Selective Inference for Time-Varying Moderated Effects

SAM 2024, Salzburg, Austria

Soham Bakshi

Department of Statistics
University of Michigan

June 23, 2025

This is a joint work with

Figure. Snigdha (my advisor) and Walter Dempsey

Background & Introduction

Micro Randomized Trials (MRTs)

MRT data for each individual: $\{O_0, O_1, A_1, O_2, A_2, \ldots, O_T, A_T, O_{T+1}\}$ where t indexes a sequence of decision points.

- Treatment is designed to impact a proximal response, denoted by $Y_{t+1} = y\left(H_t, A_t\right)$

Background & Introduction

Micro Randomized Trials (MRTs)

MRT data for each individual: $\{O_0, O_1, A_1, O_2, A_2, \ldots, O_T, A_T, O_{T+1}\}$ where t indexes a sequence of decision points.

- Treatment is designed to impact a proximal response, denoted by $Y_{t+1} = y\left(H_t, A_t\right)$
- $S_t\left(\bar{a}_{t-1}\right)$ be the potential outcome for a time-varying effect moderator
- $S_t\left(\bar{a}_{t-1}\right)$ is a summary of $H_t\left(\bar{a}_{t-1}\right)$ (the potential history up to time t)

Background & Introduction

Moderated Effects

The *causal excursion effect* estimand for linear contrast is defined as:

$$\beta_{\mathbf{p}}(t; s) = \mathbb{E}_{\mathbf{p}}\left[Y_{t+1}\left(\bar{A}_{t-1}, A_t = 1\right) - Y_{t+1}\left(\bar{A}_{t-1}, A_t = 0\right) \mid S_t\left(\bar{A}_{t-1}\right) = s\right],$$

these are time-varying effects and are function of the decision point t and a set of moderators S_t and is marginalized over all other observed and unobserved variables.

The *causal excursion effect* estimand for linear contrast is defined as:

$$\beta_{\mathbf{p}}(t;s) = \mathbb{E}_{\mathbf{p}}\left[Y_{t+1}\left(\bar{A}_{t-1}, A_t = 1\right) - Y_{t+1}\left(\bar{A}_{t-1}, A_t = 0\right) \mid S_t\left(\bar{A}_{t-1}\right) = s\right],$$

these are time-varying effects and are function of the decision point t and a set of moderators S_t and is marginalized over all other observed and unobserved variables.

The estimand can be re-expressed in terms of observable data:

$$\beta_{\mathbf{p}}(t;s) = \mathbb{E}\left[\mathbb{E}_{\mathbf{p}}\left[Y_{t+1} \mid A_t = 1, H_t\right] - \mathbb{E}_{\mathbf{p}}\left[Y_{t+1} \mid A_t = 0, H_t\right] \mid S_t = s\right] = f_t(s)^\top \beta^\star.$$

The *causal excursion effect* estimand for linear contrast is defined as:

$$\beta_{\mathbf{p}}(t; s) = \mathbb{E}_{\mathbf{p}}\left[Y_{t+1}\left(\bar{A}_{t-1}, A_t = 1\right) - Y_{t+1}\left(\bar{A}_{t-1}, A_t = 0\right) \mid S_t\left(\bar{A}_{t-1}\right) = s\right],$$

these are time-varying effects and are function of the decision point t and a set of moderators S_t and is marginalized over all other observed and unobserved variables.

The estimand can be re-expressed in terms of observable data:

$$\beta_{\mathbf{p}}(t; s) = \mathbb{E}\left[\mathbb{E}_{\mathbf{p}}\left[Y_{t+1} \mid A_t = 1, H_t\right] - \mathbb{E}_{\mathbf{p}}\left[Y_{t+1} \mid A_t = 0, H_t\right] \mid S_t = s\right] = f_t(s)^\top \beta^\star.$$

For binary outcome, the relative risk (RR):

$$\begin{aligned}\beta_{\mathbf{p}}^{(RR)}(t; s) &= \log\left(\frac{\mathbb{E}_{\mathbf{p}}\left[Y_{t+1}\left(\bar{A}_{t-1}, A_t = 1\right) \mid S_t\left(\bar{A}_{t-1}\right) = s\right]}{\mathbb{E}_{\mathbf{p}}\left[Y_{t+1}\left(\bar{A}_{t-1}, A_t = 0\right) \mid S_t\left(\bar{A}_{t-1}\right) = s\right]}\right)\\ &= \log\left(\frac{\mathbb{E}\left[\mathbb{E}\left[Y_{t+1} \mid H_t, A_t = 1\right] \mid S_t = s\right]}{\mathbb{E}\left[\mathbb{E}\left[Y_{t+1} \mid H_t, A_t = 0\right] \mid S_t = s\right]}\right) = f_t(s)^\top \beta^\star.\end{aligned}$$

