

# On the Use of Regularization Approaches in Unidentified Models (Ph.D. Thesis Defense)

LIN, Yiqi<sup>a</sup>

Supervisors: Prof. SONG, Xinyuan<sup>a</sup>, Prof. FAN, Qingliang<sup>b</sup>

Department of Statistics, The Chinese University of Hong Kong<sup>a</sup>

Department of Economics, The Chinese University of Hong Kong<sup>b</sup>

May 24, 2023

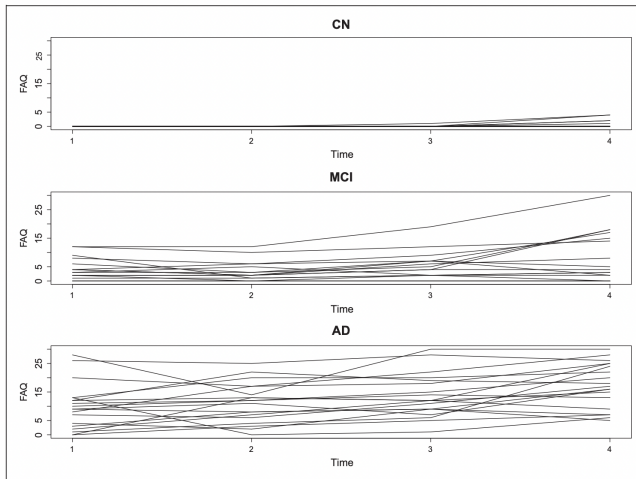
# Addressing Unidentified Models

- Unidentified models create challenges in interpreting statistical models and affect the reliability and validity of statistical analyses in various fields.
- Various methods have been developed to deal with unidentified models, including regularization techniques, Bayesian approaches, and instrumental variable (IV) methods.
- This thesis focuses on three specific unidentified models: hidden Markov models with unknown number of states, linear IV models with weak or invalid instruments, and IV regression without linearly valid IVs.
- The proposed methods aim to provide more accurate and robust estimates in the presence of unidentified components.

## 1 Order Selection for Regression-based Hidden Markov Model(RHMM)

- Motivation
- Proposed Model
  - Regression Based Hidden Markov Model
  - Estimating the order
- Order Selection via Extended-Group-Sort-Fuse (EGSF) Procedure
- Double Penalization
  - Asymptotic Result
- Estimation Via ECM-ITD method
  - ECM-ITD algorithm
- Simulation
- Application
  - ADNI data analysis
  - Results

# Motivation



**Figure 1.** ADNI-I data analysis results: individual trajectories of functional assessment questionnaire scores for 20 randomly selective samples whose baseline states are cognitive normal, mild cognitive impairment, and Alzheimer's disease, respectively.

# Motivation(Cont')

- Hidden Markov model (HMM) is a practical statistical tool to simultaneously analyze the longitudinal observation process and its dynamic transition process.
- The most existing HMMs and their extensions in the literature require a predetermined number of states (order of HMM), which is often unknown in practice. A data-driven procedure to choose the number of hidden states still remains a challenging problem.
- The most common method in the literature for model selection is information-based criteria, such as AIC (Akaike, 1974) and BIC (Schwarz et al., 1978).
- However, even though these prevailing information-based criteria have succeeded in some applications, they still lack theoretical justification for HMMs and their extensions according to some sources in the literature (MacDonald and Zucchini, 1997).

# Common Settings

$\mathbf{Y} = (\mathbf{Y}_1, \mathbf{Y}_2, \dots, \mathbf{Y}_n)$ , where  $\mathbf{Y}_i = \{y_{it}\}_{t=1}^T$  and  $y_{it}$  be the response of subject  $i$  at time  $t$ ;  $\mathbf{S} = (\mathbf{S}_1, \mathbf{S}_2, \dots, \mathbf{S}_n)$ , where  $\mathbf{S}_i = \{S_{it}\}_{t=1}^T$  is a set of hidden states associated with  $y_{it}$ , and  $S_{it}$  is assumed to be a finite-state stationary Markov chain taking values in  $\{1, \dots, K\}$ .

- ① transition between different states can be described by a homogeneous transition matrix  $P = [P_{rs}]_{K \times K}$  with  $P_{rs} = P(S_{it} = s | S_{i,t-1} = r)$  for  $\forall i$  and  $t = 2, \dots, T$  and stationary probability  $\pi_r$ , where  $r, s \in \{1, \dots, K\}$ .
- ② Let  $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n)$  be the set of covariates, where  $\mathbf{X}_i = \{\mathbf{x}_{it}\}_{t=1}^T$  and  $\mathbf{x}_{it}$  is a  $(q+1) \times 1$  vector of covariates for subject  $i$  at time  $t$ .
- ③ Conditional on hidden state  $S_{it} = k$  and covariates  $\mathbf{x}_{it}$ , a generalized linear model for response  $y_{it}$  is considered as

$$\begin{aligned} f(y_{it} | S_{it} = k, \mathbf{x}_{it}, \beta_k) &= \exp \{ (y_{it} \theta_{itk} - b(\theta_{itk})) / a(\phi) + c(y_{it}, \phi) \}, \\ \theta_{itk} &= \mathbf{x}_{it}^\top \beta_k, \end{aligned} \quad (1)$$

- ④ For ease of exposition,  $K$  and  $K_0$  are denoted as the upper bound and true value of the order, respectively. Let  $\Psi = (\pi_1, \pi_2, \dots, \pi_K; P_{11}, \dots, P_{KK}; \beta_1, \beta_2, \dots, \beta_K)$ . Then, the probability mass/density function of  $\mathbf{Y}_i$  can be written as

$$F(\mathbf{Y}_i; \mathbf{X}_i, \Psi) = \sum_{S_{i1}=1}^K \dots \sum_{S_{iT}=1}^K \left[ \prod_{t=1}^T [f(y_{it}; \mathbf{x}_{it}, \beta_{S_{it}})] \pi_{S_{i1}} P_{S_{i1}S_{i2}} \dots P_{S_{i,T-1}S_{iT}} \right]. \quad (2)$$

# Order Estimation

A natural way to estimate the order of RHMM is the maximum likelihood estimate (MLE) of overfitted log-likelihood with the upper bound of order  $K$  ( $K \geq K_0$ ):

$$l_n(\Psi) = \sum_{i=1}^n \log F(Y_i; X_i, \Psi). \quad (3)$$

However, the overfitted MLE leads to an inconsistent estimate of  $K_0$  (Chen and Khalili, 2009; Hung et al., 2013). The overfitting of MLE is of two types:

## Overfitting Types

- **Type I:** near-zero values of mixing probability.
- **Type II:** densities of some components are close to each other.

# Overfitting

$$0.98N(\mu=0, \delta^2=1) + 0.02N(\mu=1, \delta^2=1)$$

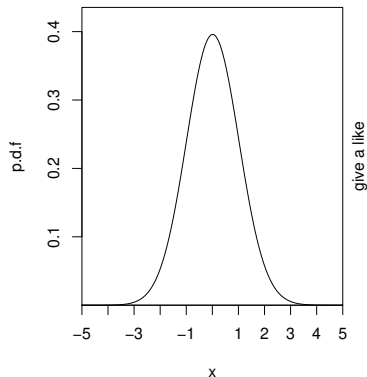


Figure: TYPE I

$$0.5N(\mu=0, \delta^2=1) + 0.5N(\mu=0.3, \delta^2=1)$$

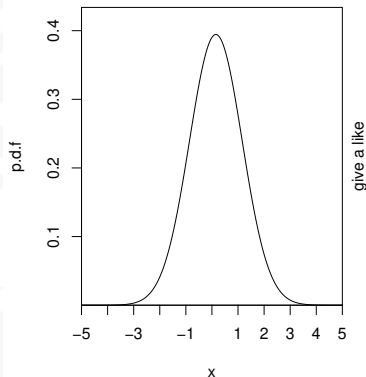


Figure: TYPE II



# Normal Type

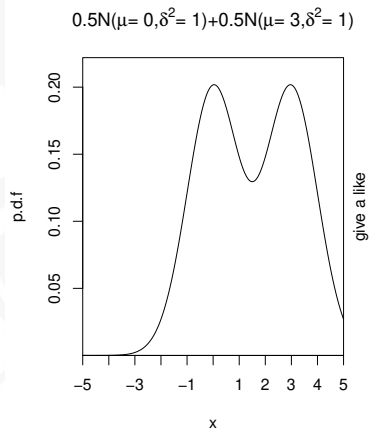


Figure: Normal type

# Penalty

The double penalized log-likelihood can be written as follows:

$$\tilde{l}_n(\Psi) = l_n(\Psi) + \underbrace{C_K \sum_{k=1}^K \log \pi_k}_{\text{Type I}} - \underbrace{n \sum_{k=2}^K p_{\lambda_n}(\|\eta_k\|_2)}_{\text{Type II}}, \quad (4)$$

where  $C_K$  is a tuning parameter,  $p_{\lambda_n}(\cdot)$  is a penalty function, and  $\|\eta_k\|_2 = \|\beta_k - \beta_{k-1}\|_2$ , which will be clarified in the subsequent section. For simplicity, we only consider the SCAD penalty for our model, and it is not essentially hard to implement the MCP and Adaptive LASSO penalties for this setting.

- we proposed a Group-Sort-Fuse procedure to sort the multidimensional parameters of the finite mixture model.

$$\beta_{(k)} = \underset{j \notin \{\beta_{(i)}: 1 \leq i \leq k-1\}}{\operatorname{argmin}} \left\| \beta_j - \beta_{(k-1)} \right\|_2, \quad k = 2, 3, \dots, K, \quad (5)$$

and  $\beta_{(1)} = \underset{k=1,2,\dots,K}{\operatorname{argmax}} \|\beta_k\|_2$ .

- $\hat{\Psi}_n = \operatorname{argmax} \tilde{l}_n(\Psi)$  as the MPLE of  $\Psi$ . Then,

$$\hat{K}_n = \text{number of distinct values of } \{\hat{\beta}_{(k)}, k = 1, \dots, K\} \quad (6)$$

is an estimator of true order  $K_0$ , and we show that  $\hat{K}_n$  converges to  $K_0$  in probability in the subsequent section.

# Asymptotic Result

## Theorem 1

Suppose that RHMM is identifiable and  $F(\mathbf{Y}; \mathbf{X}, \Psi)$  satisfies the mild regular conditions stated in Appendix A. If  $\lambda_n = cn^{-\frac{1}{4}} \log n$  for SCAD penalty and some  $c > 0$ . Then, we have the following:

- (1) For any continuous point of  $\beta^S$  of  $G_0$ , we have  $\hat{G}_n(\beta^S) \xrightarrow{p} G_0(\beta^S)$ .
- (2)  $\sum_{k=1}^K \log \hat{\pi}_k = O_p(1)$  and  $\hat{\alpha}_k = \pi_{0k} + o_p(1)$  for all  $k = 1, 2, \dots, K_0$ . Furthermore, for each  $l = 1, 2, \dots, K$ , a unique  $k = 1, 2, \dots, K_0$  exists, such that  $\|\hat{\beta}_l - \beta_{0k}\|_2 = o_p(1)$ . Thus,  $\{\hat{\nu}_k : k = 1, 2, \dots, K_0\}$  is a cluster partition of  $\{\hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_K\}$  in probability.

