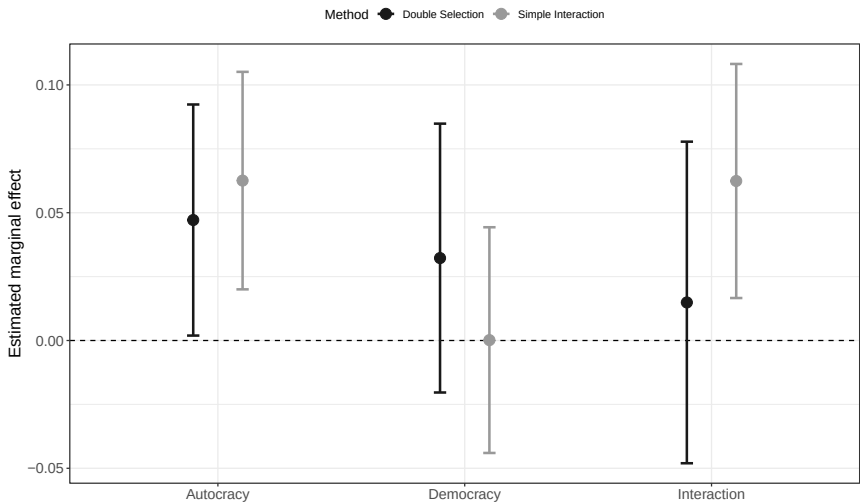
- 数値比較

	Mean	Bias	RMSE
K = 100			
Double	-2.92	0.15	0.21
Single	-1.40	1.60	1.79
K = 200			
Double	-2.86	0.21	0.31
Single	-1.02	1.98	2.07

Application I

- Blackwell and Olson (2022)
  - 交差項の選択手段としての DL を提案
    - 処置変数のみの交差項は misspecification bias を発生させることを指摘
    - 共変量にもそれぞれ交差項を作る必要があり、その変数選択のために DL を使う
  - Escribà-Folch, Meseguer, and Wright (2018): 移民労働者による本国への送金とそれが政治抗議に与える効果を検証
    - 送金の変数と「民主主義 or 権威主義」を捉えたレジーム変数で交差項を作成しレジームごとの異質効果を分析
  - Blackwell and Olson (2022) は DL を用いた交差項選択によって再検証を行った

Application I (cont)



注: 分析・データをもとに筆者作成。Source: Blackwell and Olson 2022

Application II: Demand estimation with rich covariates

- 経済学やマーケティング (ときどき政治学) では、「需要」に関する分析が盛んである

Application II: Demand estimation with rich covariates

- 経済学やマーケティング (ときどき政治学) では、「需要」に関する分析が盛んである
  - 経済・マーケティング
    - 自動車価格と需要 (Berry, Levinsohn, and Pakes 1995)
    - メディアの偏向に関する需要と供給 (Gentzkow and Shapiro 2010)
    - 学校選択 (Neilson et al. 2013)

Application II: Demand estimation with rich covariates

- 経済学やマーケティング (ときどき政治学) では、「需要」に関する分析が盛んである
  - 経済・マーケティング
    - 自動車価格と需要 (Berry, Levinsohn, and Pakes 1995)
    - メディアの偏向に関する需要と供給 (Gentzkow and Shapiro 2010)
    - 学校選択 (Neilson et al. 2013)
  - 政治学 (需要としての得票率)
    - 政治資金と得票率 (Gillen, et al. 2019)
    - 政治広告と大統領選挙 (Gordon and Hartmann 2013)
    - 政治家は市民の声を政策に反映するのか？(Iaryczower, Montero, and Kim 2022)

Application II: Demand estimation with rich covariates

- 経済学やマーケティング (ときどき政治学) では、「需要」に関する分析が盛んである
  - 経済・マーケティング
    - 自動車価格と需要 (Berry, Levinsohn, and Pakes 1995)
    - メディアの偏向に関する需要と供給 (Gentzkow and Shapiro 2010)
    - 学校選択 (Neilson et al. 2013)
  - 政治学 (需要としての得票率)
    - 政治資金と得票率 (Gillen, et al. 2019)
    - 政治広告と大統領選挙 (Gordon and Hartmann 2013)
    - 政治家は市民の声を政策に反映するのか？(Iaryczower, Montero, and Kim 2022)
- 時には大量の共変量候補があることもある