- A consistent estimator of $\beta^\star$ can be obtained by minimizing a *weighted and centered least squares* (WCLS) criterion:

$$\arg\min_{\beta} \mathbb{P}_n \left[\sum_{t=1}^{T} \left(Y_{t+1} - g_t(H_t)^\top \alpha - (A_t - \hat{p}(1|S_t))\, f_t(S_t)^\top \beta \right)^2 W_t \right], \tag{1}$$

where $W_t = \frac{\hat{p}_t(A_t|S_t)}{p_t(A_t|H_t)}$ is a weight with $\hat{p}_t(A_t|S_t) \in (0,1)$ that only depends on S_t, $g_t(H_t) \in \mathbb{R}^q$ are q control variables (hoping to explain variation in the expected outcome).

- A consistent estimator of $\beta^\star$ can be obtained by minimizing a *weighted and centered least squares* (WCLS) criterion:

$$\arg\min_{\beta} \mathbb{P}_n \left[\sum_{t=1}^{T} \left(Y_{t+1} - g_t(H_t)^\top \alpha - (A_t - \hat{p}(1|S_t))\, f_t(S_t)^\top \beta \right)^2 W_t \right], \tag{1}$$

where $W_t = \frac{\hat{p}_t(A_t|S_t)}{p_t(A_t|H_t)}$ is a weight with $\hat{p}_t(A_t|S_t) \in (0,1)$ that only depends on S_t, $g_t(H_t) \in \mathbb{R}^q$ are q control variables (hoping to explain variation in the expected outcome).

- Consistent estimator of the RR excursion effect can be obtained by solving (a set of weighted estimating equations) for EMEE:

$$\mathbb{P}_n \left[\sum_{t=1}^{T} W_t e^{-A_t f_t(S_t)^\top \beta} \left(Y_{t+1} - e^{g_t(H_t)^\top \alpha + A_t f_t(S_t)^\top \beta} \right) \begin{pmatrix} g_t(H_{i,t}) \\ (A_{i,t} - \hat{p}(1|S_{i,t})) f_t(S_{i,t}) \end{pmatrix} \right] = 0. \tag{2}$$

Background & Introduction

Broad Picture

- Treatment effects vary within levels of *effect moderators*, measured prior to the treatment.
- In presence of many potential moderators ($f_t(s) \in \mathbb{R}^p$ with large p).
- Fundamental trade-off accuracy and interpretability in the context of models with high-dimensional parameters $\beta^* \in \mathbb{R}^p$
- Use the data to select a smaller subset of *significant* moderators.
- Subsequent inference on the *selected* moderated effect $\beta^*(t; s)$? Selective bias? Data Splitting?

Selection of Moderators

Selection of Moderators

- We propose a randomized LASSO selection.

Selection of Moderators

- We propose a randomized LASSO selection.
- A randomized LASSO selection for WCLS would be:

$$\underset{\alpha,\beta}{\arg\min}\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\sum_{t=1}^{T}\left(Y_{i,t}-g_t(H_{i,t})^\top\alpha-(A_{i,t}-\hat{p}(1|S_t))\,f_t(S_{i,t})^\top\beta\right)^2 W_{i,t}+\lambda\|\beta\|_1-\sqrt{n}\omega_n^\top\beta\right\},$$

with a fixed regularization parameter $\lambda \in \mathbb{R}^+$. Say, $\widehat{\beta}_n^{(\lambda)} \in \mathbb{R}^p$ be a randomized LASSO solution, then we select the moderators in $f_t(s)$ corresponding to which the solution has non-zero coefficient, that is $\widehat{E} = \left\{j \in [p] : \widehat{\beta}_{n,j}^{(\lambda)} \neq 0\right\}$.

Selection of Moderators

Risk Formulation

- To facilitate selection using LASSO, we formulate the casual contrast estimation problems through loss minimization.
- Estimators based on ERM, for linear risk:

$$\mathbb{P}_n \left[\sum_{t=1}^{T} \tilde{\sigma}_t^2(S_t) \left\{ f_t(S_t)^\top \beta - \beta^\star(t; S_t) \right\}^2 \right]$$

Selection of Moderators
Risk Formulation

- To facilitate selection using LASSO, we formulate the casual contrast estimation problems through loss minimization.
- Estimators based on ERM, for linear risk:

$$\mathbb{P}_n \left[\sum_{t=1}^{T} \tilde{\sigma}_t^2(S_t) \left\{ f_t(S_t)^\top \beta - \beta^\star(t; S_t) \right\}^2 \right]$$

For relative risk:

$$\mathbb{P}_n \left[\sum_{t=1}^{T} \tilde{\sigma}_t^2(S_t) \left(\mu_t^\star(H_t, 1) + \mu_t^\star(H_t, 0) \right) \left\{ \log(1 + e^{f_t(S_t)^\top \beta}) - \frac{\mu_t^\star(H_t, 1)}{(\mu_t^\star(H_t, 1) + \mu_t^\star(H_t, 0))} f_t(S_t)^\top \beta \right\} \right]$$

Selection of Moderators

General Framework

Selection of Moderators

General Framework

- $(Y_i, X_i)_{i \in [n]}$ are n i.i.d sample from $\mathbb{F}_n$. Let $\psi(Y_i, X_i b)$ be a generic loss function, with response $Y_i \in \mathbb{R}^T$ and stacked temporal observations of p moderators $X_i \in \mathbb{R}^{T \times p}$.