## Theorem 2 (Consistency of order selection)

We assume that the same conditions in Theorem 1 hold. Under the true dynamic finite mixture density  $F(\mathbf{Y}; G_0)$ , if  $\hat{G}_n$  falls into an  $O(n^{-\frac{1}{4}})$  neighborhood of  $G_0$ , then  $P(\hat{K}_n = K_0) \rightarrow 1$  as  $n \rightarrow \infty$ .

# ECM-step

In the presence of hidden states, the expectation–maximization (EM) algorithm is known as an efficient statistical estimation method to obtain the maximum likelihood estimate of  $\Psi$ . In this section, we propose an ECM–ITD algorithm to obtain the MPLE  $\hat{\Psi}_n$  of  $\Psi$  in RHMM.

$$\Psi^{(p+1)} = \underset{\Psi}{\operatorname{argmax}} Q(\Psi \mid \Psi^{(p)})$$

$$Q(\Psi \mid \Psi^{(p)}) = E\left(\tilde{\ell}_n^c(\Psi; \mathbf{Y}, \mathbf{S}, \mathbf{X}) \mid \mathbf{Y}, \Psi^{(p)}\right) = \sum_{\mathbf{S}} \tilde{\ell}_n^c(\Psi; \mathbf{Y}, \mathbf{S}, \mathbf{X}) f(\mathbf{S} \mid \mathbf{Y}, \mathbf{X}, \Psi^{(p)})$$

① CM-step 1:

$$\pi_k^{(p+1)} = \frac{\sum_{i=1}^n h^p(S_{i,1} = k) + C_k}{n + KC_k}, \quad (7)$$

$$p_{r,s}^{(p+1)} = \frac{\sum_{i=1}^n \sum_{j=2}^T h^p(S_{i,j-1} = r, S_{i,j} = s)}{\sum_{i=1}^n \sum_{j=2}^T h^p(S_{i,j-1} = r)}. \quad (8)$$

② CM-step 2:

$$\beta^{(p+1)} = \underset{\beta}{\operatorname{argmax}} \sum_{k=1}^K \left[ \sum_{i=1}^n \sum_{t=1}^T \log f(y_{it} \mid S_{it} = k, \mathbf{x}_{it}, \beta) h^{(p+1)}(S_{it} = k) \right] - n \sum_{k=2}^K p_{\lambda_n}(\|\eta_k\|)$$

In order to solve the optimization problem in CM-step 2, we develop an extended ITD algorithm in our multidimensional setting.

- ① Impose a constraint  $\eta_1 = \beta_1$  to form a one-to-one mapping between  $\eta = (\eta_1, \dots, \eta_K)$  and  $\beta$ , i.e.,  $\beta_k = \sum_{l=1}^k \eta_l$  for  $k = 1, 2, \dots, K$ .
- ② we convert the optimization of updating  $\beta^{(p+1)}$  as

$$\eta^{(p+1)} = \underset{\eta}{\operatorname{argmin}} \left\{ G(\eta) = - \sum_{k=1}^K \varphi_k \left( \sum_{l=1}^k \eta_l \right) + n \sum_{k=2}^K \rho_{\lambda_n} (\|\eta_k\|_2) \right\}, \quad (10)$$

where  $\varphi_k(\beta_k) = \sum_{i=1}^n \sum_{t=1}^T \{y_{it} \theta_{itk} - b(\theta_{itk})\} h^{(p+1)}(S_{it} = k)$  and  $\theta_{itk} = \mathbf{x}_{it}^T \beta_k$ .

- ③ Inspired by the prevailing iterative shrinkage-thresholding algorithm (ISTA) for regulated convex optimization problem, we optimize a surrogate function  $\tilde{Q}(\xi; \eta^{(m)})$ :

$$\begin{aligned} \tilde{Q}(\xi; \eta) &= \rho G(\xi) + \frac{1}{2} \sum_{j=1}^K \|\xi_j - \eta_j\|_2^2 \\ &\quad - \rho \left[ \sum_{k=1}^K \sum_{i=1}^n \sum_{t=1}^T \left\{ b(\mathbf{x}_{it}^T \xi_k) - b(\mathbf{x}_{it}^T \beta_k) - b'(\mathbf{x}_{it}^T \beta_k) [\mathbf{x}_{it}^T (\beta_k - \xi_k)] \right\} h_{itk} \right], \end{aligned} \quad (11)$$

CM-step 2 could be reformulated using multivariate thresholding operator  $\vec{\mathcal{S}}(\cdot; 2n, a, \lambda_n)$  as

$$\eta_1^{(m+1)} = \eta_1^{(m)} + \rho \sum_{k=1}^K \sum_{i=1}^n \sum_{j=1}^T h_{ij,k} \left( y_{ij} - b' \left( \mathbf{x}_{ij}^T \beta_k^{(m)} \right) \right) \mathbf{x}_{ij} \quad (12)$$

$$\eta_j^{(m+1)} = \vec{\mathcal{S}} \left( \eta_j^{(m)} + \rho \sum_{k=j}^K \sum_{i=1}^n \sum_{j=1}^T h_{ij,k} \left( y_{ij} - b' \left( \mathbf{x}_{ij}^T \beta_k^{(m)} \right) \right) \mathbf{x}_{ij}; 2n\rho, a, \lambda_n \right), \quad (13)$$

for  $j = 2, 3, \dots, K$ , and then we continue iterating  $\eta^{(m)}$  as the above until it converges.

### Theorem 3

**Convergence of ITD method** Assume that sequence  $\eta^{(m)}$  is generated from (12) and  $\beta_k^{(m)} = \sum_{l=1}^k \eta_l^{(m)}$ . Let  $\tau_1$  be the maximum eigenvalue of  $\sum_{i=1}^N \sum_{t=1}^T \mathbf{x}_{it} \mathbf{x}_{it}^T$  and  $\tau_2^{(m)}$  be assigned to

$$\tau_2^{(m)} = \max_{i,t,k} \sup_{0 < \alpha < 1} b'' \left\{ \mathbf{x}_{it}^T (\alpha \beta^{(m+1)} + (1 - \alpha) \beta^{(m)}) \right\}. \quad (14)$$

If  $\rho^{-1} \geq K \tau_2^{(m)} \tau_1$ , then  $G(\eta^{(m+1)}) \leq G(\eta^{(m)})$ . Furthermore, if space

$\{\eta : G(\eta) \leq G(\eta^{(0)})\}$  is compact, then sequences  $\{\eta^{(m)}\}$  and  $\{\beta^{(m)}\}$  converge to a stationary point of  $G(\eta)$ .

Consider RHMMs with  $K_0 = 2, 3, 4$ . For each setting of  $K_0$ ,  $y_{it}$  in state  $k$  is generated from a normal distribution with mean  $\mathbf{x}_{it}^\top \boldsymbol{\beta}_k$  and standard deviation  $\sigma_k = 0.25$ . Covariates  $\mathbf{x}_{it} = (x_{it1}, x_{it2}, x_{it3})$ , where  $x_{it1} = 1$ , and  $x_{it2}$  and  $x_{it3}$  are independently generated from  $N(0, 1)$  and  $U(0, 1)$ , respectively, where  $U(0, 1)$  stands for the uniform distribution on  $(0, 1)$ . Two sample sizes,  $n = 50$  and  $100$  for normal and a transition matrix with elements  $P_{rs} = \frac{1}{K_0}$ ,  $r, s = 1, \dots, K_0$  are considered. The state-specific regression coefficients for case (1) are set as follows:

- when  $K_0 = 2$ ,  $T = 4$ ,  $\boldsymbol{\beta}_1 = (0, -0.5, 0.2)^\top$ , and  $\boldsymbol{\beta}_2 = (0.5, 0, -0.2)^\top$ ;
- when  $K_0 = 3$ ,  $T = 4$ ,  $\boldsymbol{\beta}_1 = (0, 0.5, 0.2)^\top$ ,  $\boldsymbol{\beta}_2 = (0.5, 0.5, -0.2)^\top$  and  $\boldsymbol{\beta}_3 = (1, -0.5, 0.2)^\top$ ;
- when  $K_0 = 4$ ,  $T = 6$ ,  $\boldsymbol{\beta}_1 = (0, 1, 1.25)^\top$ ,  $\boldsymbol{\beta}_2 = (1, 2, 1)^\top$ ,  $\boldsymbol{\beta}_3 = (1.5, 1.25, 0.75)^\top$ , and  $\boldsymbol{\beta}_4 = (2, 1, 1.5)^\top$ .

# Simulation Result

**Table:** Proportion of order selection for Case (1) in Simulation 1

$K_0$	$\hat{K}_n$	$n = 50$			$n = 100$		
		AIC	BIC	ECM-ITD	AIC	BIC	ECM-ITD
<b>2</b>	<b>2</b>	<b>0.488</b>	<b>0.998</b>	<b>1</b>	<b>0.636</b>	<b>1</b>	<b>1</b>
	3	0.278	0	0	0.232	0	0
	4	0.234	0.002	0	0.132	0	0
<b>3</b>	2	0.010	<b>0.578</b>	0.344	0	0.072	0.026
	<b>3</b>	<b>0.408</b>	0.422	<b>0.654</b>	<b>0.562</b>	<b>0.924</b>	<b>0.974</b>
	4	0.376	0	0.002	0.304	0.004	0
	5	0.206	0	0	0.134	0	0
<b>4</b>	3	0.002	0.474	0.190	0	0.002	0.014
	<b>4</b>	<b>0.614</b>	<b>0.524</b>	<b>0.808</b>	<b>0.690</b>	<b>0.980</b>	<b>0.984</b>
	5	0.384	0.002	0.002	0.310	0	0.02



We analyzed a dataset extracted from the ADNI study using the proposed ESGF procedure to detect the number of hidden phases in the neurodegenerative pathology.

- 1 The study focused on  $n = 616$  subjects collected from the ADNI-I, ADNI-II, and ADNI-Go studies, with four follow-up visits at baseline, 6 months, 12 months, and 24 months ( $T = 4$ ).
- 2 In this study, we treated ADAS13 as the response variable  $y_{it}$  and included some clinical and generic variables as covariates  $x_{it}$  in the proposed RHMM. The covariate vector  $\mathbf{X}_{it} = (x_{it1}, \dots, x_{it6})^T$  included the following variables:
  - $x_{it1} = 1$
  - $x_{it2}$ : age at each visit
  - $x_{it3}$ : gender (1 = female)
  - $x_{it4}$ : logarithm of the ratio of hippocampal volume over the whole brain volume (HIP)
  - $x_{it5}, x_{it6}$ : apolipoprotein E (APOE)- $\epsilon 4$ , which was coded as 0, 1, and 2, denoting the number of APOE- $\epsilon 4$  alleles.

