Technical Report

Yiqi Lin[a,b]

Supervisors: Xinyuan Song[a], Qingliang Fan[c] & Frank Windmeijer[b]

Presented by Yiqi LIN

Department of Statistics, The Chinese University of Hong Kong[a]
Department of Statistics, University of Oxford[b]
Department of Economics, The Chinese University of Hong Kong[c]

August 22, 2022

Outline

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
    - Follow up work: Bayesian order selection in heterogeneous hidden Markov models

Motivation

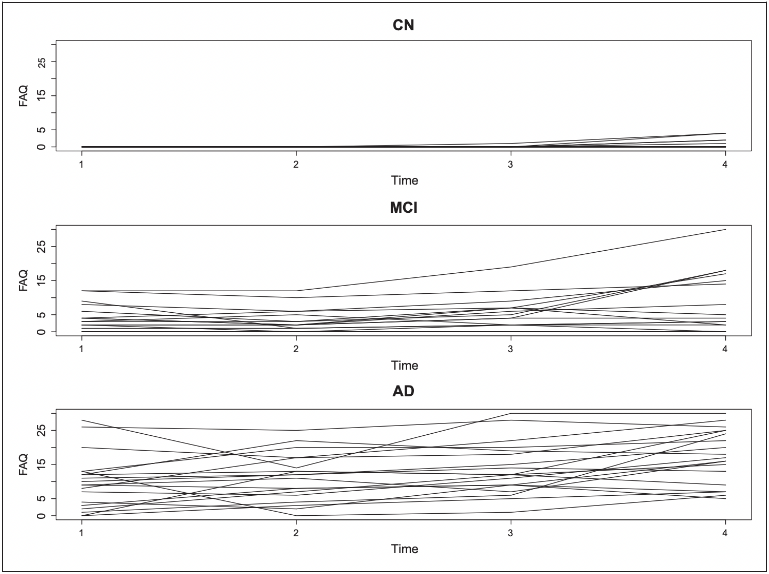


Figure 1: ADNI data analysis results: individual trajectories of functional assessment questionnaire scores for 20 randomly

Motivation(Cont')

- Hidden Markov model (HMM) is a practical statistical tool to simultaneously analysing the longitudinal observation process and its dynamic transition process.
- the most existing HMMs and its extensions in the literature suffer from the determined number of states (order of HMM) that usually are unknown in applications.
- a data-driven procedure to choose the number of hidden states still remains a challenge problem.
- The most common method in the literature for model selection is information based criterion, such as AIC (Akaike (1974)) and BIC (Schwarz et al. (1978)).
- Even though this prevailing information-based criterion succeed in applications (see e.g. Song et al. (2017) and Ip et al. (2013)), it still lack of theoretical justification for HMMs and its extensions (MacDonald and Zucchini (1997)).

Common Settings

$\boldsymbol{Y} = (\boldsymbol{Y}_1, \boldsymbol{Y}_2, \cdots, \boldsymbol{Y}_n)$, where $\boldsymbol{Y}_i = \{y_{it}\}_{t=1}^{T}$ and y_{it} be the response of subject i at time t; $\boldsymbol{S} = (\boldsymbol{S}_1, \boldsymbol{S}_2, \cdots, \boldsymbol{S}_n)$, where $\boldsymbol{S}_i = \{S_{it}\}_{t=1}^{T}$ 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 $\{1, \cdots, K\}$.

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 $t = 2, \ldots, T$ and stationary probability π_r, where $r, s \in \{1, \cdots, K\}$.
2. Let $\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_n)$ be the set of covariates, where $\boldsymbol{X}_i = \{\boldsymbol{x}_{it}\}_{t=1}^{T}$ and $\boldsymbol{x}_{it}$ 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 $\boldsymbol{x}_{it}$, a generalized linear model for response y_{it} is considered as

$$\begin{aligned} f(y_{it} | S_{it} = k, \boldsymbol{x}_{it}, \boldsymbol{\beta}_k) &= \exp\left\{\left(y_{it}\theta_{itk} - b\left(\theta_{itk}\right)\right)/a(\phi) + c\left(y_{it}, \phi\right)\right\}, \\ \theta_{itk} &= \boldsymbol{x}_{it}^{\top}\boldsymbol{\beta}_k, \end{aligned} \tag{1}$$

4. For ease of exposition, K and K_0 are denoted as the upper bound and true value of the order, respectively. Let $\boldsymbol{\Psi} = (\pi_1, \pi_2, \cdots, \pi_K; P_{11}, \cdots, P_{KK}; \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \cdots, \boldsymbol{\beta}_K)$. Then, the probability mass/density function of $\boldsymbol{Y}_i$ can be written as

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

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 ($K \geq K_0$):

$$l_n(\boldsymbol{\Psi}) = \sum_{i=1}^{n} \log F\left(\boldsymbol{Y}_i; \boldsymbol{X}_i, \boldsymbol{\Psi}\right). \tag{3}$$

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

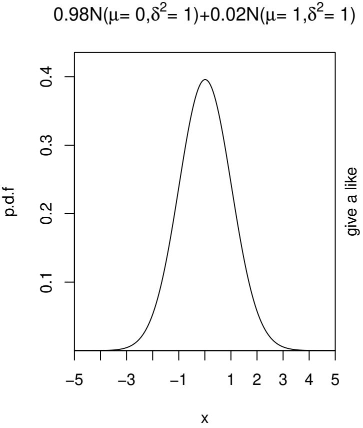


Figure: TYPE I

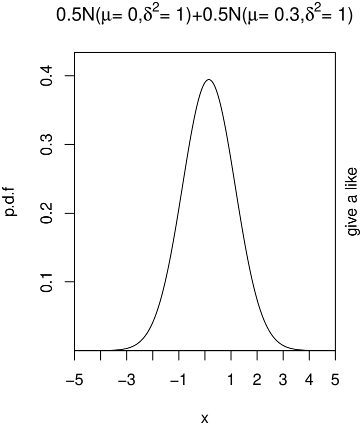


Figure: TYPE II

Normal Type

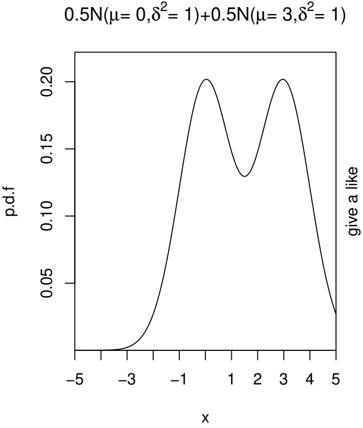


Figure: Normal type

Penalty

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

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

where C_K is a tuning parameter, $p_{\lambda_n}(\cdot)$ is a penalty function, and $\|\boldsymbol{\eta}_k\|_2 = \|\boldsymbol{\beta}_k - \boldsymbol{\beta}_{k-1}\|_2$, 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.

$$\boldsymbol{\beta}_{(k)} = \underset{j \notin \left\{\boldsymbol{\beta}_{(i)}:1 \le i \le k-1\right\}}{\operatorname{argmin}} \left\|\boldsymbol{\beta}_j - \boldsymbol{\beta}_{(k-1)}\right\|_2, \quad k = 2, 3, \ldots, K, \tag{5}$$

  and $\boldsymbol{\beta}_{(1)} = \underset{k=1,2,\ldots,K}{\operatorname{argmax}} \|\boldsymbol{\beta}_k\|_2$.

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

$$\hat{K}_n = \text{number of distinct values of } \left\{\hat{\boldsymbol{\beta}}_{(k)}, k = 1, \cdots, K\right\} \tag{6}$$

  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 $F\left(\boldsymbol{Y}; \boldsymbol{X}, \boldsymbol{\Psi}\right)$ 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 $\boldsymbol{\beta}^S$ of G_0, we have $\hat{G}_n(\boldsymbol{\beta}^S) \xrightarrow{p} G_0(\boldsymbol{\beta}^S)$.

(2) $\sum_{k=1}^{K} \log \hat{\pi}_k = O_p(1)$ and $\hat{\alpha}_k = \pi_{0k} + o_p(1)$ for all $k = 1, 2, \ldots, K_0$. Furthermore, for each $l = 1, 2, \ldots, K$, a unique $k = 1, 2, \ldots, K_0$ exists, such that $\|\hat{\boldsymbol{\beta}}_l - \boldsymbol{\beta}_{0k}\|_2 = o_p(1)$. Thus, $\left\{\hat{\mathcal{V}}_k : k = 1, 2, \ldots, K_0\right\}$ is a cluster partition of $\left\{\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2, \ldots, \hat{\boldsymbol{\beta}}_K\right\}$ in probability.

Theorem 2 (**Consitency of order selection**)

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

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 $\boldsymbol{\Psi}$. In this section, we propose an ECM–ITD algorithm to obtain the MPLE $\hat{\boldsymbol{\Psi}}_n$ of $\boldsymbol{\Psi}$ in RHMM.