- $(Y_i, X_i)_{i\in[n]}$ are n i.i.d sample from $\mathbb{F}_n$. Let $\psi(Y_i, X_i b)$ be a generic loss function, with response $Y_i \in \mathbb{R}^T$ and stacked temporal observations of p moderators $X_i \in \mathbb{R}^{T\times p}$.
- A randomized LASSO problem is :

$$\widehat{\beta}_n^{(\lambda)} \in \underset{b\in\mathbb{R}^p}{\arg\min} \left\{ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi(Y_i, X_i b) + \lambda \|b\|_1 - \sqrt{n}\omega_n' b \right\},$$

where $\lambda \in \mathbb{R}^+$ is a fixed regularization parameter and $\sqrt{n}\omega_n \sim \mathcal{N}(0_p, \Omega) \in \mathbb{R}^p$ is added randomization variable drawn independently of y.

Selection of Moderators

General Framework

- $(Y_i, X_i)_{i\in[n]}$ are n i.i.d sample from $\mathbb{F}_n$. Let $\psi(Y_i, X_i b)$ be a generic loss function, with response $Y_i \in \mathbb{R}^T$ and stacked temporal observations of p moderators $X_i \in \mathbb{R}^{T\times p}$.
- A randomized LASSO problem is :

$$\widehat{\beta}_n^{(\lambda)} \in \underset{b\in\mathbb{R}^p}{\arg\min} \left\{ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi(Y_i, X_i b) + \lambda\|b\|_1 - \sqrt{n}\omega_n' b \right\},$$

  where $\lambda \in \mathbb{R}^+$ is a fixed regularization parameter and $\sqrt{n}\omega_n \sim \mathcal{N}(0_p, \Omega) \in \mathbb{R}^p$ is added randomization variable drawn independently of y.
- We select the moderators for which the the randomized LASSO solution has non-zero coefficient, that is $E = \left\{ j \in [p] : \widehat{\beta}_{n,j}^{(\lambda)} \neq 0 \right\}$.

Post-Selection Inference

For Selected Moderated Effects

- After selection of $E \subset [p]$ our target of inference is:

$$\beta_n^E = \operatorname*{arg\,min}_{b \in \mathbb{R}^{|E|}} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathbb{E}[\psi(Y_i, X_{i,E} b)], \text{ more precisely } \beta_n^{E.j} = e_j^\top \beta_n^E, \forall j \in E.$$

Post-Selection Inference

For Selected Moderated Effects

- After selection of $E \subset [p]$ our target of inference is:

$$\beta_n^E = \arg\min_{b \in \mathbb{R}^{|E|}} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathbb{E}[\psi(Y_i, X_{i,E} b)], \text{ more precisely } \beta_n^{E.j} = e_j^\top \beta_n^E, \forall j \in E.$$

- A natural estimate of our target is what we call *refitted estimator*:

$$\widehat{\beta}_n^E = \arg\min_{b \in \mathbb{R}^{|E|}} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi(Y_i, X_{i,E} b) \text{ and } \widehat{\beta}_n^{E.j} = e_j^\top \widehat{\beta}_n^E, \forall j \in E.$$

Post-Selection Inference

For Selected Moderated Effects

- After selection of $E \subset [p]$ our target of inference is:

$$\beta_n^E = \underset{b \in \mathbb{R}^{|E|}}{\arg\min} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathbb{E}[\psi(Y_i, X_{i,E} b)], \text{ more precisely } \beta_n^{E.j} = e_j^\top \beta_n^E, \forall j \in E.$$

- A natural estimate of our target is what we call *refitted estimator*:

$$\widehat{\beta}_n^E = \underset{b \in \mathbb{R}^{|E|}}{\arg\min} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi(Y_i, X_{i,E} b) \text{ and } \widehat{\beta}_n^{E.j} = e_j^\top \widehat{\beta}_n^E, \forall j \in E.$$

- In the conditional selective inference approach, we conduct inference for $\beta_n^{E.j}$ based on the conditional distribution of $\widehat{\beta}_n^{E.j} | \{\widehat{E} = E\}$.

