Summer 7-16-2021

Three Essays on Causal Inference With Model Averaging

Guanyu Liu
Syracuse University

ABSTRACT

This dissertation contains essays on causal inference with model averaging. The first essay presents a theoretical derivation of a model-averaging-based average treatment effect estimator. The second essay provides comparison of predictability of treated counterfactual outcome between model averaging and other methods. The third essay is an empirical study evaluating the economic impact of Ukraine's 2013 conflict.

The first essay constructs a new average treatment effect estimator based on model averaging in a panel data setting. The estimator is shown to be asymptotically unbiased and consistent. Its asymptotic distribution is derived, which turns out to be non-normal and non-standard. A subsampling procedure is then applied to obtain valid inference. Simulation results show that the proposed estimator compares favourably with alternative estimators in out-of-sample prediction accuracy under a common factor structure.

The second essay further compares predictability of treated counterfactual outcome between model averaging and other methods under more general set-ups. The simulations show that the model averaging and penalized regression methods yield more accurate counterfactual prediction than the model selection methods. We also find evidences that if the predictors (e.g., control units' outcomes) are more correlated, the model averaging methods have more accurate prediction than the penalized regression, and vice versa.

The third essay evaluates the economic impact of Ukraine's 2013 conflict using a comparative case study. A modified synthetic control method is applied to account for potential spillover from the conflict on Ukraine's neighbouring countries. The results show that Ukraine's real GDP was reduced by 29.7% from late-2013 to the end of 2015. The spillover effects are detected in every quarter since the conflict began. Furthermore, negative spillover effects are found in countries selected by the modified synthetic control.

THREE ESSAYS ON CAUSAL INFERENCE WITH MODEL AVERAGING

By

Guanyu Liu

B.A., Beijing Foreign Studies University, 2012

M.A., Syracuse University, 2015

Dissertation

Submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in *Economics*

Syracuse University

July 2021

ACKNOWLEDGEMENTS

I would like to acknowledge all those who helped me complete this dissertation at Syracuse University. I am grateful for their patient guidance, friendly advice and numerous help.

In particular, I would like to express my deep and sincere gratitude to my brilliant advisors, Prof. Yoonseok Lee, Prof. Alfonso Flores-Lagunes and Prof. Hugo Jales, for their continuous encouragement, wholehearted support and immense knowledge. Coming from non-Economics undergraduate background, I have gained so much knowledge and training from their courses. I appreciate many discussions with them which not only sparked innovative research ideas but also provided enlightening suggestions.

I am also indebted to Prof. Chihwa Kao, Prof. Jan Ondrich and Prof. Pinyuen Chen who helped me with recommendation letters for graduate school applications. I would not be able to pursue my Ph.D. study without their encouragement and support.

My special thanks to my friends Yinbo Mao and Yinhan Zhang who offered me lots of rides to school and grocery stores. I would like to thank my classmates for the constructive research discussions and all the fun we have had in the past few years. I am very grateful to the department professors for their vivid classes and staff for their administrative support. I also would like to express my thanks to the financial support from the Department of Economics.

My final and most deeply felt appreciation goes to my parents and my wife, who have loved me, supported my and accompanied my through all the ups and downs during my graduate school.

TABLE OF CONTENTS

1. Introduction 1

2. Estimation and inference of average treatment effects with model averaging 4

1. Introduction . . . 5

2. Model and estimation . . . 7

2.1 Theoretical model . . . 7

2.2 Estimation . . . 9

3. MA-based ATE estimator . . . 11

4. Asymptotic properties of MA-based ATE estimator . . . 14

4.1 Consistency . . . 15

4.2 Asymptotic distribution . . . 15

5. Inference of MA-based ATE estimator . . . 17

6. Simulation study . . . 19

6.1 Comparison of different estimation methods . . . 19

6.2 The coverage probabilities of inference procedure . . . 21

7. Empirical application . . . 22

8. Conclusion . . . 28

Appendix . . . 37

References . . . 51

3. Comparing predictability between model averaging and other methods **54**

1. Introduction . . . 55

2. Description of models and estimation . . . 56

2.1 Interactive fixed-effect model . . . 56

2.2 Factor-augmented regression model . . . 58

2.3 Estimation methods . . . 59

3. Simulation studies . . . 65

4. Concluding remarks . . . 74

Appendix . . . 76

References . . . 78

4. Evaluating the Economic Impact of Conflict: A counterfactual analysis of Ukraine **81**

1. Introduction . . . 82

2. Background . . . 83

3. Methodology . . . 86

3.1 The model and estimation . . . 86

3.2 Statistical inference . . . 89

4. Evaluating the economic impact of the conflict . . . 90

4.1 Impact on real GDP . . . 90

4.2 Spillovers . . . 92

4.3 Robustness check . . . 93

5. Conclusion . . . 95

References . . . 99

5. Conclusions **100**

VITA **102**

LIST OF FIGURES

2. Estimation and inference of average treatment effects with model averaging

Figure 1: Actual and counterfactual real GDP growth rate of Hong Kong 23

Figure 2: JMA: autocorrelation of $\hat{e}_{1t}$: 1993:1-1997:1 and $\hat{\nu}_{1t}$: 1994 :2-2003:4 24

Figure 3: MMA: autocorrelation of $\hat{e}_{1t}$:1993:1-1997:1 and $\hat{\nu}_{1t}$:1994:2-2003:4 24

Figure 4: Political integration: ATE estimate based on MMA and ATE estimate based on JMA 25

Figure 5: Actual and counterfactual real GDP growth rate of Hong Kong 26

Figure 6: JMA: autocorrelation of $\hat{e}_{1t}$:1993:1-2003:4 and $\hat{\nu}_{1t}$:2004:1-2008:1 27

Figure 7: MMA: autocorrelation of $\hat{e}_{1t}$:1993:1-2003:4 and $\hat{\nu}_{1t}$:2004:1-2008:1 27

Figure 8: Economic integration: ATE estimate based on MMA and ATE estimate based on JMA 28

3. Comparing predictability between model averaging and other methods

Figure 1: Relative MSPE for $h = 1, 4, 8, 12$ ($p_{max} = 0$) 71

Figure 2: Relative MSPE for $h = 1, 4, 8, 12$ ($p_{max} = 2$) 72

Figure 3: Relative MSPE for $h = 1, 4, 8, 12$ ($p_{max} = 4$) 73

Figure 4: Relative MSPE for $h = 1, 4, 8, 12$ ($p_{max} = 9$) 74

4. Evaluating the Economic Impact of Conflict: A counterfactual analysis of Ukraine

Figure 1: Development of Ukraine's conflict: 2013.11 - 2016.01 86
Figure 2: Ukraine's actual and counterfactual log(RGDP) 92
Figure 3: Gap in log(RGDP) between Ukraine and counterfactual Ukraine (with 95% confidence interval) 94
Figure 4: Gaps in log(RGDP) with 95% confidence interval 95

LIST OF TABLES

2. Estimation and inference of average treatment effects with model averaging

Table 1: Comparison of different estimation methods: one-factor 30
Table 2: Comparison of different estimation methods: two-factor 31
Table 3: Comparison of different estimation methods: three-factor 32
Table 4: Coverage probabilities ($\alpha_0 = 0$, no treatment effects) 33
Table 5: Coverage probabilities ($\alpha_0 = 1$, positive treatment effects) 34
Table 6: Confidence intervals of MA-based ATE (political integration) 35
Table 7: Confidence intervals of MA-based ATE (economic integration) 35
Table 8: Political integration: MMA coefficients estimates 35
Table 9: Economic integration: MMA coefficients estimates 36

3. Comparing predictability between model averaging and other methods

Table 1: MSPE of different methods 67
Table A1: Sample correlation matrix of control units: factor model with three factors . . 76
Table A2: Sample correlation matrix of control units: IFE model 77

4. Evaluating the Economic Impact of Conflict: A counterfactual analysis of Ukraine

Table 1: Weights of control units for log(RGDP) 97
Table 2: Spillover effects estimates in log(RGDP) for selected control countries 97

Table 3: Weights of control units for log(RGDP) (robustness check) 98

1. Introduction

In social sciences, we often want to know the "treatment effect" of an intervention or an event d on an outcome of interest y. A panel data $\{y_{it}\}_{i=1,t=1}^{N,T}$ records rich information and it is often used to investigate the treatment effect. Under potential - treated and untreated - outcomes framework, the observed outcome y_{it} can be either treated y_{it}^1 if $d_{it} = 1$ or untreated y_{it}^0 if $d_{it} = 0$. The treatment effect is defined as the difference between the two potential outcomes, that is, $\Delta_{it} = y_{it}^1 - y_{it}^0$.

In this dissertation, we mainly consider a set-up with only a single treated unit, a fixed number of control units and large pre- and post-treatment periods. Without loss of generality, we assume the first unit receives the treatment at $T_1 + 1$. We are interested in the temporal average treatment effect

$$\Delta_1 = \mathbb{E}(\Delta_{1t}) = \mathbb{E}(y_{1t}^1 - y_{1t}^0) \quad t = T_1 + 1, \ldots, T$$

The fundamental problem in the Causal Inference is that only y_{1t}^1 is observed whereas y_{1t}^0 is not in the post-treatment periods. Therefore, we need to predict the missing counterfactual outcome y_{1t}^0. The major challenge in the current context is to accurately predict y_{1t}^0, Δ_1 and to provide measure of uncertainty for the corresponding estimators $\hat{y}_{1t}^0$, $\hat{\Delta}_1 = (T - T_1)^{-1} \sum_{T_1+1}^{T} (y_{1t} - \hat{y}_{1t}^0)$.

In the current context, synthetic control method (SCM) (Abadie et al. 2003, 2010) is perhaps one of most popular methods to investigate treatment effects and average treatment effects. It uses a convex combination of control units' outcomes - called *synthetic control* - to predict y_{1t}^0 in the post-treatment periods. The non-negativity and sum-to-one constrains on control units' weights are used to select control units from control group. For the inference, it uses placebo tests under the assumption that the treated unit is randomly assigned so that counterfactual outcome for each unit is estimated and a distribution of differences between actual outcome and counterfactual outcome is obtained. The treatment effect is significant if it is very large relative to this distribution of differences. However, the treated unit is treated for some reasons, thus the random-assignment assumption may not hold in many cases. An al-

ternative method to SCM is panel data approach (HCW) (Hsiao et al., 2012), which is a least-squares-based method. It uses a two-step procedure to select proper control units, which is essentially minimizing model selection criterion AIC or AICC. However, the formal inference procedure was not discussed in the original paper.

It is worth noting that the data-driven weight selection processes used in SCM and HCW inevitably pose the difficulty of post-model-selection inference. After model selection, the conventional inference procedure based on normal approximation is inaccurate and the distortions are potentially unbounded.

In Chapter 2, we use model averaging (MA) methods to construct the treated counterfactual outcome instead of choosing a single model with a subset of control units based on certain criterion. Within a regression framework, the MA is to firstly obtain (least squares) estimator from each candidate model; then take a weighted average of these estimators. The resulting averaging estimator is used as control units' weights to form the estimator of treated counterfactual outcome. We then derive the asymptotic distribution of MA-based ATE estimator, which turns out to be non-normal and non-standard. We instead apply a subsampling-bootstrap method to obtain valid inference result.

In Chapter 3, we conduct extensive simulations to compare predictability between model averaging and other methods in terms of mean squared prediction error. The simulation results show that our proposed MA-based ATE estimator compares favourably with alternative methods.

In Chapter 4, we examine the economic impact of Ukraine's 2013 conflict. Suspecting that the conflict could generate spillover effects on Ukraine's neighbouring countries and trade partners, we apply a modified synthetic control methods that accounts for potential spillovers. We consider Ukraine as the treated unit. We pre-specify some countries to have potential spillovers and other countries to be control countries. For each country, we obtain its synthetic control weights estimates. The weights estimates are then used to predict treatment and spillover effects.

2. Estimation and inference of average treatment effects with model averaging

1. Introduction

In many social science disciplines, it is desired to know the average treatment effect (ATE) of economic events or policy intervention (i.e., treatment). The major challenge is to accurately estimate the counterfactual outcome - the hypothetical outcome in the absence of treatment, which is the key element in estimating ATE. There are several ways to estimate the counterfactual outcome and ATE, one of them is panel data approach by Hsiao et al. (2012) (hereafter HCW). The approach is motivated by a factor model where the potential outcomes are generated by some unobserved common factors. For example, in a macroeconomic setting, the output or growth rate of different countries can be affected by common factors such as technological progress, business cycle, financial crises, etc. The information embedded in control units not subject to the treatment can help explain the counterfactual outcome of treated unit. Therefore, the observed outcomes of control units are used to construct the counterfactual outcome. Another popular approach is the synthetic control methods by Abadie et al. (2003, 2010) (hereafter SCM)[1]. Intuitively, SCM constructs a synthetic control as a convex combination of control units to maximally resemble the treated unit in terms of the outcome and a number of attributes (if available) over pre-treatment periods. The evolution of the synthetic control thereafter is considered as prediction of counterfactual outcome of treated unit.

Model selection is a common problem in HCW, SCM, and related methods. This is due to the challenge of finding proper control units to build the "optimal" model for predicting counterfactual. Here "optimal" can be defined in terms of a proper loss function such as squared error loss in prediction. HCW proposed a two-step procedure to select proper control units, which is essentially minimizing the Akaike information criterion (AIC; Akaike, 1973, 1974) or the corrected Akaike information criterion (AICC; Hurvich and Tsai, 1989). On the other hand, the non-negativity and sum-to-one weight restrictions imposed in SCM can be viewed as a regularization device to select control units from the control group (see Doudchenko and Imbens, 2016). However, these data-driven selection processes inevitably pose challenges with statistical inference after model selection, as it is well-known that after model selection,

[1] See Gardeazabal and Vega-Bayo (2017) for a comparison between HCW and SCM

the inferential procedures derived from the classical theory are inaccurate and make predictions with overoptimistic confidence (Faraway, 1992; Berk et al., 2013).

Instead of choosing one "optimal" model based on an information criterion or regularization procedure, we suggest using model averaging (MA) method to construct the treated counterfactual outcome and estimate ATE. Specifically, we replace the 2-step procedure in HCW with two frequentist MA methods: the Mallows model averaging (MMA; Hansen 2007) and the Jackknife model averaging (JMA; Hansen and Racine, 2012). We demonstrate the asymptotic unbiasedness of the MA-based ATE estimator and derive its asymptotic distribution, which is non-normal and non-standard. Nonetheless, we show that valid inference regarding ATE estimator can be obtained by using a subsampling-based procedure. On the other hand, our simulation results show that under the common factor structure, the MA-based ATE estimator achieves smaller mean squared prediction error relative to many existing estimators, though we do not provide theoretical justification of this finite-sample improvement. MA methods have long been popular within the Bayesian paradigm, in the meantime there is rapidly-growing literature on frequentist model averaging. We direct interested readers to Hoeting et al. (1999) for Bayesian model averaging and Hansen (2007) and Wan et al. (2010) for frequentist model averaging. The idea of using MA method for estimating ATE has been proposed before by Long et al. (2015), but their paper focuses on improving prediction accuracy of JMA relative to AIC and AICC. To date, there has been no formal inference theory for ATE estimator based on model averaging methods. Thus, our main contributions are: (i) we establish asymptotic properties of the MA-based ATE estimator in the panel data setting; (ii) we apply a subsampling procedure to obtain valid inference of this ATE estimator, thus avoid tackling the intricate post-selection inference problems inherent in many existing methods.

We illustrate the proposed method by revisiting two empirical examples in HCW. The first example measures the economic impact from transfer of sovereignty over Hong Kong on July 1, 1997, and the second example evaluates the impact of the implementation of the Closer Economic Partnership Arrangement (CEPA) between mainland China and Hong Kong, started on January 1, 2004. Using our proposed ATE estimator, we estimate that the transfer

of sovereignty over Hong Kong - the political integration - did not have statistically significant impact on Hong Kong's real GDP growth. On the other hand, Hong Kong's economy benefited from the implementation of the CEPA. More specifically, from the first quarter in 2004 to the first quarter in 2008, the CEPA raised the annual real GDP growth rate of Hong Kong by 3.8% based on MMA and 3.9% based on JMA. Our estimation and inference results support the original findings in HCW.

The remainder of this paper is organized as follows. Section 2 presents the theoretical model and common methods for estimating ATE. Section 3 presents MA-based ATE estimator. Section 4 presents the asymptotic properties for the proposed estimator. Section 5 presents Monte Carlo simulation results. Section 6 presents empirical applications. Section 7 concludes.

2. Model and estimation

2.1. Theoretical model

Let y_{it}^1 and y_{it}^0 denote the outcome variable for unit i at time t with and without treatment. The treatment effect for the ith unit at time t is defined as

$$\Delta_{it} = y_{it}^1 - y_{it}^0 \tag{1}$$

As we often do not observe simultaneously y_{it}^0 and y_{it}^1, the observed outcome is $y_{it} = d_{it}y_{it}^1 + (1 - d_{it})y_{it}^0$ where $d_{it} = 1$ if unit i is under treatment at time t and $d_{it} = 0$ otherwise. If the treatment effects Δ_{it} follow a stationary process, we can define the ATE as $\Delta_i = \mathbb{E}(\Delta_{it})$ and estimate it by taking the simple average of treatment effects over post-treatment periods.

Assume $\boldsymbol{f}_t$ is a $K \times 1$ vector of unobserved common factors that drive outcomes of all units to change over time. We consider the case that only one unit is treated at time $t = T_1 + 1, ..., T$. Without loss of generality, let it be the first unit. We follow HCW and consider a fac-

tor model

$$y_{it}^0 = \alpha_i + \boldsymbol{b}_i' \boldsymbol{f}_t + \epsilon_{it}, \quad i = 1, ..., N, t = 1, ..., T \tag{2}$$

where $\boldsymbol{b}_i$ is a $K \times 1$ vector of factor loadings for unit i, α_i is individual fixed effect and ϵ_{it} is the ith unit's idiosyncratic error term with $\mathbb{E}(\epsilon_{it}) = 0$. In matrix form,

$$\boldsymbol{y}_t^0 = \boldsymbol{\alpha} + \boldsymbol{B} \boldsymbol{f}_t + \boldsymbol{\epsilon}_t, \tag{3}$$

where $\boldsymbol{y}_t^0 = (y_{1t}^0, ..., y_{Nt}^0)'$, $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_N)'$, $\boldsymbol{B} = (\boldsymbol{b}_1, ..., \boldsymbol{b}_N)'$ is the $N \times K$ matrix and $\boldsymbol{\epsilon}_t = (\epsilon_{1t}, ..., \epsilon_{Nt})'$.

The first unit has been treated since $T_1 + 1$, hence $y_{1t} = y_{1t}^1$ for $t = T_1 + 1, ..., T$ and $y_{it} = y_{it}^0$ for $i = 2, ..., N$ and $t = 1, ..., T$. The counterfactual outcome y_{1t}^0 is not observed for post-treatment periods, we need to predict it. HCW proposed using $\boldsymbol{y}_{-1t} = (y_{2t}, ..., y_{Nt})'$ instead of $\boldsymbol{f}_t$ to predict y_{1t}^0[2]. Specifically, let $\boldsymbol{a} = (1, -\tilde{\boldsymbol{a}}_{-1}')'$ where $\tilde{\boldsymbol{a}}_{-1}' = (a_2, ..., a_N)'$ such that $\boldsymbol{a}'\boldsymbol{B} = 0$, i.e., $\boldsymbol{a}$ is in the null space of $\boldsymbol{B}$. Then we have $\boldsymbol{a}'\boldsymbol{y}_t^0 = y_{1t}^0 - \tilde{\boldsymbol{a}}_{-1}'\boldsymbol{y}_{-1t} = \boldsymbol{a}'\boldsymbol{\alpha} + \epsilon_{1t} - \tilde{\boldsymbol{a}}_{-1}'\boldsymbol{\epsilon}_{-1t}$ because $\boldsymbol{a}'\boldsymbol{B} = 0$. After rearranging terms, $y_{1t}^0 = \boldsymbol{a}'\boldsymbol{\alpha} + \tilde{\boldsymbol{a}}_{-1}'\boldsymbol{y}_{-1t} + \tilde{\epsilon}_{1t}$, where $\tilde{\epsilon}_{1t} = \epsilon_{1t} - \tilde{\boldsymbol{a}}_{-1}'\boldsymbol{\epsilon}_{-1t}$, $\boldsymbol{\epsilon}_{-1t} = (\epsilon_{2t}, ..., \epsilon_{Nt})'$. Because $\tilde{\epsilon}_{1t}$ depends on all $\epsilon_{1t}, ..., \epsilon_{Nt}$, $\tilde{\epsilon}_{1t}$ is correlated with $\boldsymbol{y}_{-1t}$. Denote $e_{1t} = \tilde{\epsilon}_{1t} - \mathbb{E}(\tilde{\epsilon}_{1t}|\boldsymbol{y}_{-1t})$, it is clear that $\mathbb{E}(e_{1t}|\boldsymbol{y}_{-1t}) = 0$. HCW further assumed a linear conditional mean function, i.e., $\mathbb{E}(\tilde{\epsilon}_{1t}|\boldsymbol{y}_{-1t}) = c_1 + \boldsymbol{c}'\boldsymbol{y}_{-1t}$, which leads to

$$y_{1t}^0 = \bar{\alpha} + \boldsymbol{a}_{-1}'\boldsymbol{y}_{-1t} + e_{1t} \tag{4}$$

where $\bar{\alpha} = \boldsymbol{a}'\boldsymbol{\alpha} + c_1$, $\boldsymbol{a}_{-1} = \tilde{\boldsymbol{a}}_{-1} + \boldsymbol{c}$. Let $\boldsymbol{x}_t = (1, \boldsymbol{y}_{-1t}')'$ and $\boldsymbol{\beta} = (\bar{\alpha}, \boldsymbol{a}_{-1}')'$, we rewrite model (4) as

$$y_{1t}^0 = \boldsymbol{x}_t'\boldsymbol{\beta} + e_{1t} \tag{5}$$

The least squares regression on (5) will give consistent estimator of $\boldsymbol{\beta}$.

[2]HCW argued that the information provided by $\boldsymbol{f}_t$ is embedded in $\boldsymbol{y}_{-1t}$. Also, it may be difficult to identify $\boldsymbol{f}_t$ and $\boldsymbol{B}$ in a finite sample where the number of control units or the number of time periods is small or moderate.

For the whole sample period, we can also write the observed outcome of treated unit as

$$y_{1t} = y_{1t}^0 + \Delta_{1t} d_{1t} = \boldsymbol{x}_t' \boldsymbol{\beta} + \Delta_{1t} d_{1t} + e_{1t} \tag{6}$$

where $d_{1t} = 0$ if $t < T_1$ and $d_{1t} = 1$ otherwise. The treatment effect estimator at $t = T_1 + 1, ..., T$ is defined as $\hat{\Delta}_{1t} = y_{1t} - \hat{y}_{1t}^0$. Let $T_2 = T - T_1$, the ATE estimator is given by averaging $\hat{\Delta}_{1t}$ over post-treatment periods[3]:

$$\hat{\Delta}_1 = T_2^{-1} \sum_{t=T_1+1}^{T} \hat{\Delta}_{1t} \tag{7}$$

2.2. Estimation

HCW is a least-squares-based method, the parameter $\boldsymbol{\beta}$ in (5) can be estimated via the following unconstrained minimization problem:

$$\hat{\boldsymbol{\beta}}_{\text{OLS}} = \text{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^N} \sum_{t=1}^{T_1} (y_{1t} - \boldsymbol{x}_t' \boldsymbol{\beta})^2 \tag{8}$$

However, if the mean-square criterion is adopted, it is well known that the optimal estimator is not necessarily based on the largest model. HCW suggested using Akaike Information Criterion (AIC; Akaike, 1973,1974) or the corrected Akaike information criterion (AICC; Hurvich and Tsai, 1989) to select control units. Specifically,

1) Use R^2 to select the best predictor for y_{1t}^0 using j out of $N - 1$ control units, denoted by $M(j)^*$ for $j = 1, ..., N - 1$;

2) From $M(1)^*, M(2)^*, ..., M(N-1)^*$ choose $M(m)^*$ in terms of information criterion such as AIC and AICC, etc.

The counterfactual outcome is estimated by $\hat{y}_{1t,\text{HCW}}^0 = \boldsymbol{x}_t^{*\prime} \hat{\boldsymbol{\beta}}_{\text{HCW}}^*$ and the ATE estimator is $\hat{\Delta}_{1,\text{HCW}} = T_2^{-1} \sum_{t=T_1+1}^{T} (y_{1t} - \hat{y}_{1t,\text{HCW}}^0)$.

Du and Zhang (2015) proposed an approach that is anchored by the original HCW's model

[3]To be more rigorous, it is temporal average treatment effect on the treated estimator.

and replaces above two-step procedure with a "leave-many-out" cross-validation method based on Shao (1993). Their simulation results showed that the CV method gives smaller post-treatment mean squared prediction error than that from original HCW.

A popular alternative method is the Synthetic Control methods (SCM). SCM use a weighted average of control units' outcomes to construct the counterfacual outcome of the treated unit. The weights are selected by best fitting the treated unit's outcome (and its covariates if they are available) using pre-treatment data and are assumed to be non-negative and sum-to-one. That is, $\boldsymbol{\beta}$ is estimated via the constrained minimization problem:

$$\hat{\boldsymbol{\beta}}_{\text{SC}} = \text{argmin}_{\boldsymbol{\beta}\in\Lambda_{\text{SC}}} \sum_{t=1}^{T_1} (y_{1t} - \boldsymbol{x}_t'\boldsymbol{\beta})^2 \tag{9}$$

where $\Lambda_{\text{SC}} = \{\boldsymbol{\beta} \in \mathbb{R}^{N-1} : \beta_j \geq 0 \text{ for } j = 2, ..., N \text{ and } \sum_{j=2}^{N} \beta_j = 1\}$. Then $\hat{y}^0_{1t,\text{SC}} = \boldsymbol{x}_t'\hat{\boldsymbol{\beta}}_{\text{SC}}$ and the ATE estimator is $\hat{\Delta}_{1,\text{SC}} = T_2^{-1} \sum_{t=T_1+1}^{T} (y_{1t} - \hat{y}^0_{1t,\text{SC}})$. A modified synthetic control method (MSC) was proposed by Doudchenko and Imbens (2016) and formalized by Li (2019). The proposed modifications include adding an intercept and dropping the weight sum-to-one constraint, that is,

$$\hat{\boldsymbol{\beta}}_{\text{MSC}} = \text{argmin}_{\boldsymbol{\beta}\in\Lambda_{\text{MSC}}} \sum_{t=1}^{T_1} (y_{1t} - \boldsymbol{x}_t'\boldsymbol{\beta})^2 \tag{10}$$

where $\boldsymbol{x}_t = (1, \boldsymbol{y}'_{-1t})'$ and $\Lambda_{\text{MSC}} = \{\boldsymbol{\beta} \in \mathbb{R}^{N-1} : \beta_j \geq 0 \text{ for } j = 2, ..., N\}$. The treatment effect estimator and ATE estimator are defined accordingly.

Doudchenko and Imbens (2016) also proposed using Elastic-net regularized regression (Zou and Hastie, 2005) for estimation. The idea of Elastic-net is to combine the ℓ_1 and ℓ_2 penalty terms such that the underlying model can be more flexible. The Elastic-net estimator solves the optimization problem:

$$\hat{\boldsymbol{\beta}}_{\text{en}} = \text{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^N} \sum_{t=1}^{T_1} (y_{1t} - \boldsymbol{x}_t'\boldsymbol{\beta})^2 + \lambda_R \sum_{j=1}^{N} \beta_j^2 + \lambda_L \sum_{j=1}^{N} |\beta_j|$$

with $\lambda_R = \lambda(1-\alpha)$ and $\lambda_L = \lambda\alpha$ where $\alpha \in [0, 1]$. The tuning parameters λ and α can be

selected via cross-validation or information criterion.

3. MA-based ATE estimator

For HCW and alternative methods, regardless of what criterion or constraint is used, one finally ends up with a single model among all candidate models. In addition, different criterion will favour different model given a set of candidate models. For example, the Schwarz-Bayes information criterion (BIC) will favour more parsimonious model while AIC will favour more parameterized models (Hansen and Racine, 2012). The MA method, on the other hand, avoids the reliance on a single model by averaging over the whole set of candidate models.

We seek to combine HCW with two frequentist model averaging methods: the Mallows model averaging (MMA) estimator (Hansen, 2007) and Jackknife model averaging (JMA) estimator (Hansen and Racine, 2012). In the panel data setting, because we have a single treated unit and a fixed number of control units within large pre and post-treatment periods, which is often the case with comparative case studies, we consider MMA and JMA estimators in a linear regression framework with finite number of regressors. That is, we estimate the MA-based counterfactual y_{1t}^0 using following regression model, which is the same as the model in (4):

$$y_{1t}^0 = \bar{\alpha} + \boldsymbol{a}_{-1}' \boldsymbol{y}_{-1t} + e_{1t} \tag{11}$$

where $\bar{\alpha}$ is the intercept, $\boldsymbol{y}_{-1t} = (y_{2t}, \ldots, y_{Nt})'$ is the vector of control units' outcomes and e_{1t} is idiosyncratic error with zero mean, finite variance $\mathbb{E}(e_{1t}^2 | \{\bar{\alpha}, \boldsymbol{y}_{-1t}\}) = \sigma^2(\bar{\alpha}, \boldsymbol{y}_{-1t})$. Stacking (11) over all pre-treatment periods gives:

$$\boldsymbol{y}_1^0 = \boldsymbol{\tau}\bar{\alpha} + \boldsymbol{Y}\boldsymbol{a}_{-1} + \boldsymbol{e}_1 = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}_1 \tag{12}$$

where $\boldsymbol{y}_1^0 = (y_{11}^0, ..., y_{1T_1}^0)'$, $\boldsymbol{\tau} = (1, ..., 1)'$, $\boldsymbol{Y} = (\boldsymbol{y}_{-11}, ..., \boldsymbol{y}_{-1T_1})'$ is a $T_1 \times (N-1)$ matrix, $\boldsymbol{e}_1 = (e_{11}, ..., e_{1T_1})'$, $\boldsymbol{\beta} = (\bar{\alpha}, \boldsymbol{a}_{-1}')'$ and $\boldsymbol{X} = (\boldsymbol{\tau}, \boldsymbol{Y})$ is a $T_1 \times N$ matrix of full column rank. The regression model in (12) is equivalent to the model that is commonly considered in MA

literature as in Liang et al. (2011), Liu (2015), Liu and Zhang (2018).

Suppose that we have M candidate models. We follow Hansen (2007, 2014), Liu and Zhang (2018) and consider a sequence of nested candidate models[4]. As there are $N-1$ control units, we have $M = N$ candidate models. The mth candidate model includes a constant (intercept) and the first $m-1$ control units, but excluding the remaining control units. In the set-up with finite number of regressors (control units in our case), we can define some special models. Any candidate model omitting regressors with non-zero coefficient is called under-fitted model. A candidate model is called just-fitted if the model has no omitted variables nor irrelevant variables. A candidate model is called over-fitted if it has no omitted variables but has irrelevant variables[5]. Without loss of generality, let the first M_0 candidate models be under-fitted. Clearly, $M = N > M_0 \geq 0$.

Let $\boldsymbol{X}_m = (\boldsymbol{\tau}, \boldsymbol{Y}_m)$, where $\boldsymbol{Y}_m$ contains the first $m-1$ regressors in $\boldsymbol{Y}$. Note that $\boldsymbol{X}_m$ has $K_m = m$ regressors. Denote $\boldsymbol{\Pi}_m$ as a selection matrix such that $\boldsymbol{\Pi}_m = (\boldsymbol{I}_{K_m}, \boldsymbol{0}_{K_m \times (N-K_m)})$ and thus $\boldsymbol{X}_m = \boldsymbol{X}\boldsymbol{\Pi}_m'$, where $\boldsymbol{I}_{K_m}$ is a $K_m \times K_m$ identity matrix. Under the mth candidate model, the least squares estimator is $\hat{\boldsymbol{\beta}}_m = \boldsymbol{\Pi}_m'(\boldsymbol{X}_m'\boldsymbol{X}_m)^{-1}\boldsymbol{X}_m'\boldsymbol{y}_1^0$. The model averaging estimator of $\boldsymbol{\beta}$ is $\hat{\boldsymbol{\beta}}(\boldsymbol{\omega}) = \sum_{m=1}^{M}\omega_m\hat{\boldsymbol{\beta}}_m$, where ω_m is the weight on the mth candidate model and $\boldsymbol{\omega} = (\omega_1, ..., \omega_M)' \in W = \{\boldsymbol{\omega} \in [0,1]^M : \sum_{m=1}^{M}\omega_m = 1\}$. Then $\hat{y}_{1t,\text{MA}}^0 = \boldsymbol{x}_t'\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\omega}})$ and the ATE estimator is $\hat{\Delta}_{1,\text{MA}} = T_2^{-1}\sum_{t=T_1+1}^{T}(y_{1t} - \hat{y}_{1t,\text{MA}}^0)$.