# Results

- Based on the published reports in the AD literature, we set  $K = 7$  as the upper bound for the number of hidden states to implement the proposed procedure.
- Corresponding estimated order  $\hat{K}_n = 5$  was then selected by our methods.

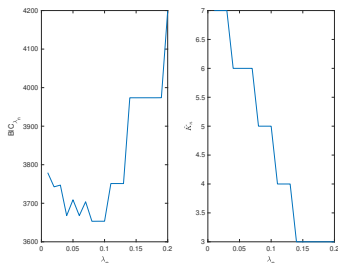


Figure: tuning selection using BIC.

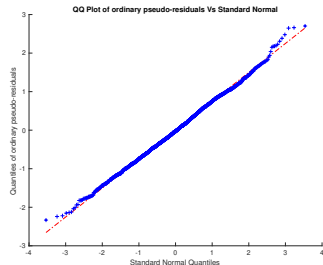


Figure: Residual Check.

**Table:** Estimated coefficients (bootstrap variability estimates) for ADNI study

Par.	State				
	1 (CN)	2 (SMC)	3 (EMCI)	4 (LMCI)	5(AD)
Intercept	-1.007(0.023)	-0.628(0.031)	-0.033(0.029)	0.876(0.035)	2.225(0.167)
$\beta_{k2}$	0.092(0.017)	0.096(0.019)	0.069(0.023)	0.076(0.027)	0.034(0.068)
$\beta_{k3}$	0.037(0.022)	0.111(0.026)	0.023(0.036)	-0.066(0.052)	-0.464(0.168)
$\beta_{k4}$	-0.146(0.013)	-0.257(0.021)	-0.378(0.017)	-0.430(0.025)	-0.347(0.105)
$\beta_{k5}$	0.057(0.036)	0.203(0.042)	0.192(0.037)	0.089(0.046)	0.519(0.215)
$\beta_{k6}$	0.278(0.214)	0.445(0.236)	0.407(0.145)	0.184(0.214)	0.240(0.409)
$\sigma_k$	0.227(0.010)	0.266(0.009)	0.315(0.011)	0.377(0.015)	0.770(0.049)

- state-specific intercept  $\beta_{k1}$  exhibits an ascending trend; patients had the lowest ADAS13 score in state 1, and the highest score in state 5.
- As ADAS13 measures cognitive impairment with a high score indicating low cognitive ability, states 1 to 5 can be explained as CN, significant memory concern (SMC), early mild cognitive impairment (EMCI), late mild cognitive impairment (LMCI), and AD accordingly.
- This classification has been reported in the public literature and ADNI study from ADNI-II to the latest phase (Jessen et al., 2014)
- Some existing studies identify four (CN, EMCI, LMCI, AD) instead of five states. The ADNI-II study suggests that SMC is highly relevant to the AD progression and introducing an additional SMC state minimizes the stratification of cognitive ability and fills the gap between CN and EMCI. Published reports also argued that the introduction of SMC could address the vague demarcation between CN and EMCI (Risacher et al., 2015).

- 2 On the instrumental variable estimation with potentially many (weak) and some invalid instruments
  - Motivation
    - Model Setting
  - Identifiability Condition
    - Example of View of Data Generating
    - Theoretical Results of Solution Structure
    - Surrogate Sparsest Penalty
  - WIT estimator
    - Theoretical Results
    - Simulation
  - Application: Trade and Growth

# Endogeneity

In traditional linear regression analysis:

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon. \quad (15)$$

The basic assumption is  $\mathbf{X}$  is exogenous that  $E(\epsilon|\mathbf{X}) = \mathbf{0}$ . But it will be violated in practice due to following problems:

- Omit Variable (unmeasurement confounder):  $\epsilon = \mathbf{X}_2 + \epsilon'$ ,  $\text{cov}(\mathbf{X}, \mathbf{X}_1) \neq 0$ , and  $E(\epsilon') = \mathbf{0}$ .  $\Rightarrow E(\epsilon^\top \mathbf{X}) = E(\mathbf{X} E(\mathbf{X}_2^\top | \mathbf{X})) \neq 0$
- measurement error in  $\mathbf{X}$ :  $\mathbf{X}^{ob} = \mathbf{X} + \mathbf{u}$ ,  $E(\mathbf{u}) = \mathbf{0}$ . Hence,  $\mathbf{Y} = \mathbf{X}^{ob}\beta - \mathbf{u}\beta + \epsilon$ . Therefore  $E((\epsilon - \mathbf{u}\beta)^\top \mathbf{X}^{ob}) = E((\epsilon - \mathbf{u}\beta)^\top (\mathbf{X} + \mathbf{u})) = -E(\mathbf{u}^\top \mathbf{u})\beta \neq 0$
- Simultaneous Equations

In short, if  $E(\epsilon|\mathbf{X}) \neq 0$  but  $\beta$  is of interest. The  $\mathbf{X}$  is endogenous variable and OLS **can't** provide the consistent estimate:

$$\hat{\beta} = E(\mathbf{X}^\top \mathbf{X})^{-1} E(\mathbf{X} \mathbf{Y}) = \beta + E(\mathbf{X}^\top \mathbf{X})^{-1} E(\mathbf{X} \epsilon) \neq \beta. \quad (16)$$

- What we need is instrumental variables (IVs).

# Requirement of IVs

Good IV should satisfy the following conditions, illustrated as follows.

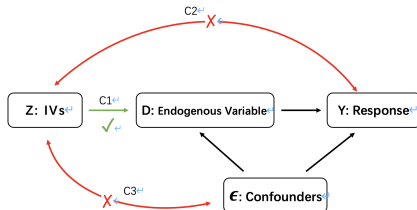


Figure: Illustration of Validity and Relevance.

## IVs Requirements

- 1 **Relevance Condition C1** : related to exposure (may strong or weak).
- 2 **Exogenous Condition C2**: not related to unmeasured variables that affect the exposure and the outcome.
- 3 **Exclusion Restriction C3**: have no direct pathway to the outcome.

## Model

Assuming the linear functional form between treatment effects  $D_i$  and instruments  $Z_i$ .

$$\begin{aligned} Y_i &= D_i\beta + \mathbf{Z}_i^\top \boldsymbol{\alpha} + \epsilon_i \\ D_i &= \mathbf{Z}_i^\top \boldsymbol{\gamma} + \eta_i. \end{aligned} \tag{17}$$

## Definition

- Relevant IV (satisfies C1): if  $\gamma_j^* \neq 0, j = 1, 2, \dots, p$ .
- Valid IV (satisfies C2 and C3): if  $\alpha_j^* = 0, j = 1, 2, \dots, p$ .



# Identifiability of Model

- Exogenous condition of  $\mathbf{Z}$ :

$$\begin{aligned}E(\mathbf{Z}^T \boldsymbol{\varepsilon}) &= E[\mathbf{Z}^T (\mathbf{Y} - \mathbf{Z}\boldsymbol{\alpha}^* - \mathbf{D}\boldsymbol{\beta}^*)] = 0 \\E(\mathbf{Z}^T \mathbf{Y}) &= E(\mathbf{Z}^T \mathbf{Z}) \boldsymbol{\alpha}^* + E(\mathbf{Z}^T \mathbf{D}) \boldsymbol{\beta}^* \\ \Rightarrow \boldsymbol{\Gamma}_{p \times 1}^* &= \boldsymbol{\alpha}_{p \times 1}^* + \boldsymbol{\gamma}_{p \times 1}^* \boldsymbol{\beta}_{1 \times 1}^*,\end{aligned}\tag{18}$$

where  $\boldsymbol{\Gamma}^* = E(\mathbf{Z}^T \mathbf{Z})^{-1} E(\mathbf{Z}^T \mathbf{Y})$  and  $\boldsymbol{\gamma}^* = E(\mathbf{Z}^T \mathbf{Z})^{-1} E(\mathbf{Z}^T \mathbf{D})$

- Both  $\boldsymbol{\Gamma}^*$  and  $\boldsymbol{\gamma}^*$  can be identified based on observed data. But (18) have  $(\boldsymbol{\alpha}_{p \times 1}^*, \boldsymbol{\beta}_{1 \times 1}^*)^T$ , i.e.  $(p+1)$  parameters, need to be determined by  $p$  equations.
- There must have some parameters in  $\boldsymbol{\alpha}^*$  is known first.

## Mixture of valid and Invalid IVs

$\boldsymbol{\alpha}^*$  must contains some 0, but we don't know exact which are. That means we are facing a mixing set of IVs, some of which are valid but some else are not.

# Assumptions

Define the valid IV set  $\mathcal{V}^* = \{j : \alpha_j^* = 0\}$  and invalid IV set  $\mathcal{V}^{c*} = \{j : \alpha_j^* \neq 0\}$ . Let  $L = |\mathcal{V}^*|$ ,  $K = |\mathcal{V}^{c*}|$  and  $p = K + L$ . Notably,  $L \geq 1$  refers to the existence of excluded IV, namely the order condition (Wooldridge, 2010). We consider many (weak) IVs cases and make the following model assumptions:

## Assumptions

**Assumption 1** (Many valid and invalid IVs):  $p < n$ ,  $p_{\mathcal{V}^{c*}}/n \rightarrow v_{p_{\mathcal{V}^{c*}}} + o(n^{-1/2})$  and  $p_{\mathcal{V}^*}/n \rightarrow v_{p_{\mathcal{V}^*}} + o(n^{-1/2})$  for some non-negative constants  $v_{p_{\mathcal{V}^{c*}}}$  and  $v_{p_{\mathcal{V}^*}}$  such that  $0 \leq v_{p_{\mathcal{V}^*}} + v_{p_{\mathcal{V}^{c*}}} < 1$ .

**Assumption 2:** Assume  $\mathbf{Z}$  is standardized. It then has full column rank and  $\|\mathbf{Z}_j\|_2^2 \leq n$  for  $j = 1, 2, \dots, p$ .

**Assumption 3:** Let  $\mathbf{u}_i = (\epsilon_i, \eta_i)^\top$ .  $\mathbf{u}_i \mid \mathbf{Z}_i$  are i.i.d. and follow a multivariate normal distribution with mean zero and positive definite covariance matrix  $\Sigma = \begin{pmatrix} \sigma_\epsilon^2 & \sigma_{\epsilon,\eta} \\ \sigma_{\epsilon,\eta} & \sigma_\eta^2 \end{pmatrix}$ .

The elements of  $\Sigma$  are finite and  $\sigma_{\epsilon,\eta} \neq 0$ .