Post-Selection Inference

- The Karush-Kuhn-Tucker (KKT) conditions of stationarity at $\widehat{\beta}_n^{(\lambda)}$:

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i^\top \nabla\psi(Y_i, X_i\widehat{\beta}_n^{(\lambda)}) + \mathcal{Z} = \sqrt{n}\omega_n, \text{ with } s_E = \text{sign}(\widehat{\beta}_{n,E}^{(\lambda)}), \|\mathcal{Z}_{-E}\| \leq \lambda \tag{3}$$

where $\mathcal{Z} = \begin{bmatrix} \lambda s_E \\ \mathcal{Z}_{-E} \end{bmatrix}$ is the observed value of the subgradient Z of the LASSO penalty at $\widehat{\beta}_n^{(\lambda)}$.

- The Karush-Kuhn-Tucker (KKT) conditions of stationarity at $\widehat{\beta}_n^{(\lambda)}$:

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i^\top \nabla\psi(Y_i, X_i\widehat{\beta}_n^{(\lambda)}) + \mathcal{Z} = \sqrt{n}\omega_n, \text{ with } s_E = \text{sign}(\widehat{\beta}_{n,E}^{(\lambda)}), \|\mathcal{Z}_{-E}\| \leq \lambda \tag{3}$$

where $\mathcal{Z} = \begin{bmatrix}\lambda s_E \\ \mathcal{Z}_{-E}\end{bmatrix}$ is the observed value of the subgradient Z of the LASSO penalty at $\widehat{\beta}_n^{(\lambda)}$.

- Define $p \times p$ matrices based on the gradient and Hessian of the loss:

$$H = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i^\top \nabla^2\psi\left(Y_i, X_i\beta_n\right)\right] \text{ and } K = \mathrm{Cov}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i^\top \nabla\psi\left(Y_i, X_i\beta_n\right)\right)$$

and the orthogonal estimator

Post-Selection Inference

- The Karush-Kuhn-Tucker (KKT) conditions of stationarity at $\widehat{\beta}_n^{(\lambda)}$:

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i^\top \nabla\psi(Y_i, X_i\widehat{\beta}_n^{(\lambda)}) + \mathcal{Z} = \sqrt{n}\omega_n, \text{ with } s_E = \text{sign}(\widehat{\beta}_{n,E}^{(\lambda)}), \|\mathcal{Z}_{-E}\| \leq \lambda \tag{3}$$

where $\mathcal{Z} = \begin{bmatrix} \lambda s_E \\ \mathcal{Z}_{-E} \end{bmatrix}$ is the observed value of the subgradient Z of the LASSO penalty at $\widehat{\beta}_n^{(\lambda)}$.

- Define $p \times p$ matrices based on the gradient and Hessian of the loss:

$$H = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i^\top \nabla^2\psi\left(Y_i, X_i\beta_n\right)\right] \text{ and } K = \mathrm{Cov}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i^\top \nabla\psi\left(Y_i, X_i\beta_n\right)\right)$$

and the orthogonal estimator

$$\widehat{\Gamma}_n^{E\cdot j} = \begin{bmatrix} \mathbb{S}_{[E]\setminus j}\left(\widehat{\beta}_n^E - \frac{1}{\sigma_j^2}\Sigma_{E,j}\widehat{\beta}_n^{E\cdot j}\right) \\ \frac{1}{n}\sum_{i=1}^{n} X_{i,-E}^\top \nabla\psi\left(Y_i, X_i\widehat{\beta}_n^E\right) - (H_{-E,E} - K_{-E,E}K_{E,E}^{-1}H_{E,E})\widehat{\beta}_n^E \end{bmatrix} \in \mathbb{R}^{p-1}.$$

Taylor expanding $\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_{i,E}^{\top}\nabla\psi\left(Y_i, X_i\widehat{\beta}_n\right) = 0$, the around β_n:

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_{i,E}^{\top}\nabla\psi\left(Y_i, X_i\beta_n^E\right) + \frac{1}{n}\sum_{i=1}^{n} X_{i,E}^{\top}\nabla^2\psi\left(Y_i, X_i\beta_n^E\right) X_{i,E}\sqrt{n}(\widehat{\beta}_n^E - \beta_n^E) = 0$$

$$\implies \sqrt{n}(\widehat{\beta}_n^E - \beta_n^E) = H_{E,E}^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_{i,E}^{\top}\nabla\psi\left(Y_i, X_i\beta_n\right) + o_p(1)$$

Thus, asymptotic covariance of $\sqrt{n}(\widehat{\beta}_n^E - \beta_n^E)$ will be $H_{E,E}^{-1}K_{E,E}H_{E,E}^{-1}$ (say $\Sigma_{E,E}$).