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

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

1. CM-step 1:

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

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

2. CM-step 2:

$$\boldsymbol{\beta}^{(p+1)} = \underset{\boldsymbol{\beta}}{\operatorname{argmax}} \sum_{k=1}^K \left[\sum_{i=1}^n \sum_{t=1}^T \log f(y_{it}|S_{it} = k, \boldsymbol{x}_{it}, \boldsymbol{\beta}) h^{(p+1)}(S_{it} = k)\right] - n \sum_{k=2}^K p_{\lambda_n}\left(\|\boldsymbol{\eta}_{\boldsymbol{k}}\|\right. \tag{9}$$

ITD

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 $\boldsymbol{\eta}_1 = \boldsymbol{\beta}_1$ to form a one-to-one mapping between $\boldsymbol{\eta} = (\boldsymbol{\eta}_1, \cdots, \boldsymbol{\eta}_K)$ and $\boldsymbol{\beta}$, i.e., $\boldsymbol{\beta}_k = \sum_{l=1}^{k} \boldsymbol{\eta}_l$ for $k = 1, 2, \ldots, K$.
2. we convert the optimization of updating $\boldsymbol{\beta}^{(p+1)}$ as

$$\boldsymbol{\eta}^{(p+1)} = \underset{\boldsymbol{\eta}}{\operatorname{argmin}} \left\{ G(\boldsymbol{\eta}) = -\sum_{k=1}^{K} \varphi_k \left(\sum_{l=1}^{k} \boldsymbol{\eta}_l \right) + n \sum_{k=2}^{K} p_{\lambda_n} \left(\|\boldsymbol{\eta}_k\|_2 \right) \right\}, \qquad (10)$$

   where $\varphi_k(\boldsymbol{\beta}_k) = \sum_{i=1}^{n} \sum_{t=1}^{T} \{y_{it}\theta_{itk} - b(\theta_{itk})\}\, h^{(p+1)}(S_{it} = k)$ and $\theta_{itk} = \boldsymbol{x}_{it}^T \boldsymbol{\beta}_k$.
3. Inspired by the prevailing iterative shrinkage-thresholding algorithm (ISTA) for regulated convex optimization problem, we optimize a surrogate function $\tilde{Q}(\boldsymbol{\xi}; \boldsymbol{\eta}^{(m)})$:

$$\tilde{Q}(\boldsymbol{\xi}; \boldsymbol{\eta}) = \rho G(\boldsymbol{\xi}) + \frac{1}{2} \sum_{j=1}^{K} \|\boldsymbol{\xi}_j - \boldsymbol{\eta}_j\|_2^2 \qquad (11)$$

$$-\rho \left[\sum_{k=1}^{K} \sum_{i=1}^{n} \sum_{t=1}^{T} \left\{ b\left(\boldsymbol{x}_{it}^T \boldsymbol{\zeta}_k\right) - b\left(\boldsymbol{x}_{it}^T \boldsymbol{\beta}_k\right) - b'\left(\boldsymbol{x}_{it}^T \boldsymbol{\beta}_k\right) \left[\boldsymbol{x}_{it}^T \left(\boldsymbol{\beta}_k - \boldsymbol{\beta}_k \right) \right] \right\} h_{itk} \right],$$

ITD

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

$$\boldsymbol{\eta}_1^{(m+1)} = \boldsymbol{\eta}_{\mathbf{1}}^{\boldsymbol{(m)}} + \rho \sum_{k=1}^{K} \sum_{i=1}^{n} \sum_{j=1}^{T} h_{ij,k} \left(y_{ij} - b' \left(\boldsymbol{x}_{ij}^T \boldsymbol{\beta}_k^{(m)} \right) \right) \boldsymbol{x}_{ij} \tag{12}$$

$$\boldsymbol{\eta}_j^{(m+1)} = \vec{\boldsymbol{S}} \left(\boldsymbol{\eta}_{\boldsymbol{j}}^{\boldsymbol{(m)}} + \rho \sum_{k=j}^{K} \sum_{i=1}^{n} \sum_{j=1}^{T} h_{ij,k} \left(y_{ij} - b' \left(\boldsymbol{x}_{ij}^T \boldsymbol{\beta}_k^{(m)} \right) \right) \boldsymbol{x}_{ij} ; 2n\rho, a, \lambda_n \right), \tag{13}$$

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

Theorem 3

Convergence of ITD method Assume that sequence $\boldsymbol{\eta}^{(m)}$ is generated from (12) and $\boldsymbol{\beta}_k^{(m)} = \sum_{l=1}^{k} \boldsymbol{\eta}_l^{(m)}$. Let τ_1 be the maximum eigenvalue of $\sum_{i=1}^{N} \sum_{t=1}^{T} \boldsymbol{x}_{it} \boldsymbol{x}_{it}^T$ and $\tau_2^{(m)}$ be assigned to

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

If $\rho^{-1} \geq \; K \tau_2^{(m)} \tau_1$, then $G\left(\boldsymbol{\eta}^{(m+1)}\right) \leq G(\boldsymbol{\eta}^{(m)})$. Furthermore, if space $\{\boldsymbol{\eta} : G(\boldsymbol{\eta}) \leq G\left(\boldsymbol{\eta}^{(0)}\right)\}$ is compact, then sequences $\{\boldsymbol{\eta}^{(m)}\}$ and $\{\boldsymbol{\beta}^{(m)}\}$ converge to a stationary point of $G(\boldsymbol{\eta})$.

Simulation

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 $\boldsymbol{x}_{it}^{\top}\boldsymbol{\beta}_k$ and standard deviation $\sigma_k = 0.25$ Covariates $\boldsymbol{x}_{it} = (x_{it1}, x_{it2}, x_{it3})$, 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 $P_{rs} = \frac{1}{K_0}$, $r, s = 1, \cdots, K_0$ are considered. The state-specific regression coefficients for case (1) are set as follows:

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

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
2	**2**	**0.488**	**0.998**	**1**	**0.636**	**1**	**1**
	3	0.278	0	0	0.232	0	0
	4	0.234	0.002	0	0.132	0	0
3	2	0.010	**0.578**	0.344	0	0.072	0.026
	3	**0.408**	0.422	**0.654**	**0.562**	**0.924**	**0.974**
	4	0.376	0	0.002	0.304	0.004	0
	5	0.206	0	0	0.134	0	0
4	3	0.002	0.474	0.190	0	0.002	0.014
	4	**0.614**	**0.524**	**0.808**	**0.690**	**0.980**	**0.984**
	5	0.384	0.002	0.002	0.310	0	0.02

Simulation Result

To mimic the scenario of the ADNI study in real data analysis, we consider a larger RHMM with $K_0 = 5$, $T = 6$, six covariates, and two different transition matrices: (1) a general transition matrix $\boldsymbol{P}_1$ and (2) a band transition matrix $\boldsymbol{P}_2$ that only allows transitions between adjacent states.

$$\boldsymbol{P}_1 = \begin{pmatrix} 0.380 & 0.120 & 0.295 & 0.066 & 0.140 \\ 0.580 & 0.237 & 0.057 & 0.066 & 0.060 \\ 0.310 & 0.090 & 0.378 & 0.083 & 0.140 \\ 0.221 & 0.056 & 0.141 & 0.533 & 0.049 \\ 0.134 & 0.121 & 0.137 & 0.065 & 0.543 \end{pmatrix}, \quad \boldsymbol{P}_2 = \begin{pmatrix} 0.5 & 0.5 & 0 & 0 & 0 \\ 0.3 & 0.4 & 0.3 & 0 & 0 \\ 0 & 0.1 & 0.7 & 0.2 & 0 \\ 0 & 0 & 0.4 & 0.4 & 0.2 \\ 0 & 0 & 0 & 0.7 & 0.3 \end{pmatrix}$$

- Here, y_{it} in state k is generated from a normal distribution with mean $\boldsymbol{x}_{it}^\top \boldsymbol{\beta}_k$ and standard deviation $\sigma_k = 0.25$.
- Covariates $\boldsymbol{x}_{it} = (x_{it1}, \cdots, x_{it6})^T$, where $x_{it1} = 1$, x_{it2} are independently generated from $N(0, 1)$, and x_{it3} to x_{it6} are independently generated from $U(0, 1)$.
- Three sample sizes, $n = 100$, 200, and 400, are considered.
- The state-specific regression coefficients are assigned as $\boldsymbol{\beta}_1 = (0, 0, 0, 0, 0, 0)^T$, $\boldsymbol{\beta}_2 = (-1.5, 2.25, -1, 0, 0.5, 0.75)^T$, $\boldsymbol{\beta}_3 = (0.25, 1.5, 0.75, 0.25, -0.5, -1)^T$, $\boldsymbol{\beta}_4 = (-0.25, 0.5, -2.5, 1.25, 0.75, 1.5)^T$, and $\boldsymbol{\beta}_5 = (-1, -1.5, -0.25, 1.75, -0.5, 2)^T$.