Since model weight is an integrated part of MA estimator, different strategies for weight selection have been proposed. We use following two strategies in this chapter:

- **MMA**

  Denote $\boldsymbol{P}_m = \boldsymbol{X}_m(\boldsymbol{X}_m'\boldsymbol{X}_m)^{-1}\boldsymbol{X}_m'$, the mth model's projection matrix; let $\boldsymbol{P}(\boldsymbol{\omega}) = \sum_{m=1}^{M}\omega_m\boldsymbol{P}_m$, and $\boldsymbol{K} = (K_1, ..., K_M)'$. In the homoscedastic error setting, Hansen

[4]Similar to Hansen (2014), Liu and Zhang (2018), we do not impose any assumptions on the ordering of regressors (control units' outcomes in the present case)

[5]We do not assume that the true model must be one of candidate models. We need that at least one candidate model is not under-fitted. If there is no true model among all candidate models, the just-fitted model is the model that has no omitted variable and the smallest number of irrelevant variables, and the over-fitted model is the model that has no omitted variable but more irrelevant variables than the just-fitted model.

(2007) proposed choosing weights by minimizing Mallows criterion

$$\mathcal{C}(\boldsymbol{\omega}) = ||\{\boldsymbol{I}_T - \boldsymbol{P}(\boldsymbol{\omega})\boldsymbol{y}\}||^2 + 2\sigma^2\boldsymbol{\omega}'\boldsymbol{K} \tag{13}$$

where $||\cdot||^2$ stands for the Euclidean norm, $\sigma^2 = \mathbb{E}(e_i^2)$. In practice, σ^2 can be estimated by $\hat{\sigma}^2 = (T-N)^{-1}||\boldsymbol{y}_1^0 - \boldsymbol{X}\hat{\boldsymbol{\beta}}_M||^2$. Let $\hat{\boldsymbol{w}}_{\text{MMA}} = \text{argmin}_{\boldsymbol{\omega}\in W}\mathcal{C}(\boldsymbol{\omega})$, so that the MMA estimator of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\omega}}_{\text{MMA}}) = \sum_{m=1}^{M} \omega_{\text{MMA,m}}\hat{\boldsymbol{\beta}}_m \tag{14}$$

- **JMA**

  Denote h_{tt}^m the tth diagonal element of $\boldsymbol{P}_m$ and $\boldsymbol{D}_m$ a diagonal matrix with $(1-h_{tt}^m)^{-1}$ being its tth diagonal element. It can be shown that $\boldsymbol{P}_m^{\text{JMA}} = \boldsymbol{D}_m(\boldsymbol{P}_m - \boldsymbol{I}_T) + \boldsymbol{I}_T$ (see Appendix A.1). Then $\boldsymbol{P}^{\text{JMA}}(\boldsymbol{\omega}) = \sum_{m=1}^{M}\omega_m\boldsymbol{P}_m^{\text{JMA}}$. In the linear regression model with heteroscedastic errors, Hansen and Racine (2012) proposed selecting weights by minimizing cross-validation or Jackknife criterion

$$\mathcal{J}(\boldsymbol{\omega}) = ||\{\boldsymbol{I}_T - \boldsymbol{P}^{\text{JMA}}(\boldsymbol{\omega})\}\boldsymbol{y}||^2. \tag{15}$$

  Let $\hat{\boldsymbol{w}}_{\text{JMA}} = \text{argmin}_{\boldsymbol{\omega}\in W}\mathcal{J}(\boldsymbol{\omega})$, so that the JMA estimator of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\omega}}_{\text{JMA}}) = \sum_{m=1}^{M} \omega_{\text{JMA,m}}\hat{\boldsymbol{\beta}}_m \tag{16}$$

Hansen (2007) and Hansen and Racine (2012) showed that MMA and JMA estimators are asymptotically optimal in the sense that they achieve the lowest possible mean squared error among the class of linear estimators in homoscedasitc and heteroscedastic setting, respectively.[6]

[6] As discussed in Liu (2015) and Liu and Zhang (2018), it is possible that the MMA and JMA estimators are not asymptotically optimal in the framework with finite number of regressors.

4. Asymptotic properties of MA-based ATE estimator

We first state some regularity conditions required for asymptotic results.

Condition C.1. The data $\{\boldsymbol{x}_t\}_{t=1}^{T}$ follows a weakly dependent stationary process with $T_1^{-1}\sum_{t=1}^{T_1}\boldsymbol{x}_t \xrightarrow{p} \mathbb{E}(\boldsymbol{x}_t)$ and $\boldsymbol{Q}_{T_1} = T_1^{-1}\boldsymbol{X}'\boldsymbol{X} \xrightarrow{p} \boldsymbol{Q}$ where $\boldsymbol{x}_t = (1, y_{2t}, \ldots, y_{Nt})', \boldsymbol{Q} = \mathbb{E}(\boldsymbol{x}_t\boldsymbol{x}_t')$ is positive definite. Let $\eta = \lim_{T_1,T_2\to\infty}\sqrt{T_2/T_1}$ be a finite non-negative constant.

Condition C.2. $\{e_{1t}\}_{t=1}^{T}$ is zero mean, serially uncorrelated and satisfies $\boldsymbol{Z}_{T_1} = T_1^{-1/2}\boldsymbol{X}'\boldsymbol{e}_1 = T_1^{-1/2}\sum_{t=1}^{T_1}\boldsymbol{x}_t e_{1t} \xrightarrow{d} \boldsymbol{Z} \sim N(\boldsymbol{0}, \boldsymbol{\Omega})$, where $\boldsymbol{\Omega} = \mathbb{E}(\boldsymbol{x}_t\boldsymbol{x}_t' e_{1t}^2)$ is a positive definite matrix and $\boldsymbol{\Omega}_{T_1} = T_1^{-1}\sum_{t=1}^{T_1}\boldsymbol{x}_t\boldsymbol{x}_t' e_{1t}^2 \xrightarrow{p} \boldsymbol{\Omega}$.

Condition C.3. $\bar{h}_{T_1} = \max_{1\le m\le M}\max_{1\le t\le T_1} h_{tt}^m = O_p(T_1^{-1})$

Condition C.4. $\{\Delta_{1t}\}_{t=1}^{T}$ has mean Δ_1 and is serially uncorrelated. We assume that $\nu_{1t} = \Delta_{1t} - \Delta_1 + e_{1t}$ satisfies a central limit theorem: $T_2^{-1/2}\sum_{t=T_1+1}^{T}\nu_{1t} \xrightarrow{d} N(0, \Omega_\nu)$ where $\Omega_\nu = \mathbb{E}(\nu_{1t}^2)$.

Condition C.5. Let $\boldsymbol{\gamma}_t = (y_{1t}, y_{2t}, ..., y_{Nt}, \Delta_{1t}d_{1t})$ for $t = 1, \ldots, T$, where $d_{1t} = 0$ if $t < T_1$ and $d_{1t} = 1$ otherwise. Assume $\{\hat{\boldsymbol{\gamma}}_t\}_{t=1}^{T}$ is weakly dependent stationary process. Define $\rho(\tau) = \max_{1\le t\le T}\max_{1\le i,j\le N}|\text{cov}(\gamma_{it}, \gamma_{j,t+\tau})/\sqrt{\text{Var}(\gamma_{it})\text{Var}(\gamma_{j,t+\tau})}|$. There exist some finite constants $C > 0$ and $0 < \lambda < 1$ such that $\rho(\tau) \le C\lambda^\tau$.

Condition C.1 is high-level. One sufficient condition for C.1 is that $\{(\boldsymbol{f}_t, \boldsymbol{\epsilon}_t)\}_{t\ge1}$ in (3) is stationary, ergodic process such that the law of large number holds. Condition C.1 and C.2 imply that $\sqrt{T_1}(\hat{\boldsymbol{\beta}}_M - \boldsymbol{\beta}) \xrightarrow{d} N(\boldsymbol{0}, \boldsymbol{Q}^{-1'}\boldsymbol{\Omega}\boldsymbol{Q}^{-1})$ where $\hat{\boldsymbol{\beta}}_M$ is the least squares estimator from the largest model, i.e., $\hat{\boldsymbol{\beta}}_M = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}_1^0$. Condition C.3 requires that h_{tt}^m be asymptotic negligible for all models considered and conditioin of this form is typical for application of cross-validation (Hansen and Racine 2012). Condition C.4 implies that ν_{1t} is serially uncorrelated and requires that a central limit theorem hold for a partial sum of ν_{1t}. Condition C.5 is related to the whole sample $t = 1, \ldots, T$. It ensures that MA estimator $\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\omega}})$ from pre-treatment sample are asymptotic independent of quantity involving the post-treatment sample average of demeaned treatment effects and idiosyncratic error. Condition C.1 and C.2 are similar to condition 1 and 2 in Liu and Zhang (2018), condition C.3 is similar to condition C.2 in

Zhang (2015). Condition C.4 and C.5 are similar to Assumption 3 and 4 in Li (2019).

4.1 Consistency

The average treatment effect for the treated unit is given by $\Delta_1 = \mathbb{E}(\Delta_{1t})$. In this subsection we show the consistency of MA-based ATE estimator $\hat{\Delta}_1$. In order to show the consistency, we need the following lemma.

Lemma 1 Under condition C.1 and C.2, $\sqrt{T_1}\{\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\omega}}_{\text{MMA}}) - \boldsymbol{\beta}\} = O_p(1)$; if condition C.3 is also satisfied, $\sqrt{T_1}\{\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\omega}}_{\text{JMA}}) - \boldsymbol{\beta}\} = O_p(1)$

We then derive the consistency result in the next proposition.

Proposition 1 Under condition C.1 - C.3, as $T_1, T_2 \to \infty$, we have

$$\hat{\Delta}_1 \xrightarrow{p} \Delta_1 \tag{17}$$

4.2 Asymptotic distribution

To study the asymptotic distribution of MA-based ATE estimator, we need to first study the asymptotic distributions of two averaging estimators $\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\omega}}_{\text{MMA}})$ and $\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\omega}}_{\text{JMA}})$, which depend on the asymptotic bahavior of weights estimator. We follow Liu and Zhang (2018) and show that MMA and JMA estimators asymptotically assign zero weights to under-fitted models, which is given in the following theorem.

Theorem 1 Under conditions C.1 - C.2, for any $m \in \{1, ..., M_0\}$,

$$\hat{\omega}_{\text{MMA,m}} = O_p(T_1^{-1}) \tag{18}$$

If condition C.3 is also satisfied,

$$\hat{\omega}_{\text{JMA,m}} = o_p(T_1^{-1/2}) \tag{19}$$

Because under-fitted models receive zero weight asymptotically, we can exclude them and

define new weight vector as $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_S) \in \mathcal{L} = \{\boldsymbol{\lambda} \in [0,1]^S : \sum_{s=1}^{S} \lambda_s = 1\}$ where $S = M - M_0$. Let $\boldsymbol{\Omega}_s = \boldsymbol{\Pi}_{M_0+s} \boldsymbol{\Omega} \boldsymbol{\Pi}'_{M_0+s}$, $\boldsymbol{Q}_s = \boldsymbol{\Pi}_{M_0+s} \boldsymbol{Q} \boldsymbol{\Pi}'_{M_0+s}$ and $\boldsymbol{V}_s = \boldsymbol{\Pi}'_{M_0+s} \boldsymbol{Q}_s^{-1} \boldsymbol{\Pi}_{M_0+s}$ denote covariance matrices based on new weight vector, where $\boldsymbol{\Omega}$ and $\boldsymbol{Q}$ are defined in condition C.1 - C.2. The following theorem summarizes the asymptotic distributions for $\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\omega}}_{\text{MMA}})$ and $\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\omega}}_{\text{JMA}})$.

Theorem 2 Under conditions C.1 - C.2,

$$\sqrt{T_1}\{\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\omega}}_{\text{MMA}}) - \boldsymbol{\beta}\} \xrightarrow{d} \sum_{s=1}^{S} \tilde{\lambda}_{\text{MMA,s}} \boldsymbol{V}_s \boldsymbol{Z} \tag{20}$$

where $\tilde{\boldsymbol{\lambda}}_{\text{MMA}} = (\tilde{\lambda}_{\text{MMA,1}}, \ldots, \tilde{\lambda}_{\text{MMA,S}})' = \text{argmin}_{\boldsymbol{\lambda} \in \mathcal{L}} \boldsymbol{\lambda}' \mathcal{T} \boldsymbol{\lambda}$ and $\mathcal{T}$ is an $S \times S$ matrix with (s,j)th element

$$\mathcal{T}_{sj} = 2\sigma^2 K_{M_0+s} - \boldsymbol{Z}' \boldsymbol{V}_{\max\{s,j\}} \boldsymbol{Z} \tag{21}$$

If condition C.3 is also satisfied, we have

$$\sqrt{T_1}\{\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\omega}}_{\text{JMA}}) - \boldsymbol{\beta}\} \xrightarrow{d} \sum_{s=1}^{S} \tilde{\lambda}_{\text{JMA,s}} \boldsymbol{V}_s \boldsymbol{Z} \tag{22}$$

where $\tilde{\boldsymbol{\lambda}}_{\text{JMA}} = (\tilde{\lambda}_{\text{JMA,1}}, \ldots, \tilde{\lambda}_{\text{JMA,S}})' = \text{argmin}_{\boldsymbol{\lambda} \in \mathcal{L}} \boldsymbol{\lambda}' \boldsymbol{\Sigma} \boldsymbol{\lambda}$ and $\boldsymbol{\Sigma}$ is an $S \times S$ matrix with (s,j)th element

$$\Sigma_{sj} = \text{tr}(\boldsymbol{Q}_s^{-1} \boldsymbol{\Omega}_s) + \text{tr}(\boldsymbol{Q}_j^{-1} \boldsymbol{\Omega}_j) - \boldsymbol{Z}' \boldsymbol{V}_{\max\{s,j\}} \boldsymbol{Z} \tag{23}$$

Theorem 2 shows that both MMA and JMA estimators have non-standard limiting distributions, which are non-linear functions of normal random vector $\boldsymbol{Z}$. Finally, the asymptotic distribution of MA-based ATE estimator is given by the next theorem:

Theorem 3 Under conditions C.1 - C.5,

$$\sqrt{T_2}(\hat{\Delta}_1 - \Delta_1) \xrightarrow{d} -\eta\, \mathbb{E}(\boldsymbol{x}'_t)(\sum_{s=1}^{S} \tilde{\lambda}_{\text{MA,s}} \boldsymbol{V}_s \boldsymbol{Z}) + Z_2 \tag{24}$$

where $\hat{\Delta}_1$ is estimated by either MMA or JMA, $\tilde{\lambda}_{\text{MA,s}}$ are corresponding weights, $\eta = \lim_{T_1, T_2 \to \infty} \sqrt{T_2/T_1}$, $\boldsymbol{Z}$ is defined in condition C.2., Z_2 is independent of $\boldsymbol{Z}$ and dis-

tributed as $N(0, \Omega_v)$, Ω_v is defined in condition C.4. It is worth noting that the ATE estimator in Equation (24) is also asymptotically unbiased. This is because both MMA and JMA estimators of regression coefficients are asymptotically unbiased since they asymptotically assign zero weights to the under-fitted models, which in turn imply that the MA-based ATE estimator is asymptotically unbiased.

5. Inference of MA-based ATE estimator

It is shown in previous section that the MA-based ATE estimator has a non-standard asymptotic distribution, which is result of non-standard distribution of MA estimators $\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\omega}}_{\text{MMA}})$ and $\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\omega}}_{\text{JMA}})$. In addition, their asymptotic distributions are not pivotal, thus they cannot be directly used for inference. To address this issue, we follow Li (2019) and consider a subsampling method for constructing confidence intervals.

From proof of Theorem 3 in the Appendix A.6,

$$\begin{aligned} \hat{A} &= \sqrt{T_2}(\hat{\Delta}_1 - \Delta_1) \\ &= -\sqrt{\frac{T_2}{T_1}}(\frac{1}{T_2}\sum_{t=T_1+1}^{T} \boldsymbol{x}_t')\sqrt{T_1}(\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\omega}}) - \boldsymbol{\beta}) + \frac{1}{\sqrt{T_2}}\sum_{t=T_1+1}^{T} \nu_{1t} \\ &= \hat{A}_1 + \hat{A}_2 \end{aligned} \tag{25}$$

where $\nu_{1t} = \Delta_{1t} - \Delta_1 + e_{1t}$, $\hat{A}_1 = -\sqrt{\frac{T_2}{T_1}}(\frac{1}{T_2}\sum_{t=T_1+1}^{T} \boldsymbol{x}_t')\sqrt{T_1}(\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\omega}}) - \boldsymbol{\beta})$ and $\hat{A}_2 = \frac{1}{\sqrt{T_2}}\sum_{t=T_1+1}^{T} \nu_{1t}$.

The expression in (25) shows that $\hat{\Delta}_1$ can be decomposed into $\hat{A}_1$ that involves $\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\omega}})$ and $\hat{A}_2$ that is not related to the averaging estimator. The inference procedure is to implement the subsampling method only to the term $\hat{A}_1$ and the regular bootstrap method to $\hat{A}_2$. Specifically, from condition C.4, ν_{1t} is serially uncorrelated, so Ω_v can be consistently estimated by $\hat{\Omega}_v = T_2^{-1}\sum_{t=T_1+1}^{T} \hat{\nu}_{1t}^2$ where $\hat{\nu}_{1t} = \hat{\Delta}_{1t} - \hat{\Delta}_1$. So we can generate ν_{1t}^* from i.i.d. $N(0, \hat{\Omega}_v)$ for post-

treatment periods. Let b be the subsample size such that $b \to \infty$ and $b/T_1 \to 0$ as $T_1 \to \infty$. For $t = 1, \ldots, b$, we randomly draw $(y_{1t}^*, \boldsymbol{x}_t^*)$ from $(y_{1t}, \boldsymbol{x}_t)_{t=1}^{T_1}$ with replacement. Then we use subsample $(y_{1t}^*, \boldsymbol{x}_t^*)_{t=1}^{b}$ to estimate $\boldsymbol{\beta}$ and obtain $\hat{\boldsymbol{\beta}}_b^*$ by either MMA or JMA method. By replacing unknown parameter $\boldsymbol{\beta}$ with its consistent MA estimator $\hat{\boldsymbol{\beta}}(\boldsymbol{\omega})$ and plugging in $\hat{\boldsymbol{\beta}}_b^*$ into $\hat{A}_1$, the subsampling-bootstrap form of $\hat{A}$ is

$$\hat{A}^* = -\sqrt{\frac{T_2}{T_1}}\left(\frac{1}{T_2}\sum_{t=T_1+1}^{T} \boldsymbol{x}_t'\right)\sqrt{b}(\hat{\boldsymbol{\beta}}_b^* - \hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\omega}})) + \frac{1}{\sqrt{T_2}}\sum_{t=T_1+1}^{T} \nu_{1t}^* \tag{26}$$

We can repeat the procedure J times and obtain subsampling-bootstrap sample $\{\hat{A}_j^*\}_{j=1}^{J}$. After sorting the statistics $\hat{A}_j^*$, the $1-\alpha$ confidence interval for $\hat{\Delta}_1$ is given by

$$[\hat{\Delta}_1 - T_2^{-1/2}\hat{A}^*_{(1-\alpha/2)J}, \hat{\Delta}_1 - T_2^{-1/2}\hat{A}^*_{(\alpha/2)J}] \tag{27}$$

The following theorem shows that the above confidence intervals are consistent for confidence intervals of Δ_1.

Theorem 4 Under conditions C.1 - C.5 and condition that $b \to \infty$ and $b/T_1 \to 0$ as $T_1 \to \infty$, the $(1-\alpha)$ confidence intervals of Δ_1 can be consistently estimated by $[\hat{\Delta}_1 - T_2^{-1/2}\hat{A}^*_{(1-\alpha/2)J}$, $\hat{\Delta}_1 - T_2^{-1/2}\hat{A}^*_{(\alpha/2)J}]$ for any $\alpha \in (0,1)$.

In summary, the following algorithm describes the entire inference procedure:

Algorithm: Subsampling - bootstrap Inference for MA-based ATE estimator

1. Use pre-treatment data to find $\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\omega}})$ based on MMA or JMA;

2. Calculate counterfactual estimate $\hat{y}_{1t}^0 = \boldsymbol{x}_t'\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\omega}})$ and treatment effect estimate $\hat{\Delta}_{1t} = y_{1t} - \hat{y}_{1t}^0$ for post-treatment periods $T_1+1, \ldots, T$; then find ATE estimate $\hat{\Delta}_1 = \frac{1}{T_2}\sum_{t=T_1+1}^{T}\hat{\Delta}_{1t}$, $T_2 = T - T_1$;

3. Calculate $\hat{\nu}_{1t} = \hat{\Delta}_{1t} - \hat{\Delta}_1$ and estimate σ_ν^2 by $\hat{\sigma}_\nu^2 = T_2^{-1}\sum_{t=T_1+1}^{T}\hat{\nu}_{1t}^2$. Sample $\nu_{1t}^* \sim i.i.d.N(0, \hat{\sigma}_\nu^2)$ for $t = T_1+1, \ldots, T$;

4. Specify the sub-sample size b[7], sample $\{y_{1t}^*, \boldsymbol{x}_t^*\}_{t=1}^{b}$ from $\{y_{1t}, \boldsymbol{x}_t\}_{t=1}^{T_1}$ with replacement;

[7]The proposed inference procedure does not perform well in general when b is too small or too large relative to T_1 (i.e., the number of pre-treatment periods). We recommend trying different b with $T_1/3 \le b \le 2T_1/3$ in

5. Use sub-sample $\{y_{1t}^*, \boldsymbol{x}_t^*\}_{t=1}^b$ to obtain $\hat{\boldsymbol{\beta}}_b^*$ based on MMA or JMA;

6. Calculate

$$\hat{A}^* = -\sqrt{\frac{T_2}{T_1}}(\frac{1}{T_2}\sum_{t=T_1+1}^{T}\boldsymbol{x}_t')\sqrt{b}(\hat{\boldsymbol{\beta}}_b^* - \hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\omega}})) + \frac{1}{\sqrt{T_2}}\sum_{t=T_1+1}^{T}\nu_{1t}^*$$

7. Repeat step 5 - 6 J times and obtain a sample $\{\hat{A}_j^*\}_{j=1}^J$. Sort $\hat{A}_j^*$, the $1-\alpha$ confidence interval for $\hat{\Delta}_1$ is

$$[\hat{\Delta}_1 - T_2^{-1/2}\hat{A}_{(1-\alpha/2)J}^*, \hat{\Delta}_1 - T_2^{-1/2}\hat{A}_{(\alpha/2)J}^*]$$

for $\alpha \in (0,1)$.

6. Simulation study

6.1 Comparison of different estimation methods

In this section we compare the predictive performance of various methods discussed in Section 2 and 3. Specifically, we consider MMA, JMA, AICC, Elastic-net (E-net), leave-many-out cross-validation (CVd) , synthetic control method (SCM) and modified synthetic control (MSC)[8]. Furthermore, we add two additional methods for comparison: the least squares estimator from the largest model (Full), i.e., no data-driven model selection or model averaging, and a simple equal-weighted average (EQ-MA). To balance the computation time and accuracy, we consider the number of units $N = 12$ and the experiment is repeated $1,000$ times. We set $T_1 = 25, 40, 60$, $T = T_1 + 10$. We generate model (2) with the same 1-factor, 2-factor and 3-factor structure as in HCW:

1-factor:

$$f_{1t} = 0.95 f_{1t-1} + u_{1t}$$

practice.

[8]Because results based on AIC are quantitatively similar to AICC in most cases, they are not reported to save space.

2-factor:

$$f_{1t} = 0.3f_{1t-1} + u_{1t}$$

$$f_{2t} = 0.6f_{2t-1} + u_{2t}$$

3-factor:

$$f_{1t} = 0.8f_{1t-1} + u_{1t}$$

$$f_{2t} = -0.6f_{2t-1} + u_{2t} + 0.8u_{2t-1}$$

$$f_{3t} = u_{3t} + 0.9u_{3t-1} + 0.4u_{3t-2}$$

where u_{it} is distributed to $N(0,1), i = 1, 2, 3$. $\{\epsilon_{it}\}$ in model (2) are generated by $\sigma^2 N(0,1)$, where $\sigma^2 = 1,\ 0.5,\ 0.1$; b_i is generated by $N(1,1)$. The various methods are compared based on post-treatment mean squared prediction error (MSPE):

$$\text{MSPE} = \frac{1}{T_2} \sum_{t=T_1+1}^{T} (y_{1t}^0 - \hat{y}_{1t}^0)^2$$

In addition to MSPE, the average number of selected control units are reported for AICC, E-net, CVd, SCM and MSC.

The simulation results are reported in Table 1-3. Firstly, the two data-driven MA methods have smaller MSPE in all cases considered. Specifically, the MMA and JMA have very similar results in most cases. This is expected as the random error $\{\epsilon_{it}\}$ in (2) are generated as homoscedastic error and the limiting distribution of the MMA and JMA estimators are the same for homoscedastic case (Liu and Zhang, 2018). Secondly, SCM and MSC perform poorly in terms of achieving smaller MSPE. As discussed in Wan et al. (2018), if constraints on synthetic control weights are invalid, then SCM could lead to biased prediction. It is likely that the constraints imposed in SCM and MSC are not satisfied, which lead to poor out-of-sample predictive performance. Thirdly, CVd generates smaller MSPE than that from AICC when the sample size is small. However, as pre-treatment sample size T_1 increases, it performs worse

than AICC in most cases, contradicting to Du et al. (2015)'s results that are obtained with smaller repetition numbers (100). Finally, in terms of number of selected control units, CVd consistently selects the most sparse model among all methods considered.

6.2 The coverage probabilities of inference procedure

We consider the same 3-factor DGP in the subsection 6.1 with the error term $\{\epsilon_{it}\}$ in model (2) being generated by $N(0,1)$. We use the DGP in Li (2019) to generate treatment effects Δ_{1t}:

$$\Delta_{1t} = \alpha_0(\frac{e^{Z_t}}{1+e^{Z_t}}), \quad t = T_1+1,\ldots,T \tag{28}$$

where $Z_t = 0.5Z_{t-1} + \phi_t$ and $\phi_t \sim i.i.d.N(0, 0.5^2)$. Therefore, $y_{1t} = y_{1t}^0 + \Delta_{1t}$ for $t = T_1+1,\ldots,T$. If $\alpha_0 = 0$, there is no treatment effect; if $\alpha_0 > 0$, there is positive treatment effect. In this simulation exercise, we set number of control units $N = 12$ and 20, $T_1 = 50$, $T_2 = 20$, $T = T_1 + T_2 = 70$. To implement the proposed inference procedure for the MA-based ATE estimator, the subsample size b = 20, 35, 50 when $N = 12$; b = 30, 40, 50 when $N = 20$. Note that when b = 50, subsampling is equivalent to the regular bootstrap. We repeat the subsampling-bootstrap procedure $1,000$ times and $J = 400$ subsamples are generated in each iteration. We also consider the following estimators for comparison:

- HCW based on the largest model with bootstrap, i.e., $b = 50$ (labeled Full).
- HCW based on the AICC model selection criterion with bootstrap (labeled AICC).
- Model averaging estimator based on the equal weights with bootstrap (labeled EQ-MA).
- Full model based on asymptotic distribution with $T_1 = 50$, $T = 70$ (labeled Asy1).[9]
- Full model based on asymptotic distribution with $T_1 = 500$, $T = 550$ (labeled Asy2).

The results are reported in Table 4 and 5. The proposed subsampling-bootstrap procedure works very well and results in coverage probability that is close to nominal level in most

[9]The asymptotic normal distribution is based on the Li and Bell (2017).

cases. However, the regular bootstrap ($b = 50$) does not work, the coverage probability are much lower than the nominal values in all cases. On the other hand, the bootstrap methods works reasonably well for the HCW based on the largest model, which is expected as there are no model selection and model averaging involved. The bootstrap does not work for model averaging estimator with equal weights or the estimator based on AICC, both estimators suffer from undercoverage problem. The ATE estimator based on the asymptotic normal distribution as in Li and Bell (2017) does not work when the sample is small. However, as we increase both the pre- and post-treatment sample size, the estimated coverage probability tends to attain the nominal level.

7. Empirical application

To demonstrate the MA-based ATE estimator, we revisit two empirical examples in HCW who investigated the impact of political integration of Hong Kong with China on July 1, 1997 and economic integration through implementation of the Closer Economic Partnership Agreement (CEPA) in 2004 Q1. We use the same dataset as in HCW, which includes quarterly real GDP growth rate of 24 control units from 1993 Q1 to 2008 Q1.

We first evaluate the impact of political integration on Hong Kong's real GDP growth. Hong Kong was a fishing village ceded to UK after the Opium War in 1842 and its sovereignty was reverted back to China on July 1, 1997. We consider the same subset containing 10 control units as in HCW due to the short pre-treatment periods. Figure 1 displays the results. The counterfactuals generated by HCW and MA method trace well the actual data in the pre-treatment period. In the post-treatment period, the predicted path from HCW (red dashed line) and MA (JMA with blue dotted line; MMA with purple dotted-dashed line) have similar shape but the paths from MA are lower than the one from HCW in most post periods. For the statistical inference, HCW fits an AR(2) model for the estimated treatment effects and the implied long-fun effect is -0.032 and statistically *insignificant*. Based on the MA method, the ATE of political integration is estimated at -0.029 by JMA and -0.024 by MMA. In order to

apply the proposed inference procedure for MA-based ATE estimator, we need to first check whether e_{1t} defined in condition C.2 and ν_{1t} defined in condition C.4 are serially uncorrelated or not.