Post-Selection Inference

Fixed E Asymptotics

Taylor expanding $\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_{i,E}^{\top}\nabla\psi\left(Y_i, X_i\widehat{\beta}_n\right) = 0$, the around β_n:

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_{i,E}^{\top}\nabla\psi\left(Y_i, X_i\beta_n^E\right) + \frac{1}{n}\sum_{i=1}^{n} X_{i,E}^{\top}\nabla^2\psi\left(Y_i, X_i\beta_n^E\right) X_{i,E}\sqrt{n}(\widehat{\beta}_n^E - \beta_n^E) = 0$$

$$\implies \sqrt{n}(\widehat{\beta}_n^E - \beta_n^E) = H_{E,E}^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_{i,E}^{\top}\nabla\psi\left(Y_i, X_i\beta_n\right) + o_p(1)$$

Thus, asymptotic covariance of $\sqrt{n}(\widehat{\beta}_n^E - \beta_n^E)$ will be $H_{E,E}^{-1}K_{E,E}H_{E,E}^{-1}$ (say $\Sigma_{E,E}$).

$$\boxed{\sqrt{n}(\widehat{\beta}_n^E - \beta_n^E) \sim \mathcal{N}_p(0, \Sigma_{E,E})}$$

Post-Selection Inference

Piece 1: Asymptotic representations of data variables for fixed E.

Piece 1: Asymptotic representations of data variables for fixed E.

Theorem 3.1

*Suppose that $E \subset [p]$ is a fixed subset, then under assumption **??**, asymptotically we would have*

$$\sqrt{n}\begin{pmatrix} \widehat{\beta}_n^{E.j} - \beta_n^{E.j} \\ \widehat{\Gamma}_n^{E.j} - \Gamma_n^{E.j} \end{pmatrix} \xrightarrow{d} \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{bmatrix} \sigma_j^2 & 0 & 0 \\ 0 & \Sigma_2 & 0 \\ 0 & 0 & \Sigma_3 \end{bmatrix}\right),$$

where $\Sigma_2 = \Sigma_{E,E} - \frac{\Sigma_{E,j}\Sigma_{j,E}}{\sigma_j^2}$, $\Sigma_3 = K_{-E,-E} + H_{-E,E}H_{E,E}^{-1}K_{-E,E}^{\top} + K_{-E,E}H_{E,E}^{-1}H_{-E,E}^{\top} - K_{-E,E}K_{E,E}^{-1}K_{-E,E}$.

Post-Selection Inference

Conditioning Set

- Start by conditioning on subgradient vector Z,

$$\{Z = \mathcal{Z}\} = \left\{\widehat{E} = E, s_E = \text{sign}(\widehat{\beta}_{n,E}^{(\lambda)}), Z_{-E} = \mathcal{Z}_{-E}\right\} \subseteq \left\{\hat{E} = E\right\}.$$

Post-Selection Inference

Conditioning Set

- Start by conditioning on subgradient vector Z,

$$\{Z = \mathcal{Z}\} = \left\{\widehat{E} = E, s_E = \text{sign}(\widehat{\beta}_{n,E}^{(\lambda)}), Z_{-E} = \mathcal{Z}_{-E}\right\} \subseteq \left\{\hat{E} = E\right\}.$$

- Condition further on a projection of $\widehat{\beta}_{n,E}^{(\lambda)}$. Define $\Lambda = (H'_{,E}\Omega^{-1}H_{,E})^{-1}$ and let

$$Q^\gamma = \frac{\Lambda\gamma}{\gamma^\top\Lambda\gamma} \text{ and } A^\gamma = \left(I - Q^\gamma\gamma^\top\right)\widehat{\beta}_{n,E}^{(\lambda)}, \text{ where } \gamma \in \mathbb{R}^{|E|}.$$

- We condition on A^γ which the orthogonal projection and consider the conditioning event $\{Z = \mathcal{Z}, A^\gamma = \mathcal{A}^\gamma\}$.

Proposition 3.2

For $j \in E$ and $\gamma = \Lambda^{-1} H_{,E}^{\top} \Omega^{-1} P_1$, define

$$I_{-}^{\gamma} = \max_{k: s_{[k]}^{\top} \Lambda\gamma < 0} \left\{ \frac{-s_{[k]}^{\top} A^{\gamma}}{s_{[k]}^{\top} Q^{\gamma}} \right\}, I_{+}^{\gamma} = \min_{k: s_{[k]}^{\top} \Lambda\gamma > 0} \left\{ \frac{-s_{[k]}^{\top} A^{\eta}}{s_{[k]}^{\top} Q^{\gamma}} \right\},$$

where $s_{[k]}$ denoting the $k-th$ column of $diag(s_E)$. We have

$$\{Z = \mathcal{Z}, A^{\gamma} = \mathcal{A}^{\gamma}\} = \left\{ \gamma^{\top} \widehat{\beta}_{n,E}^{(\lambda)} \in [I_{-}^{\gamma}, I_{+}^{\gamma}], Z_{\widehat{E}} = \mathcal{Z}_E, A^{\gamma} = \mathcal{A}^{\gamma} \right\}.$$