Simulation Result

Table: Proportion of order selection for normal case in Simulation 2

N	$\hat{K}_n$	$\boldsymbol{P}_1$ (general transition matrix)			$\boldsymbol{P}_2$ (band transition matrix)		
		AIC	BIC	ECM–ITD	AIC	BIC	ECM–ITD
100	4	0.015	0.025	0	0.025	0.03	0.005
	5	**0.400**	**0.515**	**0.740**	**0.445**	**0.540**	**0.720**
	6	0.320	0.320	0.240	0.270	0.330	0.250
	7	0.265	0.140	0.020	0.260	0.100	0.025
200	4	0.0.005	0.005	0	0	0	0
	5	**0.595**	**0.695**	**0.890**	**0.510**	**0.660**	**0.950**
	6	0.250	0.2350	0.110	0.280	0.305	0.045
	7	0.150	0.065	0	0.210	0.035	0.005
400	**5**	**0.675**	**0.785**	**0.940**	**0.635**	**0.720**	**0.975**
	6	0.230	0.190	0.060	0.225	0.225	0.025
	7	0.095	0.025	0	0.140	0.020	0

ADNI

we analyze the a data set extracted from ADNI study by employing the proposed ESGF procedure to detect the number of hidden of phases of the neuro degenerative pathology.

- We focused on $n = 616$ subjects collected from the ADNI-I, ADNI-II, and ADNI-Go study with four follow-up visits at baseline, 6 months, 12 months, and 24 months (T=4).
- In this study, we treated ADAS13 as response y_{it} and some clinical and generic variables as covariate $_{it}$ in the proposed RHMM. Covariate $\boldsymbol{X}_{it} = (x_{it1}, \ldots, x_{it6})^T$, where $x_{it1} = 1$, x_{it2}: age at each visit, x_{it3}: gender $(1 = \text{female})$, x_{it4}: logarithm of the ratio of hippocampal volume over the whole brain volume (HIP), and $\{x_{it5}, x_{it6}\}$: apolipoprotein E (APOE)-ε4, which was coded as 0, 1, and 2, denoting the number of APOE-ε4 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.

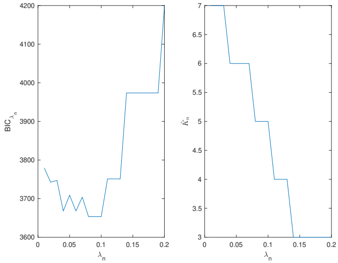


Figure: tuning selection using BIC.

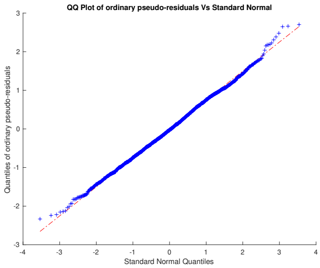


Figure: Residual Check.

Result

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

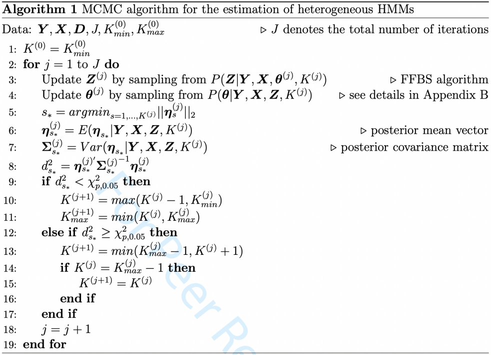
	State				
Par.	1 (CN)	2 (SMC)	3 (EMCI)	4 (LMCI)	5(AD)
Intercept	-1.007(0.023)	-0.628(0.031)	-0.033(0.029)	0.876(0.035)	2.225(0.167)
β_{k2}	0.092(0.017)	0.096(0.019)	0.069(0.023)	0.076(0.027)	0.034(0.068)
β_{k3}	0.037(0.022)	0.111(0.026)	0.023(0.036)	-0.066(0.052)	-0.464(0.168)
β_{k4}	-0.146(0.013)	-0.257(0.021)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

Result

- state-specific intercept β_{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).

Bayesian order selection in heterogeneous hidden Markov models

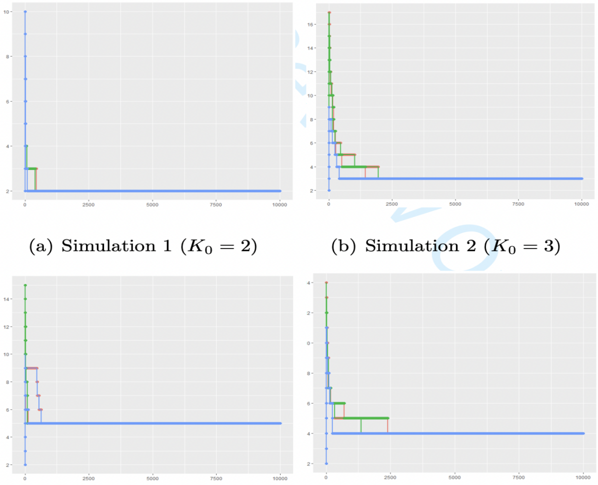
- The previous method can only apply to homogeneous transition pattern.
- we extend it to heterogeneous way that individuals characteristic can effect their own transition pattern.
- we offer full Bayesian modelling and an efficient adjust-bound reversible jump (ABRJ) sampling scheme to address the challenges of updating the order in implementing the MCMC algorithm.



Algorithm 1 MCMC algorithm for the estimation of heterogeneous HMMs

```
Data: $\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{D}, J, K_{min}^{(0)}, K_{max}^{(0)}$          ▷ $J$ denotes the total number of iterations
$K^{(0)} = K_{min}^{(0)}$
for $j = 1$ to $J$ do
    Update $\boldsymbol{Z}^{(j)}$ by sampling from $P(\boldsymbol{Z}|\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{\theta}^{(j)}, K^{(j)})$          ▷ FFBS algorithm
    Update $\boldsymbol{\theta}^{(j)}$ by sampling from $P(\boldsymbol{\theta}|\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{Z}, K^{(j)})$          ▷ see details in Appendix B
    $s_* = argmin_{s=1,\dots,K^{(j)}} ||\boldsymbol{\eta}_s^{(j)}||_2$
    $\boldsymbol{\eta}_{s_*}^{(j)} = E(\boldsymbol{\eta}_{s_*}|\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{Z}, K^{(j)})$          ▷ posterior mean vector
    $\boldsymbol{\Sigma}_{s_*}^{(j)} = Var(\boldsymbol{\eta}_{s_*}|\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{Z}, K^{(j)})$          ▷ posterior covariance matrix
    $d_{s_*}^2 = \boldsymbol{\eta}_{s_*}^{(j)'} {\boldsymbol{\Sigma}_{s_*}^{(j)}}^{-1} \boldsymbol{\eta}_{s_*}^{(j)}$
    if $d_{s_*}^2 < \chi_{p,0.05}^2$ then
        $K^{(j+1)} = max(K^{(j)} - 1, K_{min}^{(j)})$
        $K_{max}^{(j+1)} = min(K^{(j)}, K_{max}^{(j)})$
    else if $d_{s_*}^2 \geq \chi_{p,0.05}^2$ then
        $K^{(j+1)} = min(K_{max}^{(j)} - 1, K^{(j)} + 1)$
        if $K^{(j)} = K_{max}^{(j)} - 1$ then
            $K^{(j+1)} = K^{(j)}$
        end if
    end if
    $j = j + 1$
end for
```

BHMM



(a) Simulation 1 ($K_0 = 2$)

(b) Simulation 2 ($K_0 = 3$)

(c) Simulation 2 ($K_0 = 5$)

(d) real data analysis ($\hat{K} = 4$)

Figure 2: Trace plots of three MCMC chains of K in simulations and the ADNI study.

Outline

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

Endogeneity

In traditional linear regression analysis:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}. \tag{15}$$

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

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

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

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

- **What we need is instrumental variables (IVs).**

Requirement of IVs

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

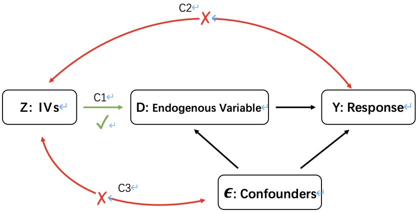


Figure: Illustration of Validity and Relevance.

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

For $i = 1, 2, \ldots, n$, we have the random sample $(Y_i, D_i, \boldsymbol{Z}_{i.})$. Let $Y_i^{(D_i, \boldsymbol{z}_{i.})}$ to be the potential outcome for the object i having exposure D_i and instrumental variables $\boldsymbol{Z}_{i.} \in \mathbb{R}^p$.

Potential Outcome Model

Given two different sets of treatment variables D_i^A, D_i^B and corresponding instruments $\boldsymbol{Z}_i^A$, $\boldsymbol{Z}_i^B$, assume

$$Y_i^{\left(D_i^B, \boldsymbol{z}_{i.}^B\right)} - Y_i^{\left(D_i^A, \boldsymbol{z}_{i.}^A\right)} = \left(\boldsymbol{Z}_{i.}^B - \boldsymbol{Z}_{i.}^A\right)^\top \phi + \left(D_i^B - D_i^A\right) \beta \text{ and } E\left(Y_i^{(0,0)} \mid \boldsymbol{Z}_{i.}\right) = \boldsymbol{Z}_{i.}^\top \boldsymbol{\theta}, \tag{17}$$

1. $D \in \mathbb{R}, \beta^* \in \mathbb{R}$ represents the constant causal parameter of interest.
2. $\boldsymbol{Z} \in \mathbb{R}^p, \boldsymbol{\phi} \in \mathbb{R}^L$ represents the violation of Exclusion Restriction.
3. $\boldsymbol{\theta} \in \mathbb{R}^p$ represents the violation of Exogenous Condition.