**Assumption 4** (Strength of valid IVs): The concentration parameter  $\mu_n$  grows at the same rate as  $n$ , i.e.,  $\mu_n \gamma_{\mathbf{Z}_{\mathcal{V}^*}}^* \mathbf{Z}_{\mathcal{V}^*}^\top \mathbf{M}_{\mathbf{Z}_{\mathcal{V}^{c*}}} \mathbf{Z}_{\mathcal{V}^*} \gamma_{\mathbf{Z}_{\mathcal{V}^*}}^* / \sigma_\eta^2 \rightarrow \mu_0 n$ , for some  $\mu_0 > 0$ .

## Example

$$(\text{Structural equation}) \mathbf{Y} = \mathbf{D}\beta^* + \mathbf{Z}_1\alpha_1^* + \mathbf{Z}_2\alpha_2^* + \epsilon \Rightarrow \boldsymbol{\alpha}^* = (\alpha_1^*, \alpha_2^*, 0)$$

$$(\text{First Stage equation}) \mathbf{D} = \mathbf{Z}_1\gamma_1^* + \mathbf{Z}_2\gamma_2^* + \mathbf{Z}_3\gamma_3^* + \eta$$

Then we rearrange first stage equation:

$$\mathbf{Z}_1\gamma_1^* = \mathbf{D} - \mathbf{Z}_2\gamma_2^* - \mathbf{Z}_3\gamma_3^* - \eta$$

$$\Rightarrow \mathbf{Z}_1\alpha_1^* = \mathbf{Z}_1\gamma_1^* \left( \frac{\alpha_1^*}{\gamma_1^*} \right) = \mathbf{D} \frac{\alpha_1^*}{\gamma_1^*} - \mathbf{Z}_2\gamma_2^* \frac{\alpha_1^*}{\gamma_1^*} - \mathbf{Z}_3\gamma_3^* \frac{\alpha_1^*}{\gamma_1^*} - \eta \frac{\alpha_1^*}{\gamma_1^*}$$

$$\Rightarrow \mathbf{Y} = \mathbf{D}\beta^* + \left( \mathbf{D} \frac{\alpha_1^*}{\gamma_1^*} - \mathbf{Z}_2\gamma_2^* \frac{\alpha_1^*}{\gamma_1^*} - \mathbf{Z}_3\gamma_3^* \frac{\alpha_1^*}{\gamma_1^*} - \eta \frac{\alpha_1^*}{\gamma_1^*} \right) + \mathbf{Z}_2\alpha_2^* + \epsilon$$

$$\Rightarrow \mathbf{Y} = \mathbf{D} \left( \beta^* + \frac{\alpha_1^*}{\gamma_1^*} \right) + \mathbf{Z}_2 \left( \alpha_2^* - \frac{\alpha_1^*}{\gamma_1^*} \right) + \mathbf{Z}_3 \left( -\frac{\alpha_1^*}{\gamma_1^*} \right) + \left( \epsilon - \frac{\alpha_1^*}{\gamma_1^*} \eta \right)$$

Then It forms a new DGP:  $\tilde{\beta} = \beta^* + \frac{\alpha_1^*}{\gamma_1^*}$ ,  $\tilde{\boldsymbol{\alpha}} = (0, \alpha_2^* - \frac{\alpha_1^*}{\gamma_1^*}, -\frac{\alpha_1^*}{\gamma_1^*})$  and  $\tilde{\epsilon} = \epsilon - \frac{\alpha_1^*}{\gamma_1^*} \eta$ .

# Identifiability of Model (Cont.)

## Theorem 1

Suppose Assumption 1-4 holds, given  $\mathcal{P}_0$  and  $\{\mathbf{D}, \mathbf{Z}, \boldsymbol{\gamma}^*, \boldsymbol{\eta}\}$ , it can only produce additional  $G = \left| \{c \neq 0 : \alpha_j^* / \gamma_j^* = c, j \in \mathcal{V}^{c*}\} \right|$  groups of different  $\mathcal{P}_c$  such that  $\mathcal{V}^* \cup \{\cup_{c \neq 0} \mathcal{I}_c\} = \{1, 2, \dots, p\}$ ,  $\mathcal{V}^* \cap \{\cup_{c \neq 0} \mathcal{I}_c\} = \emptyset$  and  $E(\mathbf{Z}^\top \tilde{\epsilon}^c) = \mathbf{0}$ . The sparsity structure regarding  $\boldsymbol{\alpha}$  is non-overlapping for different solutions.

Theorem 1 tells us there is a collection of model DGPs

$$\mathcal{Q} = \left\{ \mathcal{P} = \{\beta, \boldsymbol{\alpha}, \epsilon\} : \boldsymbol{\alpha} \text{ is sparse, } E(\mathbf{Z}^\top \epsilon) = \mathbf{0} \right\}$$

corresponding to the same observation  $\mathbf{Y}$  conditional on first stage information. Let  $\mathcal{H}$  be a collection of mappings  $h : \mathcal{Q} \rightarrow \mathcal{P} \in \mathcal{Q}$

## Theorem 2

Under same conditions in Theorem 1, let  $\mathcal{F} = \{f : \mathcal{P} \in \mathcal{Q} \rightarrow \mathbb{R}; f(\mathcal{P}_i) < f(\mathcal{P}_j) \forall j \neq i \text{ and } \exists i \in \{0, \dots, G\}\}$  and  $\mathcal{G} = \{g = \operatorname{argmin}_{\mathcal{P} \in \mathcal{Q}} f(\mathcal{P}); f \in \mathcal{F}\}$ , then we obtain:

(a)  $\mathcal{G} \subseteq \mathcal{H}$ .

(b) There never exist a necessary condition of identifying  $(\boldsymbol{\alpha}^*, \beta^*)$  unless  $\exists h \in \mathcal{H} : \mathcal{Q} \rightarrow \mathcal{P}_0$  and  $|\mathcal{H}| = 1$ .

# Explanation of Theorem 1 & 2

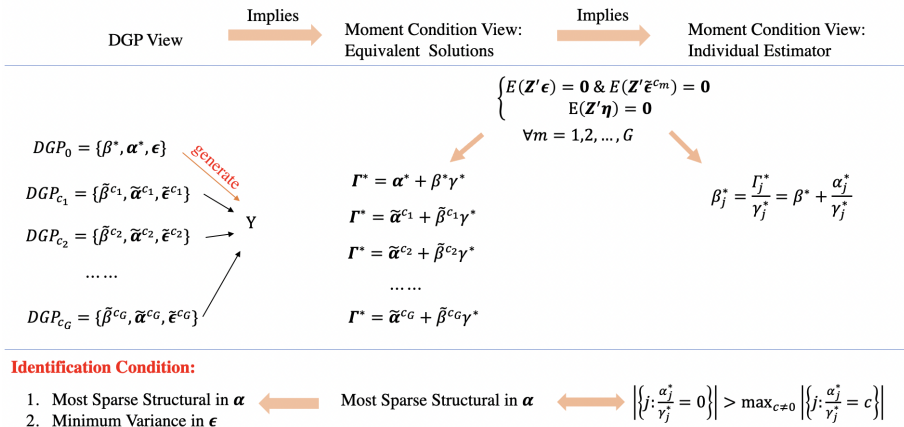


Figure: Explanation of Theorem

# Identifiability of Model

Let  $\mathcal{Q}$  to be the collections of all possible population-level solution of  $\alpha$  under the constraint, some components need to be 0.

## The Sparsest Rule

$$\alpha^* = \operatorname{argmin}_{\mathcal{P}=\{\beta, \alpha, \epsilon\} \in \mathcal{Q}} \|\alpha\|_0$$

Consider the general penalized TSLS estimator with penalty on  $\alpha$ :

$$\left( \hat{\alpha}^{\text{pen}}, \hat{\beta}^{\text{pen}} \right) = \operatorname{argmin}_{\alpha, \beta} \underbrace{\frac{1}{2n} \|P_Z(\mathbf{Y} - \mathbf{Z}\alpha - \mathbf{D}\beta)\|_2^2}_{(I)} + \underbrace{p_{\lambda}^{\text{pen}}(\alpha)}_{(II)}. \quad (19)$$

They have the two different functions: (I) approximate the structure of  $\mathcal{Q}$  and (II) imposed to ensure sparsity structure in  $\alpha$ .

## Dual Form View (Aligned with Sparsest Rule)

$$\left( \hat{\alpha}^{\text{opt}}, \hat{\beta}^{\text{opt}} \right) = \operatorname{argmin}_{\alpha, \beta} \|\alpha\|_0 \quad \text{s.t.} \quad \underbrace{\|P_Z(\mathbf{Y} - \mathbf{D}\beta - \mathbf{Z}\alpha)\|_2^2}_{\text{only possible in } \mathcal{Q}} < \delta.$$

# Surrogate Sparsest penalty

## Proposition 1 (The proper Surrogate Sparsest penalty)

Suppose Assumptions 1-7 are satisfied. If  $p_\lambda^{\text{pen}}(\alpha)$  is surrogate sparsest rule in the sense of that it gives sparse solutions and

$$\alpha^* = \underset{\mathcal{P}=\{\beta, \alpha, \epsilon\} \in \mathcal{Q}}{\operatorname{argmin}} \|\alpha\|_0 = \underset{\mathcal{P}=\{\beta, \alpha, \epsilon\} \in \mathcal{Q}}{\operatorname{argmin}} p_\lambda^{\text{pen}}(\alpha),$$

then  $p_\lambda^{\text{pen}}(\cdot)$  must be concave and  $p_\lambda^{\text{pen}}(t) = O(\lambda \kappa(n))$  for any  $t > \kappa(n)$ .

## Example

- Consider  $\alpha^* = (0, 0, 1)^\top, \gamma^* = (1, 1, 3)^\top$ . Hence, it produces another solution  $\tilde{\alpha} = (-\frac{1}{3}, -\frac{1}{3}, 0)^\top$ .
- $\mathcal{Q} = \{\alpha^*, \tilde{\alpha}\}$
- It satisfy the sparsest rule that  $\alpha^* = \underset{\mathcal{P}=\{\beta, \alpha, \epsilon\} \in \mathcal{Q}}{\operatorname{argmin}} \|\alpha\|_0$ .
- However, using  $l_1$ ,  $\|\alpha^*\|_1 = 1 > \|\tilde{\alpha}\|_1 = \frac{2}{3}$

We adopt the penalized method framework (10) and deploy a concave penalty in (11), the MCP in particular, which is nearly unbiased estimator.