Application II: Demand estimation with rich covariates

- 経済学やマーケティング (ときどき政治学) では、「需要」に関する分析が盛んである
  - 経済・マーケティング
    - 自動車価格と需要 (Berry, Levinsohn, and Pakes 1995)
    - メディアの偏向に関する需要と供給 (Gentzkow and Shapiro 2010)
    - 学校選択 (Neilson et al. 2013)
  - 政治学 (需要としての得票率)
    - 政治資金と得票率 (Gillen, et al. 2019)
    - 政治広告と大統領選挙 (Gordon and Hartmann 2013)
    - 政治家は市民の声を政策に反映するのか？(Iaryczower, Montero, and Kim 2022)
- 時には大量の共変量候補があることもある
  - ⇝ **LASSO の活用**

Demand estimation

- 市場における需要関数推定の重要性
  - 製品の価格弾力性の分析
  - 消費者の厚生分析
  - 最適価格設定
- 今回は市場シェアデータからの価格弾力性推定に着目する

BLP Logit

- **BLP Logit** (Berry, Levinsohn, and Pakes 1995)
  - 差別化財の需要を構造的に推定するモデル
  - 市場シェアデータを用いて推定
  - ランダム係数の枠組みを利用
    - IIA (independence of irrelevant assumption) 特性を排除
  - GMM (Generalized Method of Moments) を用いて価格の内生性に対処
    - 操作変数を用いた識別

BLP Logit (cont)

セットアップ

- 個人、製品、市場: $i = 1, ..., I;\ j = 0, ..., J;\ t = 1, ..., T$
- 製品情報, 価格: $x_{jt}^k,\ p_{jt}$
- ランダム係数: $\beta_i^k \equiv \bar{\beta}^k + \sigma^k v_i^k$
- 価格弾力性: α
- 観察不可能な市場-製品レベルの特性: ξ_{jt}
- 誤差項: ε_{ijt} $\cdots$ 第一種極値分布に従う (Logit 型需要)
- 効用関数: $u_{ijt} = x'_{jt}\beta_{it} - \alpha p_{jt} + \xi_{jt} + \varepsilon_{ijt} = \overbrace{x'_{jt}\bar{\beta} - \alpha p_{jt} + \xi_{jt}}^{\text{Mean utility: } \delta} + \overbrace{v'_i \sigma x_{jt}}^{\text{Deviation}} + \varepsilon_{ijt}$

BLP Logit (cont)

- 理論上の市場シェア

$$s_{jt}(x_{jt}, p_{jt}, \xi_{jt}; \theta) = \int_v \frac{\exp(x'_{jt}\bar{\beta} - \alpha p_{jt} + \xi_{jt} + v'_i \sigma x_{jt})}{1 + \sum_{l=1}^{J} \exp(x'_{lt}\bar{\beta} - \alpha p_{lt} + \xi_{lt} + v'_i \sigma x_{lt})} dF(v)$$

  - where $\theta = \{\bar{\beta}, \alpha\}$

- 観察された市場シェアデータ $\boldsymbol{S} = \{S_{11}, ..., S_{jt}, ...S_{JT}\}$
- $\hat{\boldsymbol{s}}(\boldsymbol{X}, \boldsymbol{P}, \boldsymbol{\xi}; \theta) = \boldsymbol{S}$ となるようにパラメータを求める:
  1. σ を固定して、平均効用 δ を contraction mapping によって推定
  2. 平均効用パラメータ $\{\alpha, \bar{\beta}\}$ を GMM によって推定
  3. $\boldsymbol{\xi}$ を $\hat{\boldsymbol{\delta}}$ と $\hat{\alpha}, \hat{\bar{\beta}}$ によって計算
  4. $\hat{\boldsymbol{\xi}}$ で新しい GMM の目的関数を定義し、それを最小にするような非線形パラメータの σ を準ニュートン法などで推定する

BLP with rich covariates

- 製品情報や人口統計データの豊富さ
  - 商品サイズ、形などの製品情報
  - 年齢、性別、家族構成などの市場レベルの人口統計データ
- Problem
  - どの変数を入れるべきなのか
  - そもそもモデルが解けなくなる
- Good news: **BLP-2LASSO** (Gillen, et al. 2019; Gillen, Shum, and Moon 2014)

BLP-2LASSO

- 手順 (今回はモデル簡単化のために操作変数に関する選択は省略する)
  1. $S_{jt} \sim x_{jt} + p_{jt}$ の LASSO で一つ目の共変量集合 I を得る
  2. $p_j t \sim x_{jt}$ の LASSO で二つ目の共変量集合 II を得る
  3. $I \cup II$ の変数を用いて BLP ロジットを行う