Model

A good instrument $\boldsymbol{Z}_j$ should not have a direct effect on the response and unmeasured confounders, i.e., $\phi_j = 0$ and $\theta_j = 0$.

Model

Assuming the linear functional form between treatment effects $\boldsymbol{D}_i$ and instruments $\boldsymbol{Z}_{i.}$, the above potential outcome model (17) can be rewritten as follows,

$$\begin{aligned} Y_i &= D_i \beta + \boldsymbol{Z}_{i.}^{\top} \boldsymbol{\alpha} + \epsilon_i \\ D_i &= \boldsymbol{Z}_{i.}^{\top} \boldsymbol{\gamma} + \eta_i. \end{aligned} \tag{18}$$

where $\epsilon_i = Y_i^{(0,0)} - E\left(Y_i^{(0,0)} \mid \boldsymbol{Z}_{i.}\right)$, $\boldsymbol{\alpha} = \boldsymbol{\phi} + \boldsymbol{\theta}$.

Definition

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

Identifiability of Model

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

$$
\begin{aligned}
E\left(\mathbf{Z}^T\varepsilon\right) &= E\left[\mathbf{Z}^T\left(\mathbf{Y}-\mathbf{Z}\boldsymbol{\alpha}^*-\mathbf{D}\beta^*\right)\right]=\mathbf{0} \\
E\left(\mathbf{Z}^T\mathbf{Y}\right) &= E\left(\mathbf{Z}^T\mathbf{Z}\right)\boldsymbol{\alpha}^*+E\left(\mathbf{Z}^T\mathbf{D}\right)\beta^* \\
&\Rightarrow \boldsymbol{\Gamma}^*_{p\times1}=\boldsymbol{\alpha}^*_{p\times1}+\boldsymbol{\gamma}^*_{p\times1}\beta^*_{1\times1},
\end{aligned}
\tag{19}
$$

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

- Both $\boldsymbol{\Gamma}^*$ and γ^* can be identified based on observed data. But (19) have $(\boldsymbol{\alpha}^{*\top}_{p\times1},\beta^*_{1\times1})^\top$, 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.

Identiability of Model (Cont')

- Define Expectation of individual IV Estimator: $\beta_j^* \triangleq \frac{\boldsymbol{\Gamma}_j^*}{\boldsymbol{\gamma}_j^*} = \beta^* + \frac{\alpha_j^*}{\gamma_j^*}$.

Plurality Rule (Sufficient and Necessary Conditions?)

It is proposed in (Guo et al., 2018). The parameters β^* and $\boldsymbol{\alpha}^*$ are identified **if and only if** the following hold:

$$\left|\mathcal{V}^* = \{j : \alpha_j^*/\gamma_j^* = 0\}\right| > \max_{c \neq 0} \left|\{j : \alpha_j^*/\gamma_j^* = c\}\right| \tag{20}$$

1. There are many following works take this theorem for granted!
2. However, the only if part is wrong. There never is a necessary condition for identifying this model.

In general, there is no iff condition in term of model

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): $p < n$, $p_{\mathcal{V}^{c*}}/n \to \upsilon_{p_{\mathcal{V}^{c*}}} + o(n^{-1/2})$ and $p_{\mathcal{V}^*}/n \to \upsilon_{p_{\mathcal{V}^*}} + o(n^{-1/2})$ for some non-negative constants $\upsilon_{p_{\mathcal{V}^{c*}}}$ and $\upsilon_{p_{\mathcal{V}^*}}$ such that $0 \leq \upsilon_{p_{\mathcal{V}^*}} + \upsilon_{p_{\mathcal{V}^{c*}}} < 1$.
Assumption 2: Assume $\boldsymbol{Z}$ is standardized. It then has full column rank and $\|\boldsymbol{Z}_j\|_2^2 \leq n$ for $j = 1, 2, \ldots, p$.
Assumption 3: Let $\boldsymbol{u}_i = (\epsilon_i, \eta_i)^\top$. $\boldsymbol{u}_i \mid \boldsymbol{Z}_i$ are i.i.d. and follow a multivariate normal distribution with mean zero and positive definite covariance matrix $\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_\epsilon^2 & \sigma_{\epsilon,\eta} \\ \sigma_{\epsilon,\eta} & \sigma_\eta^2 \end{pmatrix}$. The elements of $\boldsymbol{\Sigma}$ are finite and $\sigma_{\epsilon,\eta} \neq 0$.
Assumption 4 (Strength of valid IVs): The concentration parameter μ_n grows at the same rate as n, i.e., $\mu_n \boldsymbol{\gamma}_{\boldsymbol{Z}_{\mathcal{V}^*}}^{*\top} \boldsymbol{Z}_{\mathcal{V}^*}^\top M_{\boldsymbol{Z}_{\mathcal{V}^{c*}}} \boldsymbol{Z}_{\mathcal{V}^*} \boldsymbol{\gamma}_{\boldsymbol{Z}_{\mathcal{V}^*}}^* / \sigma_\eta^2 \to \mu_0 n$, for some $\mu_0 > 0$.

View of Data Generating Process (DGP)

Given first stage information: $\{\boldsymbol{D}, \boldsymbol{Z}, \gamma^*\}$, without loss of generality, we denote DGP with some $\{\beta^*, \boldsymbol{\alpha}^*, \boldsymbol{\epsilon}\}$ in (2) as DGP $\mathcal{P}_0$ that generates $\boldsymbol{Y}$.

Transformation of DGP

Given this $\mathcal{P}_0$, for $j \in \mathcal{V}^{c*}$, we have $\boldsymbol{Z}_j\alpha_j^* = \frac{\alpha_j^*}{\gamma_j^*}\left(\boldsymbol{D} - \sum_{l\neq j}\boldsymbol{Z}_l\gamma_l^* - \boldsymbol{\eta}\right)$. Denote $\mathcal{I}_c = \left\{j \in \mathcal{V}^{c*} : \alpha_j^*/\gamma_j^* = c\right\}$, where $c \neq 0$ and c could have up to K different values. For compatibility, we denote $\mathcal{I}_0 = \mathcal{V}^*$. Thus, we are able to reformulate

$$\boldsymbol{Y} = \boldsymbol{D}\beta^* + \boldsymbol{Z}\boldsymbol{\alpha}^* + \boldsymbol{\epsilon} \Rightarrow \boldsymbol{Y} = \boldsymbol{D}\tilde{\beta}^c + \boldsymbol{Z}\tilde{\boldsymbol{\alpha}}^c + \tilde{\boldsymbol{\epsilon}}^c,$$

where $\left\{\tilde{\beta}^c, \tilde{\boldsymbol{\alpha}}^c, \tilde{\boldsymbol{\epsilon}}^c\right\} = \{\beta^* + c, \boldsymbol{\alpha}^* - c\boldsymbol{\gamma}^*, \boldsymbol{\epsilon} - c\boldsymbol{\eta}\}$ and $c = \alpha_j^*/\gamma_j^*$, for some $j \in \mathcal{V}^{c*}$.

Evidently, for different $c \neq 0$, it forms different DGPs $\mathcal{P}_c = \left\{\tilde{\beta}^c, \tilde{\boldsymbol{\alpha}}^c, \tilde{\boldsymbol{\epsilon}}^c\right\}$ that can generate the same $\boldsymbol{Y}$ (given $\boldsymbol{\epsilon}$) which also satisfies the moment condition (2) as $\mathcal{P}_0$ since $E\left(\boldsymbol{Z}^\top\tilde{\boldsymbol{\epsilon}}^c\right) = \mathbf{0}$.

