ECONOMETRICS MODEL SELECTION: THEORY AND APPLICATIONS

A Dissertation

by

WEI LONG

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Qi Li
Committee Members,	Li Gan
	Dennis W. Jansen
	Ximing Wu
Head of Department,	Timothy J. Gronberg

August 2015

Major Subject: Economics

ABSTRACT

This dissertation contains two essays which examine the theory of model selection in econometrics and its applications. In the first essay, we utilize a model average approach to estimate a mixture copula. We average over the estimates of each individual copula and their composite and select their associated weights by minimizing a leave-one-group-out cross-validation criterion. We are able to prove that our model average estimator is asymptotically optimal in the sense of achieving the infeasible lowest possible squared estimation losses. Simulation results prove that our model average estimators for mixture copula exhibit smaller estimation loss than some benchmark methods. We empirically examine the dependence structures among the stock markets in U.S., United Kingdom, Japan and Hong Kong, and we show that our model average estimators give more reasonable estimations for the dependence structures among these markets.

In the second essay, we implement a panel data approach to estimate the treatment effect of the justice reform in Virginia in 1995. The fundamental idea behind this method is to exploit the dependence among cross-sectional units to construct the counterfactual analysis. This panel data method uses the outcomes of the control units to simulate the path of the treated unit during the pre-treatment period and then predict the counterfactual path of the treated unit during the post-treatment period. In order to find the control units which simulate the pre-treatment path of the treated unit best, model selection criterion such as Akaike Information Criterion (AIC) and corrected Akaike Information Criterion (AICC) are used. We confirm that both violent and property crime rates declined in Virginia after the justice reform.

ACKNOWLEDGEMENTS

This dissertation was written under the supervision of my advisor, Professor Qi Li. During my graduate studies in the Department of Economics in Texas A&M University, I have learned a lot from Professor Li in both theoretical and applied econometrics. I am deeply impressed by his master of mathematics and econometrics, his diligence and achievements in research, and his ability of critical thinking. Without his continuous guidance and encouragement, I could not achieve these research results. Besides supervising me in my academic research, I am also very grateful to Professor Li's generous help and guidance in my daily life. I will cherish his kind help and guidance to my research all my life.

I would also like to thank my committee members, Professor Li Gan, Professor Dennis Jansen and Professor Ximing Wu, for their useful comments and moral support. Their knowledge and valuable suggestions broaden my understanding of different aspects of my research. Thanks also go to all my friends, the department faculties and staffs for their continuous help along the way of the pursuit of my Ph.D degree.

In addition, I have a lot of discussions with my colleagues and would like to thank Xinyu Zhang, Hsin-Hung Yao, Daiqiang Zhang, Guannan Liu and Cheng Zhou in particular.

Finally, I am very grateful to my mother, father and wife for their persistent encouragement, love and moral support to me.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

TABLE Page

1. INTRODUCTION

Econometric methods for model selection have been extensively discussed in literature these days. To deal with the model uncertainty issue, estimation criteria such as the Akaike information criterion (AIC; Akaike, 1970), corrected Akaike information criterion (AICC; Hurvich and Tsai, 1989) and Schwarz-Bayes information criterion (SBC; Schwarz, 1978) have been introduced. Under these methods, for a set of candidate model, we calculate the estimation criteria for each candidate model and then select the one with the smallest value as the "optimal" one. However, different criterion favors different model and the model selection procedure highly relies on users' subjective judgement and experiences. For example, SBC favors more parsimonious models while AIC favors more heavily parameterized models.

Model averaging method is different from the previous methods as it handles the model uncertainty issue by averaging a set of candidate models in a particular manner rather than selecting an "optimal" one from the set. Hoeting et al. (1999) propose a Bayesian model averaging method. Buckland et al. (1997), Hansen (2007) and Wan et al. (2010) for frequentist method for model averaging. Under model average framework, we select weights for each candidate model by minimizing a cross-validation criterion function. For example, Hansen and Racine (2012) propose a Jackknife model averaging method (JMA) and show that the whole computing procedure is an application of the standard quadratic programming technique. They further prove that JMA estimator is asymptotically optimal in the sense of achieving the lowest possible expected squared estimation loss.

In this dissertation, I employ both the model selection criterion method and the model averaging method to deal with the model uncertainty issue. In the first essay, I implement the Jackknife model averaging method to select the weight for each individual copula in a mixture copula model. Weights are selected by minimizing a leave-one-group-out criterion function and the standard quadratic programming technique is employed. Simulation results show that our model averaging estimators exhibit smaller estimation losses than the benchmark method especially the mixture copula model is misspecified.

The second essay focuses on the model selection criterion method. We implement the panel data approach proposed by Hsiao et al. (2012) to estimate the treatment effect of the justice reform in Virginia in 1995. Specifically, we examine whether the tougher punishment on violent criminals has successfully deterred the violent crime rate in Virginia after 1995. Hsiao et al. (2012) use the outcome results in the control units to simulate the pre-treatment path of the treated unit and propose a two-step strategy the select the control units which could best simulate the pre-treatment path of the treated unit. Model selection criteria such as AIC and AICC are used to find out those most appropriate control units. By implementing Hsiao et al. (2012) method, we find that the violent crime rate drops immediately after the justice reform declines by 16% on average between 1995 and 2010. A series of robustness tests further confirm that the treatment effect is not driven by statistical coincidence. Non-violent property crime rate, even though declines four years after the reform, also decreases on average during the 1995 - 2010 period.