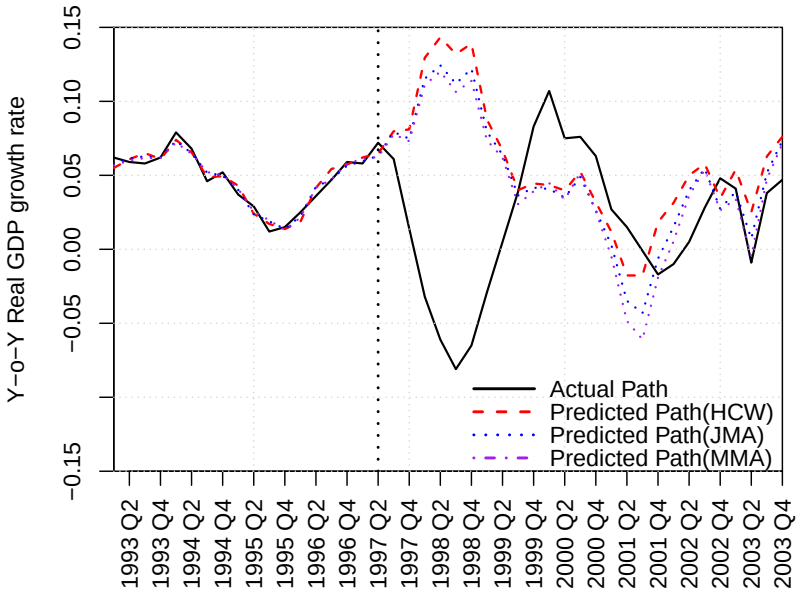

Figure 1: Actual and counterfactual real GDP growth rate of Hong Kong

Figure 2 and Figure 3 show the autocorrelation of two residuals of JMA and MMA, respectively. Because the right panels show that the residuals ν_{1t} appear to be serially correlated for both JMA and MMA, we allow for ν_{1t} to follow an AR(2) process: $\nu_{1t} = \phi_1\nu_{1,t-1} + \phi_2\nu_{1,t-2} + \xi_t$ where ξ_t is serially uncorrelated. Note that $\hat{\Delta}_1$ in (24) can be decomposed into $\hat{A}_1$ and $\hat{A}_2$. ν_{1t} only enters $\hat{A}_2$, which implies that only $\hat{A}_2$ needs to be adapted. Thus we generate the adapted value $\hat{A}_2^*$ as follows:

1) Regress $\hat{\nu}_{1t}$ on $\hat{\nu}_{1,t-1}$ and $\hat{\nu}_{1,t-2}$ to obtain $\hat{\phi}_1$ and $\hat{\phi}_2$ for $T_1+2,\ldots,T$;

2) Estimate ξ_t by $\hat{\xi}_t = \hat{\nu}_{1t} - \hat{\phi}_1\hat{\nu}_{1,t-1} - \hat{\phi}_2\hat{\nu}_{1,t-2}$ and compute $\hat{\sigma}_\xi^2 = (T-T_1-3)^{-1}\sum_{t=T_1+3}^{T}\hat{\xi}_t^2$;

3) Generate $\xi_t^* \sim i.i.d.N(0,\hat{\sigma}_\xi^2)$ and compute $\nu_{1t}^* = \hat{\phi}_1\nu_{1,t-1}^* + \hat{\phi}_2\nu_{1,t-2}^* + \xi_t^*$ for $t = T_1+2,\ldots,T$ where ν_{1,T_1}^* and ν_{1,T_1+1}^* are drawn from $i.i.d.N\big(0,(\frac{1-\hat{\phi}_2}{1+\hat{\phi}_2})\cdot\frac{\hat{\sigma}_\xi^2}{(1-\hat{\phi}_2)^2-\hat{\phi}_1^2}\big)$;

4) Compute $\hat{A}_2^* = T_2^{-1/2}\sum_{t=T_1+1}^{T}\nu_{1,t}^*$.

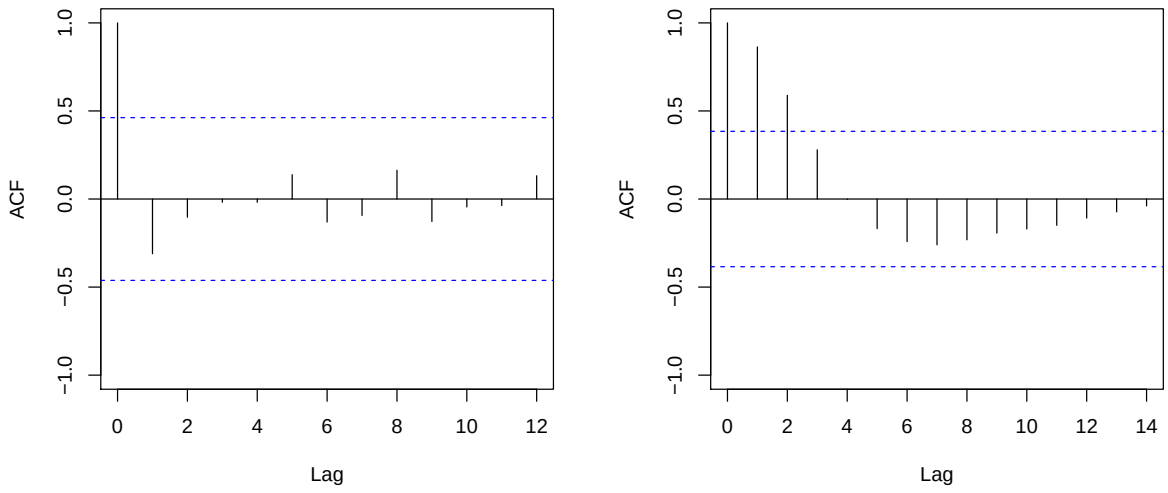

Figure 2: JMA: autocorrelation of $\hat{e}_{1t}$: 1993:1-1997:1 and $\hat{\nu}_{1t}$: 1994 :2-2003:4

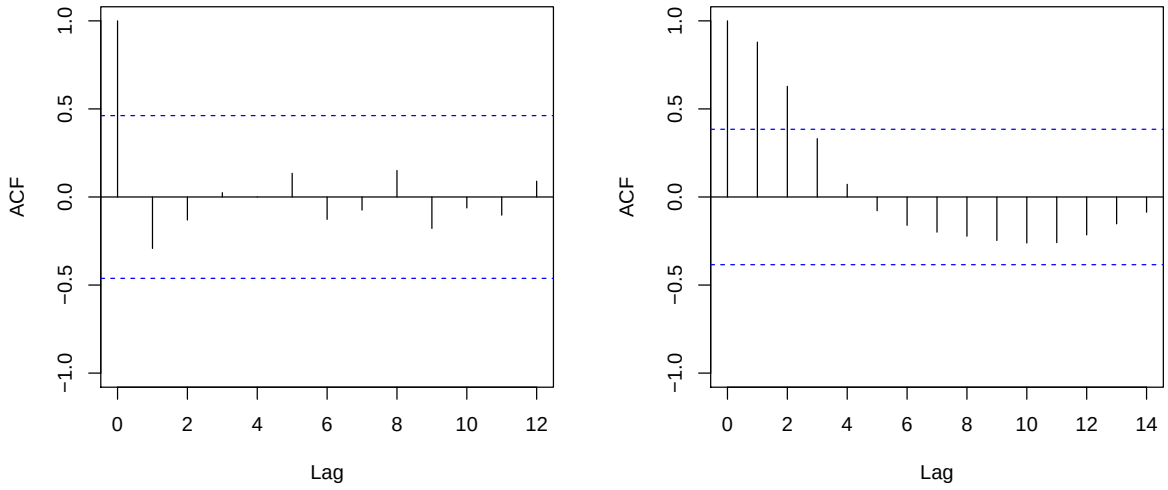

Figure 3: MMA: autocorrelation of $\hat{e}_{1t}$:1993:1-1997:1 and $\hat{\nu}_{1t}$:1994:2-2003:4

Because the simulation results in the previous section suggest that regular bootstrap doesn't work for MA-based ATE estimator, and there are only 18 pre-treatment periods, we select the subsample size $b = 11, 16$. For each b, we implement 10,000 subsampling simulations. We then sort the 10,000 statistics to obtain $\alpha/2$ and $1 - \alpha/2$ percentile for $\alpha = 0.2, 0.1, 0.05$. The confidence intervals of $\hat{\Delta}_1$ are reported in Table 6. It is noted that all these intervals con-

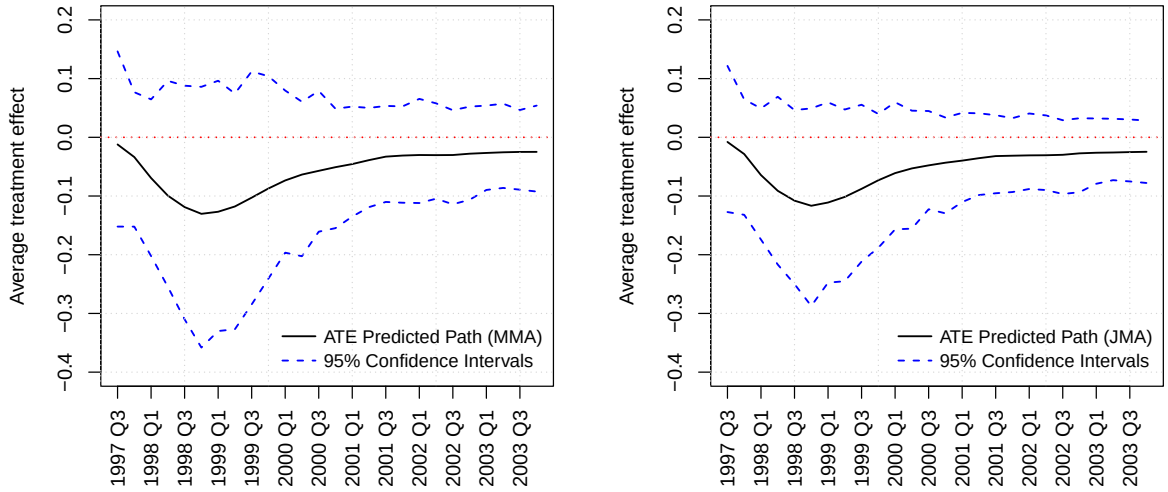

Figure 4: Political integration: ATE estimate based on MMA and ATE estimate based on JMA

tain zero. Alternatively, we can construct ATE as rolling average of treatment effect estimate starting at the first post-treatment period, that is,

$$\hat{\Delta}_t = \frac{1}{t - T_1} \sum_{i=T_1+1}^{t} \hat{\Delta}_{1i} \quad t = T_1 + 1, \ldots, T \tag{29}$$

We then use proposed inference procedure to construct the confidence intervals. Figure 4 shows the average treatment effect along with its 95% confidence intervals based on MMA and JMA. The confidence intervals based on JMA are narrower than those based on MMA. The results reported in Table 6 as well as Figure 4 suggest no significant impact of political integration on Hong Kong's real GDP growth, which support results in HCW.

We continue to investigate the impact of economic integration on Hong Kong's economy. On January 1, 2004, the CEPA, which is essentially a free trade agreement, took effect. The CEPA aimed to strengthen the linkage between mainland China and Hong Kong by liberalizing trade in services, enhancing cooperation in the area of finance, promoting trade and investments. Using 24 countries/regions not subject to the CEPA, HCW constructed the counterfactual Hong Kong to evaluate the impact of the CEPA on Hong Kong's real GDP growth.

Their results suggest that Hong Kong's real GDP growth rate was 4% higher than what it would have been in the absence of the CEPA.

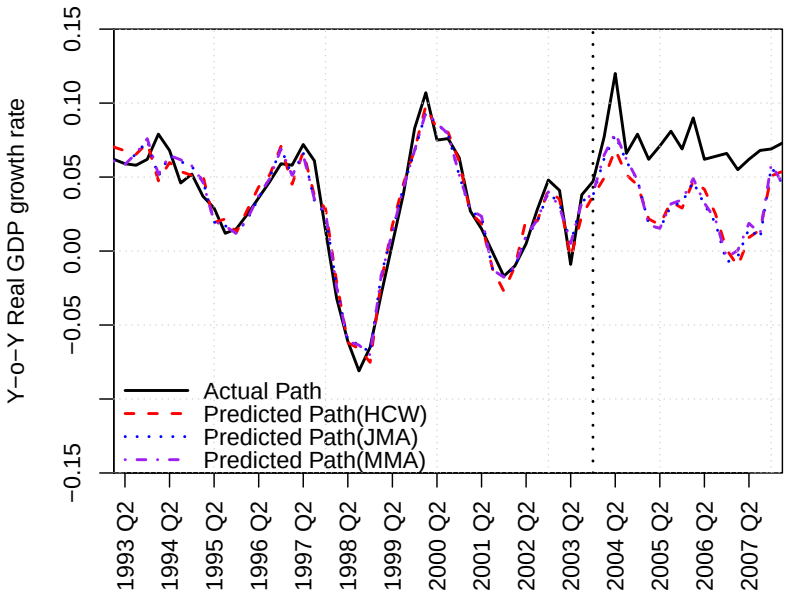

Figure 5: Actual and counterfactual real GDP growth rate of Hong Kong

In this application, the pre-treatment periods are 1993 Q1-2003 Q4 such that $T_1 = 44$, $T_2 = 17$. The ATE of the CEPA is estimated at 3.9% by JMA and 3.8% by MMA respectively, similar to the result reported in HCW (see also Figure 5). Figure 6 and Figure 7 display the autocorrelation of two residuals from JMA and MMA. All appear to be serially uncorrelated. As there are 44 pre-treatment periods, we select the subsample size $b = 30, 40$. Similarly, we implement 10,000 subsampling simulations for each b. The confidence intervals of $\hat{\Delta}_1$ are reported in Table 7. It is noted that the lower bounds of these intervals are all positive in all cases. This implies that the estimated ATE based on JMA and MMA are positive and significantly different from zero for all conventional significance levels. Figure 8 displays ATE estimates over post-treatment periods, along with its 95% confidence intervals, which also suggest significant positive impact from the CEPA on Hong Kong's real GDP growth. Table 8 and Table 9 report the model averaging estimates of regression coefficient or

"weights" in the applications of political integration and economic integration. For brevity, we only report the MMA estimates in this section. The estimates are sorted in decreasing order based on absolute value. We note that the countries with large coefficient (in absolute value) are those selected in the original HCW based on AIC or AICC.

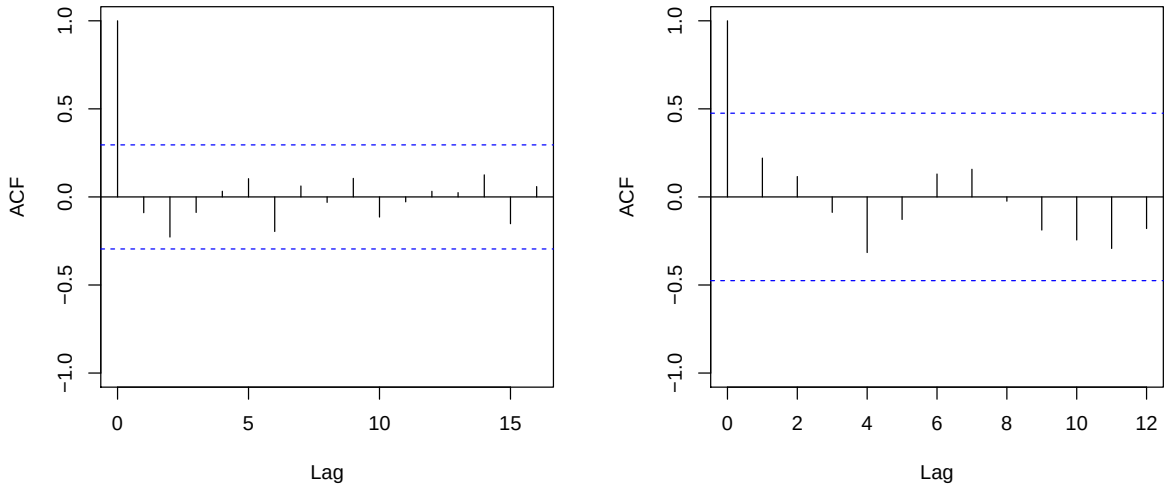

Figure 6: JMA: autocorrelation of $\hat{e}_{1t}$:1993:1-2003:4 and $\hat{\nu}_{1t}$:2004:1-2008:1

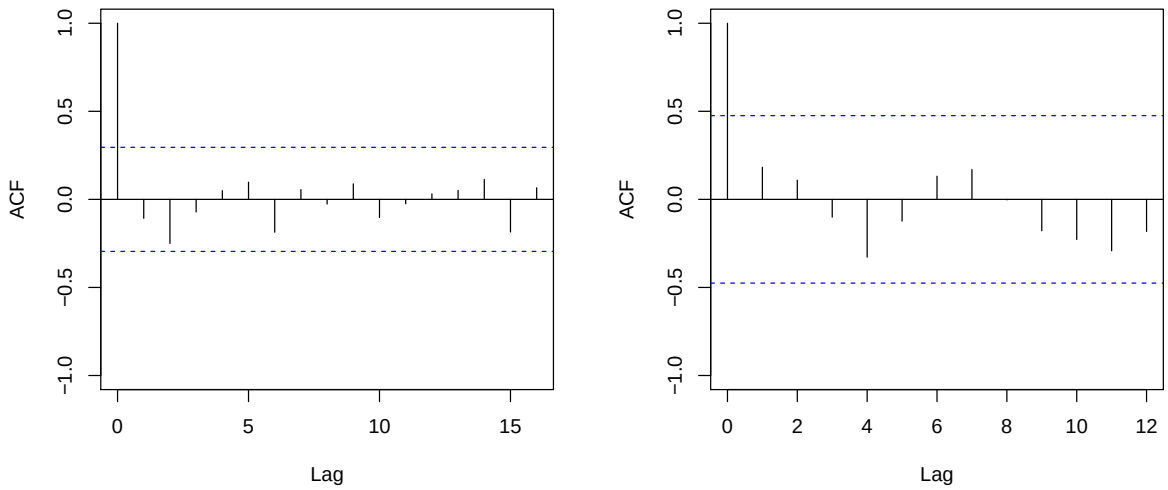

Figure 7: MMA: autocorrelation of $\hat{e}_{1t}$:1993:1-2003:4 and $\hat{\nu}_{1t}$:2004:1-2008:1

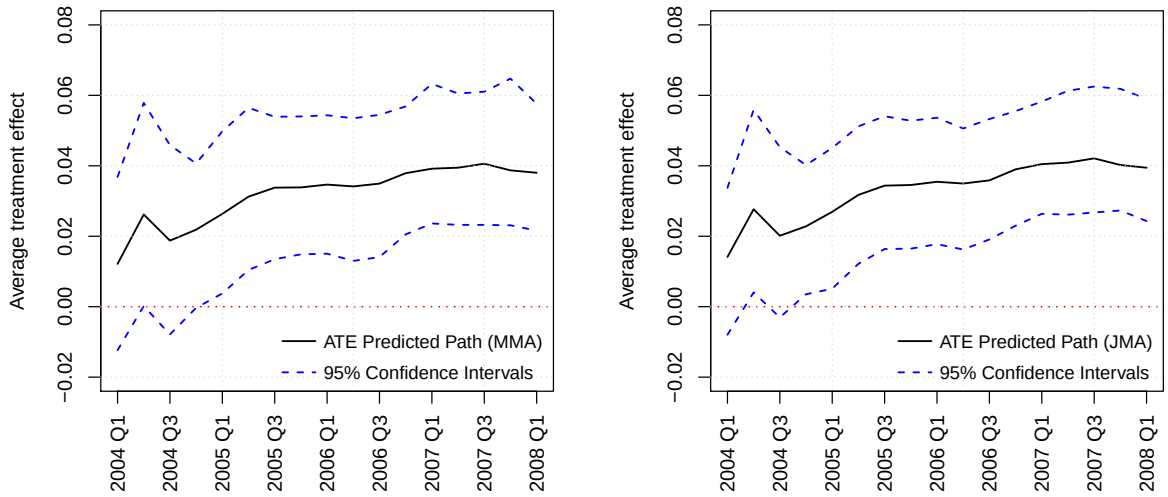

Figure 8: Economic integration: ATE estimate based on MMA and ATE estimate based on JMA

8. Conclusion

Researchers and analysts usually face the challenge of accurately estimating the average treatment effect when evaluating the impact of an economic event or a political intervention. In this chapter, we seek to use model averaging method to construct the counterfactual outcome and estimate ATE in a panel data setting. The model-averaging-based ATE estimator is shown to be asymptotically unbiased under mild assumptions. We also derive its asymptotic distribution, which turns out to be non-normal and non-standard. To address the problem of inference, we use a subsampling-based inference procedure. We assess the finite sample properties of model-averaging-based method with other commonly used methods including the Hsiao et al. (2012) and synthetic control methods, the proposed MA based-method works better in terms of small mean squared prediction error. However, we do not provide any theoretical justification of this finite-sample improvement, and it would be greatly desirable to demonstrate the formal justification in future study. Our simulation results suggest that the inference procedure based on subsampling method yields coverage probabilities close to nominal levels. Finally, we illustrate our method with the application used in Hsiao et al. (2012), who investigated the impact of mainland China - Hong Kong political and economic integration on Hong

Kong 's economy. Our results support the original findings.

Table 1: Comparison of different estimation methods: one-factor

	MMA	JMA	EQ-MA	Full	AICC	Enet	CVd	MSC	SCM
$\sigma^2 = 1$									
$T_1 = 25, T = 35$									
Avg.No					2.275	5.245	2.277	5.046	4.642
Avg.MSPE	1.5796	1.5659	1.7611	2.1327	1.8561	2.1673	1.8034	1.9255	2.5707
$T_1 = 40, T = 50$									
Avg.No					2.626	3.199	2.324	5.315	5.085
Avg.MSPE	1.3761	1.3691	1.7462	1.6406	1.5362	1.7575	1.5194	1.7481	2.6525
$T_1 = 60, T = 70$									
Avg.No					2.866	6.241	2.421	5.508	5.464
Avg.MSPE	1.2762	1.2756	1.6490	1.4056	1.3535	1.4739	1.3642	1.4970	2.5457
$\sigma^2 = 0.5$									
$T_1 = 25, T = 35$									
Avg.No					2.299	5.973	2.308	5.052	4.557
Avg.MSPE	.7896	.7829	1.0044	1.0686	.9281	1.0986	.9123	1.0761	1.8120
$T_1 = 40, T = 50$									
Avg.No					2.634	6.545	2.295	5.327	4.976
Avg.MSPE	.6862	.6833	.9989	.8188	.7654	.8897	.7535	.9874	1.9176
$T_1 = 60, T = 70$									
Avg.No					2.876	6.885	2.427	5.52	5.359
Avg.MSPE	.6380	.6377	.9427	.7019	.6504	.7414	.6795	.8326	1.8410
$\sigma^2 = 0.1$									
$T_1 = 25, T = 35$									
Avg.No					2.317	7.4	2.339	5.01	4.437
Avg.MSPE	.1577	.1565	.3709	.2145	.1853	.2178	.1817	.3515	1.1963
$T_1 = 40, T = 50$									
Avg.No					2.642	7.839	2.323	5.309	4.851
Avg.MSPE	.1370	.1363	.3726	.1633	.1533	.1755	.1508	.3493	1.3273
$T_1 = 60, T = 70$									
Avg.No					2.898	8.091	2.418	5.504	5.226
Avg.MSPE	.1275	.1275	.3502	.1401	.1353	.1464	.1358	.2753	1.2686

Notes: MMA is based on Mallow's model averaging; JMA is based on Jackknife model averaging; EQ-MA is based on model averaging with equal weights; Full stands for the full model, i.e., the Mth model; AICC is based on AICC criterion; Enet stands for Elastic-net penalized regression; CVd is based on leave-many-out cross validation; MSC stands for modified synthetic control; SCM stands for synthetic control method.

Table 2: Comparison of different estimation methods: two-factor

	MMA	JMA	EQ-MA	Full	AICC	E-net	CVd	MSC	SCM
$\sigma^2 = 1$									
$T_1 = 25, T = 35$									
Avg.No					2.726	5.416	2.529	4.860	4.558
Avg.MSPE	1.8191	1.8047	1.8865	2.3161	2.0504	1.8872	2.0278	1.9406	2.0890
$T_1 = 40, T = 50$									
Avg.No					3.176	5.687	2.608	5.151	4.813
Avg.MSPE	1.5908	1.5943	1.7473	1.7426	1.7167	1.6789	1.7213	1.7620	1.9494
$T_1 = 60, T = 70$									
Avg.No					3.577	5.916	2.828	5.390	5.090
Avg.MSPE	1.4226	1.4241	1.7043	1.5320	1.5071	1.5074	1.5302	1.6457	1.9462
$\sigma^2 = 0.5$									
$T_1 = 25, T = 35$									
Avg.No					2.818	5.892	2.576	4.749	4.260
Avg.MSPE	.9255	.9196	1.0555	1.1677	1.0322	.9593	1.0250	1.0929	1.3038
$T_1 = 40, T = 50$									
Avg.No					3.298	6.077	2.662	5.019	4.468
Avg.MSPE	.8077	.8094	.9851	.8803	.8737	.8467	.8701	.9887	1.2016
$T_1 = 60, T = 70$									
Avg.No					3.699	6.189	2.902	5.268	4.711
Avg.MSPE	.7221	.7226	.9595	.7723	.7639	.7599	.7828	.9410	1.2523
$\sigma^2 = 0.1$									
$T_1 = 25, T = 35$									
Avg.No					2.951	6.302	2.632	4.593	3.782
Avg.MSPE	.1879	.1874	.3302	.2344	.2090	0.1892	.2060	.3671	.6584
$T_1 = 40, T = 50$									
Avg.No					3.383	6.467	2.776	4.855	3.922
Avg.MSPE	.1636	.1640	.3115	.1777	.1759	.1708	.1754	.3318	.5870
$T_1 = 60, T = 70$									
Avg.No					3.768	6.483	3.022	5.12	4.134
Avg.MSPE	.1466	0.1466	.3073	.1556	0.1551	.1524	0.1576	.3303	.6736

Notes: MMA is based on Mallow's model averaging; JMA is based on Jackknife model averaging; EQ-MA is based on model averaging with equal weights; Full stands for the full model, i.e., the Mth model; AICC is based on AICC criterion; Enet stands for Elastic-net penalized regression; CVd is based on leave-many-out cross validation; MSC stands for modified synthetic control; SCM stands for synthetic control method.

Table 3: Comparison of different estimation methods: three-factor

	MMA	JMA	EQ-MA	Full	AICC	E-net	CVd	MSC	SCM
$\sigma^2 = 1$									
$T_1 = 25, T = 35$									
Avg.No					3.139	5.931	2.813	4.515	4.099
Avg.MSPE	2.2338	2.2414	2.6500	2.8193	2.5659	2.4752	.2532	2.8777	3.2075
$T_1 = 40, T = 50$									
Avg.No					3.756	4.418	3.034	4.633	4.154
Avg.MSPE	1.9009	2.0298	2.4670	2.0298	2.0551	2.1426	2.1258	2.7711	3.2642
$T_1 = 60, T = 70$									
Avg.No					4.289	6.484	3.385	4.843	4.292
Avg.MSPE	1.6879	1.6891	2.3853	1.7908	1.7862	1.8006	1.8349	2.2603	3.0024
$\sigma^2 = 0.5$									
$T_1 = 25, T = 35$									
Avg.No					3.283	6.233	2.893	4.302	3.779
Avg.MSPE	1.1459	1.1536	1.6236	1.4311	1.3045	1.2741	1.289511	1.8777	2.3393
$T_1 = 40, T = 50$									
Avg.No					3.820	6.457	3.107	4.561	3.959
Avg.MSPE	.9094	.9109	1.5355	1.0298	.9938	1.0069	1.0276	1.6656	2.2316
$T_1 = 60, T = 70$									
Avg.No					4.425	6.564	3.509	4.590	3.937
Avg.MSPE	.8589	.8594	1.4785	.9077	.9106	.9130	.9376	1.4542	2.2617
$\sigma^2 = 0.1$									
$T_1 = 25, T = 35$									
Avg.No					3.485	6.438	3.177	4.017	3.342
Avg.MSPE	.2284	.2294	.6873	.2906	.2597	.2445	.2576	.8450	1.7106
$T_1 = 40, T = 50$									
Avg.No					4.007	6.585	3.309	4.241	3.470
Avg.MSPE	.1847	.1851	.6755	.2089	.2046	.2008	.2112	.9011	1.6079
$T_1 = 60, T = 70$									
Avg.No					4.563	6.656	3.724	4.246	3.481
Avg.MSPE	.1746	.1746	.6428	.1835	.1874	.1835	.1918	.7510	1.6566

Notes: Notes: MMA is based on Mallow's model averaging; JMA is based on Jackknife model averaging; EQ-MA is based on model averaging with equal weights; Full stands for the full model, i.e., the Mth model; AICC is based on AICC criterion; Enet stands for Elastic-net penalized regression; CVd is based on leave-many-out cross validation; MSC stands for modified synthetic control; SCM stands for synthetic control method.

Table 4: Coverage probabilities ($\alpha_0 = 0$, no treatment effects)

					N = 12						
		JMA			MMA		Full	AICC	EQ-MA	Asy1	Asy2
b	20	35	50	20	35	50	50	50	50		
80%	.793	.766	.718	.801	.767	.725	.724	.661	.630	.640	.750
90%	.893	.873	.816	.903	.868	.825	.848	.767	.740	.757	.871
95%	.945	.931	.893	.951	.930	.891	.910	.844	.814	.831	.931
					N = 20						
		JMA			MMA		Full	AICC	EQ-MA	Asy1	Asy2
b	30	40	50	30	40	50	50	50	50		
80%	.781	.755	.727	.791	.755	.724	.751	.657	.673	.634	.791
90%	.887	.863	.836	.887	.863	.841	.862	.766	.790	.738	.889
95%	.938	.918	.903	.946	.924	.904	.912	.829	.860	.822	.939
					N = 30						
		JMA			MMA		Full	AICC	EQ-MA	Asy1	Asy2
b		40	50		40	50	50	50	50		
80%		.766	.711		.768	.702	.732	.631	.698	.649	.773
90%		.875	.814		.895	.812	.839	.741	.802	.761	.884
95%		.933	.889		.942	.888	.893	.817	.876	.848	.944

Notes: The pre-treatment sample size $T_1 = 50$, b is sub-sample size.
JMA estimator with subsampling-based confidence intervals (JMA);
MMA estimator with subsampling-based confidence intervals (MMA);
Least squares estimator for the largest model with bootstrap (Full);
AICC model selection estimator with bootstrap (AICC);
Model averaging estimator with equal weights with bootstrap (EQ-MA) ;
Least squares on the largest model with asymptotic normal approximation $(T_1 = 50, T = 70)$ (Asy1);
Least squares on the largest model with asymptotic normal approximation $(T_1 = 500, T = 550)$ (Asy2).

Table 5: Coverage probabilities ($\alpha_0 = 1$, positive treatment effects)

					N = 12						
		JMA			MMA		Full	AICC	EQ-MA	Asy1	Asy2
b	20	35	50	20	35	50	50	50	50		
80%	.807	.772	.720	.806	.772	.725	.726	.668	.634	.648	.763
90%	.899	.874	.819	.903	.878	.829	.852	.770	.746	.762	.883
95%	.947	.933	.903	.950	.931	.899	.903	.850	.814	.840	.942
					N = 20						
		JMA			MMA		Full	AICC	EQ-MA	Asy1	Asy2
b	30	40	50	30	40	50	50	50	50		
80%	.785	.766	.737	.793	.763	.739	.752	.656	.682	.635	.804
90%	.891	.868	.846	.896	.870	.847	.863	.771	.800	.741	.896
95%	.943	.920	.912	.949	.926	.910	.911	.832	.867	.831	.940
					N = 30						
		JMA			MMA		Full	AICC	EQ-MA	Asy1	Asy2
b		40	50		40	50	50	50	50		
80%		.776	.720		.778	.718	.738	.658	.697	.649	.792
90%		.879	.820		.894	.821	.838	.771	.808	.770	.891
95%		.937	.893		.945	.894	.902	.843	.876	.848	.951

Notes: The pre-treatment sample size $T_1 = 50$, b is sub-sample size.
JMA estimator with subsampling-based confidence intervals (JMA);
MMA estimator with subsampling-based confidence intervals (MMA);
Least squares estimator for the largest model with bootstrap (Full);
AICC model selection estimator with bootstrap (AICC);
Model averaging estimator with equal weights with bootstrap (EQ-MA) ;
Least squares on the largest model with asymptotic normal approximation $(T_1 = 50, T = 70)$ (Asy1);
Least squares on the largest model with asymptotic normal approximation $(T_1 = 500, T = 550)$ (Asy2).

Table 6: Confidence intervals of MA-based ATE (political integration)

	JMA	JMA	MMA	MMA
b	11	16	11	16
80%	[-.058, .010]	[-.056, .008]	[-.066, .017]	[-.064, .016]
90%	[-.069, .019]	[-.066, .018]	[-.078, .029]	[-.076, .027]
95%	[-.077, .028]	[-.074, .026]	[-.088, .040]	[-.085, .037]

Note: The pre-treatment sample size $T_1 = 18$, b is sub-sample size.

Table 7: Confidence intervals of MA-based ATE (economic integration)

	JMA	JMA	MMA	MMA
b	30	40	30	40
80%	[.030, .053]	[.033, .048]	[.027, .051]	[.031, .045]
90%	[.027, .058]	[.032, .050]	[.024, .055]	[.029, .047]
95%	[.023, .062]	[.030, .052]	[.021, .059]	[.027, .049]

Note: The pre-treatment sample size $T_1 = 44$, b is sub-sample size.