DGP view: Example

Example

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

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

Then we rearrange first stage equation:

$$\begin{aligned}
\boldsymbol{Z}_1\gamma_1^* =& \boldsymbol{D} - \boldsymbol{Z}_2\gamma_2^* - \boldsymbol{Z}_3\gamma_3^* - \boldsymbol{\eta} \\
\Rightarrow \boldsymbol{Z}_1\boldsymbol{\alpha}_1^* =& \boldsymbol{Z}_1\gamma_1^*(\frac{\boldsymbol{\alpha}_1^*}{\gamma_1^*}) = \boldsymbol{D}\frac{\boldsymbol{\alpha}_1^*}{\gamma_1^*} - \boldsymbol{Z}_2\gamma_2\frac{\boldsymbol{\alpha}_1^*}{\gamma_1^*} - \boldsymbol{Z}_3\gamma_3^*\frac{\boldsymbol{\alpha}_1^*}{\gamma_1^*} - \boldsymbol{\eta}\frac{\boldsymbol{\alpha}_1^*}{\gamma_1^*} \\
\Rightarrow \boldsymbol{Y} =& \boldsymbol{D}\beta^* + (\boldsymbol{D}\frac{\boldsymbol{\alpha}_1^*}{\gamma_1^*} - \boldsymbol{Z}_2\gamma_2\frac{\boldsymbol{\alpha}_1^*}{\gamma_1^*} - \boldsymbol{Z}_3\gamma_3^*\frac{\boldsymbol{\alpha}_1^*}{\gamma_1^*} - \boldsymbol{\eta}\frac{\boldsymbol{\alpha}_1^*}{\gamma_1^*}) + \boldsymbol{Z}_2\boldsymbol{\alpha}_2^* + \boldsymbol{\epsilon} \\
\Rightarrow \boldsymbol{Y} =& \boldsymbol{D}(\beta^* + \frac{\boldsymbol{\alpha}_1^*}{\gamma_1^*}) + \boldsymbol{Z}_2(\boldsymbol{\alpha}_2^* - \frac{\boldsymbol{\alpha}_1^*}{\gamma_1^*}) + \boldsymbol{Z}_3(-\frac{\boldsymbol{\alpha}_1^*}{\gamma_1^*}) + (\boldsymbol{\epsilon} - \frac{\boldsymbol{\alpha}_1^*}{\gamma_1^*}\boldsymbol{\eta})
\end{aligned}$$

Then It forms a new DGP: $\tilde{\beta} = \beta^* + \frac{\boldsymbol{\alpha}_1^*}{\gamma_1^*}$,$\tilde{\boldsymbol{\alpha}} = (0, \boldsymbol{\alpha}_2^* - \frac{\boldsymbol{\alpha}_1^*}{\gamma_1^*}, -\frac{\boldsymbol{\alpha}_1^*}{\gamma_1^*})$ and $\tilde{\boldsymbol{\epsilon}} = \boldsymbol{\epsilon} - \frac{\boldsymbol{\alpha}_1^*}{\gamma_1^*}\boldsymbol{\eta}$.

Identifiability of Model (Cont.)

Theorem 1

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

Theorem 1 tells us there is a collection of model DGPs

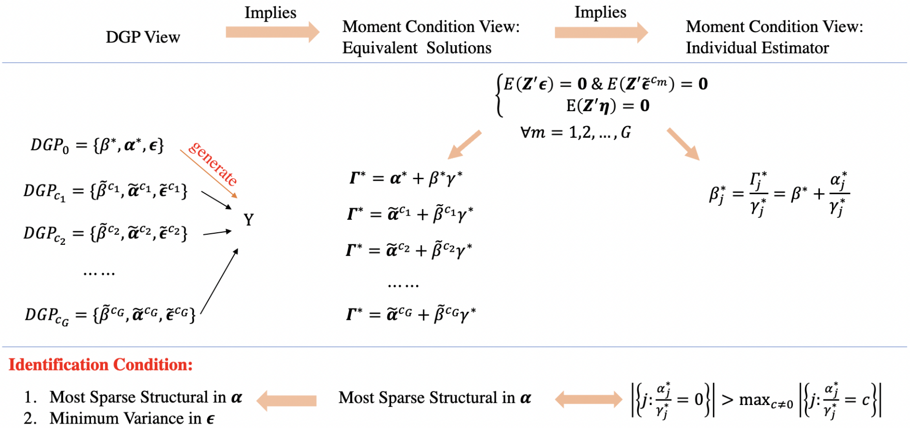
$$\mathcal{Q} = \left\{\mathcal{P} = \{\beta, \boldsymbol{\alpha}, \epsilon\} : \boldsymbol{\alpha} \text{ is sparse, } E\left(\boldsymbol{Z}^\top \boldsymbol{\epsilon}\right) = \boldsymbol{0}\right\}$$

corresponding to the same observation $\boldsymbol{Y}$ 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 $\mathcal{F} = \{f : \mathcal{P} \in \mathcal{Q} \rightarrow \mathbb{R}; f\left(\mathcal{P}_i\right) < \ f\left(\mathcal{P}_j\right) \forall j \neq i$ and $\exists i \in \{0, \ldots, G\}\}$ and $\mathcal{G} = \left\{g = \operatorname{argmin}_{\mathcal{P} \in \mathcal{Q}} f(\mathcal{P}); f \in \mathcal{F}\right\}$, 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



DGP View

Implies

Moment Condition View: Equivalent Solutions

Implies

Moment Condition View: Individual Estimator

$$DGP_0 = \{\beta^*, \boldsymbol{\alpha}^*, \boldsymbol{\epsilon}\}$$

$$DGP_{c_1} = \{\tilde{\beta}^{c_1}, \tilde{\boldsymbol{\alpha}}^{c_1}, \tilde{\boldsymbol{\epsilon}}^{c_1}\}$$

$$DGP_{c_2} = \{\tilde{\beta}^{c_2}, \tilde{\boldsymbol{\alpha}}^{c_2}, \tilde{\boldsymbol{\epsilon}}^{c_2}\}$$

……

$$DGP_{c_G} = \{\tilde{\beta}^{c_G}, \tilde{\boldsymbol{\alpha}}^{c_G}, \tilde{\boldsymbol{\epsilon}}^{c_G}\}$$

generate

Y

$$\begin{cases} E(\boldsymbol{Z}'\boldsymbol{\epsilon}) = \mathbf{0} \;\&\; E(\boldsymbol{Z}'\tilde{\boldsymbol{\epsilon}}^{c_m}) = \mathbf{0} \\ \mathrm{E}(\boldsymbol{Z}'\boldsymbol{\eta}) = \mathbf{0} \end{cases}$$

$$\forall m = 1, 2, \dots, G$$

$$\boldsymbol{\Gamma}^* = \boldsymbol{\alpha}^* + \beta^* \gamma^*$$

$$\boldsymbol{\Gamma}^* = \tilde{\boldsymbol{\alpha}}^{c_1} + \tilde{\beta}^{c_1} \gamma^*$$

$$\boldsymbol{\Gamma}^* = \tilde{\boldsymbol{\alpha}}^{c_2} + \tilde{\beta}^{c_2} \gamma^*$$

……

$$\boldsymbol{\Gamma}^* = \tilde{\boldsymbol{\alpha}}^{c_G} + \tilde{\beta}^{c_G} \gamma^*$$

$$\beta_j^* = \frac{\Gamma_j^*}{\gamma_j^*} = \beta^* + \frac{\alpha_j^*}{\gamma_j^*}$$

Identification Condition:

1. Most Sparse Structural in $\boldsymbol{\alpha}$
2. Minimum Variance in $\boldsymbol{\epsilon}$

Most Sparse Structural in $\boldsymbol{\alpha}$

$$\left|\left\{j : \frac{\alpha_j^*}{\gamma_j^*} = 0\right\}\right| > \max_{c \neq 0} \left|\left\{j : \frac{\alpha_j^*}{\gamma_j^*} = c\right\}\right|$$

Figure: Explanation of Theorem

Individual IV based approaches

The sparsest rule is conceptually equivalent to plurality rule, since the non-overlapping sparse solutions given by Theorem 1.

Estimation method in Literature

Recall $\beta_j^* \triangleq \frac{\Gamma_j^*}{\gamma_j^*} = \beta^* + \frac{\alpha_j^*}{\gamma_j^*}$. They estimate the individual IV estimate $\hat{\beta}_j = \frac{\hat{\Gamma}_j}{\hat{\gamma}_j}$ first, then explicit utilize plurality rule:

$$\left|\mathcal{V}^* = \{j : \alpha_j^*/\gamma_j^* = 0\}\right| > \max_{c \neq 0} \left|\{j : \alpha_j^*/\gamma_j^* = c\}\right|. \tag{21}$$

1. For the sake of stable estimation, Guo et al. (2018) proposed to use plurality rule based on relevant IV set:

$$\left|\mathcal{V}_{\mathcal{S}^*}^* = \{j \in \mathcal{S}^* : \alpha_j^*/\gamma_j^* = 0\}\right| > \max_{c \neq 0} \left|\{j \in \mathcal{S}^* : \alpha_j^*/\gamma_j^* = c\}\right|.$$

2. $\mathcal{S}^*$ is the relevant IVs estimated via first-step hard thresholding (Guo et al., 2018) and the individual margin to select relevant IVs is $\sqrt{\widehat{\mathrm{Var}}\left(\widehat{\gamma}_j\right)} \cdot \sqrt{2.01 \log p \vee n} \asymp \sigma_\eta \sqrt{2.01 \log p \vee n / n}$.

Thus, TSHT proposed by Guo et al. (2018) and CIIV proposed by Windmeijer et al. (2021) all explicitly leverage $\mathcal{S}^*$-based plurality rule to estimate $\mathcal{V}_{\mathcal{S}^*}^*$ and β^*.

Drawback of individual IV estimator based method

However, weak IVs sometimes is determinant in identification while limited in estimation.

Example

Consider a toy example with mixed weak IV. Let $\gamma^* = (\mathbf{0.04}_3, \mathbf{0.5}_2, 0.2, \mathbf{0.1}_4)^\top$ and $\boldsymbol{\alpha}^* = (\mathbf{0}_5, 1, \mathbf{0.7}_4)^\top$. Obviously it forms three groups: $\mathcal{I}_0 = \mathcal{V}^* = \{1, 2, 3, 4, 5\}$, $\mathcal{I}_5 = \{6\}$, $\mathcal{I}_7 = \{7, 8, 9, 10\}$ and plurality rule $|\mathcal{I}_0| > \max_{c=5,7} |\mathcal{I}_c|$ holds in whole IVs set.