Monte Carlo Simulation of BLP-2LASSO

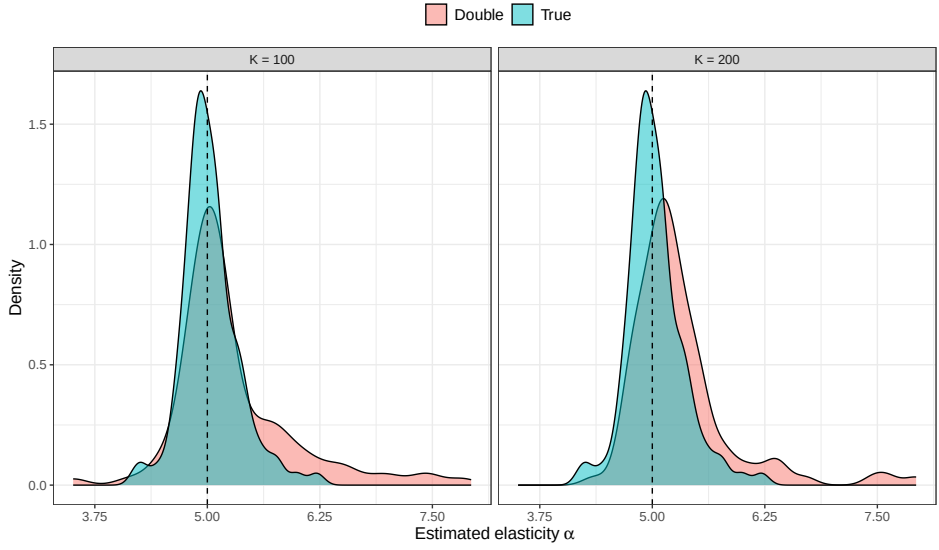
- $J = 10, T = 10, K = \{100, 200\}$
- $X \sim N(0, \Sigma)$
  - $\Sigma_{ij} = \left(\frac{1}{32}\right)^{|i-j|}$
- $\boldsymbol{\xi} \sim N(0, 1/16)$
- 操作変数: $Z^1 \sim |N(1, 0.04)|, \quad Z^2 \sim U(0, 1)$
- $p = |X\eta + \boldsymbol{\xi} + Z^1 + Z^2|$
  - $\eta = [1, 1, 1, 0, \cdots, 0]'$
- $\bar{\beta} = [-2, 4, 4, 2, 2, 0, \cdots, 0]'$
- $\alpha = 5 \leftarrow$ これを推定!
- $\sigma^k = 0.5$

Monte Carlo Simulation of BLP-2LASSO (cont)

- 各 K ごとに Double および Single モデルを 100 回推定
- 高速化のために、Mathematical Programming with Equilibrium Constraint による推定を行った (Su and Judd 2012)
- `R`, `C++` および `Python` を利用した

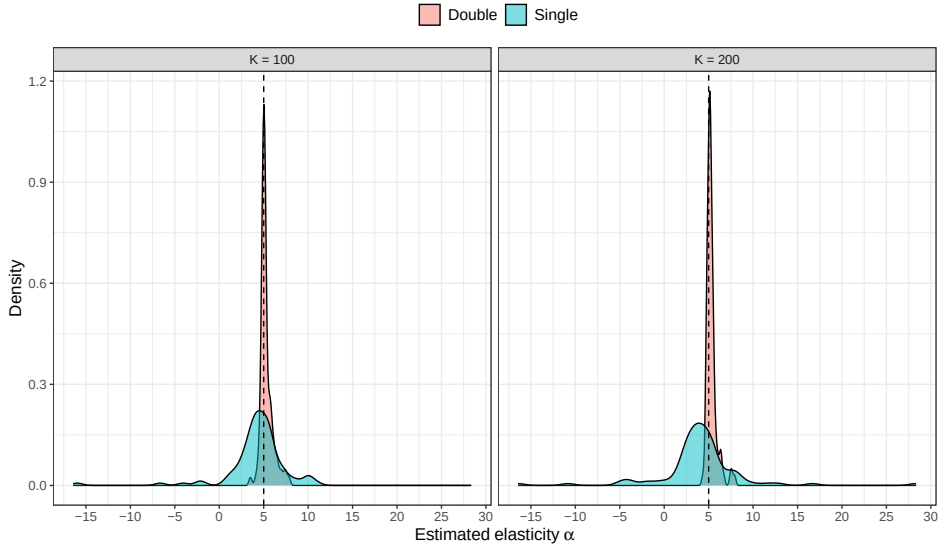
Simulation results of BLP-2LASSO

- Debiased と真の DGP モデルの比較



Simulation results of BLP-2LASSO (cont)

- Debiased と Post-Single モデルの比較



Simulation results of BLP-2LASSO (cont)

- 数値比較

	Mean	Bias	RMSE
K = 100			
Double	5.29	0.45	0.73
Single	4.31	2.00	3.38
K = 200			
Double	5.31	0.40	0.66
Single	3.98	2.85	4.82

Replication codes

シミュレーションの再現コードは GitHub に保存してあります

https://github.com/vkyo23/DoubleLASSO_Simulation