Table 8: Political integration: MMA coefficients estimates

Country	coefficient estimate
(Intercept)	.018
Taiwan	1.053
Japan	- .536
Korea	-.464
US	.105
Singapore	.067
Malaysia	-.062
China	.019
Philippines	.013
Thailand	-.008
Indonesia	-.002

Table 9: Economic integration: MMA coefficients estimates

Country	coefficient estimate
(Intercept)	-.001
Austria	- 1.304
Germany	.329
Italy	-.328
Mexico	.323
Singapore	.293
Norway	.257
Korea	.253
Philippines	.169
China	-.166
Switzerland	.154
Denmark	-.135
Thailand	.089
Japan	-.054
France	-.028
Finland	-.018
Indonesia	-.008
Netherlands	-.008
Taiwan	.006
US	.002
Australia	.000
Canada	.000
New Zealand	.000
Malaysia	.000
UK	.000

Appendix

A.1 Leave-one-out formula

For notational simplicity, we consider a generic candidate model so that subscript m can be omitted. That is, $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ where $\boldsymbol{y}, \boldsymbol{e}$ are $T \times 1$ vector, $\boldsymbol{X}$ is an $T \times N$ matrix and $\boldsymbol{\beta}$ is an $N \times 1$ vector. Denote $\boldsymbol{X}_{(-t)}$, $\boldsymbol{y}_{(-t)}$ with tth row being deleted. Similarly, $\hat{\boldsymbol{\beta}}_{(-t)}$ and $\hat{e}_{(-t)}$ are the estimate and residual from the sample when leaving out tth observation $\boldsymbol{x}_t$. Let $\hat{y}_t = \boldsymbol{x}_t'\hat{\boldsymbol{\beta}}$ and $\hat{e}_t = y_t - \hat{y}_t$, $h_{tt} = \boldsymbol{x}_t'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_t$.

We first show the following :

$$\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(-t)} = ((1 - h_{tt})^{-1}\hat{e}_t)(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_t \tag{A.1}$$

The proof uses 'Sherman-Marrison' formula that states: Let $\boldsymbol{A}$ be a non-singular matrix, $\boldsymbol{b}$ a vector and λ a scalar. If $\lambda \neq -(\boldsymbol{b}'\boldsymbol{A}^{-1}\boldsymbol{b})^{-1}$, then

$$(\boldsymbol{A} + \lambda\boldsymbol{b}\boldsymbol{b}')^{-1} = \boldsymbol{A}^{-1} - \big(\lambda(1 + \lambda\boldsymbol{b}'\boldsymbol{A}^{-1}\boldsymbol{b})^{-1}\big)\boldsymbol{A}^{-1}\boldsymbol{b}\boldsymbol{b}'\boldsymbol{A}^{-1} \tag{A.2}$$

Let $\boldsymbol{A} = \boldsymbol{X}'\boldsymbol{X}$, $\lambda = -1$, $\boldsymbol{b} = \boldsymbol{x}_t$, the above formula implies

$$(\boldsymbol{X}_{(-t)}'\boldsymbol{X}_{(-t)})^{-1}\boldsymbol{x}_t = ((1 - h_{tt})^{-1})(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_t \tag{A.3}$$

From the normal equation $\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}'\boldsymbol{y}$,

$$\begin{aligned}
(\boldsymbol{x}_t\boldsymbol{x}_t' + \boldsymbol{X}_{(-t)}'\boldsymbol{X}_{(-t)})\hat{\boldsymbol{\beta}} &= \boldsymbol{X}_{(-t)}'\boldsymbol{y}_{(-t)} + \boldsymbol{x}_t y_t \\
\{(\boldsymbol{X}_{(-t)}'\boldsymbol{X}_{(-t)})^{-1}\boldsymbol{x}_t\boldsymbol{x}_t' + \boldsymbol{I}_N\}\hat{\boldsymbol{\beta}} &= (\boldsymbol{X}_{(-t)}'\boldsymbol{X}_{(-t)})^{-1}\boldsymbol{X}_{(-t)}'\boldsymbol{y}_{(-t)} \\
&\quad + (\boldsymbol{X}_{(-t)}'\boldsymbol{X}_{(-t)})^{-1}\boldsymbol{x}_t(\boldsymbol{x}_t'\hat{\boldsymbol{\beta}} + \hat{e}_t) \\
\hat{\boldsymbol{\beta}} &= \hat{\boldsymbol{\beta}}_{(-t)} + (\boldsymbol{X}_{-t}'\boldsymbol{X}_{(-t)})^{-1}\boldsymbol{x}_t\hat{e}_t
\end{aligned}$$

A.1 follows by substituting A.3 into the above equation. Given A.1,

$$\boldsymbol{x}_t'\hat{\boldsymbol{\beta}} - y_t + y_t - \boldsymbol{x}_t'\hat{\boldsymbol{\beta}}_{(-t)} = ((1-h_{tt})^{-1})\boldsymbol{x}_t'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_t\hat{e}_t$$

That is, $\hat{e}_{(-t)} = ((1-h_{tt})^{-1})\hat{e}_t$. Let $\boldsymbol{D}$ be an $T \times T$ diagonal matrix with tth diagonal element $(1-h_{tt})^{-1}$. The leave-one-out or jackknife residual vector is $\tilde{\boldsymbol{e}} = \boldsymbol{D}\hat{\boldsymbol{e}}$, then $(\boldsymbol{I}_T - \boldsymbol{P}^{\text{JMA}})\boldsymbol{y} = \boldsymbol{D}(\boldsymbol{y} - \boldsymbol{P}\boldsymbol{y})$. Therefore, we have $\boldsymbol{P}^{\text{JMA}} = \boldsymbol{D}(\boldsymbol{P} - \boldsymbol{I}_T) + \boldsymbol{I}_T$.

A.2 Proof of Lemma 1

Suppose at least one candidate model is not under-fitted. We denote $\boldsymbol{X}_{m^c}$ as a matrix containing columns of $\boldsymbol{X}$ not in $\boldsymbol{X}_m$, $\boldsymbol{\Pi}_{m^c}$ as the corresponding selection matrix such that $\boldsymbol{X}_{m^c} = \boldsymbol{X}\boldsymbol{\Pi}_{m^c}'$. Similarly, $\boldsymbol{\beta}_m = \boldsymbol{\Pi}_m\boldsymbol{\beta}$ and $\boldsymbol{\beta}_{m^c} = \boldsymbol{\Pi}_{m^c}\boldsymbol{\beta}$.

Since $\boldsymbol{Q}$ is a positive definite matrix, it is well known that

$$|\boldsymbol{Q}| = \begin{vmatrix} \boldsymbol{Q}_{11} & \boldsymbol{Q}_{12} \\ \boldsymbol{Q}_{21} & \boldsymbol{Q}_{22} \end{vmatrix} = |\boldsymbol{Q}_{11}||\boldsymbol{Q}_{22} - \boldsymbol{Q}_{21}\boldsymbol{Q}_{11}^{-1}\boldsymbol{Q}_{12}|$$

which implies $|\boldsymbol{Q}_{22} - \boldsymbol{Q}_{21}\boldsymbol{Q}_{11}^{-1}\boldsymbol{Q}_{12}| > 0$. Thus, for any candidate model m, there exists a positive definite matrix $\boldsymbol{Q}_m$ such that

$$T_1^{-1}\boldsymbol{X}_{m^c}'(\boldsymbol{I}_{T_1} - \boldsymbol{P}_m)\boldsymbol{X}_{m^c} = T_1^{-1}\boldsymbol{X}_{m^c}'\boldsymbol{X}_{m^c} - T_1^{-1}\boldsymbol{X}_{m^c}'\boldsymbol{X}_m(\boldsymbol{X}_m'\boldsymbol{X}_m)^{-1}\boldsymbol{X}_m'\boldsymbol{X}_{m^c}$$

$$\xrightarrow{p} \boldsymbol{Q}_m \tag{A.4}$$

We denote $q_m = \boldsymbol{\beta}_{m^c}'\boldsymbol{Q}_m\boldsymbol{\beta}_{m^c}$. For any under-fitted model $m \in \{1, \ldots, M_0\}$, from equation

(11),

$$\begin{aligned} T_1^{-1}||\boldsymbol{y}_1^0 - \boldsymbol{P}_m\boldsymbol{y}_1^0|| &= T_1^{-1}(\boldsymbol{e}_1 + \boldsymbol{X}_{m^c}\boldsymbol{\beta}_{m^c})'(\boldsymbol{I}_{T_1} - \boldsymbol{P}_m)(\boldsymbol{e}_1 + \boldsymbol{X}_{m^c}\boldsymbol{\beta}_{m^c}) \\ &= T_1^{-1}\boldsymbol{\beta}'_{m^c}\boldsymbol{X}'_{m^c}(\boldsymbol{I}_{T_1} - \boldsymbol{P}_m)\boldsymbol{X}_{m^c}\boldsymbol{\beta}_{m^c} \\ &+ T_1^{-1}||\boldsymbol{e}_1||^2 + 2T_1^{-1}\boldsymbol{e}_1{}'(\boldsymbol{I}_{T_1} - \boldsymbol{P}_m)\boldsymbol{X}_{m^c}\boldsymbol{\beta}_{m^c} \\ &- T_1^{-1}\boldsymbol{e}_1{}'\boldsymbol{P}_m\boldsymbol{e}_1 \\ &= q_m + T_1^{-1}||\boldsymbol{e}_1||^2 + o_p(1) \end{aligned} \tag{A.5}$$

where the term q_m comes from omitted variables. From condition C.1 and C.2, for any $m \in \{1, \ldots, M\}$,

$$\boldsymbol{e}_1'\boldsymbol{P}_m\boldsymbol{e}_1 = O_p(1), \quad \boldsymbol{e}_1'(\boldsymbol{I}_{T_1} - \boldsymbol{P}_m)\boldsymbol{X}_{m^c}\boldsymbol{\beta}_{m^c} = O_p(1) \tag{A.6}$$

Therefore, A.5 follows from A.6 as $T_1 \to \infty$. For any $m \notin \{1, ..., M_0\}$,

$$T_1^{-1}||\boldsymbol{y}_1^0 - \boldsymbol{P}_m\boldsymbol{y}_1^0|| = T_1^{-1}\boldsymbol{e}_1{}'(\boldsymbol{I}_{T_1} - \boldsymbol{P}_m)\boldsymbol{e_1} = T_1^{-1}||\boldsymbol{e}_1||^2 + o_p(1) \tag{A.7}$$

Let $\boldsymbol{\mu}$ denote linear conditional mean function, the Mallows criterion is

$$\begin{aligned} \mathcal{C}(\boldsymbol{\omega}) &= ||\{\boldsymbol{I}_{T_1} - \boldsymbol{P}(\boldsymbol{\omega})\}\boldsymbol{y}_1^0||^2 + 2\hat{\sigma}^2\boldsymbol{\omega}'\boldsymbol{K} \\ &= ||X\hat{\boldsymbol{\beta}}(\boldsymbol{\omega}) - \boldsymbol{\mu} - \boldsymbol{e}_1||^2 + 2\hat{\sigma}^2\boldsymbol{\omega}'\boldsymbol{K} \\ &= ||\boldsymbol{e}_1||^2 + (\hat{\boldsymbol{\beta}}(\boldsymbol{\omega}) - \boldsymbol{\beta})'\boldsymbol{X}'\boldsymbol{X}(\hat{\boldsymbol{\beta}}(\boldsymbol{\omega}) - \boldsymbol{\beta}) \\ &- 2\boldsymbol{e}_1{}'\boldsymbol{X}(\hat{\boldsymbol{\beta}}(\boldsymbol{\omega}) - \boldsymbol{\beta}) + 2\hat{\sigma}^2\boldsymbol{\omega}'\boldsymbol{K} \end{aligned} \tag{A.8}$$

Condition C.1 and C.2 also imply that

$$\hat{\sigma}^2 = O_p(1) \tag{A.9}$$

Given $j \notin \{1, ..., M_0\}$, suppose $\boldsymbol{\omega} = \{0, ..., \omega_j, ..., 0\} = \{0, ..., 1, ..., 0\}$, by condition C.1-C.2, and A.9, $\mathcal{C}(\boldsymbol{\omega}) = ||\boldsymbol{e}_1||^2 + \eta_{T_1}$ where $\eta_{T_1} = O_p(1)$, then $\mathcal{C}(\hat{\boldsymbol{\omega}}_{\text{MMA}}) \leq ||\boldsymbol{e}_1||^2 + \eta_{T_1}$.

From A.8

$$(\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\omega}}_{\text{MMA}}) - \boldsymbol{\beta})'\boldsymbol{X}'\boldsymbol{X}(\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\omega}}_{\text{MMA}}) - \boldsymbol{\beta})$$
$$-2\boldsymbol{e}_1'\boldsymbol{X}(\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\omega}}_{\text{MMA}}) - \hat{\boldsymbol{\beta}}) + 2\hat{\sigma}^2\boldsymbol{\omega}_{\text{MMA}}'\boldsymbol{K} \leq \eta_{T_1} \tag{A.10}$$

Let $\lambda_{\text{min}}(\boldsymbol{X})$ be the smallest eigenvalue of matrix $\boldsymbol{X}$. From A.10,

$$\begin{aligned}
\lambda_{\text{min}}(\boldsymbol{Q}_{T_1})||\sqrt{T_1}(\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\omega}}_{\text{MMA}}) - \boldsymbol{\beta})||^2 &\leq (\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\omega}}_{\text{MMA}}) - \boldsymbol{\beta})'\boldsymbol{X}'\boldsymbol{X}(\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\omega}}_{\text{MMA}}) - \boldsymbol{\beta}) \\
&\leq \eta_{T_1} + 2\boldsymbol{e}_1'\boldsymbol{X}(\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\omega}}_{\text{MMA}}) - \boldsymbol{\beta}) - 2\hat{\sigma}^2\hat{\boldsymbol{\omega}}_{\text{MMA}}'\boldsymbol{K} \\
&\leq \eta_{T_1} + 2||T_1^{-1/2}\boldsymbol{e}_1'\boldsymbol{X}||\ ||\sqrt{T_1}(\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\omega}}_{\text{MMA}}) - \boldsymbol{\beta})||
\end{aligned}$$

Therefore, we have

$||\sqrt{T_1}(\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\omega}}_{\text{MMA}}) - \boldsymbol{\beta})||$

$$\begin{aligned}
\in [&-\{\lambda_{\text{min}}(\boldsymbol{Q}_{T_1})^{-1}(\eta_{T_1} + \lambda_{\text{min}}(\boldsymbol{Q}_{T_1})^{-1}||T_1^{-1/2}\boldsymbol{e}_1'\boldsymbol{X}||^2)\}^{1/2} + \lambda_{\text{min}}(\boldsymbol{Q}_{T_1})^{-1}||T^{-1/2}\boldsymbol{e}_1'\boldsymbol{X}||, \\
&\{\lambda_{\text{min}}(\boldsymbol{Q}_{T_1})^{-1}(\eta_{T_1} + \lambda_{\text{min}}(\boldsymbol{Q}_{T_1})^{-1}||T_1^{-1/2}\boldsymbol{e}_1'\boldsymbol{X}||^2)\}^{1/2} + \lambda_{\text{min}}(\boldsymbol{Q}_{T_1})^{-1}||T_1^{-1/2}\boldsymbol{e}_1'\boldsymbol{X}||]
\end{aligned}$$

The above relationship, together with condition C.2 and $\eta_{T_1} = O_p(1)$, imply that $\sqrt{T_1}(\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\omega}}_{\text{MMA}}) - \boldsymbol{\beta}) = O_p(1)$.

Before showing the consistency of JMA estimator, we rewrite $\mathcal{C}(\boldsymbol{\omega})$ as

$$\mathcal{C}(\boldsymbol{\omega}) = \boldsymbol{\omega}'\boldsymbol{\Phi}\boldsymbol{\omega} \quad \text{for any} \quad \boldsymbol{\omega} \in \mathcal{W} \tag{A.11}$$

where $\boldsymbol{\Phi}$ is an $M \times M$ matrix with (m, j)th element $\boldsymbol{\Phi}_{mj} = a_{\max\{m,j\}} + \hat{\sigma}^2(K_m + K_j)$ and $a_m = \boldsymbol{y}_1^{0\prime}(\boldsymbol{I}_{T_1} - \boldsymbol{P}_m)\boldsymbol{y}_1^0$. To see this, we stack residual vectors by column, so we have $\hat{\boldsymbol{E}} = (\hat{\boldsymbol{e}}_1, ..., \hat{\boldsymbol{e}}_M)$ where $\hat{\boldsymbol{e}}_m = (\hat{e}_{1,m}, ..., \hat{e}_{T_1,m})'$. Let $\boldsymbol{I} = (1, ..., 1)'$ be an M-dimensional

vector, then

$$\begin{aligned}
\mathcal{C}(\boldsymbol{\omega}) &= \boldsymbol{\omega}'\hat{\boldsymbol{E}}'\hat{\boldsymbol{E}}\boldsymbol{\omega} + 2\hat{\sigma}^2\boldsymbol{\omega}'\boldsymbol{K} \\
&= \boldsymbol{\omega}'\hat{\boldsymbol{E}}'\hat{\boldsymbol{E}}\boldsymbol{\omega} + \hat{\sigma}^2\boldsymbol{\omega}'(\boldsymbol{K}\boldsymbol{I}' + \boldsymbol{I}\boldsymbol{K}')\boldsymbol{\omega} \\
&= \boldsymbol{\omega}'(\hat{\boldsymbol{E}}'\hat{\boldsymbol{E}} + \hat{\sigma}^2(\boldsymbol{K}'\boldsymbol{I} + \boldsymbol{I}\boldsymbol{K}'))\boldsymbol{\omega} \\
&= \boldsymbol{\omega}'\boldsymbol{\Phi}\boldsymbol{\omega}
\end{aligned}$$

where

$$\begin{aligned}
\boldsymbol{\Phi}_{mj} &= \hat{\boldsymbol{e}}'_m\hat{\boldsymbol{e}}'_j + \hat{\sigma}^2(K_m + K_j) \\
&= \boldsymbol{y}_1^{0\prime}(\boldsymbol{I}_{T_1} - \boldsymbol{P}_m)(\boldsymbol{I}_{T_1} - \boldsymbol{P}_j)\boldsymbol{y}_1^0 + \hat{\sigma}^2(K_m + K_j) \\
&= a_{\max\{m,j\}} + \hat{\sigma}^2(K_m + K_j) \qquad \text{(A.12)}
\end{aligned}$$

Now we take a look at JMA criterion function. Denote $\boldsymbol{C}_m$ as a $T_1 \times T_1$ diagonal matrix with the tth diagonal element $\boldsymbol{C}_{m,tt} = h_{tt}^m/(1 - h_{tt}^m)$ such that $\boldsymbol{D}_m = \boldsymbol{C}_m + \boldsymbol{I}_{T_1}$. Based on A.11, we can write $\mathcal{J}(\boldsymbol{\omega}) = \mathcal{C}(\boldsymbol{\omega}) + \boldsymbol{\omega}'\boldsymbol{\Psi}\boldsymbol{\omega}$ where $\boldsymbol{\Psi}$ is an $M \times M$ matrix with the (mj)th element $\boldsymbol{\Psi}_{mj} = (\boldsymbol{e}_1 + \boldsymbol{X}_{m^c}\boldsymbol{\beta}_{m^c})'(\boldsymbol{I}_{T_1} - \boldsymbol{P}_m)(\boldsymbol{C}_m + \boldsymbol{C}_j + \boldsymbol{C}_m\boldsymbol{C}_j)(\boldsymbol{I}_{T_1} - \boldsymbol{P}_j)(\boldsymbol{e}_1 + \boldsymbol{X}_{j^c}\boldsymbol{\beta}_{j^c}) - 2K_m\hat{\sigma}^2$. To derive this, note that the model averaging leave-one-out residual $\tilde{e}_{1t}(\boldsymbol{\omega}) = y_{1t}^0 - \sum_{m=1}^M \omega_m \boldsymbol{x}'_{tm}\hat{\boldsymbol{\beta}}_{(-t),m}$. For candidate model m, denote residual vector $\tilde{\boldsymbol{e}}_{1,m} = (\tilde{e}_{11,m}, ...\tilde{e}_{1T_1,m})'$. Similarly, we stack the residual vectors by column and obtain $\tilde{\boldsymbol{E}} = (\tilde{\boldsymbol{e}}_{1,1}, ..., \tilde{\boldsymbol{e}}_{1,M})$. Therefore, $\tilde{\boldsymbol{e}}_1(\boldsymbol{\omega}) = \tilde{\boldsymbol{E}}\boldsymbol{\omega}$ and $\mathcal{J}(\boldsymbol{\omega}) = \tilde{\boldsymbol{e}}_1(\boldsymbol{\omega})'\tilde{\boldsymbol{e}}_1(\boldsymbol{\omega}) = \boldsymbol{\omega}'\tilde{\boldsymbol{E}}'\tilde{\boldsymbol{E}}\boldsymbol{\omega} =$

$\boldsymbol{\omega}'(\tilde{\boldsymbol{e}}'_{1,m}\tilde{\boldsymbol{e}}_{1,j})_{m,j\in\{1,\ldots,M\}}\boldsymbol{\omega}$. From previous section A.1, $\tilde{\boldsymbol{e}} = \boldsymbol{D}\hat{\boldsymbol{e}}$, thus,

$$\begin{aligned}
\tilde{\boldsymbol{e}}'_{1,m}\tilde{\boldsymbol{e}}_{1,j} &= \hat{\boldsymbol{e}}'_{1,m}\boldsymbol{D}_m\boldsymbol{D}_j\hat{\boldsymbol{e}}'_{1,j} \\
&= \boldsymbol{e}'_{1,m}(\boldsymbol{I}_{T_1} - \boldsymbol{P}_m)\boldsymbol{D}_m\boldsymbol{D}_j(\boldsymbol{I}_{T_1} - \boldsymbol{P}_j)\boldsymbol{e}_{1,j} \\
&= \boldsymbol{e}'_{1,m}(\boldsymbol{I}_{T_1} - \boldsymbol{P}_m)(\boldsymbol{C}_m + \boldsymbol{C}_j + \boldsymbol{C}_m\boldsymbol{C}_j + \boldsymbol{I}_{T_1})(\boldsymbol{I}_{T_1} - \boldsymbol{P}_j)\boldsymbol{e}_{1,j} \\
&= \boldsymbol{y}_1^{0\prime}(\boldsymbol{I}_{T_1} - \boldsymbol{P}_m)(\boldsymbol{C}_m + \boldsymbol{C}_j + \boldsymbol{C}_m\boldsymbol{C}_j + \boldsymbol{I}_{T_1})(\boldsymbol{I}_{T_1} - \boldsymbol{P}_j)\boldsymbol{y}_1^0 \\
&= a_{\max\{m,j\}} + \boldsymbol{y}_1^{0\prime}(\boldsymbol{I}_{T_1} - \boldsymbol{P}_m)(\boldsymbol{C}_m + \boldsymbol{C}_j + \boldsymbol{C}_m\boldsymbol{C}_j)(\boldsymbol{I}_{T_1} - \boldsymbol{P}_j)\boldsymbol{y}_1^0 \\
&= a_{\max\{m,j\}} \\
&\quad +(\boldsymbol{e}_1 + \boldsymbol{X}_{m^c}\boldsymbol{\beta}_{m^c})'(\boldsymbol{I}_{T_1} - \boldsymbol{P}_m)(\boldsymbol{C}_m + \boldsymbol{C}_j + \boldsymbol{C}_m\boldsymbol{C}_j)(\boldsymbol{I}_{T_1} - \boldsymbol{P}_j)(\boldsymbol{e}_1 + \boldsymbol{X}_{j^c}\boldsymbol{\beta}_{j^c}) \\
&= \boldsymbol{\Phi}_{mj} \\
&\quad +(\boldsymbol{e}_1 + \boldsymbol{X}_{m^c}\boldsymbol{\beta}_{m^c})'(\boldsymbol{I}_{T_1} - \boldsymbol{P}_m)(\boldsymbol{C}_m + \boldsymbol{C}_j + \boldsymbol{C}_m\boldsymbol{C}_j)(\boldsymbol{I}_{T_1} - \boldsymbol{P}_j)(\boldsymbol{e}_1 + \boldsymbol{X}_{j^c}\boldsymbol{\beta}_{j^c}) \\
&\quad -\hat{\sigma}^2(K_m + K_j) \\
&= \boldsymbol{\Phi}_{mj} + \boldsymbol{\Psi}_{mj} - \hat{\sigma}^2(K_m - K_j)
\end{aligned}$$

So we write

$$\mathcal{J}(\boldsymbol{\omega}) = \mathcal{C}(\boldsymbol{\omega}) + \boldsymbol{\omega}'\boldsymbol{\Psi}\boldsymbol{\omega} \tag{A.13}$$

Now let $S(\boldsymbol{X})$ denote the largest singular value of a matrix $\boldsymbol{X}$. It is known that for any two $n \times n$ matrix $\boldsymbol{A}$ and $\boldsymbol{B}$, $S(\boldsymbol{AB}) \leq S(\boldsymbol{A})S(\boldsymbol{B})$ and $S(\boldsymbol{A} + \boldsymbol{B}) \leq S(\boldsymbol{A}) + S(\boldsymbol{B})$,

$$\begin{aligned}
&(\boldsymbol{e}_1 + \boldsymbol{X}_{m^c}\boldsymbol{\beta}_{m^c})'(\boldsymbol{I}_{T_1} - \boldsymbol{P}_m)(\boldsymbol{C}_m + \boldsymbol{C}_j + \boldsymbol{C}_m\boldsymbol{C}_j)(\boldsymbol{I}_{T_1} - \boldsymbol{P}_j)(\boldsymbol{e}_1 + \boldsymbol{X}_{j^c}\boldsymbol{\beta}_{j^c}) \\
\leq\ &||\boldsymbol{e_1} + \boldsymbol{X}_{m^c}\boldsymbol{\beta}_{m^c}||\ ||\boldsymbol{e}_1 + \boldsymbol{X}_{j^c}\boldsymbol{\beta}_{j^c}||S((\boldsymbol{I}_{T_1} - \boldsymbol{P}_m)(\boldsymbol{C}_m + \boldsymbol{C}_j + \boldsymbol{C}_m\boldsymbol{C}_j)(\boldsymbol{I}_{T_1} - \boldsymbol{P}_j)) \\
&\leq ||\boldsymbol{e_1} + \boldsymbol{X}_{m^c}\boldsymbol{\beta}_{m^c}||\ ||\boldsymbol{e}_1 + \boldsymbol{X}_{j^c}\boldsymbol{\beta}_{j^c}||\left(2\bar{h}_{T_1} + \bar{h}_{T_1}^2\right) = O_p(1)
\end{aligned} \tag{A.14}$$

The first inequality follows from Schwarz inequality and spectral norm definition. Condition C.3 implies $1/(1-h_{tt}^m) = 1 + h_{tt}^m + o_p(1)$, so $S(\boldsymbol{C}_m) \leq \bar{h}_{T_1}$ and the second inequality follows. Therefore, for any $\boldsymbol{\omega} \in \mathcal{W}$, $\boldsymbol{\omega}'\boldsymbol{\Phi}\boldsymbol{\omega} = O_p(1)$. Then the consistency follows the arguments in

above proof for $\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\omega}}_{\text{MMA}})$ and we obtain $\sqrt{T_1}(\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\omega}}_{\text{JMA}}) - \boldsymbol{\beta}) = O_p(1)$.

A.3 Proof of Proposition 1

The argument works for both JMA and MMA-based estimator. For $t = T_1 + 1, T_1 + 2, \ldots T$,

$$\begin{aligned}
\hat{\Delta}_{1t} &= y_{1t}^1 - \hat{y}_{1t}^0(\hat{\boldsymbol{\omega}}) \\
&= y_{1t}^1 - y_{1t}^0 + y_{1t}^0 - \hat{y}_{1t}^0(\hat{\boldsymbol{\omega}}) \\
&= \Delta_{1t} + \boldsymbol{x}_t'\boldsymbol{\beta} + e_{1t} - \boldsymbol{x}_t'\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\omega}}) \\
&= \Delta_{1t} + e_{1t} + O_p(T_1^{-1/2})
\end{aligned}$$

where the last equality is from Lemma 1. As $T_1, T_2 \to \infty$,

$$\begin{aligned}
\hat{\Delta}_1 - \bar{\Delta}_1 &= T_2^{-1} \sum_{t=T_1+1}^{T} \hat{\Delta}_{1t} - T_2^{-1} \sum_{t=T_1+1}^{T} \Delta_{1t} \\
&= T_2^{-1} \sum_{t=T_1+1}^{T} e_{1t} + O_p(T_1^{-1/2}) \\
&= o_p(1) \qquad \text{(A.15)}
\end{aligned}$$

The last equality follows from condition C.1-C.2. Therefore, by A.15 and the assumption that Δ_{1t} follows a stationary process,

$$\hat{\Delta}_1 - \Delta_1 = \hat{\Delta}_1 - \bar{\Delta}_1 + \bar{\Delta}_1 - \Delta_1 = o_p(1) \qquad \text{(A.16)}$$

A.4 Proof of Theorem 1

By definition under (A.11), $a_m = \boldsymbol{y}_1^0(\boldsymbol{I}_{T_1} - \boldsymbol{P}_m)\boldsymbol{y}_1^0$, so $a_j - a_m = \boldsymbol{y}_1^0(\boldsymbol{P}_m - \boldsymbol{P}_j)\boldsymbol{y}_1^0$ and for sequence of nested model, $\boldsymbol{P}_m - \boldsymbol{P}_j$ is a projection matrix. Therefore, $a_j \geq a_m$ for $m > j$. For $m \in \{1, ..., M_0\}$, we define a weight

$$\tilde{\boldsymbol{\omega}}_m = \left(\hat{\omega}_{\text{MMA},1}, ..., \hat{\omega}_{\text{MMA},m-1}, 0, \hat{\omega}_{\text{MMA},m+1} ..., \hat{\omega}_{\text{MMA},M_0}, ..., \hat{\omega}_{\text{MMA},M} + \hat{\omega}_{\text{MMA},m}\right)'$$

Clearly, $\tilde{\boldsymbol{\omega}}_m \in \mathcal{W}$. In the proof of Theorem 1(Liu and Zhang, 2018), they show the following

$$\begin{aligned} 0 &\leq \mathcal{C}(\tilde{\boldsymbol{\omega}}_m) - \mathcal{C}(\hat{\boldsymbol{\omega}}_{\text{MMA}}) \\ &\leq \hat{\omega}^2_{\text{MMA,m}}(a_m - a_M) + 2\hat{\omega}^2_{\text{MMA,m}}(a_M - a_m) + 2\hat{\omega}_{\text{MMA,m}}\hat{\sigma}^2(K_M - K_m) \end{aligned} \tag{A.17}$$

From A.17, when $\hat{\omega}_{\text{MMA,m}} \neq 0$

$$\hat{\omega}_{\text{MMA,m}} \leq (a_m - a_M)^{-1} 2\hat{\sigma}^2(K_M - K_m) \tag{A.18}$$

For $m \in \{1, ..., M_0\}$,

$$\begin{aligned} a_m - a_M &= (\boldsymbol{e}_1 + \boldsymbol{X}_{m^c}\boldsymbol{\beta}_{m^c})'(\boldsymbol{I}_{T_1} - \boldsymbol{P}_m)(\boldsymbol{e}_1 + \boldsymbol{X}_{m^c}\boldsymbol{\beta}_{m^c}) - \boldsymbol{e}_1'(\boldsymbol{I}_{T_1} - \boldsymbol{P}_M)\boldsymbol{e}_1 \\ &= \boldsymbol{e}_1'(\boldsymbol{P}_M - \boldsymbol{P}_m)\boldsymbol{e}_1 + 2\boldsymbol{e}_1'(\boldsymbol{I}_{T_1} - \boldsymbol{P}_m)\boldsymbol{X}_{m^c}\boldsymbol{\beta}_{m^c} \\ &+ (\boldsymbol{X}_{m^c}\boldsymbol{\beta}_{m^c})'(\boldsymbol{I}_{T_1} - \boldsymbol{P}_m)\boldsymbol{X}_{m^c}\boldsymbol{\beta}_{m^c} \end{aligned} \tag{A.19}$$

Similar to A.4, we have

$$\begin{aligned} & T_1^{-1}(\boldsymbol{X}_{m^c}\boldsymbol{\beta}_{m^c})'(\boldsymbol{I}_{T_1} - \boldsymbol{P}_m)(\boldsymbol{X}_{m^c}\boldsymbol{\beta}_{m^c}) \\ = T_1^{-1}\boldsymbol{\beta}_{m^c}'(\boldsymbol{X}_{m^c}'\boldsymbol{X}_{m^c} - \boldsymbol{X}_{m^c}'\boldsymbol{X}_m(\boldsymbol{X}_m'\boldsymbol{X}_m)^{-1}\boldsymbol{X}_m'\boldsymbol{X}_{m^c})\boldsymbol{\beta}_{m^c} \\ & \xrightarrow{p} c \end{aligned}$$

where c is a positive constant. The above result, along with A.6 and A.19, imply

$$T_1(a_m - a_M)^{-1} = O_p(1) \tag{A.20}$$

Then $\hat{\omega}_{\text{MMA,m}} = O_p(T_1^{-1})$ follows from A.9, A.18 and A.20.