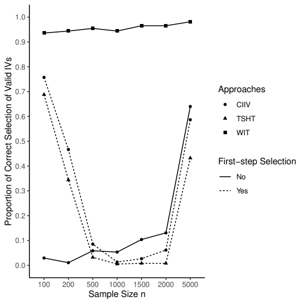
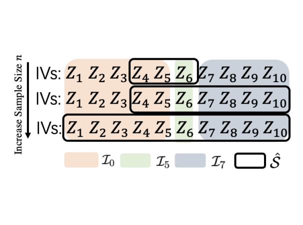


Figure: Proportion of correct selection of (subset) valid IVs.



Increase Sample Size n
IVs: Z_1 Z_2 Z_3 Z_4 Z_5 Z_6 Z_7 Z_8 Z_9 Z_{10}
IVs: Z_1 Z_2 Z_3 Z_4 Z_5 Z_6 Z_7 Z_8 Z_9 Z_{10}
IVs: Z_1 Z_2 Z_3 Z_4 Z_5 Z_6 Z_7 Z_8 Z_9 Z_{10}
$\mathcal{I}_0$
$\mathcal{I}_5$
$\mathcal{I}_7$
$\hat{\mathcal{S}}$

Figure: Plurality based on first stage selection.

The Sparsest Rule (cont.)

- Mixture of weak and invalid IVs is ubiquitous in practice, especially in many IVs.
- For using weak IVs information and improving finite sample performance, it motivates us to turn to the sparsest rule that is also operational in computation algorithms.

Recall $\mathcal{P}_c = \left\{\tilde{\beta}^c, \tilde{\boldsymbol{\alpha}}^c, \tilde{\boldsymbol{\epsilon}}^c\right\} = \{\beta^* + c, \boldsymbol{\alpha}^* - \ c\boldsymbol{\gamma}^*, \boldsymbol{\epsilon} - c\boldsymbol{\eta}\}$, where $\tilde{\boldsymbol{\alpha}}^c_{\mathcal{I}_c} = \mathbf{0}$. For other elements in $\tilde{\boldsymbol{\alpha}}^c$ (corresponds to a different DGP in $\mathcal{Q}$) and $j \in \left\{j : \alpha_j^*/\gamma_j^* = \tilde{c} \neq c\right\}$, we obtain,

$$\left|\tilde{\alpha}_j^c\right| = \left|\alpha_j^* - c\gamma_j^*\right| = \left|\alpha_j^*/\gamma_j^* - c\right| \cdot \left|\gamma_j^*\right| = \left|\tilde{c} - c\right| \cdot \left|\gamma_j^*\right|.$$

The above $\left|\tilde{\alpha}_j^c\right|$ needs to be distinguished with 0 on the ground of non-overlap structure stated in Theorem 1. To facilitate the discovery of all solutions in $\mathcal{Q}$, we assume

Additional Assumptions

Assumption 5. $|\tilde{\alpha}_j^c| > \kappa^c(n)$ for $j \in \{j : \alpha_j^*/\gamma_j^* = \tilde{c} \neq c\}$ and $|\boldsymbol{\alpha}^*_{\mathcal{V}^{c*}}|_{\min} > \kappa(n)$, where $\kappa(n)$ and $\kappa^c(n)$ are a generally vanishing rate specified by some estimator under consideration to separate zero and non-zero terms.

Assumption 6. (The Sparsest Rule): $\boldsymbol{\alpha}^* = \text{argmin}_{\mathcal{P}=\{\beta,\boldsymbol{\alpha},\boldsymbol{\epsilon}\}\in\mathcal{Q}} \|\boldsymbol{\alpha}\|_0$

The Sparsest Rule (Cont.)

For penalized regression, this condition is known as "beta-min" condition. Notably $\left|\tilde{\alpha}_j^c\right| = |\tilde{c} - c| \cdot \left|\gamma_j^*\right| > \kappa(n)$ depends on product of $|\tilde{c} - c|$ and $\left|\gamma_j^*\right|$.

1. As discussed in Guo et al. (2018), $|\tilde{c} - c|$ can not be too small to separate different solutions, but the larger gap $|\tilde{c} - c|$ are helpful to mitigate the small or local to zero $\left|\gamma_j^*\right|$ in favor of our model

Example (continued)

1. Follow the procedure, we are able to reformulate total three solutions of (19): $\boldsymbol{\alpha}^* = (\mathbf{0}_5, 1, \mathbf{0.7}_4)^\top$, $\tilde{\boldsymbol{\alpha}}^5 = (\mathbf{-0.2}_3, \mathbf{-2.5}_2, 0, \mathbf{0.2}_4)^\top$ and $\tilde{\boldsymbol{\alpha}}^7 = (\mathbf{-0.28}_3, \mathbf{-3.5}_2, -0.4, \mathbf{0}_4)^\top$.
2. Thus, the sparsest rule $\operatorname{argmin}_{\boldsymbol{\alpha} \in \{\boldsymbol{\alpha}^*, \tilde{\boldsymbol{\alpha}}^5, \tilde{\boldsymbol{\alpha}}^7\}} \|\boldsymbol{\alpha}\|_0$ picks $\boldsymbol{\alpha}^*$ up, and Assumption 6 is easy to satisfy since fixed minimum absolute value except 0 are $0.7, 0.2, 0.28$ in $\boldsymbol{\alpha}^*, \tilde{\boldsymbol{\alpha}}^5, \tilde{\boldsymbol{\alpha}}^7$, respectively.

Penalization Approaches with Embedded Surrogate Sparsest Rule

Consider the general penalized TSLS estimator based on moment conditions:

$$\left(\widehat{\boldsymbol{\alpha}}^{\text{pen}}, \hat{\beta}^{\text{pen}}\right) = \underset{\boldsymbol{\alpha},\beta}{\operatorname{argmin}} \underbrace{\frac{1}{2n}\|P_{\boldsymbol{Z}}(\boldsymbol{Y}-\boldsymbol{Z}\boldsymbol{\alpha}-\boldsymbol{D}\beta)\|_2^2}_{(I)} + \underbrace{p_\lambda^{\text{pen}}(\boldsymbol{\alpha})}_{(II)}. \tag{22}$$

They have the two different functions:

1. Approximate $\mathcal{Q}$: (I) is a scaled finite sample version of $E\left(\left[\boldsymbol{Z}^\top\boldsymbol{\epsilon}\right]^\top \left[\boldsymbol{Z}^\top\boldsymbol{Z}\right]^{-1}\left[\boldsymbol{Z}^\top\boldsymbol{\epsilon}\right]\right)$, which is a $\left[\boldsymbol{Z}^\top\boldsymbol{Z}\right]^{-1}$ weighed quadratic term of condition $E\left(\boldsymbol{Z}^\top\boldsymbol{\epsilon}\right) = \mathbf{0}$, and (II) is imposed to ensure sparsity structure in $\boldsymbol{\alpha}$.
2. Rewrite equivalent constrained objective function form with the optimal penalty $\|\boldsymbol{\alpha}\|_0$ with respect to the sparest rule:

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

The constraint above narrows the feasible solutions into $\mathcal{Q}$ because Sargan test statistics $\|P_{\boldsymbol{Z}}(\boldsymbol{Y}-\boldsymbol{D}\beta-\boldsymbol{Z}\boldsymbol{\alpha})\|_2^2/\|(\boldsymbol{Y}-\boldsymbol{D}\beta-\boldsymbol{Z}\boldsymbol{\alpha})/\sqrt{n}\|_2^2 = O_p(1)$ (Sargan, 1958) under null hypothesis $E\left(\boldsymbol{Z}^\top\boldsymbol{\epsilon}\right) = \mathbf{0}$ as required in $\mathcal{Q}$, otherwise $O_p(n)$ that cannot be bounded by δ.

Penalized method in literature

Kang et al. (2016) propose to use Lasso in (4), call sisVIVE. It has three problems:

1. Failure in consistent variable selection under some deterministic conditions, namely the sign-aware invalid IV strength (SAIS) condition:

$$\left|\widehat{\gamma}_{\mathcal{V}^{c*}}^{\top} \operatorname{sgn}\left(\boldsymbol{\alpha}_{\mathcal{V}^{c*}}\right)\right| > \|\widehat{\gamma}_{\mathcal{V}^{*}}\|_1 \text{ (Windmeijer et al., 2019, Proposition 2)}$$

   The SAIS is a rather common situation in practice, under which sisVIVE cannot achieve $\mathcal{P}_0$.