## WIT Estimator (Two step method)

$$\text{Selection Stage: } \hat{\alpha}^{\text{MCP}} = \underset{\alpha}{\operatorname{argmin}} \frac{1}{2n} \|\mathbf{Y} - \tilde{\mathbf{Z}}\alpha\|_2^2 + p_{\lambda}^{\text{MCP}}(\alpha),$$

$$\hat{\mathcal{V}} = \{j : \hat{\alpha}_j^{\text{MCP}} = 0\}, \text{ with } \Pr(\hat{\mathcal{V}} = \mathcal{V}^*) \xrightarrow{P} 1,$$

$$\begin{aligned} \text{Estimation Stage : } \hat{\beta}^{\text{WIT}} &= \left( \mathbf{D}_{\perp}^{\top} (\mathbf{I} - \hat{\kappa}_{\text{liml}} \mathbf{M}_{\mathbf{Z}_{\hat{\mathcal{V}}_{\perp}}}) \mathbf{D}_{\perp} \right)^{-1} \left( \mathbf{D}_{\perp}^{\top} (\mathbf{I} - \hat{\kappa}_{\text{liml}} \mathbf{M}_{\mathbf{Z}_{\hat{\mathcal{V}}_{\perp}}}) \mathbf{Y}_{\perp} \right) \\ \hat{\kappa}_{\text{liml}} &= \lambda_{\min} \left( \{[\mathbf{Y}_{\perp}, \mathbf{D}_{\perp}]^{\top} \mathbf{M}_{\mathbf{Z}_{\hat{\mathcal{V}}_{\perp}}} [\mathbf{Y}_{\perp}, \mathbf{D}_{\perp}]\}^{-1} \{[\mathbf{Y}_{\perp}, \mathbf{D}_{\perp}]^{\top} [\mathbf{Y}_{\perp}, \mathbf{D}_{\perp}]\} \right) \end{aligned}$$

- 1 The above estimation is derived based on residual model  $\mathbf{Y}, \mathbf{D}, \mathbf{Z}_{\hat{\mathcal{V}}}$  on  $\mathbf{Z}_{\hat{\mathcal{V}}^c}$ .
- 2 Let  $\mathbf{Y}_{\perp} = \mathbf{M}_{\mathbf{Z}_{\hat{\mathcal{V}}^c}} \mathbf{Y}$ ,  $\mathbf{D}_{\perp} = \mathbf{M}_{\mathbf{Z}_{\hat{\mathcal{V}}^c}} \mathbf{D}$  and  $\mathbf{Z}_{\hat{\mathcal{V}}_{\perp}} = \mathbf{M}_{\mathbf{Z}_{\hat{\mathcal{V}}^c}} \mathbf{Z}_{\hat{\mathcal{V}}}$  and notice  $\mathbf{M}_{\mathbf{Z}_{\hat{\mathcal{V}}^c}} \mathbf{M}_{\mathbf{Z}_{\hat{\mathcal{V}}_{\perp}}} = \mathbf{M}_{\mathbf{Z}}$ .



# Selection Consistency

## Theorem 3 (Selection Consistency of Valid IVs)

Specify  $\kappa(n)$  and  $\kappa^c(n)$  in Assumption 5 as

$$\kappa(n) \asymp \underbrace{\sqrt{\frac{\log p_{\mathcal{V}^*}}{n}}}_{T_1} + \underbrace{\frac{p_{\mathcal{V}^*}}{n} \cdot \frac{\|\tilde{\mathbf{Q}}_n \gamma_{\mathcal{V}^*}^*\|_\infty}{\gamma_{\mathcal{V}^*}^{*\top} \tilde{\mathbf{Q}}_n \gamma_{\mathcal{V}^*}^*}}_{T_2} + \underbrace{|\text{Bias}(\hat{\beta}_{\text{or}}^{\text{TSLs}})| \|\tilde{\gamma}_{\mathcal{V}^*}^*\|_\infty}_{T_3}, \quad (20)$$

$$\kappa^c(n) \asymp (1+c) \left\{ \sqrt{\frac{\log |\mathcal{I}_c|}{n}} + \frac{|\mathcal{I}_c|}{n} \cdot \frac{\|\tilde{\mathbf{Q}}_n^c \gamma_{\mathcal{I}_c}^*\|_\infty}{\gamma_{\mathcal{I}_c}^{*\top} \tilde{\mathbf{Q}}_n^c \gamma_{\mathcal{I}_c}^*} \right\} + |\text{Bias}(\hat{\beta}_{\text{or}}^{c, \text{TSLs}})| \|\tilde{\gamma}_{\mathcal{I}_c}^*\|_\infty, \quad (21)$$

Moreover, under Assumption 1-6, consider computable local solutions, then

$$\hat{\alpha}^{\text{MCP}} = \underset{\hat{\alpha} \in \mathcal{B}_0(\lambda, \rho)}{\text{argmin}} \|\hat{\alpha}\|_0, \quad \Pr(\hat{\nu} = \nu^*, \hat{\alpha}^{\text{MCP}} = \hat{\alpha}^{\text{or}}) \xrightarrow{p} 1. \quad (22)$$

- ① T1: Common rate in Lasso/non-convex penalty in ordinal linear regression.
- ② T2 and T3: Additional difficulty in penalized regression within IV contents:
  - T2: many IVs risk,
  - T3: bias in weak IVs problem.

## Remark of $\kappa(n)$

### Proposition 1 (Magnitude of $T_2$ )

If there does not exist dominant scaled  $\gamma_j^*$ , i.e.

$\|\tilde{\mathbf{Q}}_n^{1/2} \gamma_{\mathcal{V}^*}^*\|_\infty / \|\tilde{\mathbf{Q}}_n^{1/2} \gamma_{\mathcal{V}^*}^*\|_1 = o\left(\|\tilde{\mathbf{Q}}_n^{1/2} \gamma_{\mathcal{V}^*}^*\|_1 / (p_{\mathcal{V}^*} \|\tilde{\mathbf{Q}}_n^{1/2}\|_\infty)\right)$ , then  $T_2 \rightarrow 0$ .

### Proposition 2 (Approximation of Bias( $\hat{\beta}_{or}^{TSLs}$ ))

Let  $s = \max(\mu_n, p_{\mathcal{V}^*})$ , under the Assumptions 1-4, we obtain

$$E \left[ \text{Bias}(\hat{\beta}_{or}^{TSLs}) \right] = \frac{\sigma_{\epsilon\eta}}{\sigma_\eta^2} \left( \frac{p_{\mathcal{V}^*}}{(\mu_n + p_{\mathcal{V}^*})} - \frac{2\mu_n^2}{(\mu_n + p_{\mathcal{V}^*})^3} \right) + o\left(s^{-1}\right). \quad (23)$$

### Discussion on $T_3$

The rate of concentration parameter  $\mu_n$  will affect  $T_3$  through  $|\text{Bias}(\hat{\beta}_{or}^{TSLs})|$  under many IVs setting. Suppose Assumption 4 holds, that  $\mu_n \xrightarrow{P} \mu_0 n$ , the leading term in (23) is

$$\frac{\sigma_{\epsilon\eta}}{\sigma_\eta^2} \frac{\nu_{p_{\mathcal{V}^*}}}{\mu_0 + \nu_{p_{\mathcal{V}^*}}} \ll \frac{\sigma_{\epsilon\eta}}{\sigma_\eta^2} \text{ for moderate } \mu_0.$$

## Theorem 4 (Consistency and Asymptotic Normality)

Under same condition in Theorem 3, together with Assumption A5 and conditional on  $\mathbf{Z}$ , we obtain:

- ① (Consistency):  $\hat{\beta}^{\text{WIT}} \xrightarrow{p} \beta^*$  with  $\hat{\kappa}_{\text{liml}} = \frac{1-v_L}{1-v_K-v_L} + o_p(1)$ .
- ② (Asymptotic normality):  $\sqrt{n}(\hat{\beta}^{\text{WIT}} - \beta^*) \xrightarrow{d} \mathcal{N}\left(0, \mu_0^{-2} [\sigma_\epsilon^2 \mu_0 + \frac{v_K(1-v_L)}{1-v_K-v_L} |\boldsymbol{\Sigma}|]\right)$ .
- ③ (Consistent variance estimator):

$$\widehat{\text{Var}}(\hat{\beta}^{\text{WIT}}) = \frac{\hat{\mathbf{b}}^\top \hat{\boldsymbol{\Omega}} \hat{\mathbf{b}} (\hat{\mu}_n + L/n)}{-\hat{\mu}_n} \left( \hat{Q}_S \hat{\boldsymbol{\Omega}}_{22} - \boldsymbol{\tau}_{22} + \frac{\hat{c}}{1 - \hat{c}} \frac{\hat{Q}_S}{\hat{\mathbf{a}}^\top \hat{\boldsymbol{\Omega}}^{-1} \hat{\mathbf{a}}} \right)^{-1} \\ \xrightarrow{p} \mu_0^{-2} \left[ \sigma_\epsilon^2 \mu_0 + \frac{v_K(1-v_L)}{1-v_K-v_L} |\boldsymbol{\Sigma}| \right],$$

where  $\hat{\mathbf{b}} = (1, -\hat{\beta}^{\text{WIT}})$  and  $\hat{Q}_S = \frac{\hat{\mathbf{b}}^\top \boldsymbol{\tau} \hat{\mathbf{b}}}{\hat{\mathbf{b}}^\top \hat{\boldsymbol{\Omega}} \hat{\mathbf{b}}}$ .

# Simulation

Further, we present a replication of simulation design in literature and its variant:

Case 1( III ) :  $\gamma^* = (\mathbf{0.4}_{21})^\top$  and  $\alpha^* = (\mathbf{0}_9, \mathbf{0.4}_6, \mathbf{0.2}_6)^\top$ .

Case 1(IV) :  $\gamma^* = (\mathbf{0.15}_{21})^\top$  and  $\alpha^* = (\mathbf{0}_9, \mathbf{0.4}_6, \mathbf{0.2}_6)^\top$ .

**Table:** Simulation results in low dimension: A replication of experiment

Case	Approaches	$n = 500$				$n = 1000$			
		MAD	CP	FPR	FNR	MAD	CP	FPR	FNR
1(III)	TSLs	0.436	0	-	-	0.435	0	-	-
	oracle-LIML	0.021	0.932	-	-	0.014	0.944	-	-
	TSHT	0.142	0.404	0.398	0.150	0.016	0.924	0.023	0.004
	CIIV	0.037	0.710	0.125	0.032	0.017	0.894	0.031	0.002
	sisVIVE	0.445	-	0.463	0.972	0.465	-	0.482	0.999
	Post-Alasso	0.436	0	1	0	0.435	0	0.999	0
	WIT	0.036	0.708	0.121	0.099	0.016	0.910	0.020	0.027
1(IV)	TSLs	1.124	0	-	-	1.144	0	-	-
	oracle-LIML	0.056	0.948	-	-	0.042	0.948	-	-
	TSHT	0.532	0.058	0.342	0.457	0.155	0.660	0.310	0.208
	CIIV	1.213	0.224	0.337	0.670	0.100	0.526	0.300	0.426
	sisVIVE	1.101	-	0.392	0.936	1.175	-	0.428	0.996
	Post-Alasso	1.112	0	0.945	0.010	1.029	0	0.652	0.205
	WIT	0.102	0.634	0.198	0.220	0.048	0.844	0.068	0.090