Proposition 3.2

For $j \in E$ and $\gamma = \Lambda^{-1} H_{,E}^{\top} \Omega^{-1} P_1$, define

$$I_{-}^{\gamma} = \max_{k: s_{[k]}^{\top} \Lambda \gamma < 0} \left\{ \frac{-s_{[k]}^{\top} A^{\gamma}}{s_{[k]}^{\top} Q^{\gamma}} \right\}, I_{+}^{\gamma} = \min_{k: s_{[k]}^{\top} \Lambda \gamma > 0} \left\{ \frac{-s_{[k]}^{\top} A^{\eta}}{s_{[k]}^{\top} Q^{\gamma}} \right\},$$

where $s_{[k]}$ denoting the $k-th$ column of $diag(s_E)$. We have

$$\{Z = \mathcal{Z}, A^{\gamma} = \mathcal{A}^{\gamma}\} = \left\{ \gamma^{\top} \widehat{\beta}_{n,E}^{(\lambda)} \in [I_{-}^{\gamma}, I_{+}^{\gamma}], Z_{\widehat{E}} = \mathcal{Z}_E, A^{\gamma} = \mathcal{A}^{\gamma} \right\}.$$

Choice of γ? We choose $\gamma = \Lambda^{-1} H_{,E}^{\top} \Omega^{-1} P_1$, so that we don't pay any extra price for conditioning at least in the situations when selection has no impact.

Post-Selection Inference

Piece 2: Connect the data variables $\widehat{\beta}_n^{E.j}, \widehat{\Gamma}_n^{E.j}$ with the optimization variables $\widehat{\beta}_{n,E}^{(\lambda)}, Z_E$ through the KKT.

Piece 2: Connect the data variables $\widehat{\beta}_n^{E.j}, \widehat{\Gamma}_n^{E.j}$ with the optimization variables $\widehat{\beta}_{n,E}^{(\lambda)}, Z_E$ through the KKT.

Lemma 3.3

The KKT condition at $\widehat{\beta}_{n,E}^{(\lambda)}$ can be written as

$$P_1\sqrt{n}\widehat{\beta}_n^{E.j} + P_2\sqrt{n}\widehat{\Gamma}_n^{E.j} + H_{,E}\sqrt{n}\widehat{\beta}_{n,E}^{(\lambda)} + \lambda \begin{bmatrix} s_E \\ Z_{-E} \end{bmatrix} + o_p(1) = \sqrt{n}w_n$$

where

$$P_1 = -\frac{1}{\sigma_j^2}K_{,E}H_{E,E}^{-1}\mathbb{S}_j^\top \in \mathbb{R}^p \text{ and } P_2 = \begin{bmatrix} -H_{E,E}\mathbb{S}_{[E]\setminus j}^\top & 0 \\ -K_{-E,E}K_{E,E}^{-1}H_{E,E}\mathbb{S}_{[E]\setminus j}^\top & I \end{bmatrix} \in \mathbb{R}^{p\times(p-1)}.$$

Piece 3: Derive the joint density of $\widehat{\beta}^{(\lambda)}_{n,E}, Z_{-E}$ and $\widehat{\beta}^{E.j}_n, \widehat{\Gamma}^{E.j}_n$.

Lemma 3.4

For a fixed subset E and fixed signs s_E, asymptotically the joint density of the variables $\sqrt{n}\widehat{\beta}^{E.j}_n, \sqrt{n}\widehat{\Gamma}^{E.j}_n, \sqrt{n}\widehat{\beta}^{(\lambda)}_{n,E}, Z_{-E}$ is equal to

$$\phi(\sqrt{n}\widehat{\beta}^{E.j}_n; \sqrt{n}\beta^{E.j}_n, \sigma_j^2)\phi\left(\sqrt{n}\Gamma^{E.j}_n; \sqrt{n}\widehat{\Gamma}^{E.j}_n, A_2A_2^\top\right)$$
$$\times\,\phi\left(P_1\sqrt{n}\widehat{\beta}^{E.j}_n + P_2\sqrt{n}\widehat{\Gamma}^{E.j}_n + H_{,E}\sqrt{n}\widehat{\beta}^{(\lambda)}_{n,E} + \lambda\begin{bmatrix} s_E \\ Z_{-E}\end{bmatrix}; 0_p, \Omega\right)|J|$$

where $J = \begin{bmatrix} H_{E,E} & 0 \\ H_{-E,E} & \lambda I\end{bmatrix}$.