2. Unclear dependency of regularization condition of $\tilde{Z}$: (Kang et al., 2016, Theorem 2) proposed an non-asymptotic error bound $\left|\hat{\beta}^{\text{sis}} - \beta^*\right|$ for sisVIVE. Under some regularity of restricted isometry property (RIP) constants of $\boldsymbol{Z}$ and $P_{\widehat{\boldsymbol{D}}}\boldsymbol{Z}$,

$$\left\|\hat{\beta}^{\text{sis}} - \beta^*\right\|_2 \leq \frac{\left|\widehat{\boldsymbol{D}}^{\top}\boldsymbol{\epsilon}\right|}{\|\widehat{\boldsymbol{D}}\|_2^2} + \frac{1}{\|\widehat{\boldsymbol{D}}\|_2}\left(\frac{(4/3\sqrt{5})\lambda\sqrt{L\delta_{2L}^{+}\left(P_{\widehat{\boldsymbol{D}}}\boldsymbol{Z}\right)}}{2\delta_{2L}^{-}(\boldsymbol{Z}) - \delta_{2L}^{+}(\boldsymbol{Z}) - 2\delta_{2L}^{+}\left(P_{\widehat{\boldsymbol{D}}}\boldsymbol{Z}\right)}\right)$$

   where $\delta_k^{+/-}(\boldsymbol{H})$ refers to upper and lower RIP constant of matrix $\boldsymbol{H}$.

The Surrogate Sparsest Penalty

3. Objective function deviates from the original sparsest rule: As shown in Theorem 2 and Remark 1, $g_1(\mathcal{P}) = \|\boldsymbol{\alpha}\|_0$ and $g_2(\mathcal{P}) = \|\boldsymbol{\alpha}\|_1$ correspond to incompatible identification conditions unless satisfying an additional strong requirement

$$\boldsymbol{\alpha}^* = \operatorname{argmin}_{\mathcal{P}=\{\beta,\boldsymbol{\alpha},\boldsymbol{\epsilon}\}\in\mathcal{Q}} g_j(\mathcal{P}), \forall j = 1,2 \Longleftrightarrow \|\boldsymbol{\alpha}^* - c\boldsymbol{\gamma}^*\|_1 > \|\boldsymbol{\alpha}^*\|_1, \forall c \neq 0$$

It further impedes sisVIVE in estimating of $\beta^* \in \mathcal{P}_0$.

How to solve?

- 1 and 2 corresponds to non-igonrable bias in Lasso and RIP conditions, respectively. It could be avoided by other penalties.
- 3 reveals the root of the problem: a proper surrogate penalty in (4) should align with identification condition.

One solution: Windmeijer et al. (2019) proposed to use Adaptive Lasso (Zou, 2006) with proper constructed initial estimator through median estimator. However, it requires the more stringent majority rule (than the sparsest rule, see Remark 1) and suffers from the same sensitivity issue on weak IVs as TSHT and CIIV.

Surrogate Sparsest penalty

Proposition 1 (The proper Surrogate Sparsest penalty)

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

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

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

Example

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

WIT Estimator

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)

Selection Stage: $\widehat{\boldsymbol{\alpha}}^{\mathrm{MCP}} = \underset{\boldsymbol{\alpha}}{\operatorname{argmin}} \frac{1}{2n} \|\boldsymbol{Y} - \widetilde{\boldsymbol{Z}}\boldsymbol{\alpha}\|_2^2 + p_{\lambda}^{\mathrm{MCP}}(\boldsymbol{\alpha}),$

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

Estimation Stage : $\hat{\beta}^{\mathrm{WIT}} = \left(\boldsymbol{D}_{\perp}^{\top}\left(\boldsymbol{I} - \hat{\kappa}_{\mathrm{liml}} M_{\boldsymbol{Z}_{\hat{\mathcal{V}}\perp}}\right)\boldsymbol{D}_{\perp}\right)^{-1}\left(\boldsymbol{D}_{\perp}^{\top}\left(\boldsymbol{I} - \hat{\kappa}_{\mathrm{liml}} M_{\boldsymbol{Z}_{\hat{\mathcal{V}}\perp}}\right)\boldsymbol{Y}_{\perp}\right)$

$$\hat{\kappa}_{\mathrm{liml}} = \lambda_{\min}\left(\{[\boldsymbol{Y}_{\perp}, \boldsymbol{D}_{\perp}]^{\top} M_{\boldsymbol{Z}_{\hat{\mathcal{V}}\perp}} [\boldsymbol{Y}_{\perp}, \boldsymbol{D}_{\perp}]\}^{-1}\{[\boldsymbol{Y}_{\perp}, \boldsymbol{D}_{\perp}]^{\top}[\boldsymbol{Y}_{\perp}, \boldsymbol{D}_{\perp}]\}\right)$$

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

Regularization condition of WIT

- restricted cone $\mathscr{C}\left(\mathcal{V}^{*};\xi\right)=\left\{\boldsymbol{u}:\left\|\boldsymbol{u}_{\mathcal{V}^{*}}\right\|_{1}\leq\xi\left\|\boldsymbol{u}_{\mathcal{V}^{c*}}\right\|_{1}\right\}$ for some $\xi>0$ that estimation error $\widehat{\boldsymbol{\alpha}}-\boldsymbol{\alpha}^{*}$ belongs to.
- The restricted eigenvalue $K_{\mathscr{C}}$ for $\tilde{\boldsymbol{Z}}$ formally is defined as $K_{\mathscr{C}}=K_{\mathscr{C}}\left(\mathcal{V}^{*},\xi\right):=\inf_{\boldsymbol{u}}\left\{\|\tilde{\boldsymbol{Z}}\boldsymbol{u}\|_{2}/\left(\|\boldsymbol{u}\|_{2}n^{1/2}\right):\boldsymbol{u}\in\mathscr{C}\left(\mathcal{V}^{*};\xi\right)\right\}$ and RE condition refers to that $K_{\mathscr{C}}$ for $\tilde{\boldsymbol{Z}}$ should be bounded away from zero.

LEMMA 2. (RE condition of $\tilde{Z}$)

Under assumption $A1-4$, For any given $\gamma^{*}\neq\mathbf{0}$, there always exists a constant $\xi\in\left(0,\left\|\widehat{\gamma}_{\mathcal{V}^{*}}\right\|_{1}/\left\|\widehat{\gamma}_{\mathcal{V}^{c*}}\right\|_{1}\right)$ and further define the restricted cone $\mathscr{C}\left(\mathcal{V}^{*};\xi\right)$ such that $K_{\mathscr{C}}^{2}>0$ holds strictly.

Lemma 2 elaborates RE condition of $\tilde{\boldsymbol{Z}}$ holds without any additional assumptions on $\tilde{\boldsymbol{Z}}$, unlike sisVIVE. Moreover, this restricted cone is invariant of scaling and, thus, indicates accommodation of many weak IVs cases.

Selection Consistency

Theorem 3 (Selection Consistency of Valid IVs)

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

$$\kappa(n) \asymp \underbrace{\sqrt{\frac{\log p_{\mathcal{V}^*}}{n}}}_{T_1} + \underbrace{\frac{p_{\mathcal{V}^*}}{n} \cdot \frac{\|\tilde{\tilde{\boldsymbol{Q}}}_n \gamma^*_{\mathcal{V}^*}\|_\infty}{\gamma^{*\top}_{\mathcal{V}^*} \tilde{\tilde{\boldsymbol{Q}}}_n \gamma^*_{\mathcal{V}^*}}}_{T_2} + \underbrace{|\,\mathrm{Bias}(\hat{\beta}^{\mathrm{TSLS}}_{\mathrm{or}})|\|\bar{\gamma}^*_{\mathcal{V}^{c*}}\|_\infty}_{T_3}, \tag{23}$$

$$\kappa^c(n) \asymp (1+c)\left\{\sqrt{\frac{\log|\mathcal{I}_c|}{n}} + \frac{|\mathcal{I}_c|}{n} \cdot \frac{\|\tilde{\tilde{\boldsymbol{Q}}}^c_n \gamma^*_{\mathcal{I}_c}\|_\infty}{\gamma^{*\top}_{\mathcal{I}_c} \tilde{\tilde{\boldsymbol{Q}}}^c_n \gamma^*_{\mathcal{I}_c}}\right\} + |\,\mathrm{Bias}(\hat{\beta}^{c\ \mathrm{TSLS}}_{\mathrm{or}})|\|\bar{\gamma}^*_{\mathcal{I}^c_c}\|_\infty, \tag{24}$$

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

$$\widehat{\boldsymbol{\alpha}}^{\mathrm{MCP}} = \underset{\widehat{\boldsymbol{\alpha}} \in \mathscr{B}_0(\lambda,\rho)}{\mathrm{argmin}} \|\widehat{\boldsymbol{\alpha}}\|_0, \;\; \Pr(\hat{\mathcal{V}} = \mathcal{V}^*, \widehat{\boldsymbol{\alpha}}^{\mathrm{MCP}} = \widehat{\boldsymbol{\alpha}}^{\mathrm{or}}) \xrightarrow{p} 1. \tag{25}$$

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 γ_j^*, i.e.
$\|\boldsymbol{\tilde{\tilde{Q}}}_n^{1/2}\gamma^*_{\mathcal{V}^*}\|_\infty / \|\boldsymbol{\tilde{\tilde{Q}}}_n^{1/2}\gamma^*_{\mathcal{V}^*}\|_1 = o\left(\|\boldsymbol{\tilde{\tilde{Q}}}_n^{1/2}\gamma^*_{\mathcal{V}^*}\|_1 / (p_{\mathcal{V}^*}\|\boldsymbol{\tilde{\tilde{Q}}}_n^{1/2}\|_\infty)\right)$, then $T_2 \to 0$.