# Simulation (Cont.)

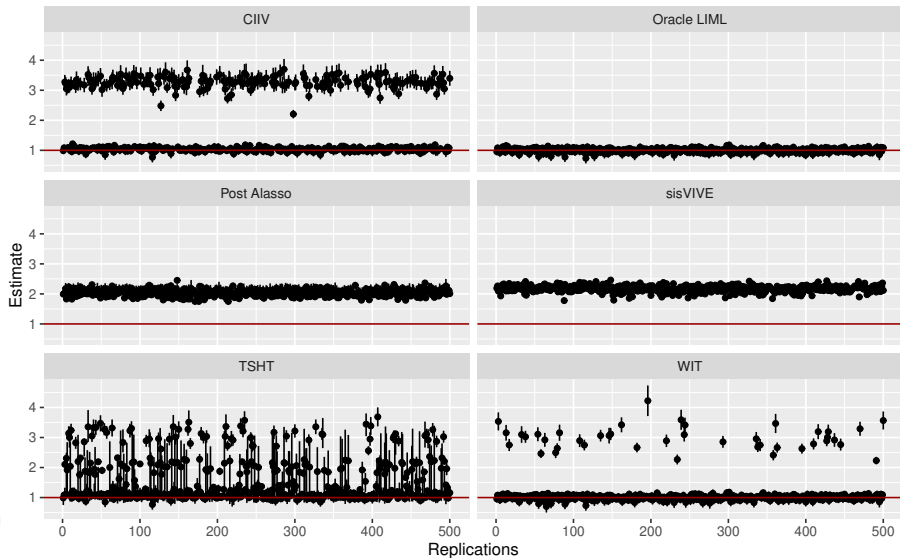


Figure: Scatter plot of estimations of  $\beta^*$  with confidence intervals of Case 1 (IV)

# Real Application

- We revisit the classic empirical study in trade and growth (Frankel and Romer, 1999, FR99 henceforth).
- We investigate the causal effect of trade on income using a more comprehensive and updated data, taking into account that trade is an endogenous variable (it correlates with unobserved common factors driving both trade and growth) and some instruments might be invalid.
- The structural equation considered in FR99 is,

$$\log(Y_i) = \alpha + \beta T_i + \psi S_i + \epsilon_i \quad (24)$$

where for each country  $i$ ,  $Y_i$  is the GDP per worker,  $T_i$  is the share of international trade to GDP,  $S_i$  is the size of the country, such as area, population, and  $\epsilon_i$  is the error term.

- FR99 proposed to construct an IV (called a proxy for trade) based on the celebrated gravity theory of trade (Anderson, 1979). The logic of IV validity in aggregate level is that the geographical variables, such as common border and distance between countries, indirectly affect growth through the channel of convenience for trade.

# Trade on GDP

Following the same logic, Fan and Zhong (2018) extended the IV set to include more geographic and meteorological variables. The reduced form equation is

$$T_i = \gamma^\top \mathbf{Z}_i + \nu_i, \quad (25)$$

where  $\mathbf{Z}_i$  is a vector of instruments.

**Table:** Summary statistics of main variables

	Notation	Type	Mean	Std	Median	Min	Max
log(GDP)	log(Y)	Response	10.177	1.0102	10.416	7.463	12.026
Trade	$T$	Endogenous Variable	0.866	0.520	0.758	0.198	4.128
log(Population)	$S_1$	Control Variable	1.382	1.803	1.480	-3.037	6.674
log(Land Area)	$S_2$	Control Variable	11.726	2.260	12.015	5.680	16.611
$\hat{T}$ (proxy for trade)	$Z_1$	IV	0.093	0.052	0.079	0.015	0.297
log(Water Area)	$Z_2$	IV	6.756	3.654	7.768	0	13.700
log(Land Boundaries)	$Z_3$	IV	6.507	2.920	7.549	0	10.005
% Forest	$Z_4$	IV	29.89	22.380	30.62	0	98.26
% Arable Land	$Z_5$	IV	40.947	21.549	42.062	0.558	82.560
Languages	$Z_6$	IV	1.873	2.129	1	1	16
Annual Freshwater	$Z_7$	IV	2.190	2.129	2.155	-2.968	8.767

Source: FR99, the World Bank, and CIA world Factbook.

**Table:** Empirical Results of Various Estimators

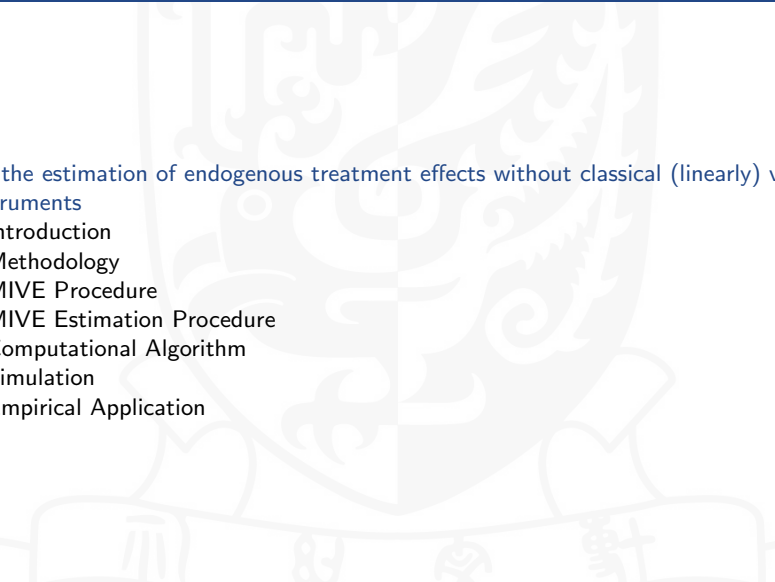
	$\hat{\beta} \left( \widehat{\text{Var}}^{1/2}(\hat{\beta}) \right)$	95% CI	Valid IVs $\hat{\mathcal{V}}$	Relevant IVs $\hat{\mathcal{S}}$	Sargan Test
OLS	0.413(0.084)	(0.246, 0.581)	-	-	-
FR99	0.673(0.220)	(0.228, 1.117)	-	-	0.999
LIML	2.969(1.503)	(0.023, 5.916)	-	-	0.001
TSHT	0.861(0.245)	(0.380, 1.342)	{1}	{1}	0.999
CIIV*	2.635(1.974)	(-1.233, 6.504)	{2,4,5,6,7}	-	0.385
sisVIVE	0.819(-)	-	{1,2,4}	-	0.418
Post-Alasso	0.964(0.251)	(0.471, 1.457)	{1,2,4,5,6}	-	0.086
WIT	0.974(0.323)	(0.340, 1.609)	{1,2,4,6}	-	0.275

Note: CIIV\* stands for CIIV method without first stage IVs selection because it reports that “Less than two IVs are individually relevant, treat all IVs as strong”. Sargan test means  $p$ -value of Sargan test and selection of relevant IVs  $\hat{\mathcal{S}}$  is only be implemented in TSHT and CIIV.



## Results:

- $p$ -value of the Hausman test for endogeneity is 0.000181 using the proxy for trade as IV.
- LIML using all potential IVs (without distinguishing the invalid ones) likely overestimates the treatment effect. The 0.001  $p$ -value of Sargan test strongly reject the null of all potential IVs are valid.
- $Z_5$  should be a invalid IV:
  - 1 Concerning the Sargan Test,  $p$ -value of  $0.086 < 0.1$  in Post-Alasso indicates  $Z_5$  is not very credible to be valid.
  - 2 In the economic perspective, more arable land generates higher crop yields and maintains a higher agriculture sector labor force, which directly affects GDP.
- Strong IVs based method fails.

- 
- 3 On the estimation of endogenous treatment effects without classical (linearly) valid instruments
    - Introduction
    - Methodology
    - MIVE Procedure
    - MIVE Estimation Procedure
    - Computational Algorithm
    - Simulation
    - Empirical Application

# Background

## Background and Introduction:

- Empirical research often faces endogeneity issues due to unmeasured confounders
- Difficulties in verifying strict assumptions for valid IV inference
- Non-linear relationships between IVs, treatment, and outcome variables

## Proposed Approach:

- MIVE identifies **locally** valid IVs that satisfy the conditions:
  - Have a non-zero effect on the treatment variable locally
  - Have zero or negligible effect on the outcome variable locally
- MIVE selects the IV that **maximizes** a concentration parameter which measures the information contained in the IV for identifying causal effects.

## Advantages:

- **Accuracy:** Improved causal inference by considering non-linear and local effects
- **Interpretability:** Identify relevant and valid local IVs, examine functional forms and interactions
- **Insights:** Contribute to the accurate and explainable machine learning in empirical econometric studies.

# Model

Consider the following model:

$$\begin{aligned}Y_i &= D_i \beta^* + f^*(Z_i) + \epsilon_i, \\ D_i &= g^*(Z_i) + \eta_i,\end{aligned}$$

where

- $(Y_i, D_i, Z_i) \in \mathbb{R}^{1 \times 1 \times p}$  are i.i.d. observed samples,  $f^*$  and  $g^*$  are unknown function mapping  $\mathbb{R}^p$  to  $\mathbb{R}^1$  and  $\epsilon_i$  and  $\eta_i$  are error terms.
- Error terms satisfy the following conditions:

$$\mathbb{E} \left( \begin{pmatrix} \epsilon_i \\ \eta_i \end{pmatrix} \mid Z_i \right) = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \quad \text{and} \quad \text{Var} \left( \begin{pmatrix} \epsilon_i \\ \eta_i \end{pmatrix} \mid Z_i \right) = \begin{pmatrix} \sigma_\epsilon^2 & \rho \sigma_\epsilon \sigma_\eta \\ \rho \sigma_\epsilon \sigma_\eta & \sigma_\eta^2 \end{pmatrix}.$$

The above  $\rho \neq 0$  measures the endogeneity problem inherent in  $D_i$ .

# General Additive Model

To simplify the identification problem, we assume sufficient smoothness of  $f^*(Z_i)$  and  $g^*(Z_i)$ , and employ a general additive model with suitable spline or basis functions for approximation.

## General Additive Model

For  $j = 1, 2, \dots, p$ , express the model as follows:

$$Y_i = D_i \beta^* + \sum_{j=1}^p \sum_{k=1}^K \psi_{jk}(Z_{ij}) \alpha_{jk} + \epsilon_i,$$

$$D_i = \sum_{j=1}^p \sum_{k=1}^K \psi_{jk}(Z_{ij}) \gamma_{jk} + \eta_i.$$

For convenience and conciseness, re-express the equations in matrix form:

$$Y = D \beta^* + \Psi \alpha + \epsilon,$$

$$D = \Psi \gamma + \eta.$$

where  $\Psi = (\psi_{11}(Z_1), \psi_{12}(Z_1), \dots, \psi_{pK}(Z_p)) \in \mathbb{R}^{n \times pK}$ ,  $\alpha = (\alpha_{11}, \alpha_{12}, \dots, \alpha_{pK})^T \in \mathbb{R}^n$  and  $\gamma = (\gamma_{11}, \gamma_{12}, \dots, \gamma_{pK})^T \in \mathbb{R}^n$ .