Given the conditioning event $\left\{Z = \mathcal{Z}, A^j = \mathcal{A}^j\right\}$, define

$$W_0(\sqrt{n}\widehat{\beta}_n^{E.j}, \sqrt{n}\widehat{\Gamma}_n^{E.j}) = \int_{I_-^j}^{I_+^j} \phi\left(H_{,E}Q^j t + P_1\sqrt{n}\widehat{\beta}_n^{E.j} + P_2\sqrt{n}\widehat{\Gamma}_n^{E.j} + T^j; 0_p, \Omega\right) dt.$$

where $T^j = H_{,E}\sqrt{n}\mathcal{A}^j + \lambda \begin{bmatrix} s_E \\ \mathcal{Z}_{-E} \end{bmatrix}$.

Given the conditioning event $\left\{Z = \mathcal{Z}, A^j = \mathcal{A}^j\right\}$, define

$$W_0(\sqrt{n}\widehat{\beta}_n^{E\cdot j}, \sqrt{n}\widehat{\Gamma}_n^{E\cdot j}) = \int_{I_-^j}^{I_+^j} \phi\left(H_{,E}Q^j t + P_1\sqrt{n}\widehat{\beta}_n^{E\cdot j} + P_2\sqrt{n}\widehat{\Gamma}_n^{E\cdot j} + T^j; 0_p, \Omega\right) dt.$$

where $T^j = H_{,E}\sqrt{n}\mathcal{A}^j + \lambda \begin{bmatrix} s_E \\ \mathcal{Z}_{-E} \end{bmatrix}$.

Using these notations, we now propose our pivot as:

$$\boxed{P^{E\cdot j}(\widehat{\beta}_n^{E\cdot j}, \widehat{\Gamma}_n^{E\cdot j}) = \frac{\int_{-\infty}^{\sqrt{n}\widehat{\beta}_n^{E\cdot j}} \phi(x; \sqrt{n}\beta_n^{E\cdot j}, \sigma_j^2) W_0(x, \sqrt{n}\widehat{\Gamma}_n^{E\cdot j}) dx}{\int_{-\infty}^{\infty} \phi(x; \sqrt{n}\beta_n^{E\cdot j}, \sigma_j^2) W_0(x, \sqrt{n}\widehat{\Gamma}_n^{E\cdot j}) dx}}$$

Post-Selection Inference

- Two-sided confidence intervals at level α can be constructed as

$$\left(L^j_{\alpha.n}, U^j_{\alpha.n}\right) = \left\{\beta_n^{E.j} : P^{E.j}(\widehat{\beta}_n^{E.j}) \in \left[\frac{\alpha}{2}, 1 - \frac{\alpha}{2}\right]\right\}.$$

- Two-sided confidence intervals at level α can be constructed as

$$\left(L^{j}_{\alpha.n}, U^{j}_{\alpha.n}\right) = \left\{\beta^{E.j}_{n} : P^{E.j}(\widehat{\beta}^{E.j}_{n}) \in \left[\frac{\alpha}{2}, 1 - \frac{\alpha}{2}\right]\right\}.$$

- For each $n \in \mathbb{N}$, say $\mathcal{F}_n = \{\mathbb{F}_n\}$ denotes a collection of data-generating distributions. Then, under *some* conditions we have

$$\lim_{n} \sup_{\mathbb{F}_n \in \mathcal{F}} \left| \mathbb{P}_{\mathbb{F}_n}\left[\beta^{E.j}_{n} \in \left(L^{j}_{\alpha.n}, U^{j}_{\alpha.n}\right) \mid \{Z = \mathcal{Z}, A^{\gamma} = \mathcal{A}^{\gamma}\}\right] - (1-\alpha)\right| = 0.$$

- In other words, for any fixed $\epsilon > 0$, we can find a $N(\epsilon) \in \mathbb{N}$ such that $\left(L^{j}_{\alpha.n}, U^{j}_{\alpha.n}\right)$ attains the selective coverage of $1 - \alpha - \epsilon$, for any $n \geq N(\epsilon)$ and for any data-generating distribution in $\mathcal{F}_n$.

Simulation Study
Settings

- A simple generative model for Y_{t+1}, which is linear in $(A_t, S_t, A_{t-1}, S_{t-1}, A_{t-2}, A_t S_t, A_{t-1} S_t, A_{t-2} S_{t-1})$.
- Let $S_t \in \mathbb{R}^d$ satisfy $S_t = \Gamma_0 S_{t-1} + A_t \Gamma_1 + e_t$, where $\Gamma_0 \in \mathbb{R}^{p \times p}$ and $\Gamma_1 \in \mathbb{R}^p$ and $e_t \sim N(0, \sigma^2 I_p)$ is an iid white-noise term.
- The outcome model is

$$Y_{t+1} = \theta_1^\top \tilde{S}_t + \theta_2 (A_{t-1} - p_t(1|H_{t-1})) + (A_t - p_t(1|H_t)) \left(\beta_{10} + \beta_{11}^\top \tilde{S}_{t,E} \right) + \epsilon_t,$$

  where ϵ_t is a Gaussian process with covariance function $\mathrm{Cov}(\epsilon_t, \epsilon_s) = \rho^{|t-s|}$.
- $\boxed{\text{Set } T = 30,\ p = 50,\ E \text{ to be the first 5 moderators in } S_t.}$
- Further, we set $\rho = 0.5$, $\theta_1 = 0.8\mathbf{1}_p$ and $\sigma^2 = 1$. The target of inference is the sparse p-dimensional vector $\boxed{\beta^* = (\beta_{11}^\top, \mathbf{0}_{p-|E|})^\top.}$

Simulation Study

To evaluate our method's performance across different settings, we vary our sample size through n and signal strength through $\beta_{11} = c\mathbf{1}_5$.