For JMA, given any $m \in \{1, ..., M_0\}$, we can define a similar weight as

$$\bar{\boldsymbol{\omega}}_m = \left(\hat{\omega}_{\text{JMA},1}, ..., \hat{\omega}_{\text{JMA},m-1}, 0, \hat{\omega}_{\text{JMA},m+1} ..., \hat{\omega}_{\text{JMA},M_0}, ..., , \hat{\omega}_{\text{JMA},M} + \hat{\omega}_{\text{JMA},m}\right)'$$

From A.13 and similar arguments for A.17,

$$\begin{aligned} 0 &\leq \mathcal{J}(\bar{\boldsymbol{\omega}}_m) - \mathcal{J}(\hat{\boldsymbol{\omega}}_{\text{JMA}}) \\ &= \mathcal{C}(\bar{\boldsymbol{\omega}}_m) - \mathcal{C}(\hat{\boldsymbol{\omega}}_{\text{JMA}}) + \hat{\omega}^2_{\text{JMA,m}}(\Psi_{MM} + \Psi_{mm} - \Psi_{Mm} - \Psi_{mM}) \\ &+ 2\hat{\omega}_{\text{JMA,m}} \sum_{j=1}^{M} \hat{\omega}_{\text{JMA,j}} \end{aligned}$$

When $\hat{\omega}_{\text{JMA,m}} \neq 0$

$$\begin{aligned} \hat{\omega}_{\text{JMA,m}} \leq\ & (a_m - a_M)^{-1}[2\hat{\sigma}^2(K_M - K_m) + \hat{\omega}_{\text{JMA,m}}(\Psi_{MM} + \Psi_{mm} - \Psi_{Mm} - \Psi_{mM}) \\ & +2\sum_{j=1}^{M} \hat{\omega}_{\text{JMA,j}}(\Psi_{Mj} - \Psi_{mj})] \end{aligned} \tag{A.21}$$

As $T_1 \to \infty$, $T_1^{-1}/T_1^{-1/2} \to 0$, condition C.3 implies $\bar{h}_{T_1} = o_p(T_1^{-1/2})$. Following similar arguments for A.14, we have

$$\begin{aligned} &(\boldsymbol{e}_1 + \boldsymbol{X}_{m^c}\boldsymbol{\beta}_{m^c})'(\boldsymbol{I}_{T_1} - \boldsymbol{P}_m)(\boldsymbol{C}_m + \boldsymbol{C}_j + \boldsymbol{C}_m\boldsymbol{C}_j)(\boldsymbol{I}_{T_1} - \boldsymbol{P}_j)(\boldsymbol{e}_1 + \boldsymbol{X}_{j^c}\boldsymbol{\beta}_{j^c}) \\ &\qquad \leq ||\boldsymbol{e_1} + \boldsymbol{X}_{m^c}\boldsymbol{\beta}_{m^c}||\ ||\boldsymbol{e}_1 + \boldsymbol{X}_{j^c}\boldsymbol{\beta}_{j^c}||(2\bar{h}_{T_1} + \bar{h}^2_{T_1}) = o_p(T_1^{1/2}) \end{aligned} \tag{A.22}$$

The result in A.22, along with A.13 and observation that $\hat{\sigma}^2 = \sigma^2 + o_p(1)$, imply that

$$\Psi_{mj} = o_p(T_1^{1/2}) \quad \text{for any} \quad m, j \in \{1, ..., M\} \tag{A.23}$$

Then $\hat{\omega}_{\text{JMA,m}} = o_p(T_1^{-1/2})$ follows from A.9, A.20 ,A.21 and A.23.

A.5 Proof of Theorem 2

The proof mainly follows Liu and Zhang(2018). For MMA, we denote $\boldsymbol{\Phi}^* = \boldsymbol{\Phi} - ||\boldsymbol{e}_1||^2 \boldsymbol{I}\boldsymbol{I}'$ where $\boldsymbol{I} = (1, ..., 1)'$ and the last term is unrelated to $\boldsymbol{\omega}$. Therefore,

$$\hat{\boldsymbol{\omega}}_{\text{MMA}} = \text{argmin}_{\boldsymbol{\omega} \in \mathcal{W}} \boldsymbol{\omega}' \boldsymbol{\Phi}^* \boldsymbol{\omega}$$

Let $\hat{\boldsymbol{\omega}}_{\text{MMA}} = (\hat{\boldsymbol{\omega}}_1', \hat{\boldsymbol{\omega}}_2')'$ with $\hat{\boldsymbol{\omega}}_1$ containing weights of under-fitted models only. We write

$$|\boldsymbol{\Phi}^*| = \begin{vmatrix} \boldsymbol{\Phi}^*_{11} & \boldsymbol{\Phi}^*_{12} \\ \boldsymbol{\Phi}^*_{21} & \boldsymbol{\Phi}^*_{22} \end{vmatrix}$$

such that $\hat{\boldsymbol{\omega}}'_{\text{MMA}} \boldsymbol{\Phi}^* \hat{\boldsymbol{\omega}}_{\text{MMA}} = \hat{\boldsymbol{\omega}}_1' \boldsymbol{\Phi}^*_{11} \hat{\boldsymbol{\omega}}_1 + \hat{\boldsymbol{\omega}}_2' \boldsymbol{\Phi}^*_{21} \hat{\boldsymbol{\omega}}_1 + \hat{\boldsymbol{\omega}}_1' \boldsymbol{\Phi}^*_{12} \hat{\boldsymbol{\omega}}_2 + \hat{\boldsymbol{\omega}}_2' \boldsymbol{\Phi}^*_{22} \hat{\boldsymbol{\omega}}_2$. Condition C.1 and C.2 imply that $T_1^{-1}(a_m - ||\boldsymbol{e}_1||^2) = O_p(1)$ when $1 \leq m \leq M_0$; $a_m - ||\boldsymbol{e}_1||^2 = O_p(1)$ when $M_0 < m \leq M$. From previous section, we know that $\hat{\omega}_{\text{MMA,m}} = O_p(T_1^{-1})$ when $1 \leq m \leq M_0$. These results imply

$$\hat{\boldsymbol{\omega}}_1' \boldsymbol{\Phi}^*_{11} \hat{\boldsymbol{\omega}}_1 = o_p(1) \quad \hat{\boldsymbol{\omega}}_1' \boldsymbol{\Phi}^*_{12} \hat{\boldsymbol{\omega}}_2 = o_p(1) \tag{A.24}$$

Let $S = M - M_0$ such that $\boldsymbol{\Phi}^*_{22}$ is an $S \times S$ matrix, then (s, j)th element of $\boldsymbol{\Phi}^*_{22}$

$$\begin{aligned} \Phi^*_{22,sj} &= \hat{\boldsymbol{e}}_{1s}' \hat{\boldsymbol{e}}_{1j} + \hat{\sigma}^2 (K_{M_0+s} + K_{M_0+j}) - ||\boldsymbol{e}_1||^2 \\ &= \hat{\sigma}^2 (K_{M_0+s} + K_{M_0+j}) - \boldsymbol{e}_1' \boldsymbol{P}_{M_0+\max(s,j)} \boldsymbol{e}_1 \\ &\xrightarrow{p} \mathcal{T}_{sj} \end{aligned} \tag{A.25}$$

where $\mathcal{T}_{sj} = 2\hat{\sigma}^2 K_{M_0+s} - \boldsymbol{Z}' \boldsymbol{V}_{\max(\{s,j\}} \boldsymbol{Z}$. As in the proof of Theorem 3 in Liu(2015), $\hat{\boldsymbol{\omega}}_2 \xrightarrow{d} \tilde{\boldsymbol{\lambda}}_{\text{MMA}}$. From condition C.1-C.2, for any $m \in \{1, ..., M_0\}$

$$\hat{\boldsymbol{\beta}}_m = \boldsymbol{\Pi}'_m (\boldsymbol{X}'_m \boldsymbol{X}_m)^{-1} \boldsymbol{X}'_m \boldsymbol{y}_1^0 = O_p(1) \tag{A.26}$$

Then we have

$$\begin{aligned}
\sqrt{T_1}\{\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\omega}}_{\text{MMA}}) - \boldsymbol{\beta}\} &= \sum_{m=1}^{M_0} \hat{\omega}_{\text{MMA,m}} \sqrt{T_1}(\hat{\boldsymbol{\beta}}_m - \boldsymbol{\beta}) \\
&\quad + \sum_{m=M_0+1}^{M} \hat{\omega}_{\text{MMA,m}} \sqrt{T_1} \boldsymbol{\Pi}_m' (\boldsymbol{X}_m' \boldsymbol{X}_m)^{-1} \boldsymbol{X}_m' \boldsymbol{e}_1 \\
&= O_p(T_1^{-1/2}) + \sum_{m=M_0+1}^{M} \hat{\omega}_{\text{MMA,m}} \sqrt{T_1} \boldsymbol{\Pi}_m' (\boldsymbol{\Pi}_m \boldsymbol{Q}_{T_1} \boldsymbol{\Pi}_m')^{-1} \boldsymbol{\Pi}_m \boldsymbol{Z}_{T_1} \\
&\xrightarrow{d} \sum_{s=1}^{S} \tilde{\lambda}_{\text{MMA,s}} \boldsymbol{V}_s \boldsymbol{Z} \qquad \text{(A.27)}
\end{aligned}$$

For JMA, we similarly define $\boldsymbol{\Xi}^* = \boldsymbol{\Phi} + \boldsymbol{\Psi} - ||\boldsymbol{e}_1||^2 \boldsymbol{I}\boldsymbol{I}'$ so that

$$\hat{\boldsymbol{\omega}}_{\text{JMA}} = \text{argmin}_{\boldsymbol{\omega} \in \mathcal{W}} \boldsymbol{\omega}' \boldsymbol{\Xi}^* \boldsymbol{\omega}$$

Following above proof, we need to look at term $\boldsymbol{\Psi}$. Note that for any $m \in \{1, ..., M\}$,

$$\begin{aligned}
\boldsymbol{e}_1' \text{diag}(h_{11}^m, ..., h_{T_1 T_1}^m) \boldsymbol{e}_1 &= \sum_{t=1}^{T_1} e_{1t}^2 \boldsymbol{x}_{m,t}' (\boldsymbol{X}_m' \boldsymbol{X}_m) \boldsymbol{x}_{m,t} \\
&= \text{tr}((T_1^{-1} \boldsymbol{X}_m' \boldsymbol{X}_m)^{-1} T_1^{-1} \sum_{t=1}^{T_1} e_{1t}^2 \boldsymbol{x}_{m,t} \boldsymbol{x}_{m,t}') \\
&= \text{tr}((\boldsymbol{\Pi}_m \boldsymbol{Q}_{T_1} \boldsymbol{\Pi}_m')^{-1} \boldsymbol{\Pi}_m \boldsymbol{\Omega}_{T_1} \boldsymbol{\Pi}_m')
\end{aligned}$$

Recall $\boldsymbol{C}_m$ is a $T_1 \times T_1$ diagonal matrix with tth diagonal element $\boldsymbol{C}_{m,tt} = h_{tt}^m/(1-h_{tt}^m)$. From condition C.3, $1/(1-h_{tt}^m) = 1+h_{tt}^m+o_p(1)$, so we have $\boldsymbol{e}_1' \boldsymbol{C}_m \boldsymbol{e}_1 = \text{tr}((\boldsymbol{\Pi}_m \boldsymbol{Q}_{T_1} \boldsymbol{\Pi}_m')^{-1} \boldsymbol{\Pi}_m \boldsymbol{\Omega}_{T_1} \boldsymbol{\Pi}_m')$. By condition C.3 and following arguments for deriving A.22, we have

$$\begin{aligned}
\boldsymbol{e}_1' \boldsymbol{P}_m (\boldsymbol{C}_m + \boldsymbol{C}_j + \boldsymbol{C}_m \boldsymbol{C}_j)(\boldsymbol{I}_{T_1} - \boldsymbol{P}_j) \boldsymbol{e}_1 &\leq ||\boldsymbol{P}_m \boldsymbol{e}_1|| S((\boldsymbol{C}_m + \boldsymbol{C}_j + \boldsymbol{C}_m \boldsymbol{C}_j)(\boldsymbol{I}_{T_1} - \boldsymbol{P}_j)) ||\boldsymbol{e}_1|| \\
&\leq ||\boldsymbol{P}_m \boldsymbol{e}_1|| \bar{h}_{T_1} ||\boldsymbol{e}_1|| \\
&= o_p(1)
\end{aligned}$$

and

$$\boldsymbol{e}_1'\boldsymbol{P}_m(\boldsymbol{C}_m+\boldsymbol{C}_j+\boldsymbol{C}_m\boldsymbol{C}_j)\boldsymbol{P}_j\boldsymbol{e}_1 \leq ||\boldsymbol{P}_m\boldsymbol{e}_1||\bar{h}_{T_1}||\boldsymbol{P}_j\boldsymbol{e}_1|| = o_p(1)$$

Based on the above results,

$$\begin{aligned}\Psi_{mj} &= \boldsymbol{e}_1'(\boldsymbol{I}_{T_1}-\boldsymbol{P}_m)(\boldsymbol{C}_m+\boldsymbol{C}_j+\boldsymbol{C}_m\boldsymbol{C}_j)(\boldsymbol{I}_{T_1}-\boldsymbol{P}_j)\boldsymbol{e}_1 - 2K_m\sigma^2 \\ &= \text{tr}\big((\boldsymbol{\Pi}_m\boldsymbol{Q}_{T_1}\boldsymbol{\Pi}_m')^{-1}\boldsymbol{\Pi}_m\boldsymbol{\Omega}_{T_1}\boldsymbol{\Pi}_m'\big) + \text{tr}\big((\boldsymbol{\Pi}_j\boldsymbol{Q}_{T_1}\boldsymbol{\Pi}_j')^{-1}\boldsymbol{\Pi}_j\boldsymbol{\Omega}_{T_1}\boldsymbol{\Pi}_j'\big) \\ &\quad +\hat{\gamma}_{mj} - 2K_m\hat{\sigma}^2\end{aligned}$$

where $\hat{\gamma}_{mj} = o_p(1)$. We can similarly define $\hat{\boldsymbol{\omega}}_{\text{JMA}} = (\hat{\boldsymbol{\omega}}_1', \hat{\boldsymbol{\omega}}_2')'$ with $\hat{\boldsymbol{\omega}}_1$ including only weights of under-fitted models. Let $\boldsymbol{\Xi}^*_{22}$ be a $S \times S$ matrix which is bottom-right block of $\boldsymbol{\Xi}^*$. The (s,j)th element of $\boldsymbol{\Xi}^*_{22}$ is

$$\Xi^*_{22,sj} = \boldsymbol{\Phi}^*_{22,sj} + \boldsymbol{\Psi}_{22,sj} \xrightarrow{d} \Sigma_{sj}$$

where $\Sigma_{sj} = \text{tr}(\boldsymbol{Q}_s^{-1}\boldsymbol{\Omega}_s) + \text{tr}(\boldsymbol{Q}_j^{-1}\boldsymbol{\Omega}_j) - \boldsymbol{Z}'\boldsymbol{V}_{\max\{s,j\}}\boldsymbol{Z}$. Similarly, $\hat{\boldsymbol{\omega}}_2 \xrightarrow{d} \tilde{\boldsymbol{\lambda}}_{\text{JMA}}$. Finally,

$$\begin{aligned}\sqrt{T_1}\{\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\omega}}_{\text{JMA}}) - \boldsymbol{\beta}\} &= \sum_{m=1}^{M_0}\hat{\omega}_{\text{JMA,m}}\sqrt{T_1}(\hat{\boldsymbol{\beta}}_m - \boldsymbol{\beta}) \\ &\quad + \sum_{m=M_0+1}^{M}\hat{\omega}_{\text{JMA,m}}\sqrt{T_1}\boldsymbol{\Pi}_m'(\boldsymbol{X}_m'\boldsymbol{X}_m)^{-1}\boldsymbol{X}_m'\boldsymbol{e}_1 \\ &= o_p(1) + \sum_{m=M_0+1}^{M}\hat{\omega}_{\text{JMA,m}}\sqrt{T_1}\boldsymbol{\Pi}_m'(\boldsymbol{\Pi}_m\boldsymbol{Q}_{T_1}\boldsymbol{\Pi}_m')^{-1}\boldsymbol{\Pi}_m\boldsymbol{Z}_{T_1} \\ &\xrightarrow{d} \sum_{s=1}^{S}\tilde{\lambda}_{\text{JMA,s}}\boldsymbol{V}_s\boldsymbol{Z} \qquad \text{(A.28)}\end{aligned}$$

A.6 Proof of Theorem 3

We can write

$$\begin{aligned}
\hat{A} &= \sqrt{T_2}(\hat{\Delta}_1 - \Delta_1) \\
&= \frac{1}{\sqrt{T_2}} \sum_{t=T_1+1}^{T} (y_{1t} - \hat{y}_{1t}^0 - \Delta_1) \\
&= \frac{1}{\sqrt{T_2}} \sum_{t=T_1+1}^{T} (\boldsymbol{x}_t'\boldsymbol{\beta} + \Delta_{1t} + e_{1t} - \boldsymbol{x}_t'\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\omega}}) - \Delta_1) \\
&= -\sqrt{\frac{T_2}{T_1}}(\frac{1}{T_2} \sum_{t=T_1+1}^{T} \boldsymbol{x}_t')\sqrt{T_1}(\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\omega}}) - \boldsymbol{\beta}) + \frac{1}{\sqrt{T_2}} \sum_{t=T_1+1}^{T} (\Delta_{1t} - \Delta_1 + e_{1t}) \\
&= -\sqrt{\frac{T_2}{T_1}}(\frac{1}{T_2} \sum_{t=T_1+1}^{T} \boldsymbol{x}_t')\sqrt{T_1}(\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\omega}}) - \boldsymbol{\beta}) + \frac{1}{\sqrt{T_2}} \sum_{t=T_1+1}^{T} \nu_{1t} \\
&= \hat{A}_1 + \hat{A}_2 \qquad \text{(A.29)}
\end{aligned}$$

where $\hat{A}_1 = -\sqrt{\frac{T_2}{T_1}}(\frac{1}{T_2} \sum_{t=T_1+1}^{T} \boldsymbol{x}_t')\sqrt{T_1}(\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\omega}}) - \boldsymbol{\beta})$, $\hat{A}_2 = \frac{1}{\sqrt{T_2}} \sum_{t=T_1+1}^{T} \nu_{1t}$.

$\hat{A}_2 \xrightarrow{d} Z_2$ by condition C.4, where $Z_2 \sim N(0, \Omega_v)$. By Theorem 2 and condition C.1, $\hat{A}_1 \xrightarrow{d} A_1 = -\eta\, \mathbb{E}(\boldsymbol{x}_t')(\sum_{s=1}^{S} \tilde{\lambda}_{\text{MA,s}} \boldsymbol{V}_s \boldsymbol{Z})$ and $\tilde{\lambda}_{\text{MA}}$ can be either MMA or JMA weights. Let $\boldsymbol{Z}_1$ denote the asymptotic distribution of $\sqrt{T_1}(\hat{\boldsymbol{\beta}}_M - \boldsymbol{\beta})$, that is, the asymptotic distribution of the full model. In Theorem 3.2 and Lemma A.1 of Li and Bell (2017), they showed that $\boldsymbol{Z}_1$ and Z_2 are asymptotically uncorrelated and therefore asymptotically independent, which suggests the independence of $\boldsymbol{Z}$ and Z_2 . As a result, A_1, as a function of $\boldsymbol{Z}$, is asymptotically independent of Z_2 and expression (24) holds true.

A.7 Proof of Theorem 4

The proof mainly follows Li (2019). From derivation of Proposition 1, we know that $\hat{\Delta}_{1t} = \Delta_{1t} + e_{1t} + O_p(T_1^{-1/2})$. Also, $\hat{\Delta}_1 = \bar{\boldsymbol{x}}'(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}(\boldsymbol{\omega})) + \bar{\Delta}_1 + \bar{\boldsymbol{e}}_1 = \Delta_1 + O_p(T_1^{-1/2} + T_2^{-1/2})$.

Therefore,

$$\begin{aligned}\hat{\Omega}_v &= \frac{1}{T_2}\sum_{t=T_1+1}^{T}(\Delta_{1t}+e_{1t}-\Delta_1)^2 + O_p(T_1^{-1/2}+T_2^{-1/2}) \\ &= \Omega_v + O_p(T_1^{-1/2}+T_2^{-1/2})\end{aligned}$$

That is, Ω_v can be consistently estimated by $\hat{\Omega}_v$. Then $T_2^{-1/2}\sum_{t=T_1+1}^{T}\nu_{1t}^* \sim T_2^{-1/2}\sum_{t=T_1+1}^{T}\nu_{1t}$ $\xrightarrow{d} Z_2$, where $\sim$ represents asymptotic equivalence. From condition that $b \to \infty,\ b/T_1 \to 0$ as $T_1 \to \infty$ and Theorem 3, by Theorem 2.2.1 of Politis, Ramano, Wolf(1999), $\sqrt{b}(\hat{\boldsymbol{\beta}}_b^* - \hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\omega}})) \sim \sqrt{T_1}(\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\omega}}) - \boldsymbol{\beta})$. Therefore, $\hat{A}^*$ in (25). and $\hat{A}$ in (24). have the same asymptotic distribution.

References

[1] Abadie, A., and Gardeazabal, J. 2003. "The Economic costs of conflict: A case study of the Basque Country." *American Economic Review*, 93, 113-132.

[2] Abadie, A., Diamond, A., and Hainmueller, J. 2010. "Synthetic control methods for comparative case studies: estimating the effect of California's Tobacco Control Program." *Journal of the American Statistical Association*, 105, 493-505.

[3] Akaike, H. 1973. "Information theory and an extension of the maximum likelihood principle". In *Proceedings of the 2nd international Symposium on Information Theory*, Petrov BN, Csaki F (eds). Akadémiai Kiadó: Budapest, 267-281

[4] Akaike, H. 1974. "A new look at the statistical model identification". *IEEE Transactions on Automatic Control*, AC19, 716-723

[5] Birk, R., Brown, L., Buja, A., Zhang, K. and Zhao, L. 2013. "Valid post-selection inference". *The Annals of Statistics*, 41, 802-837

[6] Doudchenko, N., and Imbens, G.W. 2016. "Balancing, Regression, Difference-in-Differences and Synthetic Control Methods: A Synthesis". Discussion paper, National Bureau of Economic Research.

[7] Du, Z., and Zhang, L. 2015. "Home-purchase restriction, property tax and housing price in China: A Counterfactual Analysis". *Journal of Econometrics*, 188, 558-568.

[8] Faraway, J. 1992. "On the cost of data analysis". *Journal of Computational and Graphical Statistics*, 1, 213-229.

[9] Gobillon, L., and Thierry, M. 2016. “Regional policy evaluation: Interactive fixed effects and synthetic controls”. *The Review of Economics and Statistics*, 98, 535-551.

[10] Gardeazabal, J., and Vega-Bayo, A. 2016. “An empirical comparison between the synthetic control method and Hsiao et al.’s panel data approach to program evaluation”. *Journal of Applied Econometrics*, 32, 983-1002.

[11] Hansen, B.E. 2007. “Least squares model averaging”. *Econometrica*, 75, 1175-1189.

[12] Hansen, B.E., and Racine, J. 2012. “Jackknife model averaging”. *Journal of Econometrics*, 167, 38-46.

[13] Hansen, B.E. 2014. “Model averaging, asymptotic risk, and regressor groups”. *Quantitative Economics*, 5, 495-530.

[14] Hsiao, C., Ching, S., and Wan, K.S. 2012. “A panel data approach for program evaluation: measuring the benefits of political and economic integration of Hong Kong with mainland China”. *Journal of Applied Econometrics*, 27, 705-740.

[15] Hoeting, J., Madigan, D., Raftery, A., and Volinsky, C. 1999. “Bayesian model averaging: a tutorial”. *Statistical Science*, 14, 382-417.

[16] Jales, H., Kang, T.H., Stein, G., and Ribeiro, F.G. 2018. “Measuring the role of the 1959 revolution on Cuba’s economic performance”. *World Economy*, 1-32.

[17] Hurvich, C., and Tsai, C. 1989. “Regression and time series model selection in small samples”. *Biometrica*, 76, 297-307.

[18] Li, K., and Bell, D. 2017. “Estimation of average treatment effects with panel data: Asymptotic theory and implementation”. *Journal of Econometrics*, 186, 142-159.

[19] Li, K. 2019. “Statistical inference for average treatment effects estimated by synthetic control methods”. *Journal of the American Statistical Association*, 0, 1-16.

[20] Liang, H., Zou, G., Wan, A.T.K and Zhang, X. 2011. “Optiamal weight choice for frequentist model average estimators”. *Journal of the American Statistical Association*, 106, 1053-1066.

[21] Liu, C. 2015. “Distribution theory of the least squares averaging estimator”. *Journal of Econometrics*, 186, 142-159.

[22] Liu, C., and Zhang, X. 2018. “Inference after model averaging in linear regression models”. *Econometric Theory*, 0, 1-26.

[23] Long, W., Gao, Y., and Wang, Z. 2015. “Estimating average treatment effect by model averaging”. *Economics Letters*, 135, 42-45.

[24] Politis, D.N., Ramano, J.P., and Wolf, M. 1999. *Subsampling*, Springer Series in Statistics, Berlin: Springer.

[25] Shao, J. 1993. “Linear model selection by cross-validation”. *Journal of the American Statistical Association*, 422, 486-495.

[26] Wan, S., Xie, Y., and Hsiao, C. 2018. “Panel data approach vs synthetic control method”. *Economics letters*, 25, 121-123.

[27] Wan, A.T.K., Zhang, X., and Zou, G. 2010. “Least squares model average by Mallows criterion”. *Journal of Econometrics*, 156, 277-283.

[28] Zhang, X. 2015. “Consistency of model averaging estimators”. *Economics Letters*, 130, 120-123.

[29] Zou, H., and Hastie, T. 2005. “Regularization and variable selection via the elastic net”. *Journal of Royal Statistical Society: Series B*, 67, 301-320.

3. Comparing predictability between model averaging and other methods

1. Introduction

An important problem in social sciences is to infer the causal impact of an intervention or an event on an outcome variable of interest. The causal impact of an intervention on the treated unit is also known as the treatment effects on the treated. It is measured as the difference between the observed outcome and the unobserved counterfactual outcome that would have been obtained under the alternative treatment[1]. Therefore, measuring treatment effect can be transformed into a problem of predicting the counterfactual outcome.

One common strategy in practice is to characterize the counterfactual outcome for the treated unit as a linear combination of outcomes for the control units. A recent summary regarding this strategy can be found in Doudchenko and Imbens (2016). However, as with other domains, one often does not know exactly which control units (as predictors) should be included in the model for constructing counterfactual outcome. When predictors with zero effect are included, they cause loss in predictive performance of the model.

In Chapter 2, we proposed using model averaging approach other than model selection or Elastic-net (Zou and Hastie, 2005) in the above framework. Since the true data generating process (DGP) is typically unknown, we conduct simulation experiments to compare which approach is more likely to yield more accurate prediction of counterfactual outcome. When the DGP of the outcome variables follows a common factor structure, we show that the model averaging approach achieves smaller mean squared prediction error (MSPE) than other methods. It is also worth further investigating whether such finite-sample performance can be extended to cases other than pure common factors structure.

To this end, we conduct simulation experiments to compare the predictability between model averaging and other methods under two set-ups. In the first set-up, the counterfactual outcome is assumed to follow an interactive fixed effects (IFE) model (Bai, 2009), which incorporates covariates and common factor structure. References of using the IFE model for causal inference include Gobillon and Magnac (2016), Xu (2017). In the second set-up, the treated counterfactual outcome is time series. We assume that it follows factor-augmented

[1]We consider binary treatment in this chapter.

regression model (Bai and Ng, 2006; Cheng and Hansen, 2015), which adds dynamics by allowing lagged outcome variables and factors[2]. The idea of using time series behaviour of the outcome to measure causal impact of an intervention or an event is explored in Carter and Smith (2007), who estimated the price impact of a market event (a food scare of genetically modified corn) on corn price. Broadersen et al.(2015) proposed a structural Bayesian time series model that estimates the effect of an intervention on a target time-series by comparing the differences between the observed and counterfactual outcome.

In this chapter, the predictability is measured by the MSPE over post-treatment periods. We compare several (frequentist) model averaging approaches with model selection methods such as AIC and BIC. Since penalized regression methods are also discussed for predicting counterfactual outcome (e.g., Doudchenko and Imbens, 2016; Li and Bell, 2017), we include LASSO (Tibshirani, 1996), Elastic-net (Zou and Hastie, 2005) and Adaptive LASSO (Zou, 2006) for comparison.

The rest of the chapter is organized as follows. In section 2, we describe the two models underlying the DGPs for the counterfactual outcome and review different estimation methods. In section 3, we conduct simulations that explore the finite-sample performance of model averaging and other methods. The last section concludes.

2. Description of models and estimation

2.1 Interactive fixed-effect model

Bai (2009) proposed the interactive fixed effects (IFE) model, which can be implemented in prediction of counterfactual outcome. We consider the following model:

$$y_{it} = \delta D_{it} + \boldsymbol{x}_{it}'\boldsymbol{\beta} + \boldsymbol{\lambda}_i'\boldsymbol{f}_t + \epsilon_{it} \tag{1}$$

[2]We use the factor-augmented regression model as DGP to generate outcome variables and compare predictability of different methods, the formal treatment of using the model for causal inference is left for future research.

For simplicity, the treatment effect is assumed to be homogeneous, which is represented by δ. D_{it} is the treatment indicator such that $D_{it} = 1$ if unit i has been exposed to the treatment at t and $D_{it} = 0$ otherwise, $\boldsymbol{x}_{it}$ is a $(k \times 1)$ vector of observed covariates and $\boldsymbol{\beta}$ is a $(k \times 1)$ vector of unknown parameters. The factor component of the model $\boldsymbol{\lambda}_i'\boldsymbol{f}_t = \lambda_{i1}f_{1t} + \lambda_{i2}f_{2t} + \cdots + \lambda_{ir}f_{rt}$, is assumed to take a linear additive form, the number of factors r is assumed to be known. The underlying counterfactual outcome is then given by

$$y_{it}^0 = \boldsymbol{x}_{it}'\boldsymbol{\beta} + \boldsymbol{\lambda}_i'\boldsymbol{f}_t + \epsilon_{it} \tag{2}$$

The model (1) is called the IFE model, in which the unobserved factors $\boldsymbol{f}_t$ and factor loadings $\boldsymbol{\lambda}_i$ are regarded as unknown parameters. The IFE model is a generalization of the common factor model considered in Chapter 1, as it allows for the causal impact of observed covariates $\boldsymbol{x}_{it}$. In the present setting, one has observations $(y_{it}, D_{it}, \boldsymbol{x}_{it})$ for $i = 1, \ldots, N$ and $t = 1, \ldots, T$. We assume that $D_{1t} = 0$ for $t = 1, \ldots, T_0$, $D_{1t} = 1$ for $T_0 + 1, \ldots, T$ and $D_{it} = 0$ for $i = 2, \ldots, N$, $t = 1, \ldots, T$. That is, only first unit receives treatment. If N and T (and T_0) are large, one can estimate the counterfactual outcome of treated unit using the method proposed in Bai (2009), which is to perform principal component analysis and least squares estimation in iterations. However, the large N and T set-up may be a luxury in many applications. When neither N nor T is large, a more practical strategy is to use control units' outcomes $(y_{2t}, \ldots, y_{Nt})$ to predict counterfactual outcome of treated unit y_{1t}^0:

$$y_{1t}^0 = \mu + \sum_{i=2}^{N} \omega_i y_{it} + \nu_{1t} = \boldsymbol{z}_t'\boldsymbol{\pi} + \nu_{1t} \tag{3}$$

where $\boldsymbol{\pi} = (\mu, \omega_2, \ldots, \omega_N)'$, $\boldsymbol{z}_t = (1, y_{2t}, \ldots, y_{Nt})'$, ν_{1t} satisfies $\mathbb{E}(\nu_{1t}) = 0, \mathbb{E}(\nu_{1t}\boldsymbol{z}_t) = \boldsymbol{0}$. However, there is an issue of using (3) directly for predicting treated counterfactuals, as including more control units leads to larger estimation variance and the resulting estimator may suffer from lack of precision. Several methods will be discussed in later section to address such issue.