# Local Identification

- To address the problem of invalid or linearly valid IVs, a local identification condition is proposed to identify  $\beta$  using heterogeneity in local exposure conditional means of  $Y_i$  and  $D_i$  based on the local structure of potential IVs.
- A subset  $\Xi \subseteq \sigma(\mathbf{Z}_i)$  that satisfies:

$$g^*(\Xi) \neq 0, \quad f^*(\Xi) = 0 \quad \text{and} \quad g^*(\Xi) \neq \mathbb{E}(g^*(\Xi)|\Xi^c)$$

is referred to as locally identifying  $\beta$ .

- If the set  $\Xi$  is known in advance,  $\beta$  can be identified as:

$$\beta^* = \frac{\mathbb{E}(Y_i|\mathbf{Z}_i) - \mathbb{E}(Y_i|\Xi^c)}{g^*(\Xi) - \mathbb{E}(g^*(\Xi)|\Xi^c)}$$

- The local identification condition is linked to B-spline expansion by assuming  $\Xi$  can be decomposed into disjoint subsets corresponding to knot intervals which can be used to expand  $f$  and  $g$  via B-splines.

# Most Informative Principle

- **Problem:** Identifying valid IVs without specific information on which IVs are valid.
- **Goal:** Estimate  $\beta$  by identifying the true data-generating process ( $\mathcal{P}$ ) in the collection  $\mathcal{Q}$ .

$$\mathcal{Q} = \{\mathcal{P} = \{\beta, \alpha\} : \alpha_j = 0 \text{ for some } j, E(\Psi^\top \epsilon) = \mathbf{0}\} \quad (26)$$

- **Approaches:** From majority rule to sparsest rule (plurality rule), conditions have been gradually relaxed.
- **New Mechanism:** Allows for greater flexibility, expanding the scope of potential IVs with varying exposure to  $Y$  and  $D$  at a local and non-linear level.

## Concentration Parameter (CP)

- **Definition:**  $\mathbb{E}\delta(\mathcal{P}) = \frac{\gamma_{\mathcal{V}^{\mathcal{P}}}^\top \Psi_{\mathcal{V}^{\mathcal{P}}}^\top \mathbb{M}_{\Psi_{\mathcal{V}^{\mathcal{P}}}} \Psi_{\mathcal{V}^{\mathcal{P}}} \gamma_{\mathcal{V}^{\mathcal{P}}}}{\sigma_\eta^2}$
- **Importance:** Measures the available information for estimating  $\beta^* \in \mathcal{P}$  via TSLS estimator and plays a crucial role in determining the bias of such estimator.

# Most Informative Principle (Cont')

## Definition

The true data-generating process  $\mathcal{P} \in \mathcal{Q}$  satisfy the most informative principle if

$$\mathcal{P}^* = \operatorname{argmax}_{\mathcal{P} \in \mathcal{Q}} \mathbb{E} \delta(\mathcal{P}). \quad (27)$$

- **Advantages:** More reliable and robust results, not contingent on choices of the number of expansions used.
- **Connection to Sparsest Rule:** Can be viewed as an extension, reduces to Sparsest Rule under certain conditions.



# Construction of $\hat{\mathcal{Q}}$

- **Objective:** Check the validity of IV regression models.
- **Moment Condition:**  $\mathbb{H}_0 : \mathbb{E}(\Psi^\top \epsilon) = \mathbf{0}$  vs  $\mathbb{H}_1 : \mathbb{E}(\Psi^\top \epsilon) \neq \mathbf{0}$ .

## Construct $\hat{\mathcal{Q}}$

$$\hat{\mathcal{Q}} = \{(\hat{\beta}, \hat{\alpha}) : \|\mathbb{P}_{\mathbf{Z}}(\mathbf{Y} - \mathbf{D}\hat{\beta} - \mathbf{Z}\hat{\alpha})\|_2^2 \leq \min_{(\beta, \alpha) \in \mathcal{B}_\epsilon(\hat{\beta}, \hat{\alpha})} \|\mathbb{P}_{\mathbf{Z}}(\mathbf{Y} - \mathbf{D}\beta - \mathbf{Z}\alpha)\|_2^2 \text{ s.t. } \|\alpha\|_0 < pK\}$$

Why elements in  $\hat{\mathcal{Q}}$  satisfy the Sargan Test :

$$S(\mathcal{V}^{\mathcal{P}}) = \frac{\|\mathbb{P}_{\mathbf{Z}}(\mathbf{Y} - \mathbf{D}\hat{\beta}_{\mathcal{V}^{\mathcal{P}}}^{\text{TSLs}} - \mathbf{Z}_{\mathcal{V}_c^{\mathcal{P}}} \hat{\alpha}_{\mathcal{V}_c^{\mathcal{P}}}^{\text{TSLs}})\|_2^2}{\|(\mathbf{Y} - \mathbf{D}\hat{\beta}_{\mathcal{V}^{\mathcal{P}}}^{\text{TSLs}} - \mathbf{Z}_{\mathcal{V}_c^{\mathcal{P}}} \hat{\alpha}_{\mathcal{V}_c^{\mathcal{P}}}^{\text{TSLs}})/\sqrt{n}\|_2^2}. \quad (28)$$

If  $\mathcal{V}^{\mathcal{P}}$  matches support of  $\alpha$  for someone in  $\mathcal{Q}$ , then  $S(\mathcal{V}^{\mathcal{P}}) = o_p(1)$ . Otherwise,  $S(\mathcal{V}^{\mathcal{P}}) = O_p(n)$  and cannot be asymptotically bounded by any finite constant.

# CP Approximation & Tied Constraint

## CP Approximation

- The most informative principle (MIP) is expressed as:

$$\hat{\mathcal{P}} = \operatorname{argmax}_{\mathcal{P} \in \hat{\mathcal{Q}}} \gamma_{\mathcal{V}^{\mathcal{P}}}^{\top} \Psi_{\mathcal{V}^{\mathcal{P}}}^{\top} \mathbb{M}_{\Psi_{\mathcal{V}^{\mathcal{C}}}} \Psi_{\mathcal{V}^{\mathcal{P}}} \gamma_{\mathcal{V}^{\mathcal{P}}} = \operatorname{argmax}_{\mathcal{P} \in \hat{\mathcal{Q}}} \left( \min_{\omega_{\mathcal{V}^{\mathcal{C}}}} \|\Psi \gamma - \Psi_{\mathcal{V}^{\mathcal{C}}} \omega_{\mathcal{V}^{\mathcal{C}}}\|_2^2 \right),$$

where projection matrix  $\mathbb{M}_{\Psi_{\mathcal{V}^{\mathcal{C}}}}$  is replaced with  $\Psi \gamma - \Psi_{\mathcal{V}^{\mathcal{C}}} \hat{\omega}_{\mathcal{V}^{\mathcal{C}}}$ .

## Tied Constraint

- Reformulating the minimization in (50) to a constrained version:

$$(\hat{\beta}, \hat{\alpha}) = \operatorname{argmax}_{(\beta, \alpha) \in \hat{\mathcal{Q}}} \left( \min_{\omega} \|\Psi \gamma - \Psi \omega\|_2^2 \right) \quad \text{s.t.} \quad \alpha \odot \omega = \mathbf{0},$$

- Introducing the **tied constraint** to relax the original constraint:

$$\text{Tied constraint:} \quad \|\alpha \odot \omega\|_1 \leq R$$

- Tied constraint approximates subspace spanned by columns  $\Psi_{\mathcal{V}^{\mathcal{C}}}$  and facilitates optimization algorithm design.

# Computation of MIVE

The MIVE procedure consists of three steps:

- **Step 1:** Choosing proper spline basis  $\Psi$ , conduct first stage sparse regression using MCP penalty to identify candidate IV set:

$$\hat{\gamma} = \underset{\gamma}{\operatorname{argmin}} \frac{1}{2n} \|D - \Psi\gamma\|_2^2 + p_\lambda(\gamma), \quad (29)$$

- **Step 2:** Combine construction of  $\hat{Q}$ , tied constraint and CP approximation to solve constrained max-min optimization problem from

$$(\hat{\beta}, \hat{\alpha}) = \underset{(\beta, \alpha)}{\operatorname{argmax}} \left( \min_{\omega} \|\bar{\Psi}\bar{\gamma} - \bar{\Psi}\omega\|_2^2 \right) \quad (30)$$

$$\text{s.t. } \|\alpha \odot \omega\|_1 \leq R, \|\mathbb{P}_{\bar{\Psi}}(\bar{Y} - \bar{D}\beta - \bar{\Psi}\alpha)\|_2^2 < C, \|\alpha\|_0 < pK.$$

to

$$\hat{\alpha} = \underset{\alpha}{\operatorname{argmax}} \left[ \min_{\omega} \|\bar{\Psi}\bar{\gamma} - \bar{\Psi}\omega\|_2^2 \right] - \lambda_1 \|\bar{Y} - \bar{\Psi}\alpha\|_2^2, \text{ s.t. } \|\alpha \odot \omega\|_1 \leq R. \quad (31)$$

- **Step 3:** Post- TSLS estimation:  $\hat{\mathcal{V}} = \{j : \hat{\alpha}_j = 0\}$ ,

$$\hat{\beta}^{\text{MIVE}} = (\bar{D}^\top \bar{\Psi}_{\hat{\mathcal{V}}}^\top \mathbb{M}_{\bar{\Psi}_{\hat{\mathcal{V}}_c}} \bar{\Psi}_{\hat{\mathcal{V}}} \bar{D})^{-1} (\bar{D}^\top \bar{\Psi}_{\hat{\mathcal{V}}}^\top \mathbb{M}_{\bar{\Psi}_{\hat{\mathcal{V}}_c}} \bar{\Psi}_{\hat{\mathcal{V}}} \bar{Y}).$$

as our final MIVE estimator of  $\beta^*$ .