Proposition 2 (**Approximation of** $\text{Bias}(\hat{\beta}_{or}^{TSLS})$)

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

$$E\left[\text{Bias}(\hat{\beta}_{\text{or}}^{\text{TSLS}})\right] = \frac{\sigma_{\epsilon\eta}}{\sigma_\eta^2}\left(\frac{p_{\mathcal{V}^*}}{(\mu_n + p_{\mathcal{V}^*})} - \frac{2\mu_n^2}{(\mu_n + p_{\mathcal{V}^*})^3}\right) + o\left(s^{-1}\right). \qquad (26)$$

Discussion on T3

The rate of concentration parameter μ_n will affect T_3 through $|\text{Bias}(\hat{\beta}_{\text{or}}^{\text{TSLS}})|$ under many IVs setting. Suppose Assumption 4 holds, that $\mu_n \xrightarrow{p} \mu_0 n$, the leading term in (26) is $\frac{\sigma_{\epsilon\eta}}{\sigma_\eta^2}\frac{\nu_{p_{\mathcal{V}^*}}}{\mu_0 + \nu_{p_{\mathcal{V}^*}}} \ll \frac{\sigma_{\epsilon\eta}}{\sigma_\eta^2}$ for moderate μ_0.

Asymptotic of WIT

Theorem 4 (Consistency and Asymptotic Normality)

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

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

$$\widehat{\text{Var}}(\hat{\beta}^{\text{WIT}}) = \frac{\hat{\boldsymbol{b}}^\top \widehat{\boldsymbol{\Omega}} \hat{\boldsymbol{b}}(\hat{\mu}_n + L/n)}{-\hat{\mu}_n}\left(\hat{Q}_S \widehat{\boldsymbol{\Omega}}_{22} - \boldsymbol{T}_{22} + \frac{\hat{c}}{1-\hat{c}} \frac{\hat{Q}_S}{\hat{\boldsymbol{a}}^\top \widehat{\boldsymbol{\Omega}}^{-1} \hat{\boldsymbol{a}}}\right)^{-1}$$

$$\xrightarrow{p} \mu_0^{-2}\big[\sigma_\epsilon^2\mu_0 + \frac{\upsilon_K(1-\upsilon_L)}{1-\upsilon_K-\upsilon_L}|\boldsymbol{\Sigma}|\big],$$

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

Simulation

Further, we present a replication of simulation design in literature and its variant:
Case 1(III) : $\boldsymbol{\gamma}^* = (\mathbf{0.4}_{21})^\top$ and $\boldsymbol{\alpha}^* = (\mathbf{0}_9, \mathbf{0.4}_6, \mathbf{0.2}_6)^\top$.
Case 1(IV) : $\boldsymbol{\gamma}^* = (\mathbf{0.15}_{21})^\top$ and $\boldsymbol{\alpha}^* = (\mathbf{0}_9, \mathbf{0.4}_6, \mathbf{0.2}_6)^\top$.

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.)

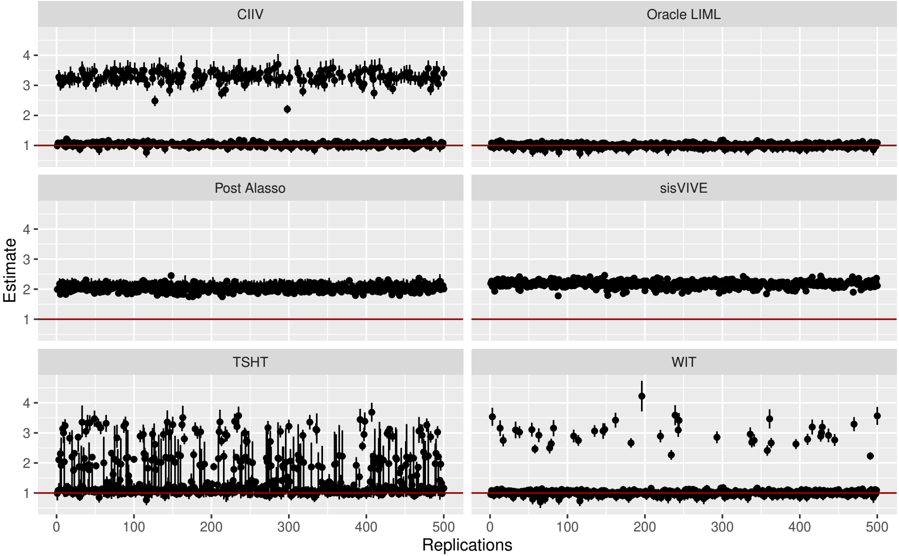


Figure: Scatter plot of estimations of β^* with confidence intervals of Case 1 (IV)

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,

  $$\log(Y_i) = \alpha + \beta T_i + \psi S_i + \epsilon_i \tag{27}$$

  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 ϵ_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

$$T_i = \gamma^\top \boldsymbol{Z_i} + \nu_i, \tag{28}$$

where $\boldsymbol{Z_i}$ 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)	$\boldsymbol{S}_1$	Control Variable	1.382	1.803	1.480	-3.037	6.674
log(Land Area)	$\boldsymbol{S}_2$	Control Variable	11.726	2.260	12.015	5.680	16.611
$\widehat{T}$ (proxy for trade)	$\boldsymbol{Z}_1$	IV	0.093	0.052	0.079	0.015	0.297
log(Water Area)	$\boldsymbol{Z}_2$	IV	6.756	3.654	7.768	0	13.700
log(Land Boundaries)	$\boldsymbol{Z}_3$	IV	6.507	2.920	7.549	0	10.005
% Forest	$\boldsymbol{Z}_4$	IV	29.89	22.380	30.62	0	98.26
% Arable Land	$\boldsymbol{Z}_5$	IV	40.947	21.549	42.062	0.558	82.560
Languages	$\boldsymbol{Z}_6$	IV	1.873	2.129	1	1	16
Annual Freshwater	$\boldsymbol{Z}_7$	IV	2.190	2.129	2.155	-2.968	8.767

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

Trade on GDP

We first standardize all the variables, then we formulate the structural equation as:

$$\log(Y_i) = T_i\beta + \boldsymbol{Z}_{i.}^\top \boldsymbol{\alpha} + \boldsymbol{S}_{i.}^\top \psi + \epsilon_i \quad \text{for } i = 1, 2, \ldots, 158, \tag{29}$$

Partial out of Control: $\ddot{Y}_i = \ddot{T}_i\beta + \ddot{\boldsymbol{Z}}_{i.}^\top \boldsymbol{\alpha} + \ddot{\epsilon}_i, \quad \ddot{T}_i = \ddot{\boldsymbol{Z}}_{i.}^\top \phi + \ddot{\nu}_i.$

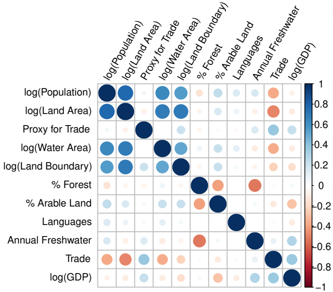


Figure: Correlation of all variables

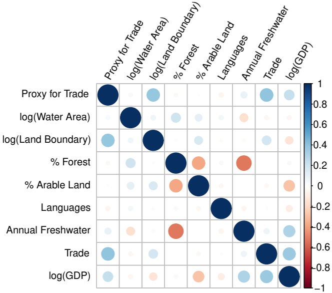


Figure: Correlation of transformed variables

Estimation

Table: Empirical Results of Various Estimators

	$\hat{\beta}\left(\widehat{\text{Var}}^{1/2}(\hat{\beta})\right)$	95% CI	Valid IVs $\widehat{\mathcal{V}}$	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.

Observations

Observation:

- 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.
- $\boldsymbol{Z}_5$ should be a invalid IV:
  1. In view of marginal correlation in Fig. 12, $\boldsymbol{Z}_5$ is nearly uncorelated to trade but significantly correlated with log(GDP).
  2. Concerning the Sargen Test, p-value of $0.086 < 0.1$ in Post-Alasso indicates $\boldsymbol{Z}_5$ is not very credible to be valid.
  3. 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.

Reference

Anderson, J. E. (1979). A theoretical foundation for the gravity equation. *Am. Econ. Rev.*, 69(1):106–116.

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.

Guo, Z., Kang, H., Tony Cai, T., and Small, D. S. (2018). Confidence intervals for causal effects with invalid instruments by using two-stage hard thresholding with voting. *J. R. Statist. Soc. B*, 80(4):793–815.

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.

Kang, H., Zhang, A., Cai, T. T., and Small, D. S. (2016). Instrumental variables estimation with some invalid instruments and its application to mendelian randomization. *J. Am. Statist. Ass.*, 111(513):132–144.

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