2.2 Factor-augmented regression model

Inferring the causal impact of an intervention can often be transformed into a problem of predicting the missing counterfactual outcome. When the outcome is a time series, this observation motivates the use of approaches that explores the time-series behaviour of an outcome of interest. In this section, we assume that in the absence of treatment, the outcome is generated by the factor-augmented regression model (Bai and Ng, 2006; Cheng and Hansen, 2015).

Suppose we have observations $\{y_t, x_{it}\}$ for $t = 1, \ldots, T_0, T_0 + 1, \ldots, T_0 + h$ and $i = 1, \ldots, N$, where T_0 is the number of pre-treatment periods, $h \geq 1$ is number of post-treatment periods or forecast horizon. A linear form of factor-augmented regression model for prediction is

$$y_{t+h} = \alpha_0 + \alpha(L)y_t + \boldsymbol{\beta}(L)'\boldsymbol{f}_t + \epsilon_{t+h} \tag{4}$$

$$x_{it} = \boldsymbol{\lambda}_i'\boldsymbol{f}_t + u_{it} \tag{5}$$

where $(\alpha(L), \boldsymbol{\beta}(L))$ are polynomials of lags of order p and q, for some $0 \leq p \leq p_{max}$ and $0 \leq q \leq q_{max}$, the factors $\boldsymbol{f}_t$ in (4) are assumed to satisfy the factor model in (5). We use pre-treatment data to estimate (4) and (5). Hence, (5) can be written in matrix form as $\boldsymbol{X} = \boldsymbol{F}\boldsymbol{\Lambda}' + \boldsymbol{u}$, where $\boldsymbol{X}$ is $T_0 \times N$. $\boldsymbol{F} = (\boldsymbol{f}_1, \ldots, \boldsymbol{f}_{T_0})'$ is $T_0 \times r$, $\boldsymbol{\Lambda} = (\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_N)'$ is $N \times r$ and $\boldsymbol{u}$ is a $T_0 \times N$ error matrix. The number of factors r is assumed to be fixed and known. In our simulations, r can be estimated by the information criterion in Bai and Ng (2002), namely,

$$IC_{p2}(k) = \log(V(k)) + k(\frac{N + T_0}{NT_0})\log(\min\{N, T_0\})$$

where $V(k) = \min_{\boldsymbol{\Lambda} \in \mathbb{R}^{N \times k}, \boldsymbol{F} \in \mathbb{R}^{T_0 \times k}} (1/NT_0) \sum_{i=1}^{N} \sum_{t=1}^{T_0} (x_{it} - \boldsymbol{\lambda}_i'\boldsymbol{f}_t)^2$ for $0 \leq k \leq \bar{k}$ ($\bar{k}$ is an upper bound of the number of factors). Let $\hat{k} = \text{argmin}_{0 \leq k \leq \bar{k}} IC_{p2}(k)$, $P(\hat{k} = r) \to 1$ as $N, T_0 \to \infty$ under some conditions.

The factor-augmented regression model allows the impact from lagged outcome and lagged factors so that the underlying model becomes more dynamic. Our goal is to predict y_{T_0+h} for

$h \geq 1$. However, we cannot use (4) directly since $\boldsymbol{f}_t$ is not observed. Instead, we can first estimate the factor model in (5) via principle component methods and obtain common factor estimates $\tilde{\boldsymbol{f}}_t$. As a result, the application using this factor-augmented regression model requires large number of predictors and pre-treatment periods. We can then regress y_{t+h} on the intercept, y_t, $\tilde{\boldsymbol{f}}_t$ and obtain the least squares estimates $\hat{\alpha}_0$, $\hat{\alpha}(L)$ and $\hat{\boldsymbol{\beta}}(L)$. The prediction for y_{T_0+h} can be written as

$$\hat{y}_{T_0+h|T_0} = \hat{\alpha}_0 + \hat{\alpha}(L)y_{T_0} + \hat{\boldsymbol{\beta}}(L)'\tilde{\boldsymbol{f}}_{T_0} \tag{6}$$

Given p_{max}, q_{max} and estimated factors, the largest possible model for (4) includes predictors

$$\tilde{\boldsymbol{z}}_t = (1, y_t, \ldots, y_{t-p_{max}}, \tilde{\boldsymbol{f}}_t', \ldots, \tilde{\boldsymbol{f}}_{t-q_{max}}')' \tag{7}$$

and it can be written as

$$y_{t+h} = \tilde{\boldsymbol{z}}_t'\boldsymbol{\pi} + \epsilon_{t+h} \tag{8}$$

where $\boldsymbol{\pi}$ includes all coefficients from (4). Similar to model (3), the regression on full model (8) may not deliver accurate out-of-sample prediction.

2.3 Estimation methods

Model averaging

Given linear regression models in (3) and (8), different combinations of predictors (e.g., control units' outcomes in (3), lagged outcomes and factors in (8)) constitute different candidate models. Unlike model selection, which selects a single model among candidate models, model averaging incorporates all available information by averaging over all candidate models. The motivation of model averaging is originated from addressing model uncertainty, it also leads to reduced prediction variance and good finite-sample performance (smaller MSPE) (see Zhang, Wan and Zou, 2013; Hansen, 2014).

Denote M the total number of candidate models. Suppose that one is considering M can-

didate models indexed by $m = 1, \ldots, M$, where each candidate model m specifies a subset $\boldsymbol{z}_t(m)$ of the predictors $\boldsymbol{z}_t$. The mth candidate model based on (3) is given by

$$y_{1t}^0 = \boldsymbol{z}_t(m)'\boldsymbol{\pi}(m) + \nu_{1t}(m) \tag{9}$$

The mth candidate model based on (8) is

$$y_{t+h} = \tilde{\boldsymbol{z}}_t(m)'\boldsymbol{\pi}(m) + \epsilon_{t+h}(m) \tag{10}$$

Following Hansen (2007), we do not place any restrictions on the candidate models, that is, they can be nested or non-nested. Nevertheless, we consider nested models in our simulations as they are computationally feasible for moderate or even large number of predictors. In terms of candidate model (9), we set $\boldsymbol{z}_t(m) = (1, y_{2t}, \ldots, y_{mt})'$ for $2 \leq m \leq N$ and $z_t(1)$ contains only one. Let $\boldsymbol{Z}(m) = (\boldsymbol{z}_1(m), \boldsymbol{z}_2(m), \ldots, \boldsymbol{z}_{T_0}(m))'$, the least squares estimate of $\boldsymbol{\pi}(m)$ is $\hat{\boldsymbol{\pi}}(m) = (\boldsymbol{Z}(m)\boldsymbol{Z}(m))^{-1}\boldsymbol{Z}(m)'\boldsymbol{y}$ with residual $\hat{\nu}_{1t}(m) = y_{1t} - \boldsymbol{z}_t(m)'\hat{\boldsymbol{\pi}}(m)$. The prediction based on the mth candidate model for $t \geq T_0 + 1$ is

$$\hat{y}_{1t}^0(m) = \boldsymbol{z}_t(m)'\hat{\boldsymbol{\pi}}(m) \tag{11}$$

In terms of factor-augmented regression candidate model (10), we set $\tilde{\boldsymbol{z}}_t(m) = (1, y_t, y_{t-1}$ $\ldots, y_{t-p(m)}, \tilde{\boldsymbol{f}}_t^m, \ldots, \tilde{\boldsymbol{f}}_{t-q(m)}^m)$ where $0 \leq p(m) \leq p_{max}$ and $0 \leq q(m) \leq q_{max}$. Let $\tilde{\boldsymbol{Z}}(m) = (\tilde{\boldsymbol{z}}_1(m), \ldots, \tilde{\boldsymbol{z}}_{T_0}(m))'$. The least squares estimate of $\boldsymbol{\pi}(m)$ is given by $\hat{\boldsymbol{\pi}}(m) = (\tilde{\boldsymbol{Z}}(m)'\tilde{\boldsymbol{Z}}(m))^{-1}\tilde{\boldsymbol{Z}}(m)'\boldsymbol{y}$ with $\hat{\epsilon}_{t+h}(m) = y_{t+h} - \hat{\boldsymbol{\pi}}_t(m)'\tilde{\boldsymbol{z}}_t(m)$. The least squares prediction by the mth candidate model is then given by

$$\hat{y}_{T_0+h|T_0}(m) = \tilde{\boldsymbol{z}}_{T_0}(m)'\hat{\boldsymbol{\pi}}(m) \tag{12}$$

Once we obtain prediction from each candidate model given in (11) and (12), we construct

the combination of predictions by taking weighted average in the form:

$$\hat{y}_{1t}^{0}(\boldsymbol{\omega}) = \sum_{m=1}^{M} \omega_m \hat{y}_{1t}^{0}(m) \tag{13}$$

and

$$\hat{y}_{T_0+h|T_0}(\boldsymbol{\omega}) = \sum_{m=1}^{M} \omega_m \hat{y}_{T_0+h|T_0}(m) \tag{14}$$

where $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_M)$ is the weights vector with $0 \leq \omega_m \leq 1$ and $\sum_{m=1}^{M} \omega_m = 1$. The averaging residuals are given by $\hat{\nu}_{1t}(\boldsymbol{\omega}) = \sum_{m=1}^{M} \omega_m \hat{\nu}_{1t}(m)$ and $\hat{\epsilon}_{t+h}(\boldsymbol{\omega}) = \sum_{m=1}^{M} \omega_m \hat{\epsilon}_{t+h}(m)$, respectively.

The application of model averaging requires selecting model weights that are subject to non-negativity and sum-to-one constraints. In this chapter, we consider three criteria for weights selection: Mallows criterion, leave-one-out cross-validation criterion and leave-h-out cross-validation criterion.

Let k_m denote the number of predictors in the mth candidate model. The Mallows criterion for weight selection in (13) is

$$C(\boldsymbol{\omega}) = \frac{1}{T_0} \sum_{t=1}^{T_0} (\sum_{m=1}^{M} \omega_m \hat{\nu}_{1t}(m))^2 + \frac{2\hat{\sigma}_{T_0}^2}{T_0} \sum_{m=1}^{M} \omega_m k_m \tag{15}$$

where $\hat{\sigma}_{T_0}^2 = (T_0 - k_M)^{-1} \sum_{t=1}^{T_0} \hat{\nu}_{1t}(M)^2$ using the largest model M. Then the Mallows weight vector is

$$\hat{\boldsymbol{\omega}}_{\text{MMA}} = \text{argmin}_{\{0 \leq \omega_i \leq 1; \sum_{i=1}^{M} \omega_i = 1\}} C(\boldsymbol{\omega}) \tag{16}$$

By replacing $\hat{\nu}_{1t}(m)$ with $\hat{\epsilon}_{t+h}(m)$, we can use Mallows criterion to obtain weights in (14). The weight vector $\hat{\boldsymbol{\omega}}_{\text{MMA}}$ is called Mallows model averaging (MMA) weights. Hansen (2008) showed that the MMA criterion is an asymptotically unbiased estimate of the both in-sample MSE and the out-of-sample one-step mean-squared forecast error for stationary dependent observations under homoscedastic regression model.

In the case of heteroscedastic linear regression model, Hansen and Racine (2012) pro-

posed using leave-one-out cross validation criterion to select weights. To obtain averaging weights in (13), the criterion is

$$CV(\boldsymbol{\omega}) = \frac{1}{T_0}\sum_{t=1}^{T_0}(\sum_{m=1}^{M}\omega_m\tilde{\nu}_{1t}(m))^2 \tag{17}$$

where $\tilde{\nu}_{1t}(m)$ is the residual from model m obtained by least squares estimation without the tth observation. The leave-one-out cross-validation choice for weight vector is

$$\hat{\boldsymbol{\omega}}_{\text{JMA}} = \text{argmin}_{\{0\leq\omega_i\leq 1;\sum_{i=1}^{M}\omega_i=1\}}CV(\boldsymbol{\omega}) \tag{18}$$

Hansen and Racine (2012) also called the above weights the Jackknife model averaging (JMA) weights. The JMA weights for (14) can be obtained similarly.

Although the leave-one-out cross-validation criterion allows heteroscedasticity, it requires the errors to be serially uncorrelated. Regarding the factor-augmented regression model in (8), the errors ϵ_{t+h} can be serially correlated when $h > 1$. In this case, the leave-h-out cross-validation (or h-block cross-validation) can be used for weight selection. The idea is to remove $h-1$ observations before and after the tth observation[3] and only use the remaining observations for estimation. We denote the leave-h-out residual $\tilde{\epsilon}_{t+h,h}(m) = y_{t+h} - \tilde{\boldsymbol{z}}_t(m)'\tilde{\boldsymbol{\pi}}_{t,h}(m)$ where $\tilde{\boldsymbol{\pi}}_{t,h}(m)$ is the least squares estimate from regression of y_{t+h} on $\tilde{\boldsymbol{z}}_t(m)$ with the observations $\{y_{j+h}, \tilde{\boldsymbol{z}}_j(m) : j = t-h+1,\ldots,t+h-1\}$ removed. For $h > 1$, Hansen (2010) showed that it can be computed as

$$\tilde{\epsilon}_{t+h,h}(m) = \hat{\epsilon}_{t+h}(m) + \tilde{\boldsymbol{z}}_t(m)'(\sum_{|j-t|\geq h}\tilde{\boldsymbol{z}}_j(m)\tilde{\boldsymbol{z}}_j(m)')^{-1}\times(\sum_{|j-t|<h}\tilde{\boldsymbol{z}}_j(m)\hat{\epsilon}_{j+h}(m)) \tag{19}$$

where $\hat{\epsilon}_{t+h}(m)$ is the residual from least squares regression using all observations. The averaging residual is then $\tilde{\epsilon}_{t+h,h}(\boldsymbol{\omega}) = \sum_{m=1}^{M}\omega_m\tilde{\epsilon}_{t+h,h}(m)$. The leave-$h$-out cross-validation

[3]By doing so, CVA-h is equivalent to leave-one-out cross validation when $h = 1$.

criterion is

$$\begin{aligned} CV_h(\boldsymbol{\omega}) &= \frac{1}{T_0}\sum_{t=1}^{T_0}\tilde{\epsilon}_{t+h,h}(\boldsymbol{\omega})^2 \\ &= \frac{1}{T_0}\sum_{t=1}^{T_0}(\sum_{m=1}^{M}\omega_m\tilde{\epsilon}_{t+h,h}(m))^2 \end{aligned} \tag{20}$$

The selected weight vector is the minimizer to $CV_h(\boldsymbol{\omega})$ as

$$\hat{\boldsymbol{\omega}}_{\text{CVA-h}} = \text{argmin}_{\{0\leq\omega_i\leq 1;\sum_{i=1}^{M}\omega_i=1\}}CV_h(\boldsymbol{\omega}) \tag{21}$$

Penalized regression methods

When the accuracy of predicting the counterfactual outcome is of primary interest, including all available predictors in (3) and (8) is generally not optimal choice as it comes with the danger of over-fitting. Thus, some level of constraints or regularization can be used to prevent the models from being overly complex. Penalized regression methods are often designed such that the degree of model complexity is tuned towards more accurate prediction. This type of methods reduce the model complexity by incorporating a term that penalizes the size of predictors (i.e., control units' outcomes or lagged treated outcome and factors) to the usual measure of model fit (e.g., the sum of squared residuals). Among various penalized regression methods, we pay special attention to LASSO (Tibshirani, 1996), Elastic-net (Zou and Hastie, 2005) and Adaptive LASSO (Zou, 2006). For brevity, we use model (3), the regression on control units' outcomes, to define penalized regression estimators in the following section. The penalized estimators from regression on lagged outcomes and factors in (8) can be defined similarly.

We denote the LASSO estimator of $\boldsymbol{\pi}$ as $\hat{\boldsymbol{\pi}}_{\text{LASSO}}$. It solves the following minimization problem:

$$\text{min}_{\boldsymbol{\pi}\in\mathbb{R}^N}\frac{1}{T_0}\sum_{t=1}^{T_0}(y_{1t}^0 - \boldsymbol{z}_t'\boldsymbol{\pi})^2 + \lambda\sum_{i=1}^{N}|\pi_i| \tag{22}$$

where $\lambda > 0$ is tuning parameter. This is the classic LASSO (Tibshirani, 1996). The LASSO

fits a model containing all N predictors (e.g., control units' outcomes) at once. The ℓ_1 penalty term $\sum_{i=1}^{N} |\pi_i|$ can force some of coefficient estimates to be exactly equal to zero when λ is sufficiently large. By doing so, it avoids over-fitting problem that worsens out-of-sample prediction performance. In practice, the tuning parameter λ can be selected by cross-validation method. However, the cross-validation method can be computationally intensive if the sample is large and there are many sample splits. Alternatively, one can use information criteria such as AIC or BIC to select λ as it is directly related to the degrees of freedom of the model. The prediction based on LASSO estimator is then $\hat{y}_{1t}^0 = \boldsymbol{z}_t' \hat{\boldsymbol{\pi}}_{\text{LASSO}}$.

The elastic-net estimator solves the minimization problem:

$$\min_{\boldsymbol{\pi} \in \mathbb{R}^N} \frac{1}{T_0} \sum_{t=1}^{T_0} (y_{1t}^0 - \boldsymbol{z}_t' \boldsymbol{\pi})^2 + \lambda_R \sum_{i=1}^{N} \pi_i^2 + \lambda_L \sum_{i=1}^{N} |\pi_i| \tag{23}$$

where $\lambda_R = \lambda(1-\alpha)$ and $\lambda_L = \lambda\alpha$ for $\alpha \in [0,1]$. Thus, the elastic-net penalty is a convex combination of ridge penalty (when $\alpha = 0$) and LASSO penalty (when $\alpha = 1$). The idea of elastic-net is to combine the strength of both ridge and LASSO regression and make the underlying model more flexible. In practice, one can choose the value of α while λ can be selected by cross-validation or information criteria. The prediction based on elastic-net estimator is $\hat{y}_{1t}^0 = \boldsymbol{z}_t' \hat{\boldsymbol{\pi}}_{\text{Enet}}$ where $\hat{\boldsymbol{\pi}}_{\text{Enet}}$ is the minimizer of (23).

The adaptive LASSO estimator solves the minimization problem:

$$\min_{\boldsymbol{\pi} \in \mathbb{R}^N} \frac{1}{T_0} \sum_{t=1}^{T_0} (y_{1t}^0 - \boldsymbol{z}_t' \boldsymbol{\pi})^2 + \lambda \sum_{i=1}^{N} \frac{1}{|\hat{\pi}_i|^\alpha} |\pi_i| \tag{24}$$

where $\lambda > 0$ and $\alpha > 0$ are tuning parameters. The adaptive LASSO can be viewed as a more "tailored " version of classic LASSO in the sense that it scales the individual weight by initial estimator such as least square estimator whereas the classic LASSO penalizes all parameters by the same λ. It is worth noting that the initial estimator is not limited to the least squares estimator, one can choose other estimators depending on the applications. In the following simulation, we use LASSO as initial estimator (excluding all variables for which $\hat{\beta}_{i,\text{LASSO}} = 0$). In

practice, one chooses the value of α, and λ is determined by cross-validation or information criteria. We denote the adaptive LASSO estimator $\hat{\boldsymbol{\pi}}_{\text{AdaLASSO}}$. The corresponding prediction is given by $\hat{y}_{1t}^0 = \boldsymbol{z}_t' \hat{\boldsymbol{\pi}}_{\text{AdaLASSO}}$.

3. Simulation studies

Because the counterfactual outcome is unobserved, we carry out computer simulations to generate it and compare the predictability of model averaging and other methods discussed in this chapter. In order to conduct the simulation we use DGPs following the IFE model specified in equation (2) and the factor-augmented regression model specified in equation (4) and (5).

DGP1: IFE model

The first design is modified based on Xu (2017). We assume the DGP that includes two observed time-varying covariates, two unobserved factors and additive individual and time effects:

$$y_{it}^0 = x_{it,1} \cdot 2 + x_{it,2} \cdot 3 + \boldsymbol{\lambda}_i' \boldsymbol{f}_t + \alpha_i + \xi_t + 5 + \epsilon_{it} \tag{25}$$

where $\boldsymbol{f}_t = (f_{1t}, f_{2t})'$ are time-varying factors, $\boldsymbol{\lambda}_i = (\lambda_{i1}, \lambda_{i2})'$ are individual-specific factor loadings. The two covariates are positively correlated with factors and factor loadings, that is, $x_{it,k} = 1 + \boldsymbol{\lambda}_i' \boldsymbol{f}_t + \lambda_{i1} + \lambda_{i2} + f_{1t} + f_{2t} + \eta_{it,k}, k = 1, 2$. The idiosyncratic error term ϵ_{it} and disturbances in covariates $\eta_{it,1}$ and $\eta_{it,2}$ are generated from $i.i.d.N(0, 1)$. We assume factors f_{1t}, f_{2t}, ξ_t are $i.i.d.N(0, 1)$. The factor loadings λ_{i1} and λ_{i2} as well as individual fixed effects α_i are $i.i.d.U(-\sqrt{3}, \sqrt{3})$. We consider two cases of number of units $N = 20, 40$; for each N, we set pre-treatment periods $T_0 = 50, 70, 90$. The number of post-treatment periods is set to 10 for all cases such that $T = 60, 80, 100$. The data are replicated $5,000$ times according to equation (25). The predictability is measured by mean squared prediction error (MSPE), which is computed as MSPE $= (1/S) \sum_{s=1}^{S} \{ \frac{1}{T - T_0} \sum_{t=T_0+1}^{T} (y_{1t,s}^0 - \hat{y}_{1t,s}^0)^2 \}$, where $S = 5,000$.

In this simulation experiment, we consider the following approaches for comparisons: Mallows model averaging (MMA), Jackknife model averaging (JMA), the simple averaging

with equal weights (EQ-MA), LASSO, adaptive LASSO (AdaLASSO), Elastic-net (Enet), AIC, BIC and the least squares on all control units (Full). For each approach, we assume the first unit is treated and remaining $N-1$ units are controls. The counterfactual prediction $\hat{y}_{1t}^0$ is generated based on model (3) as a linear combination of an intercept and control units' outcomes. Regarding model averaging approaches, we consider nested models, hence the number of models is $M = 20$ when $N = 20$, $M = 40$ when $N = 40$. For three penalized regression methods, we follow Medeiros and Mendes (2016) to apply BIC to select the tuning paramter λ. Regarding the adaptive LASSO, we choose LASSO as initial estimator and set $\alpha = 1$, which is a common value in practice. For Elastic-net, we set $\alpha = 0.5$ to avoid selection of an additional tuning parameter.

The main results from the first simulation are summarized in Table 1. Firstly, as expected, the MSPEs of all different methods decrease as the pre-treatment periods increase and number of control units (predictors) increases. In almost all cases, the model averaging with data-driven model weights and penalized regression methods outperform the simple (equal) averaging. On the other hand, the model selection methods AIC and BIC perform worse than the model averaging methods and three penalized methods[4]. Among three penalized regression methods, Elastic-net performs better than LASSO and both yield smaller MSPE than the adaptive LASSO. As discussed in the end of Section 2.1, the least squares on the full model does not deliver good out-of-sample prediction. Here it has worst predictive performance in all cases considered.

In this simulation experiment, the three penalized regression methods yield more accurate prediction than MMA and JMA. This is different from our previous result in Chapter 1 when MMA and JMA are compared to the Elastic-net. We suspect that the difference is due to the correlation among control units' outcomes, which are used as predictors for treated counterfactual outcome. Table A1 in the appendix shows a sample correlation matrix of control units' outcomes generated by three common factor structure used in simulations in Chapter 1. The

[4]We also compare the MSPE of a new model selection criterion — the bridge criterion (BC) that has benefits of both AIC and BIC (Ding, Tarokh and Yang, 2018). Because it is still outperformed by model averaging and three penalized methods in this simulation, we do not report it here to save space.

Table 1: MSPE of different methods

	MMA	JMA	EQ-MA	AIC	BIC	LASSO	Enet	AdaLASSO	Full
$N = 20$									
$T_1 = 50$ $T = 60$	43.69	43.36	46.56	49.03	47.28	39.93	39.14	41.81	473.57
$T_1 = 70$ $T = 80$	40.04	39.92	44.38	43.08	43.10	37.74	37.00	39.56	470.34
$T_1 = 90$ $T = 100$	38.52	38.47	43.89	40.68	41.52	36.79	36.09	38.11	465.75
$N = 40$									
$T_1 = 50$ $T = 60$	26.27	22.98	24.33	62.12	25.35	20.75	22.00	21.14	305.20
$T_1 = 70$ $T = 80$	20.73	20.24	20.84	27.05	21.75	18.24	17.77	19.18	256.51
$T_1 = 90$ $T = 100$	19.53	19.36	19.80	22.39	20.99	17.93	17.43	18.88	238.85

Notes: MMA is Mallows model averaging; JMA is Jackknife model averaging; EQ-MA is simple averaging with equal weights; Enet is Elastic-net penalized regression; AdaLASSO is adaptive LASSO; Full is the least squares using all control units.

control units' outcomes are highly-correlated as many pairwise correlations are greater than 0.6. On the other hand, the IFE model considered in this simulation experiment generates less correlated control units' outcomes, as shown in Table A2 in the appendix[5]. In the presence of highly correlated predictors, the LASSO-type methods can end up selecting arbitrarily one of them (variable selection instability), which could reduce prediction accuracy. The model averaging approaches, however, could give more accurate prediction in this case as they do not select variables for prediction.

[5]One possible reason for the decreased correlation is that by including individual and time fixed effects, which are random draw form $i.i.d.$ uniform distribution and normal distribution (both are of same magnitude as the normal distribution generating common factors), the correlations induced by common factors are reduced.

DGP2: Factor-augmented regression model

In this section we follow Cheng and Hansen (2015) and consider the following DGP for generating time series outcome. Let f_{jt} denote the jth component of $\boldsymbol{f}_t$. For $j = 1, \ldots, r$, $i = 1, \ldots, N$ and $t = 1, \ldots, T$, we assume that the factor model follows:

$$\begin{aligned} x_{it} &= \boldsymbol{\lambda}_i' \boldsymbol{f}_t + \sqrt{r} e_{it} \\ f_{jt} &= \alpha_j f_{jt-1} + u_{jt} \\ e_{it} &= \rho_i e_{it-1} + \epsilon_{it} \end{aligned} \tag{26}$$

where $r = 4$, $\boldsymbol{\lambda}_i \sim N(\mathbf{0}, r\boldsymbol{I}_r)$, $\alpha_j \sim U[0.2, 0.8]$, $\rho_i \sim U[0.3, 0.8]$, $(u_{jt}, \epsilon_{it}) \sim N(\mathbf{0}, \boldsymbol{I_2})$, $i.i.d.$ over t, for j and i. α_j and ρ_i are drawn once and held fixed over all repetitions. Hence, factors f_{jt} and error terms e_{it} are assumed to follow AR(1) process. The outcome is generated by

$$y_{t+h} = \pi_1 f_{2t} + \pi_2 f_{4t} + \pi_3 f_{2t-1} + \pi_4 f_{4t-1} + \pi_5 f_{2t-2} + \pi_6 f_{4t-2} + \epsilon_{t+h} \tag{27}$$

$$\epsilon_{t+h} = \sum_{j=1}^{h-1} \beta^j \nu_{t+h-j} \tag{28}$$

where $\nu_t \sim i.i.dN(0, 1)$, $\{\nu_t\}$ is independent of $\{u_{js}\}$ and $\{\epsilon_{is}\}$ for any t and s. That is, the outcome is determined by the second factor and the fourth factor with their corresponding lags. The error term ϵ_{t+h} follows a moving average process. The parameters are $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_6) = c[0.5, 0.5, 0.2, 0.2, 0.1, 0.1]$, where the scaling parameter c controls the magnitude of the coefficients and is varied from 0.2 to 1.2 when $h = 1$. For $h > 1$, we fix $c = 1$ and vary the moving average parameter β in (28) from 0.1 to 0.9. We set the sample size $N, T = 100$. The total number of replications is 5,000.

We follow Cheng and Hansen (2015) and treat the number of factors r as unknown. We use the information criterion IC_{p2} to select the number of factors $\tilde{r}$, where the number of feasible factors ranges from 0 to 10. Then the first $\tilde{r}$ factors are placed in $\tilde{\boldsymbol{f}}_t$. Given the estimated factors, the set of all predictors are $\boldsymbol{z}_t = (1, y_t, \ldots, y_{t-p_{max}}, \tilde{\boldsymbol{f}}_t', \ldots, \tilde{\boldsymbol{f}}_{t-p_{max}}')'$. Similar to the first simulation experiment, we consider the nested models such that the first candidate model

has $z_t(1) = 1$, the second candidate model has $\boldsymbol{z}_t(2) = (1, y_t)', etc.$ Hence, we construct the candidate models based on the order of predictors in $\boldsymbol{z}_t$. The total number of model M is $(1 + p_{max})(1 + \tilde{r})$. We set possible lags $p_{max} = 0, 2, 4, 9$ and forecast horizon $h = 1, 4, 8, 12$.

For this simulation exercise, we add leave-h-out cross validation averaging (CVA-h) for comparison. The predictability is measured by MSPE = $(1/S)\sum_{s=1}^{S}(y^0_{T_0+h,s} - \hat{y}^0_{T_0+h|T_0,s})^2$ with $S = 5,000$. To more compactly report the comparisons, we normalize the MSPE by the MSPE of simple averaging with equal weights, hence a value smaller than one indicates the superior predictive performance relative to the simple averaging with equal weights.

Figure 1 displays the results for p_{max} = 0, corresponding to the case where the largest candidate model has predictor set $(1, y_t, \tilde{\boldsymbol{f}}'_t)$. Each panel corresponds to the relative MSPE for the forecast horizon $h = 1, 4, 8, 12$. When $h = 1$ (upper left panel), it is worth noting that as c increases, the parameters π_i in (27) increases and the overall signal-to-noise ratio in (27) increases. In the meantime, the parameter c affects the persistence of $\{f_{2t-1}, f_{4t-1}, f_{2t-2}, f_{4t-2}\}$. The lagged factors have more persistent effect on the outcome variable as c increases. The upper left panel shows the relative MSPE of all methods considered decrease as c increases and are well below one, indicating their superior predictive performance over simple equal averaging.

For multiple forecast horizons, the relative MSPE of all methods are again below one in most ranges of moving averaging coefficient β. However, as β increases or equivalently the serial dependence of errors gets stronger, the difference in MSPE between those methods and simple equal averaging shrinks. It is noted that when the forecast horizon is large (h = 8 and 12) and serial dependence of errors is strong ($\beta > 0.8$), only model averaging with leave-h-out CV (CVA-h) outperforms the simple equal averaging benchmark. In general, the model averaging and penalized regression methods have smaller relative MSPE than that of AIC and BIC.

Figure 2 - 4 display the results for largest possible lags $p_{max} = 2, 4, 9$. In the case of $h = 1$, the model averaging approaches outperform penalized regression methods and model selection methods in most range of c. When $h > 1$, the model averaging approaches have

smaller MSPE than that of penalized regression methods when β is large (strong serial dependence of errors ϵ_{t+h}). Again, the advantage of CVA-h becomes prominent for large β. The model selection methods AIC and BIC, however, have worse predictive performance in most range of β.

As we include more lags of y_t and $\tilde{\boldsymbol{f}}_t'$, we also increase the number of correlated predictors in the model. The result that model averaging approaches generally have lower MSPE than the penalized regression methods also supports previous observation that averaging approaches tend to perform better than the penalized regression methods when correlated predictors are present .