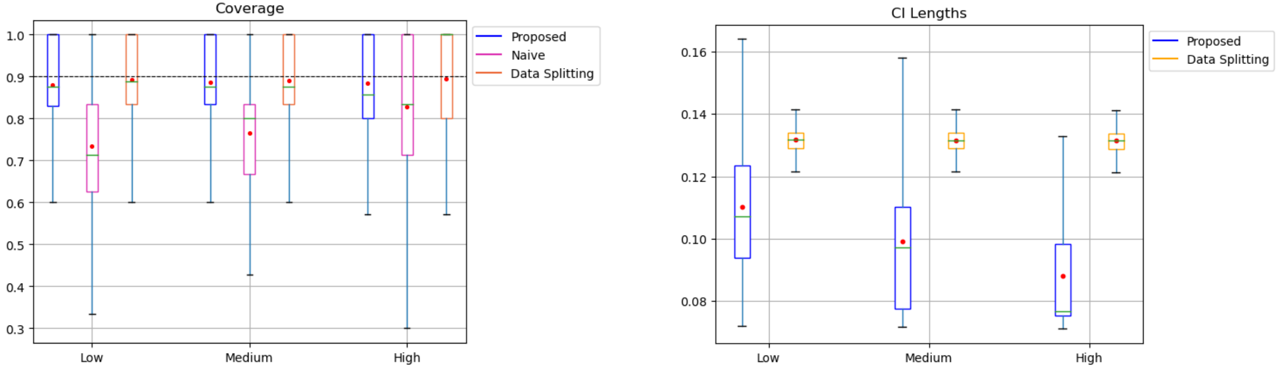

Figure. Coverage (left) and Length (right) for varying signal strength

Simulation Study

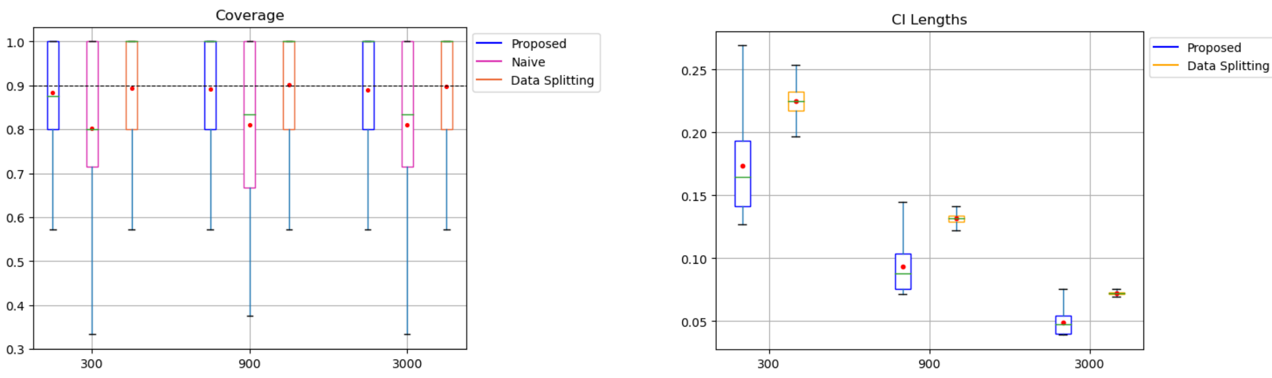

Figure. Coverage (left) and Length (right) for varying sample size

Simulation Study

Findings

- Not accounting for selection results in failure to attain right coverage. Both splitting and our selective method achieves 90% coverage.
- The proposed selective confidence intervals are consistently narrower than those obtained by data splitting. Specifically, for low signal strength and low sample sizes.

Discussion
And Future Work

- We propose a LASSO based causal model selection and conditional inference framework (beats data-splitting, powerful inference?).
- Directly applicable to a broad class of moderated effects, allowing any generic loss based estimation.
- We provide a selective inference method based on an easy to compute pivotal quantity and make asymptotic guarantees without requiring stringent assumptions.
- How does the choice of Ω effect our inference? In practice, what should we do?
- Can we use cross-validation to tune λ? How do you choose λ in practice?