# Computational Algorithm

- Projected Gradient Descending Ascending (PGDA) algorithm is proposed to solve the constrained max-min problem in Step 2 of MIVE procedure.
- In each iteration ( $t$ ), PGDA:
  - Majorizes  $\mathcal{L}(\alpha^{(t)}, \omega)$  at  $\omega^{(t)}$ , and minimizes this to obtain the updating for  $\omega^{(t+1)}$
  - Projects it onto the feasible set  $|\alpha^{(t)} \odot \omega|_1 \leq R$  to ensure feasibility
  - Updates  $\alpha^{(t+1)}$  similarly using gradient ascent
- Selecting the step size also impacts the convergence of gradient descent algorithms.
- Solution: replace vanilla gradients with adaptive moment estimation (Adam) to escape local minima

# Computational Algorithm (continued)

- Updating for  $\omega^{(t+1)}$

$$\begin{aligned}\omega^{(t+1)} &= P_{\Omega(\alpha^{(t)}, \omega)} \left( \operatorname{argmin}_{\omega} \left\{ (\omega - \omega^{(t)})^\top \frac{\partial}{\partial \omega} \mathcal{L}(\alpha^{(t)}, \omega) \Big|_{\omega^{(t)}} + \frac{\phi_\omega}{2} \|\omega - \omega^{(t)}\|_2^2 \right\} \right) \\ &= P_{\Omega(\alpha^{(t)}, \omega)} \left( \operatorname{argmin}_{\omega} \left\{ \frac{\phi_\omega}{2} \left\| \omega - \left[ \omega^{(t)} - \frac{1}{\phi_\omega} \frac{\partial}{\partial \omega} \mathcal{L}(\alpha^{(t)}, \omega) \Big|_{\omega^{(t)}} \right] \right\|_2^2 \right\} \right) \\ &= \operatorname{argmin}_{\xi} \frac{1}{2} \left\| \xi - \left[ \omega^{(t)} - \frac{1}{\phi_\omega} \frac{\partial}{\partial \omega} \mathcal{L}(\alpha^{(t)}, \omega) \Big|_{\omega^{(t)}} \right] \right\|_2^2 + \lambda_3 \|\alpha^{(t)} \odot \xi\|_1 \\ &= S \left( \omega^{(t)} - \frac{1}{\phi_\omega} \frac{\partial}{\partial \omega} \mathcal{L}(\alpha^{(t)}, \omega) \Big|_{\omega^{(t)}}, \lambda_1 |\alpha^{(t)}| \right),\end{aligned}\tag{32}$$

where  $\phi_\omega \geq \Lambda_{\max}(\bar{\Psi}^\top \bar{\Psi})$ .  $S(x, a)_j = \operatorname{sgn}(x_j)(|x_j| - a)_+$ .

- A similar procedure is applied to update  $\alpha^{(t+1)}$  using:

$$\alpha^{(t+1)} = S \left( \alpha^{(t)} + \frac{1}{\phi_\alpha} \frac{\partial}{\partial \alpha} \mathcal{L}(\alpha, \omega^{(t+1)}) \Big|_{\alpha^{(t)}}, \lambda_2 |\omega^{(t+1)}| \right).\tag{33}$$

- The updating consisting **gradient descend/ascend** and then **projected to ensure feasibility**.

# Simulation setup

- Throughout all settings, we fix:
  - The true treatment effect  $\beta^* = 1$
  - Sample sizes  $n = 500, 1000, 2000$
- The potential IV matrix  $\mathbf{Z}$  is generated as:  $\mathbf{Z}_i \stackrel{i.i.d.}{\sim} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}^{\mathbf{Z}})$  where  $\boldsymbol{\Sigma}_{jj}^{\mathbf{Z}} = 0.3$  and  $\boldsymbol{\Sigma}_{jk}^{\mathbf{Z}} = 0.5|j - k|^{0.8}$
- The error terms  $(\epsilon_i, \eta_i)^\top$  are generated from a bivariate normal distribution with:
$$\text{corr}(\epsilon_i, \eta_i) = 0.6$$
- We include sisVIVE and TSHT for comparison and extend them using the generated B-splines to enable local identification.

## Model 2

$$\begin{aligned} \mathbf{Y} &= \mathbf{D} + \mathbf{Z}_1^2 + \phi(\mathbf{Z}_2, -1, 0.25) + \phi(\mathbf{Z}_3, 0, 0.04) + 0.5\mathbf{Z}_4^2 + \epsilon, \\ \mathbf{D} &= 1 + \mathbf{Z}_1^2 + 2\mathbf{Z}_2^3 + 2\sin(\mathbf{Z}_3) + 0.5\mathbf{Z}_4^2 + \eta, \end{aligned} \tag{34}$$

where  $\phi(\cdot, \mu, \sigma^2)$  represents the density function of the standard normal distribution w.r.t mean  $\mu$  and variance  $\sigma^2$ .

# Simulation

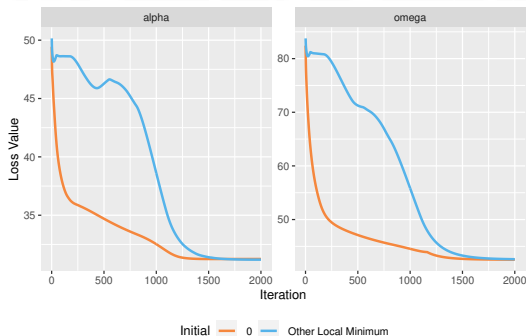
We use Monte Carlo simulation to demonstrate MIVE's performance in two models. We fix the true effect  $\beta = 1$ ,  $p = 4$  and vary  $n = 500, 1000, 2000$ .

Table: Simulation results in Model 2

Estimator	Sample Size	$K = 10$			$K = \lceil 1.5n^{1/3} \rceil$		
		Bias	MAD	EmpSE	Bias	MAD	EmpSE
MIVE	$n = 500$	0.005	0.002	0.058	0.000	0.002	0.029
	$n = 1000$	0.002	0.000	0.050	0.001	0.001	0.026
	$n = 2000$	0.002	-0.001	0.050	0.000	0.000	0.008
TSHT	$n = 500$	0.877	0.982	0.316	0.806	0.977	0.395
	$n = 1000$	0.967	0.995	0.167	0.881	0.986	0.314
	$n = 2000$	0.993	0.999	0.080	0.917	0.994	0.272
sisVIVE	$n = 500$	0.047	0.047	0.020	0.047	0.044	0.021
	$n = 1000$	0.042	0.038	0.020	0.042	0.040	0.015
	$n = 2000$	0.038	0.033	0.020	0.036	0.034	0.015

- MIVE exhibits negligible bias and variance, indicating robustness for varying  $K$  and  $n$ . In contrast, TSHT and sisVIVE perform worse, possibly failing to capture the most informative solution.

# Performance



**Figure:** Trace plots of the loss value with different initial values of  $\alpha^{(0)}$ . A lower loss value indicates a higher CP value, representing the optimal solution.

Figure 9 shows PGDA's behavior.

- Non-informative initials lead to the optimal solution.
- Local solution initials cause early convergence to a local minimum, but eventually reach the optimal solution, suggesting PGDA's capability of escaping local minima.



- We re-examine the study by Card (1993) who estimated the return to education (*educ*) using data of  $n = 3010$  men in 1976.
- He employed a dummy variable for growing up near a 4-year college (*nearc4*) as an IV for education.
- The log(wage) equation included controls:
  - Experience (*exper*) and its square
  - Race dummy (*race*)
  - Region dummies (*reg661-reg668*)
- We apply MIVE to explore the local IV mechanism:
  - Included *nearc4* and its interaction with controls except *expersq*
  - Only *nearc4*·*exper* was continuous, captured using B-spline with degree 5 and degrees of freedom 7

**Table:** Empirical Results of Effect of Education on Wages

	TSLS(Card)	TSCI	MIVE
$\hat{\beta} \left( \widehat{\text{Var}}^{1/2}(\hat{\beta}) \right)$	0.129(0.053)	0.057(0.015)	0.052(0.018)
95% CI	(0.024,0.242)	(0.027,0.087)	(0.015,0.089)
Sargan Test	-	-	0.660

Sargan test p-value is shown in the last row.

- Table 7 reports the results of MIVE and other methods.
- The MIVE estimate of the treatment effect is slightly smaller than TSCI, but with a comparable standard deviation. It also provides an improvement over OLS by correcting for positive “ability bias”.
- The Sargan test p-value of 0.66 supports validity of relevant IVs in MIVE.

# Stage Exposure

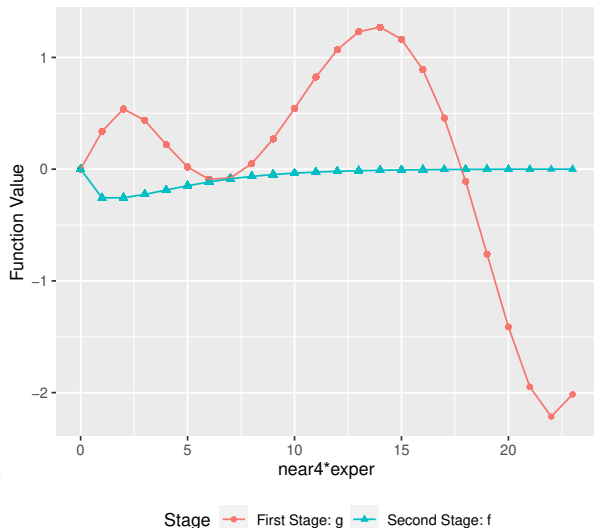


Figure: estimated function value of  $\text{nearc4} \cdot \text{exper}$  in two stages.  $f$  measures the non-linear exposure to  $\log(\text{wage})$  directly and  $g$  measures the non-linear exposure to  $\text{educ}$ .

## Stage exposure (Cont')

- Figure 10 displays the estimated functional forms of  $\hat{f}(\text{nearc4} \cdot \text{exper})$  and  $\hat{g}(\text{nearc4} \cdot \text{exper})$ .
- The first stage  $\hat{g}(\text{nearc4} \cdot \text{exper})$  appears more informative than the second stage  $\hat{f}(\text{nearc4} \cdot \text{exper})$ .
- The range corresponding to local valid IVs in MIVE is when  $\text{nearc4} \cdot \text{exper}$  exceeds 10, where the second stage approximates zero.
- This finding highlights the importance of considering non-linear and local effects in IV estimation, as these effects can greatly influence IV validity and relevance.

# Reference

- Anderson, J. E. (1979). A theoretical foundation for the gravity equation. *Am. Econ. Rev.*, 69(1):106–116.
- Card, D. (1993). Using geographic variation in college proximity to estimate the return to schooling.
- Chen, J. and Khalili, A. (2009). Order selection in finite mixture models with a nonsmooth penalty. *Journal of the American Statistical Association*, 104(485):187–196.
- Fan, Q. and Zhong, W. (2018). Nonparametric additive instrumental variable estimator: A group shrinkage estimation perspective. *J. Bus. Econ. Statist.*, 36(3):388–399.
- Hung, Y., Wang, Y., Zarnitsyna, V., Zhu, C., and Wu, C. J. (2013). Hidden markov models with applications in cell adhesion experiments. *Journal of the American Statistical Association*, 108(504):1469–1479.
- Jessen, F., Wolfsgruber, S., Wiese, B., Bickel, H., Mösch, E., Kaduszkiewicz, H., Pentzek, M., Riedel-Heller, S. G., Luck, T., Fuchs, A., et al. (2014). Ad dementia risk in late mci, in early mci, and in subjective memory impairment. *Alzheimer's & Dementia*, 10(1):76–83.
- Risacher, S. L., Kim, S., Nho, K., Foroud, T., Shen, L., Petersen, R. C., Jack Jr, C. R., Beckett, L. A., Aisen, P. S., Koeppe, R. A., et al. (2015). Apoe effect on alzheimer's disease biomarkers in older adults with significant memory concern. *Alzheimer's & Dementia*, 11(12):1417–1429.
- Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data*. MIT press