Another interesting result is that as lagged predictors are included for prediction , the difference in MSPE between simple equal averaging and those data-driven methods become smaller. Compared to Figure 1, the MSPE in Figure 2-4 from simple equal averaging is comparable to the MSPEs of model averaging with data-driven weights. In many cases, the MSPE of simple equal averaging is smaller than the MSPE of penalized methods

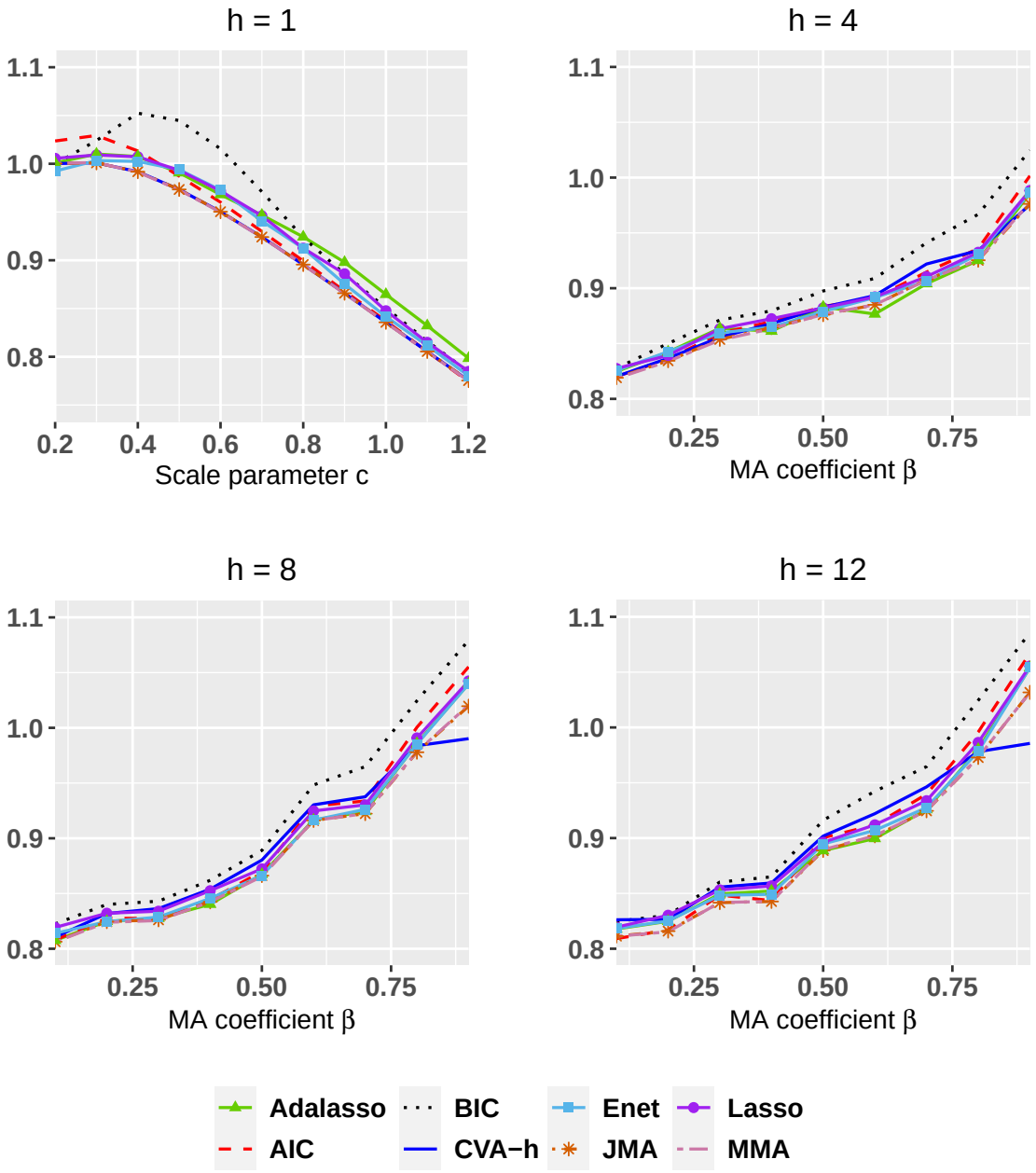

Figure 1: Relative MSPE for $h = 1, 4, 8, 12$ ($p_{max} = 0$)

Note: p_{max} = 0 is the case that no lags of y_t and $\tilde{\boldsymbol{f}}_t$ are used. The MSPE is normalized by that of simple averaging with equal weights. Adalasso is adaptive Lasso. CVA-h is leave-h-out cross-validation averaging. Enet is elastic-net penalized regression. JMA is Jackknife model averaging. MMA is Mallows model averaging.

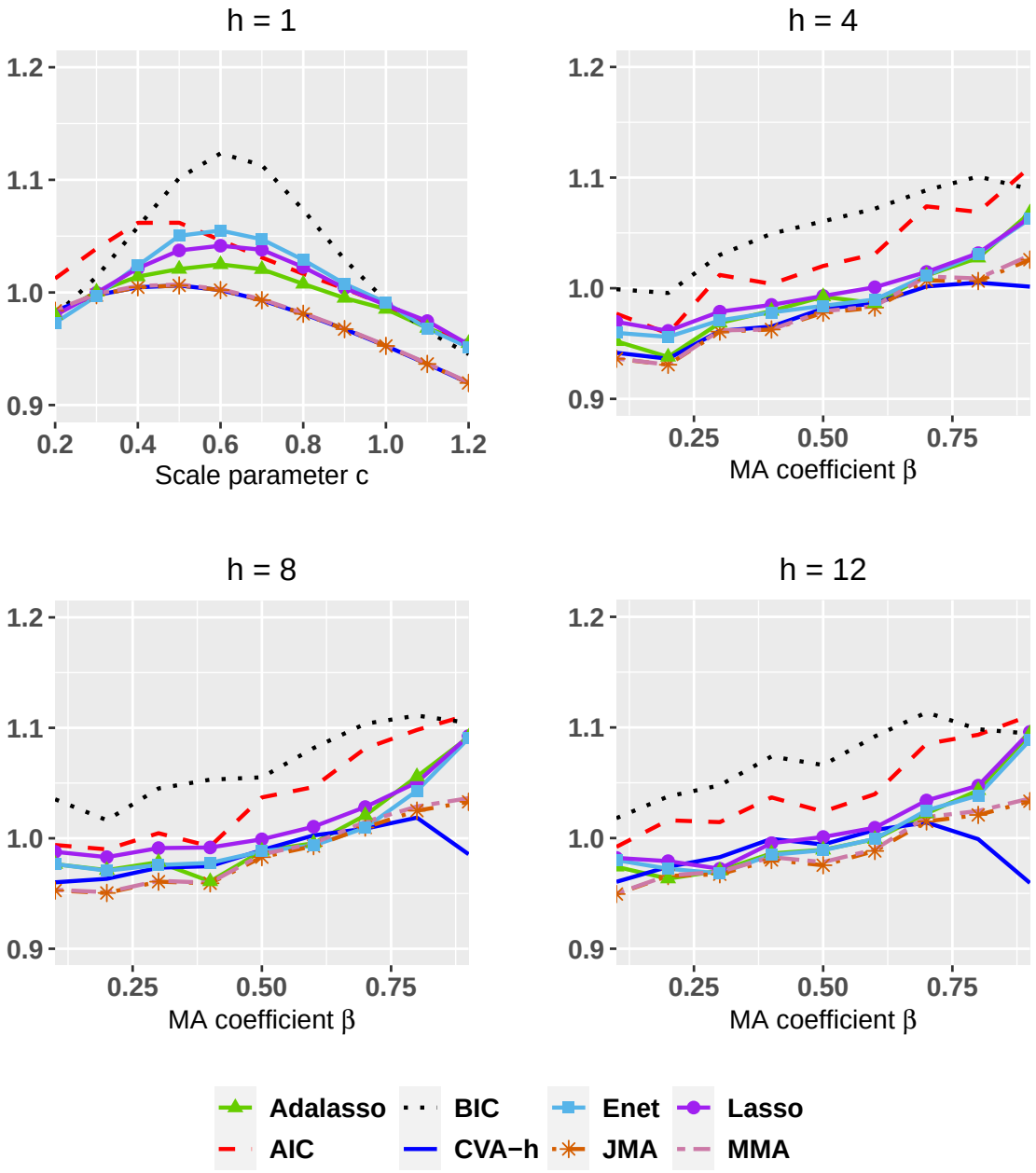

Figure 2: Relative MSPE for $h = 1, 4, 8, 12$ ($p_{max} = 2$)

Note: p_{max} = 2 is the case that two lags of y_t and $\tilde{\boldsymbol{f}}_t$ are used. The MSPE is normalized by that of simple averaging with equal weights. Adalasso is adaptive Lasso. CVA-h is leave-h-out cross-validation averaging. Enet is elastic-net penalized regression. JMA is Jackknife model averaging. MMA is Mallows model averaging.

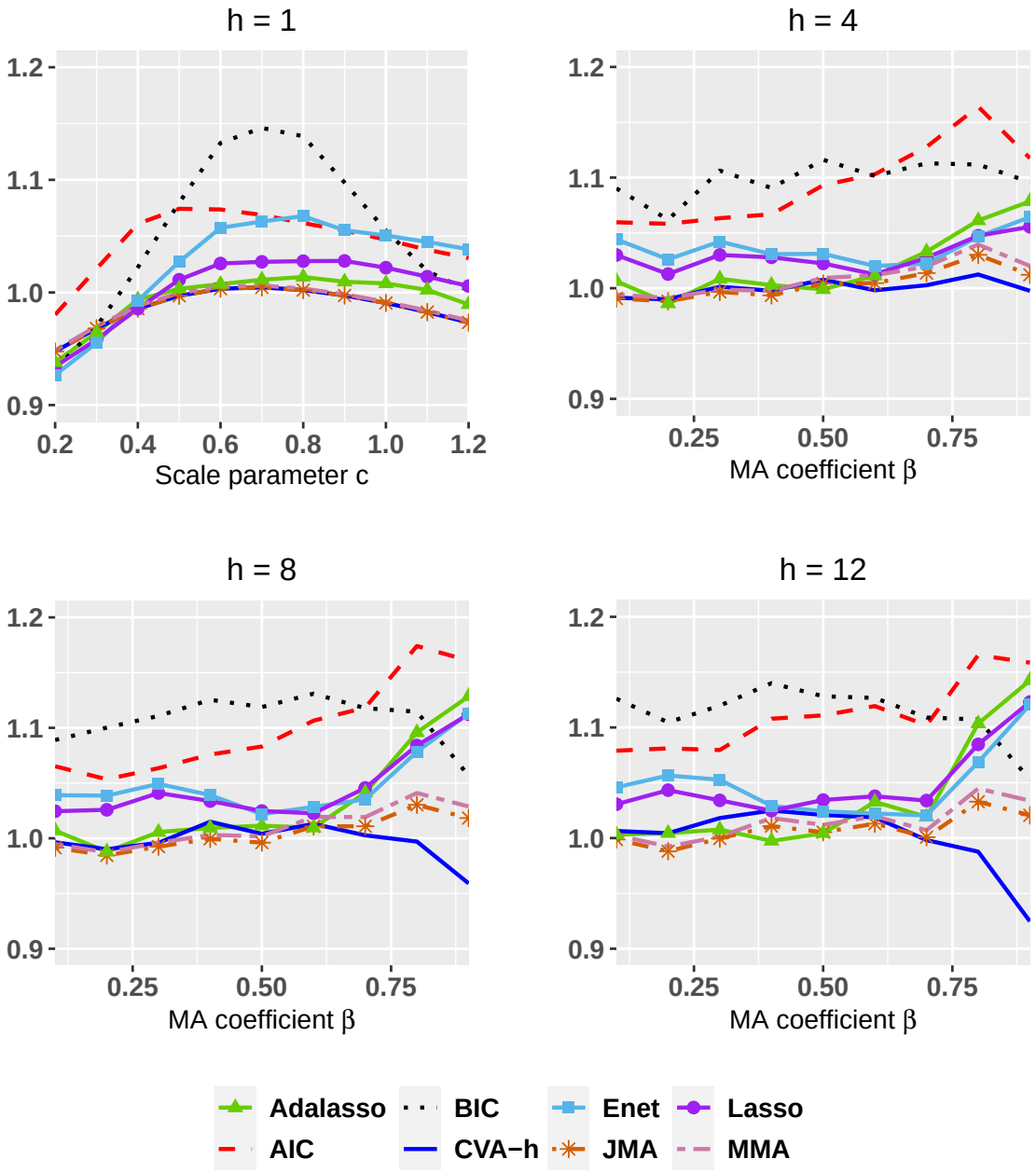

Figure 3: Relative MSPE for $h = 1, 4, 8, 12$ ($p_{max} = 4$)

Note: p_{max} = 4 is the case that four lags of y_t and $\tilde{\boldsymbol{f}}_t$ are used. The MSPE is normalized by that of simple averaging with equal weights. Adalasso is adaptive Lasso. CVA-h is leave-h-out cross-validation averaging. Enet is elastic-net penalized regression. JMA is Jackknife model averaging. MMA is Mallows model averaging.

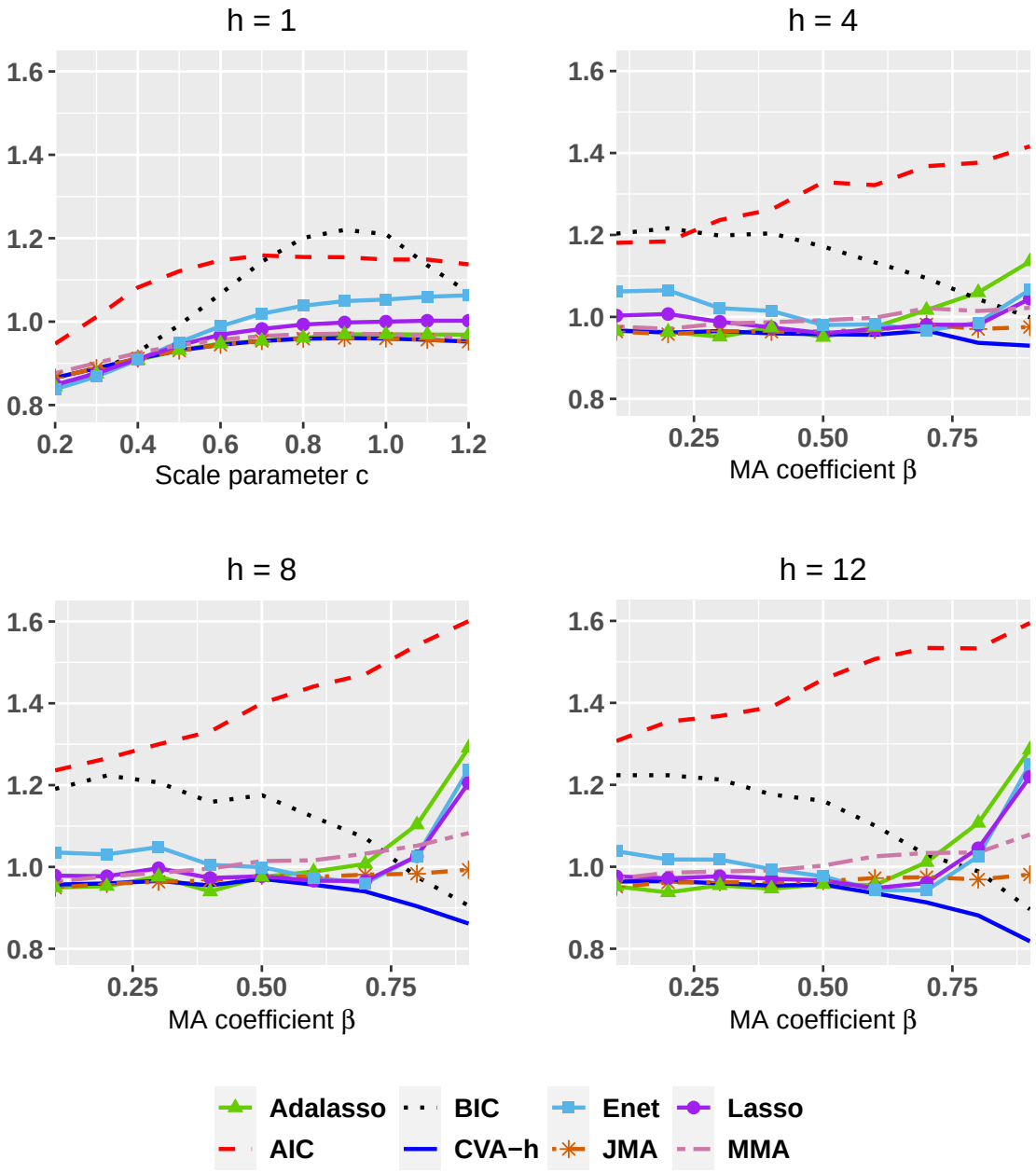

Figure 4: Relative MSPE for $h = 1, 4, 8, 12$ ($p_{max} = 9$)

Note: p_{max} = 9 is the case that nine lags of y_t and $\tilde{\boldsymbol{f}}_t$ are used. The MSPE is normalized by that of simple averaging with equal weights. Adalasso is adaptive Lasso. CVA-h is leave-h-out cross-validation averaging. Enet is elastic-net penalized regression. JMA is Jackknife model averaging. MMA is Mallows model averaging.

4. Concluding remarks

The treatment effect for a treated unit is measured as the difference between the outcome under the treatment and the outcome in the absence of treatment. The outcome under treatment

is observed but the latter is unobserved in post-treatment periods. Therefore, the problem of evaluating the causal impact of an intervention turns into a problem of predicting the missing treated counterfactual outcome. In this chapter, we compare the predictability of of model averaging approaches, penalized regression methods and model selection methods. To this end, we conduct simulation experiments. In the first simulation experiment, we use a linear combination of control units' outcomes to predict the treated counterfactual outcome that is generated by interactive fixed effect model. In the second simulation experiment, we assume that the DGP for treated counterfactual outcome follows a factor-augmented regression model. We predict the counterfactual outcome under this model specification with different lags of outcome variable and factors. Our simulation results show that the model averaging approaches and penalized regression methods have more accurate counterfactual prediction than the model selection methods such as AIC and BIC. In terms of predictability comparison between model averaging approaches and penalized regression methods, neither dominates uniformly the other. If the predictors (e.g., control units' outcomes) are more correlated, the model averaging approaches have more accurate prediction than the penalized regression methods, and vice versa.

Appendix

Table A1: Sample correlation matrix of control units: factor model with three factors

$\sigma^2 = 0.1$											
	y_2	y_3	y_4	y_5	y_6	y_7	y_8	y_9	y_{10}	y_{11}	y_{12}
y_2	1.00										
y_3	-.45	1.00									
y_4	-.37	.98	1.00								
y_5	-.68	.96	.91	1.00							
y_6	.18	.78	.84	.58	1.00						
y_7	-.11	.80	.74	.69	.76	1.00					
y_8	.19	.75	.76	.56	.94	.90	1.00				
y_9	-.50	.99	.95	.96	.73	.84	.74	1.00			
y_{10}	-.26	.90	.84	.82	.77	.97	.87	.93	1.00		
y_{11}	-.58	.98	.96	.98	.68	.69	.62	.97	.81	1.00	
y_{12}	-.10	.93	.94	.79	.93	.87	.92	.90	.91	.85	1.00

Notes: The three-factor structure is the one considered on page 16 of Chapter 1. The sample correlation matrix is calculated using pre-treatment observations ($T_0 = 25$) of 11 control units' outcomes.

Table A2: Sample correlation matrix of control units: IFE model

	y_2	y_3	y_4	y_5	y_6	y_7	y_8	y_9	y_{10}	y_{11}	y_{12}	y_{13}	y_{14}	y_{15}	y_{16}	y_{17}	y_{18}	y_{19}	y_{20}
y_2	1.00																		
y_3	.75	1.00																	
y_4	.17	-.08	1.00																
y_5	-.54	-.05	-.50	1.00															
y_6	-.29	-.71	.20	-.36	1.00														
y_7	.58	.46	.19	-.57	-.17	1.00													
y_8	-.59	-.17	-.54	.92	-.18	-.59	1.00												
y_9	.44	.20	.38	-.55	.17	.47	-.57	1.00											
y_{10}	-.60	-.24	-.40	.87	-.09	-.57	.89	-.44	1.00										
y_{11}	-.57	-.12	-.50	.94	-.24	-.58	.94	-.52	.87	1.00									
y_{12}	.80	.82	.11	-.29	-.50	.42	-.36	.24	-.44	-.34	1.00								
y_{13}	-.20	.29	-.53	.86	-.53	-.34	.79	-.39	.73	.84	.07	1.00							
y_{14}	.05	.40	-.38	.69	-.54	-.20	.64	-.19	.58	.62	.22	.77	1.00						
y_{15}	.87	.85	.15	-.39	-.43	.61	-.45	.44	-.49	-.42	.84	-.01	.11	1.00					
y_{16}	.39	.81	-.29	.44	-.76	.10	.35	-.15	.24	.39	.62	.71	.70	.57	1.00				
y_{17}	.60	.90	-.21	.21	-.72	.28	.11	.04	.02	.14	.75	.53	.59	.73	.93	1.00			
y_{18}	.46	.84	-.29	.37	-.77	.18	.26	-.06	.13	.31	.68	.67	.67	.64	.95	.93	1.00		
y_{19}	-.68	-.36	-.32	.81	-.13	-.57	.83	-.50	.82	.81	-.50	.61	.53	-.59	.14	-.09	.05	1.00	
y_{20}	.78	.61	.36	-.51	-.18	.48	-.56	.46	-.56	-.54	.74	-.26	-.02	.77	.27	.47	.33	-.63	1.00

Notes: The IFE model is based on equation (25). The sample correlation matrix is calculated using pre-treatment observations ($T_0 = 50$) of 19 control units' outcomes.

References

[1] Bai, J. 2009. "Panel data models with interactive fixed effects". *Econometrica*, 77(4), 1229-1279

[2] Bai, J., and Ng, S. 2002. "Determining the number of factors in approximate factor models". *Econometrica*, 70(1), 191-221.

[3] Bai, J., and Ng, S. 2006. "Confidence intervals for diffusion index forecasts and inference for factor-augmented regressions". *Econometrica*, 74, 1133-1150.

[4] Brodersen, K., Fabian, G., Koehler, J., Nicolas, R., and Steven., S. 2015. "Inferring causal impact using Bayesian structural time-series models". *Annals of Applied Statistics*, 9(1), 247-274

[5] Carter, C., and Smith A. 2007. "Estimating the market effect of a food scare: the case of genetically modified Starlink corn". *The Review of Economcis and Statistics*, 89(3), 522-533

[6] Cheng, X., and Hansen, B. 2015. "Forecasting with factor -augmented regression: A frequentist model averaging approach". *Journal of Econometrics*, 186, 280-293

[7] Ding, J., Tarokh, V. and Yang, Y. 2018. "Bridging AIC and BIC: a new criterion for autoregression". *IEEE Transactions on Information Theory*, 64(6), 4024-4043

[8] Doudchenko, N., and Imbens, G. 2016. "Balancing, regression, difference-in-differences and synthetic control methods: a synthesis". Working Paper 22791, National Bureau of Economic Research, Cambridge, MA.

[9] Gobillon, L., and Thierry, M. 2016. "Regional policy evaluation: interactive fixed effects and synthetic controls". *The Review of Economics and Statistics*, 98(3), 535-551

[10] Hansen, B. 2007. "Least squares model averaging". *Econometrica*, 75, 1175-1189.

[11] Hansen, B. 2008. "Least squares forecast averaging". *Journal of Econometrics*, 146, 342-50.

[12] Hansen, B. 2010. "Multi-step forecast model selection". University of Wisconsin, Working paper.

[13] Hansen, B., and Racine, J. 2012. "Jackknife model averaging". *Journal of Econometrics*, 167, 38-46.

[14] Hansen, B. 2014. "Model averaging, asymptotic risk, and regressor groups". *Quantitative Economics*, 5, 495-530.

[15] Li, K., and Bell, D. 2017. "Estimation of average treatment effects with penal data: asymptotic theory and implementation". *Journal of Econometrics*, 197, 65-75.

[16] Medeiros, M., and Mendes, E. 2016. "ℓ_1-regularization of high-dimensional time-series models with non-Gaussian and heteroskedastic errors". *Journal of Econometrics*, 191(1), 255-271.

[17] Tibshirani, R. 1996. "Regression shrinkage and selection via the lasso". *Journal of the Royal Statistical Society Series B*, 58(1), 267-288.

[18] Xu, Y. 2017. "Generalized synthetic control method: causal inference with interactive fixed effects model". *Political Analysis*, 25(1), 57-76.

[19] Zhang, X., Wan, A.T.K, and Zou, G. 2013. "Model averaging by jackknife criterion in models with dependent data". *Journal of Econometrics*, 67(2), 301-320.

[20] Zou, H., and Hastie, T. 2005. "Regularization and variable selection via the elastic net". *Journal of the Royal Statistical Society, Series B*, 67(2), 301-320.

[21] Zou, H. 2006. “The adaptive LASSO and its oracle properties”. *Journal of the American Statistical Association*, 101, 1418-1429.

4. Evaluating the Economic Impact of Conflict: A counterfactual analysis of Ukraine

1. Introduction

It has been observed that violent conflict adversely affects a country or region's economic performance at the aggregate level. However, to evaluate the impact of a violent conflict is a challenging task. Ideally, one wishes to compare the economic outcome under conflict and the one in the absence of conflict. However, it is rarely the case that researchers can observe these data simultaneously in both states. Instead, researchers often rely on the construction of counterfactual outcome - the outcome that would have been observed in the absence of conflict - to carry out comparison. In a seminal work, Abadie and Gardeazabal (2003) introduced the synthetic control method (hereafter SCM) to measure the economic cost of the Basque conflict between Spain and terrorist organization ETA. A weighted average of potential control countries/regions is constructed to approximate the pre-treatment outcome and available characteristics of a country/region of interest. Then the impact of conflict is measured as difference between the actual and counterfactual outome in the post-treatment periods. A fundamental assumption of SCM and its variants is that the treatment must remain exogenous to the control group. Once the assumption is violated, the SCM will induce a biased treatment effect estimate. However, when the treated unit is a single entity and the treatment/shock is at aggregate level, which is often the case with comparative case study, imposing this assumption is not always realistic. For instance, when researchers study impact of a violent conflict on a country/region's economy, such assumption would imply that the conflict generates no externalities or spillovers to its neighbouring countries/regions or even further. In contrast, there are lots of evidence about the existence of externalities and in most cases negative spillovers from a violent conflict on proximate countries' economic performance (Ades and Chua, 1997; De Groot, 2010; Qureshi, 2013).

In this chapter, we study the economic impact of conflict on Ukraine's Real GDP, accounting for potential spillovers to other countries. In particular, we apply the modified synthetic control method with spillover effects (hereafter SCM-SP) (Cao and Dowd, 2019) to perform a comparative case study on Ukraine's economy. Within this framework, we can evaluate the

impact of the armed conflict that started in late 2013 using a group of countries that are geographically close to and share economic, cultural linkage with Ukraine, but were not directly involved in the conflict. The advantage of this approach is to estimate the impact of conflict in the presence of spillovers. It also sheds lights on the direction and size of potential spillover effects. A remaining limitation, however, is that the approach requires a spillover structure to be pre-specified. This is potentially a weaker assumption in the current setting because knowledge of geographic distance, social-political and economic relationships, and history should enable one to construct a spillover structure that is reasonable.

Applying the SCM-SP approach, we estimate that the conflict decreased Ukraine's real GDP by 29.7% from late-2013 to the end of 2015, that is, the estimated counterfactual of real GDP is 29.7% higher than the actual real GDP. In comparison, the results based on original SCM indicates that the conflict reduced Ukraine's real GDP by 16.3%. We also find the existence of spillovers in all nine post-treatment periods. The spillover effects seem to have significant impact on the treatment effect estimates. Moreover, the estimates of spillover effects for the countries selected by SCM-SP are negative, suggesting adverse impact of the conflict on these countries.

The rest of the chapter is organized as follows: in Section 2, we review the background of the violent conflict in Ukraine since late-2013. In Section 3 we introduce our empirical methodology. In Section 4 we assess the economic impact on Ukraine's real GDP and adverse spillover effects on control countries. Section 5 concludes.

2. Background

This section presents a brief review of the origin and development of the 2013-2015 Ukrainian conflict. The divisions within Ukraine can go back much further than the recent conflict. In fact, Ukraine has been torn between West and East, which is reflected in its cultural and linguistic divisions. From the mid-seventeenth through the nineteenth centuries, the western territories of Ukraine belonged to the Polish administration, Austrian and Austro-Hungarian

Empire. The majority of western part was back under Polish rule between two World Wars. Meanwhile, the eastern Ukraine were governed by the Russian Empire since the seventeenth century. Accordingly, there have been significant developments in Ukrainian language education in the west, while the education was restricted to the Russian language in the east (Bilaniuk and Melnyk, 2008). The divisions continue since the collapse of Soviet Union. For example, a 2006 survey found that 38% of the population reported speaking Ukrainian only, 30% only Russian and 31% reported using both, depending on the situation (Bilaniuk and Melnyk, 2008). On the other hand, most Ukrainians in the west see Ukraine as part of Europe. Those in the east are more pro-Russia and see the two countries are more historically linked. A conflict over such disagreement may have been inevitable. In the mean time, there has been a competition between the European Union (EU) and Russia for the future economic and foreign policy orientation of Ukraine.

The trigger of the conflict is the fact that the Ukraine's former president Yanokovch refused to sign the Ukraine European Union Association Agreement on November 21, 2013. The agreement committed Ukraine to economic, judicial and financial reforms so that Ukraine could converge its policies and legislation to those of EU, which has been long opposed by Russia. Figure 1 shows several stages in the development of the conflict.

The rejection of the deal for greater integration with EU soon sparked mass protests called *Euromaidan* protests, which the Ukrainian government attempted to put down. However, the protests escalated into a series of riots and in the February 2014, the anti-government protests toppled the government and the former president Yanokovch fled to Russia. In the following month, Russia sent military force into Crimea, an autonomous region of southern Ukraine with strong Russian connections. Russia completed its annexation of Crimea in a referendum not recognized by Ukraine and most of the world. In April, pro-Russia separatist rebels began seizing territories in eastern Ukraine. In Donbass, the conflict between the rebels and newly-formed Ukrainian government escalated into an open warfare. Then two self-declared states were formed: Donetsk and Luhansk People's Republics. On July 17, 2014, a Malaysia Airline flight was shot down in eastern Ukraine and 298 people on the plane were killed. Two

months later, the first cease-fire agreement , Minsk Protocol, was signed between Ukraine and rebels. However, the violation of the cease-fire on both sides were common and the fighting in eastern Ukraine intensified in the following months. In February 2015, the 2nd cease-fire agreement was signed, but minor violation of the cease-fire continued. In January 2016, the Deep and Comprehensive Free Trade Agreement (DCFTA) between EU and Ukraine took effect. This agreement meant both sides would mutually open their markets for goods and services based on enforceable trade rules and it had significant impact on Ukraine's trade. EU, as a single market, became Ukraine's most important trading partner, accounting for more than 40% of Ukraine's total trade in 2016[1]. Meanwhile, Russia lost its position as the dominant trade partner of Ukraine. In 2013, it accounted for 30.19% of Ukraine's imports and 23.81% of Exports. In 2016, these numbers reduced to 13.12% and 9.88%, respectively[2].

In addition to the transformation of political power and change of trade relationship, the Ukrainian conflict resulted in a humanitarian crisis. From mid-April 2014 to November 2015, the United Nations Human Rights Monitoring Mission in Ukraine (HRMMU) recorded approximately thirty thousand casualties, including nine thousand killed and twenty-one thousand injured. HRMMU had also registered one and half million internally displaced individuals throughout the Ukraine.

Last, it is worth noting that EU, US, and several other western countries have imposed economic sanctions on Russia in response to its invasion and occupation of Ukraine's Crimea region and parts of eastern Ukraine. For instance, EU's sanctions restricted Russian state access to Western loans, blocked exports of defense-related equipment to Russia, banned exports of the oil industry technology to Russia. The series of sanctions were pushing the Russian economy to the brink of recession. Russia reacted to economic sanctions by banning a wide range of imported western food. The sanctions and counter-sanctions from Russia would have large impact on the regional economy, which will be discussed in Section 4.

[1]Data source: European Commission
[2]Data source: World Bank WITS database

Figure 1: Development of Ukraine's conflict: 2013.11 - 2016.01

3. Methodology

3.1 The model and estimation

Let y_{it}^1 and y_{it}^0 denote unit i's outcome in time t with and without treatment. As we often do not observe simultaneously y_{it}^1 and y_{it}^0. the observed outcome is

$$y_{it} = d_{it}y_{it}^1 + (1 - d_{it})y_{it}^0 \tag{1}$$

where $d_{it} = 1$ if the ith unit is under treatment at time t and $d_{it} = 0$ otherwise. Let N and T denote the total number of units and the total number of pre-treatment periods, we first consider the case with one post-treatment period so that we observe a panel of $N \times (T + 1)$ of outcomes y_{it} and treatment assignments d_{it}. We assume that only one unit receives treatment at time $T + 1$. Without loss of generality, let the first unit receive a treatment, but all other units y_{jt}, $j = 2, \ldots, N$ do not experience any treatment at $T + 1$. Throughout this paper, N is

fixed and T can be sufficiently large.

Let $\Delta_i = y^1_{iT+1} - y^0_{iT+1}$, which can be the treatment effect or spillover effect depending on whether $i = 1$. One popular approach to estimate the treatment effect Δ_1 is SCM (Abadie et al. 2003, 2010). Let $\boldsymbol{x}_{\text{t,SCM}} = (y_{1t}, \ldots, y_{Nt})'$, then a synthetic control weight estimator is

$$\hat{\boldsymbol{\beta}}_{\Lambda_{\text{SCM}}} = \text{argmin}_{\boldsymbol{\beta} \in \Lambda_{\text{SCM}}} \sum_{t=1}^{T} (y_{1t} - \boldsymbol{x}'_{t,\text{SCM}} \boldsymbol{\beta})^2 \tag{2}$$

where $\Lambda_{\text{SCM}} = \{\boldsymbol{\beta} \in \mathbb{R}^N : \beta_1 = 0, \beta_j \geq 0 \text{ for } j = 2, \ldots, N, \sum_{j=1}^{N} \beta_j = 1\}$. The estimator for treatment effect Δ_1 is given by

$$\hat{\Delta}_1 = y_{1T+1} - \boldsymbol{x}'_{T+1,\text{SCM}} \hat{\boldsymbol{\beta}}_{\Lambda_{\text{SCM}}}$$

To account for potential spillovers, we follow Cao and Dowd (2019) and assume that some units are likely to experience spillover effects but not remaining units, while the sizes of spillover effects are allowed to vary across those affected units. Suppose J units experience spillovers. Then we can write a vector of treatment and spillover effects as a linear transformation of some unknown parameter $\boldsymbol{\gamma} \in \mathbb{R}^k : \boldsymbol{\Delta} = \boldsymbol{A}\boldsymbol{\gamma}$. Specifically,

$$\boldsymbol{A} = \begin{bmatrix} 1 & \mathbf{0}_{1 \times J} \\ \mathbf{0}_{J \times 1} & \boldsymbol{I}_J \\ \mathbf{0}_{(N-J-1) \times 1} & \mathbf{0}_{(N-J-1) \times J} \end{bmatrix}, \boldsymbol{\gamma} = \begin{bmatrix} \Delta_1 \\ \Delta_{k_1} \\ \vdots \\ \Delta_{k_J} \end{bmatrix} \tag{3}$$

where the matrix $\boldsymbol{A}$ specifies the structure of treatment and spillover, the vector $\boldsymbol{\gamma}$ includes magnitude of treatment effect and spillover effects on J units. Under this specification, $\boldsymbol{\Delta} = (\Delta_1, \Delta_{k_1}, \ldots, \Delta_{k_J}, 0, \ldots, 0)'$. In other words, at $T + 1$, the first unit receives treatment, the units indexed by $k_1, \ldots, k_J$ experience each own spillover effects and the remaining units are not exposed to either treatment effects or spillover effects.

We need to estimate an $N \times 1$ vector $\boldsymbol{\Delta}$. To this end, let $\boldsymbol{Y}_t = (y_{1t}, y_{2t}, \ldots, y_{Nt})'$, the

vector of all units' outcomes at time t. We can define unit i's synthetic control weights as following:

$$\begin{bmatrix} \hat{a}_i \\ \hat{\boldsymbol{b}}_i \end{bmatrix} = \text{argmin}_{\tilde{a}\in\mathbb{R},\tilde{\boldsymbol{b}}\in W^{(i)}} \sum_{t=1}^{T}(y_{it} - \tilde{a} - \boldsymbol{Y}_t\tilde{\boldsymbol{b}}')^2. \tag{4}$$

where $W^{(i)} = \{(w_1, \ldots, w_N)' \in \mathbb{R}_+^N : w_i = 0, \sum_{j=1}^N w_j = 1\}$. They are weights used for SCM-SP method, which include an unrestricted intercept. Let a_i = plim$\hat{a}_i$, $\boldsymbol{b}_i$ = plim$\hat{\boldsymbol{b}}_i$, it can be shown that a_i and $\boldsymbol{b}_i$ are well-defined under a factor model (Lemma 1 in Cao and Dowd, 2019). For unit i at time t, the specification error is given by $u_{it} = y_{it}^0 - (a_i + \boldsymbol{Y}_t^{0\prime}\boldsymbol{b}_i)$. Stacking u_{it} for all i's gives

$$\boldsymbol{u}_t = \boldsymbol{Y}_t^0 - (\boldsymbol{a} + \boldsymbol{B}\boldsymbol{Y}_t^0) \tag{5}$$

where $\boldsymbol{u}_t = (u_{1t}, \ldots, u_{Nt})'$, $\boldsymbol{a} = (a_1, \ldots, a_N)'$, $\boldsymbol{B} = (\boldsymbol{b}_1, \ldots, \boldsymbol{b}_N)'$. As in the post-treatment period $T+1$, $\boldsymbol{Y}_{T+1} = \boldsymbol{Y}_{T+1}^0 + \boldsymbol{\Delta}$, the equation (5) indicates that

$$\begin{aligned} \boldsymbol{u}_{T+1} &= (\boldsymbol{Y}_{T+1} - \boldsymbol{\Delta}) - (\boldsymbol{a} + \boldsymbol{B}(\boldsymbol{Y}_{T+1} - \boldsymbol{\Delta})) \\ &= (\boldsymbol{I} - \boldsymbol{B})(\boldsymbol{Y}_{T+1} - \boldsymbol{\Delta}) - \boldsymbol{a} \end{aligned} \tag{6}$$

By imposing $\boldsymbol{\Delta} = \boldsymbol{A}\boldsymbol{\gamma}$, we have

$$\boldsymbol{u}_{T+1} = (\boldsymbol{I} - \boldsymbol{B})(\boldsymbol{Y}_{T+1} - \boldsymbol{A}\boldsymbol{\gamma}) - \boldsymbol{a} \tag{7}$$

We use (7) to estimate $\boldsymbol{\gamma}$ and therefore $\boldsymbol{\Delta}$. In particular, we could estimate weights using (4) for each $i = 1, \ldots, N$. That is, we pretend each unit i to be treated and other units to be controls. The estimator for $\boldsymbol{a}$ is then given by $\hat{\boldsymbol{a}} = (\hat{a}_1, \ldots, \hat{a}_N)'$, the estimator for $\boldsymbol{B}$ is $\hat{\boldsymbol{B}} = (\hat{\boldsymbol{b}}_1, \ldots, \hat{\boldsymbol{b}}_N)'$. Let $\boldsymbol{M} = (\boldsymbol{I} - \boldsymbol{B})'(\boldsymbol{I} - \boldsymbol{B})$ whose estimator is $\hat{\boldsymbol{M}} = (\boldsymbol{I} - \hat{\boldsymbol{B}})'(\boldsymbol{I} - \hat{\boldsymbol{B}})$,

then

$$\begin{aligned}\hat{\boldsymbol{\gamma}} &= \text{argmin}_{\boldsymbol{h}\in\mathbb{R}^k}||(\boldsymbol{I}-\hat{\boldsymbol{B}})(\boldsymbol{Y}_{T+1}-\boldsymbol{A}\boldsymbol{h})-\hat{\boldsymbol{a}}|| \\ &= (\boldsymbol{A}'\hat{\boldsymbol{M}}\boldsymbol{A})^{-1}\boldsymbol{A}'(\boldsymbol{I}-\hat{\boldsymbol{B}})'((\boldsymbol{I}-\hat{\boldsymbol{B}})\boldsymbol{Y}_{T+1}-\hat{\boldsymbol{a}}) \end{aligned} \tag{8}$$

and the parameter of interest $\boldsymbol{\Delta}$ can be estimated by $\hat{\boldsymbol{\Delta}} = \boldsymbol{A}\hat{\boldsymbol{\gamma}}$. Under a factor model, it can also be shown that as $T \to \infty$,

$$\hat{\boldsymbol{\Delta}} - (\boldsymbol{\Delta} + \boldsymbol{G}\boldsymbol{u}_{T+1}) \xrightarrow{p} \boldsymbol{0} \tag{9}$$

, where $\boldsymbol{G} = \boldsymbol{A}(\boldsymbol{A}\boldsymbol{M}\boldsymbol{A})^{-1}\boldsymbol{A}'(\boldsymbol{I}-\boldsymbol{B})'$ and $\mathbb{E}(\boldsymbol{G}\boldsymbol{u}_{T+1}) = \boldsymbol{0}$. That is, $\hat{\boldsymbol{\Delta}}$ is asymptotically unbiased for $\boldsymbol{\Delta}$.

In the case with multiple post-treatment periods, let $\boldsymbol{Y}_t = \boldsymbol{Y}_t^0$ if $t \leq T$, $\boldsymbol{Y}_t = \boldsymbol{Y}_t^0 + \boldsymbol{\Delta}_t$ for $t \geq T+1$. For each post-treatment period, we can similarly specify a treatment and spillover structure $\boldsymbol{A}_t$. Therefore, an estimator for $\boldsymbol{\Delta}_t$ is defined as $\hat{\boldsymbol{\Delta}}_t = \boldsymbol{A}_t(\boldsymbol{A}_t'\hat{\boldsymbol{M}}\boldsymbol{A}_t)^{-1}\boldsymbol{A}_t'$ $(\boldsymbol{I}-\hat{\boldsymbol{B}})'\big((\boldsymbol{I}-\hat{\boldsymbol{B}})\boldsymbol{Y}_t - \hat{\boldsymbol{a}}\big)$ for $t \geq T+1$.

3.2 Statistical inference

The SCM-SP testing procedure is based on Andrews' end-of-sample instability test (Andrews, 2003) which is introduced to assess the instability at the end of a single series. At a high level, the test uses pre-treatment sample to form the null distribution of post-treatment quantity, which is the discrepancy between treated group (those under direct treatment and indirect spillovers) and synthetic control group in the present case. There will be instability if the treatment effect and/or spillover effects exist. Similar to the previous section, we first consider the case with one post-treatment period. A general test regarding the vector of treatment and spillover effects $\boldsymbol{\Delta}$ is H_0: $\boldsymbol{C}\boldsymbol{\Delta} = \boldsymbol{d}$ v.s. H_1: $\boldsymbol{C}\boldsymbol{\Delta} \neq \boldsymbol{d}$ where $\boldsymbol{C}$ and $\boldsymbol{d}$ are known. If we want to test the hypothesis that there is no treatment effect on the treated unit (e.g., the first unit in our case), we can set $\boldsymbol{C} = (1, 0, \ldots, 0) \in \mathbb{R}^{1\times N}$ and $d = 0$. If we want to test

if there are spillover effects or not, we then set $\boldsymbol{C} = (\boldsymbol{0}_{(N-1)\times 1}, \boldsymbol{I}_{N-1}) \in \mathbb{R}^{(N-1)\times N}$ and $\boldsymbol{d} = (0, \ldots, 0)' \in \mathbb{R}^{(N-1)\times 1}$.

We define the test statistic as

$$S = (\boldsymbol{C}\hat{\boldsymbol{\Delta}} - \boldsymbol{d})'(\boldsymbol{C}\hat{\boldsymbol{\Delta}} - \boldsymbol{d}) \tag{10}$$

From expression (9), S is asymptotically equivalent to $\boldsymbol{u}'_{T+1}\boldsymbol{G}'\boldsymbol{C}'\boldsymbol{C}\boldsymbol{G}\boldsymbol{u}_{T+1}$, where $\boldsymbol{G} = \boldsymbol{A}(\boldsymbol{A}'\boldsymbol{M}\boldsymbol{A})^{-1}\boldsymbol{A}'(\boldsymbol{I} - \boldsymbol{B})'$. Denote $\boldsymbol{x}_t = (1, \boldsymbol{Y}'_t)'$ and $\boldsymbol{\theta} = [\boldsymbol{a}\ \boldsymbol{B}] \in \mathbb{R}^{N\times(N+1)}$. The null distribution consists of $S(t) = \boldsymbol{u}'_t\boldsymbol{G}'\boldsymbol{C}'\boldsymbol{C}\boldsymbol{G}\boldsymbol{u}_t = (\boldsymbol{Y}_t - \boldsymbol{\theta}\boldsymbol{x}_t)'\boldsymbol{G}'\boldsymbol{C}'\boldsymbol{C}\boldsymbol{G}(\boldsymbol{Y}_t - \boldsymbol{\theta}\boldsymbol{x}_t)$ for $t = 1, \ldots, T$. The sample analogue is given by $\hat{S}(t) = (\boldsymbol{Y}_t - \hat{\boldsymbol{\theta}}\boldsymbol{x}_t)'\hat{\boldsymbol{G}}'\boldsymbol{C}'\boldsymbol{C}\hat{\boldsymbol{G}}(\boldsymbol{Y}_t - \hat{\boldsymbol{\theta}}\boldsymbol{x}_t)$ for each period t, where $\hat{\boldsymbol{\theta}}$ contains synthetic control weights obtained from Equation (4). Then we can calculate the p-value of the Andrews' test as

$$p = 1/T \sum_{t=1}^{T} \mathbb{1}_{\{\hat{S}(t)\geq S\}} \tag{11}$$

In the case with multiple post-treatment periods, we perform separate tests as the above procedure for each $t \geq T + 1$.

4. Evaluating the economic impact of the conflict

In this section, we assess the economic impact of Ukraine's conflict. First we apply both SCM and SCM-SP to evaluate its impact on Ukraine's real GDP. Then we examine the potential spillover effects. Last we do robustness checks of our main result.

4.1 Impact on real GDP

We use quarterly data from 2001 Q1 to study the impact of the conflict on Ukraine's real GDP. As Russia has been directly involved and impacted by the conflict, we do not include it in the control countries. Considering Ukraine is located in Eastern Europe and a member state of former Soviet Union, we include Poland, Slovak Republic, Hungary, Romania, Estonia,

Latvia, Lithuania, Bulgaria, Slovenia, Czech Republic, Turkey[3]. Aware of growing political influence and closer economic relationship with EU, we include Finland, France,Germany, Italy, Netherland, Sweden and Switzerland. We also include US, UK and China. The real GDP are from World Bank's Global Economic Monitor database.

Our in-sample period ends at 2013 Q3 because the conflict started in November 2013. As discussed in Section 2, the DCFTA took effect in the beginning of 2016 and had large impact on Ukraine's foreign trade. It is not easy to disentangle the impact of the conflict from the trade agreement without strong assumptions. Therefore, we restrict out-of-sample evaluation period to 2015 Q4 to obtain more clear effect of the conflict.

The countries that are likely to experience spillovers are selected by investigating the geometric proximity, cultural/ historical linkage and trade relationship with Ukraine. In particular, we include Ukraine's four contiguous countries: Poland, Slovak Republic, Hungary, Romania and countries that are a country away from Ukraine: Czech Republic, Slovenia and Bulgaria. We include three Baltic states that are member states of former Soviet Union: Estonia, Latvia and Lithuania. Finally, we include Ukraine's top trading countries in the region: Germany and Turkey[4].

We use in-sample data to estimate control countries' weights via equation (2) and (4). The estimated weights of the control countries selected by SCM and SCM-SP for logRGDP are shown in Table 1. The SCM selects Estonia, Latvia, Poland, Slovak Republic and UK. The SCM-SP selects Bulgaria, Estonia, Hungary, Lithuania and Slovenia with intercept equal to 1.3054. In order to interpret the results more clearly, we convert the treatment effect estimates back to percentage change. The estimated average treatment effects on Ukraine's real GDP is -16.4% based on SCM, which is the average difference between the actual Real GDP and the counterfactual estimates from 2013 Q4 to 2015 Q4. When the spillover effects are allowed, the average treatment effect is -29.7% during the same time periods. In other words, the conflict decreased the real GDP of Ukraine by 29.7% since late-2013.

[3]Two neighbouring countries Belarus and Moldova are not included because their data are unavailable.

[4]Based on the WITS data from World Bank, Germany had the 2nd and 3rd highest partner share in Ukraine's overall exports in 2014 and 2015 while Turkey had 2nd highest partner share in overall imports of Ukraine in both 2014 and 2015.

Figure 2. plots the actual and counterfactual logRGDP paths for the pre-conflict and post-conflict time periods. The left panel shows that the counterfactual path estimated by SCM (dotted line) and SCM-SP (dashed line) are similar to each other and both trace closely the actual paths. The right panel shows that both counterfactual paths lie above the actual path in all post-conflict periods. Moreover, the counterfactual path estimated by SCM-SP is higher than the estimate by SCM in all post-conflict periods.

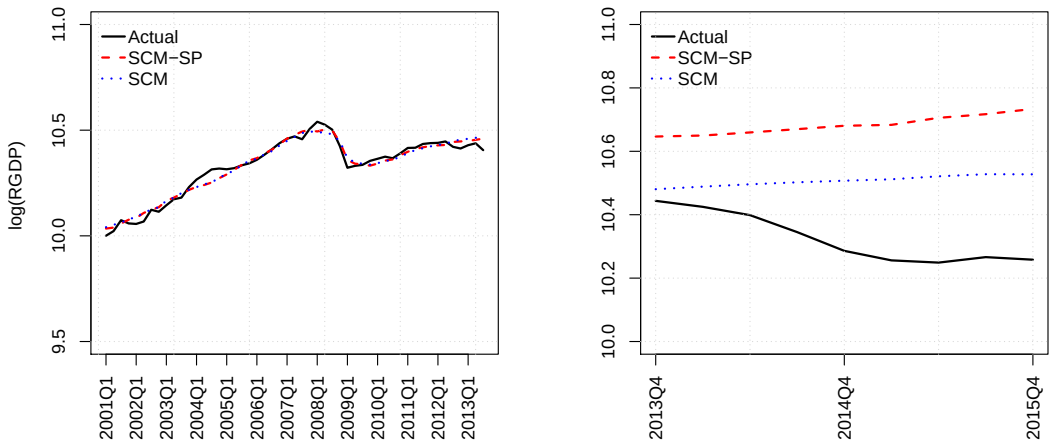

Figure 2: Ukraine's actual and counterfactual log(RGDP)

4.2 Spillovers

The inference procedure discussed in Section 3.2 allows us to test the existence of spillover effects. We use the spillover structure specified in Section 4.1. The test result shows that there were spillovers in every quarter after the conflict broke out. Figure 3 shows the treatment effect (as gaps in logRGDP between actual and counterfactual path). The left panel shows that the treatment effects are close to zero based on both SCM and SCM-SP, which is expected as there should be no effect from conflict before 2013 Q4. In the right panel, the error bars around the SCM-SP estimates are 95% confidence intervals. We can see that there were substantial spillover effects, which in all post-conflict periods result in significant changes in treatment effect estimates compared to SCM. In particular, in the presence of spillovers, the treatment effects estimates based on SCM are attenuated. The result is also supported by the

estimated spillover effects. First note that the five countries selected by SCM-SP: Bulgaria, Estonia, Hungary, Lithuania and Slovenia are included in the specification of spillover structure. Table 2. shows their spillover effects estimates and weights estimates. It is not surprising that all five countries experienced adverse spillover effects on their real GDP. In terms of possible channel of the negative spillover effects, reasoning that Russia exerted great influence in the region's economic growth and sanctions imposed by western countries in response to its involvement in the conflict would disrupt the trade flows from/to Russia, the spillover effects may arise from such disruptions. To explore this hypothesis, we examine World Bank's the world integrated trade solution (WITS) dataset which includes bilateral trading share. We find that in 2014 and 2015, Russia was among top 3 trade partners in terms of export share and/or import share for the four of five countries selected by SCM-SP: Bulgaria, Estonia, Hungary and Lithuania. For Estonia and Lithuania, the two countries that were hit hard by negative spillovers, Russia was the most important trade partners. For example, in 2014, Russia had highest import share (10.7%) and second highest export share (14.16%) for Estonia; while it ranked the first in both import (21.64%) and export share (20.85%) for Lithuania. In 2015, Russia ranked second and third among Estonia's top trade partners, with 9.81% of import share and 9.68% of export share; Russia still dominated Lithuania's trade, though with reduced import share (16.31%) and export share (13.69%).

4.3 Robustness check

In this section we conduct two robustness checks. In the first robustness check, we consider a smaller control group than the one used in the Section 4.1. Table 1 shows that Latvia and Lithuania receive substantial weight in the SCM-SP, we exclude all three Baltic countries: Estonia, Latvia and Lithuania from the control group to see whether our result is sensitive to their inclusion/exclusion. The specified spillover structure is the same as in the Section 4.1 except the three Baltic countries. In the second robustness check, we consider a spillover structure that includes more countries than the one specified in Section 4.1 to see whether our

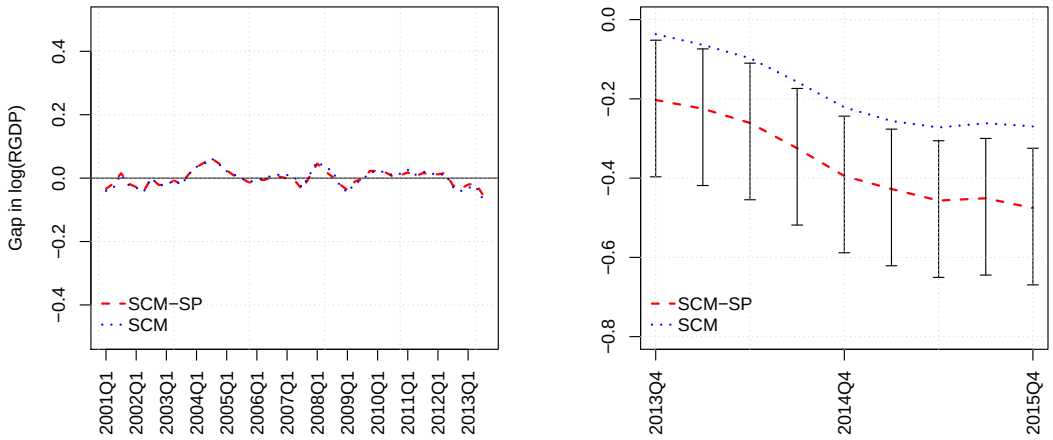

Figure 3: Gap in log(RGDP) between Ukraine and counterfactual Ukraine (with 95% confidence interval)

result is sensitive to the specification of spillover structure. Specifically, we add five more EU countries: France, Italy, Netherlands, Sweden and Switzerland so that seventeen countries are specified to have potential spillovers from the Ukraine's conflict.

The results are reported in Figure 4. The left panel shows the treatment effects estimates from SCM-SP on nineteen control countries, among which nine countries are specified to have spillovers (SCM-SP (9)). The new treatment effect estimates are also negative and statistically different from zero, though not statistically different from our main result (SCM-SP). The right panel displays the treatment effect estimates from SCM-SP with seventeen countries affected by spillovers (SCM-SP(17)). The estimate is similar to the estimate obtained in Section 4.1. Table 3. reports the weights of countries selected by SCM-SP excluding Estonia, Latvia and Lithuania[5]. It is worth noting that the four selected countries - Turkey, Slovenia, Romania and Hungary are included in the specified spillover structure.

[5]As we use the same control group as in Section 4.1 in the second robustness check, the countries selected by SCM-SP are the same as shown in Table 1.

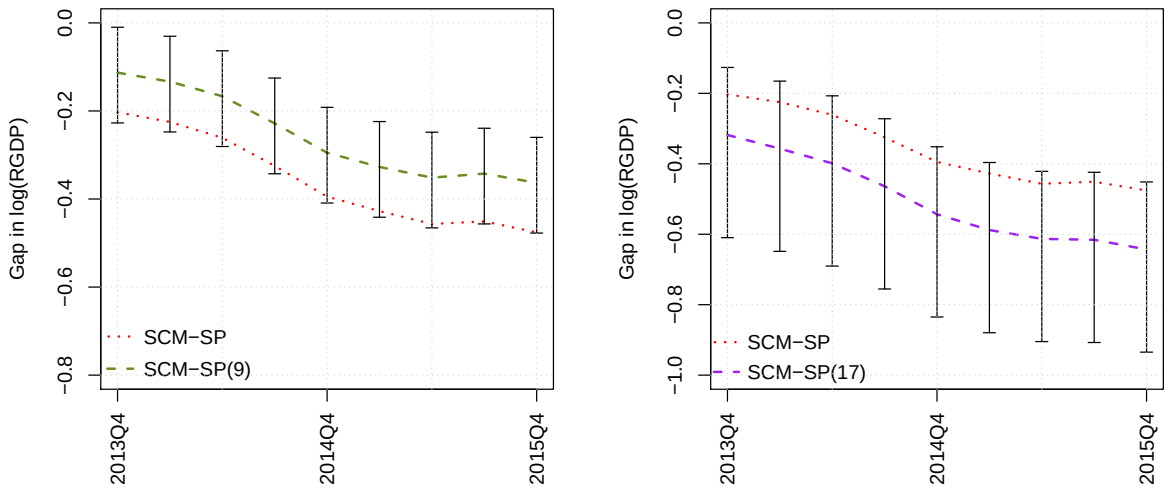

Figure 4: Gaps in log(RGDP) with 95% confidence interval

5. Conclusion

In this chapter, we study the economic impact of the conflict on Ukraine's real GDP since late-2013. Considering the interdependence between the economies across the region and the magnitude of the conflict, other countries within the same region of Ukraine are susceptible to the external spillover effects from the conflict. In order to accommodate this situation, we apply the synthetic control method that is adapted to the potential spillovers. We find that the conflict reduced the Ukraine's real GDP by -29.7% during 2014 and 2015, a much larger estimate of negative effects than the estimate based on the original synthetic control method (-16.4%). According to the test results regarding the existence of spillovers, we detect the spillovers in every quarter since the conflict broke out. We also find negative spillovers on all the countries with positive weights. Our empirical results are robust to different specified control group and spillover structure.

Given the magnitude of negative impact from the conflict and evolving humanitarian crisis in such a short time, the international communities who would like a resolution of Ukraine's conflict should take actions faster so that the economic future remains for Ukraine. The huge estimated economic cost also requires those communities to take a tougher stand and actions.

On the other hand, the conflict is costly for all parties, not only Ukraine and Russia, but also those countries in the same region. Picking sides between western countries and Russia is difficult and finding that balance between the two sides is even more challenging. As the Ukraine's conflict is still under way, those countries should remain vigilant and actively participate in resolving the conflict to mitigate the long-run negative impact on the whole region.

Table 1: Weights of control units for log(RGDP)

Country	SCM weights	SCM-SP weights
(Intercept)		1.3054
Bulgaria		.0056
China		
Czech Republic		
Estonia	.1531	.3257
Finland		
France		
Germany		
Hungary		.0468
Italy		
Latvia	.4946	
Lithuania		.4835
Netherland		
Poland	.1248	
Romania		
Slovak Republic	.0037	
Slovenia		.1385
Sweden		
Switzerland		
Turkey		
UK	.2239	
USA		

Table 2: Spillover effects estimates in log(RGDP) for selected control countries

Country	spillover effects	SCM-SP weights
Bulgaria	-.0825	.0056
Estonia	-.2655	.3257
Hungary	-.006	.0468
Lithuania	-.1910	.4835
Slovenia	-.0662	.1385

Table 3: Weights of control units for log(RGDP) (robustness check)

Country	SCM-SP weights
(Intercept)	-.3335
Bulgaria	
China	
Czech Republic	
Finland	
France	
Germany	
Hungary	.3334
Italy	
Netherland	
Poland	
Romania	.4201
Slovak Republic	
Slovenia	.0842
Sweden	
Switzerland	
Turkey	.1623
UK	
USA	

References

[1] Abadie, A., and Gardeazabal, J. 2003. "The Economic costs of conflict: A case study of the Basque Country". *American Economic Review*, 93, 113-132.

[2] Abadie, A., Diamond, A., and Hainmueller, J. 2010. "Synthetic control methods for comparative case studies: estimating the effect of California's Tobacco Control Program". *Journal of the American Statistical Association*, 105, 493-505.

[3] Ades, A., and Chua, H. 1997. "Thy Neighbor's Curse: Regional Instability and Economic Growth". *Journal of Economic Growth*, 2, 279-304.

[4] Andrews, D.W.K. 2003. "End-of-Sample Instability Tests". *Econometrica*, 71, 1661-1694.

[5] Bilaniuk, L., and Melnyk, S. 2008. "A Tense and Shifting Balance: Bilingualism and Education in Ukraine". *International Journal of Bilingual Education and Bilingualism*, 11, 340-372.

[6] Cao, J. and Dowd, C. 2019. "Estimation and Inference for Synthetic Control Methods with Spillover effects". *arXiv preprint arXiv:1902.07343.*

[7] De Groot, O. 2010. "The Spillover Effects of Conflict on Economic Growth in Neighbouring Countries in Africa". *Defense and Peace Economics*, 21, 149-164.

[8] Qureshi, M. 2013. "Trade and thy neighbor's war". *Journal of Development Economics*. 105, 178-195.

5. Conclusions

In this dissertation, we first propose using the model averaging (MA) method to estimate treated counterfactual outcome and temporal average treatment effect (ATE) in a panel data setting. We derive the asymptotic distribution of proposed MA-based ATE estimator when N is fixed and T is large. We leave the asymptotic analysis of the proposed estimator for both large N and T case in future work. The derived asymptotic distribution turns out to be non-normal and non-standard, and cannot be directly used for inference. We instead apply a subsampling-bootstrap inference procedure for the MA-based ATE estimator. Monte Carlo simulations show that the proposed inference procedure results in reasonably good estimated coverage probabilities. In the empirical application, we revisit Hsiao et al. (2012)'s evaluation of impact of Hong Kong's reunion with mainland China, using our proposed estimator and inference procedure. Our results suggest that the political integration did not have significant impact on Hong Kong' s real GDP growth whereas the economic integration had significant positive impact on Hong Kong's economy, which support original findings.

Secondly, we conduct extensive simulations to compare the predictability of counterfactual outcome between model averaging and other methods in terms of mean squared prediction errors. The results show that the model averaging methods compare favourably with alternative methods, such as AIC, BIC, synthetic control methods and the penalized regression methods Elastic-net.

In the third part, we examine the economic impact of Ukraine's 2013 conflict, using a modified synthetic control method that accounts for potential spillover effects proposed by Cao and Dowd (2019). Using the data of 21 countries consisting of Ukraine's neighbouring countries and trade partners, we find that the conflict decreased Ukraine's real GDP by 29.7% from late-2013 to the end of 2015, which is larger (in absolute value) than the estimate of original synthetic control (- 16.3%). We also conduct statistical test on the existence of spillovers and find evidence of spillovers in all 9 post-conflict periods.

VITA

NAME OF AUTHOR: Guanyu Liu

PLACE OF BIRTH: Changsha, China

DATE OF BIRTH: October, 1989

EDUCATION:

2012 B.A. Beijing Foreign Studies University, Beijing, China

2015 M.A. Syracuse University, Syracuse, U.S.

PROFESSIONAL EXPERIENCE:

Teaching assistant, Department of Economics, Syracuse University, 2016 - 2021

AWARDS AND HONORS:

2016 - 2021 Syracuse University Graudate Assistantship, Syracuse University

2020 Eggers Fund Graduate Fellowship, Syracuse University

2016 - 2020 Maxwell School Summer Fellowship, Syracuse University