2. DETECTING FINANCIAL DATA DEPENDENT STRUCTURE BY AVERAGING MIXTURE COPULA

In this article we propose to use a model average approach to estimate mixture copula models, which is a linear combination of multiple individual copulas. Nelsen (1999) provides a thorough introduction about copula and he defines copula as "functions that join or couple multivariate distribution functions to their one-dimensional marginal distribution function" (Nelsen, 1999, page 1). Specifically, let $\boldsymbol{X} = (X_1, ..., X_n)^{\mathrm{T}}$ be a random vector and the respective marginal cumulative distribution functions are defined as F_i, where $i \in \{1, ..., n\}$. Then there exists a copula $C : [0, 1]^n \to [0, 1]$ such that $\forall\ \boldsymbol{x} = (x_1, ..., x_n) \in \mathbb{R}^n$, $F(\boldsymbol{x}) = C\{F_1(x_1), ..., F_n(x_n)\}$ (see Sklar, 1959). Therefore, copula is flexible as it does not constrain the selection of marginal distributions so that one could always couple various margins together via a copula.

Copula model is primarily used to study the dependence patterns among variables, e.g., the co-movements among the international equity markets. The empirical application of copula in finance started from Li (2000), who proposes to use copula to estimate default correlations. After that, copula models have been extensively applied in many empirical studies such as examining the difference in dependence structure between the developed and developing economies (Chollete, Peña and Lu, 2005, Chollete et al., 2009, and Aloui et al., 2010), the structure break in exchange rates (Patton, 2006), financial contagion (Rodriguez, 2007), and the cross-state housing prices during the subprime mortgage crisis (Zimmer, 2012).

One fundamental issue for empirical studies is how to select an appropriate copula to satisfactorily describe the dependence structure among variables. Almost all the

the works mentioned above presumptively build a candidate set and rely on certain statistical criterion to pin down one copula, and then estimate the parameter(s) associated with the selection to evaluate the degree of dependence. For example, Zimmer (2012) uses both Bayesian Information Criterion (BIC) and Vuong test to show that a Clayton-Gumbel mixture copula provides a better estimate of dependence compared with Gaussian, Clayton and Gumbel.[1]

In practice, the varieties of copula are large. Ideally, one might fit the data under analysis to each existing copula family and pin down the most appropriate one. But this strategy could be less efficient, especially considering that one can always create a new copula by making certain transformations on an existed copula.[2] Thus, most empirical users only consider several commonly used copulas, e.g., Gaussian, Clayton and Gumbel, to construct their candidate set. The argument is that the candidate set should be general enough to capture most of the possible dependence patterns in the real world. Even though this strategy becomes relatively easy to implement, its cost is that one needs to assume that the observations are generated from one of the copula included in the candidate set: One pins down the most "appropriate" copula from the candidate set, estimates its parameter via maximum likelihood method, and then describes the dependence pattern and evaluates the degree of dependence. If one's candidate set does include the true copula that generates the observations, the estimating procedures discussed above should be effective and efficient. However, the true data generating copula model is always unknown to econometricians and in practice, it is highly probable that one's candidate set fails to include the true copula or all the candidates are far from the true dependence structure. Under this

[1] Other methods include comparing which copula gives the largest log-likelihood function value. Interested readers are referred to Manner and Reznikova (2011), Patton (2012) and Fan and Patton (2014) for details.

[2] For example, Patton (2006) introduce an symmetrized Joe-Clayton copula by taking a particular Laplace transformation on the BB7 copula of Joe (1997).

circumstance, the candidate copula that exhibits smallestr BIC may fail to provide reasonable description about the true dependence structure.

To take advantage of different copula shapes, Chollete, Peña and Lu (2005) and Hu (2006) introduce mixture copula models. In their analysis, a mixture copula is formulated as a weighted average of several individual copulas with the weights constrained between 0 and 1 and the weights sum to 1. Comparing with the individual copula model, mixture copula is more flexible as it nests various individual copulas that exhibit quite different dependence structures. As we shall see, a mixture copula is able to generate dependence structures that do not belong to any existing copulas. By combining several widely used individual copulas, one can build a parsimonious but flexible mixture copula to capture various dependence patterns in the financial data, e.g., zero and non-zero tail dependence, symmetric and asymmetric tail dependence. In their analysis, Chollete et al. (2005) and Hu (2006) both consider the mixture model including Gaussian, Gumbel and rotated Gumbel copula to evaluate the dependence structures among stock indexes in developed economies. They find strong left tail dependence as weight associated with the rotated Gumbel copula tends to be non-zero, while Gumbel tends to be filtered out due to its small weight. They thus conclude that stock markets in developed economies tend to go down simultaneously. This finding is consistent with Longin and Solnik (2001) who find that equity returns tend to take on joint negative extremes. In a more recent work, Cai and Wang (2014) introduce a penalized likelihood with a shrinkage operator method to estimate weighting and copula parameters simultaneously. This data-driven method is similar to the Least Absolute Shrinkage and Selection Operator (LASSO) due to Tibshirani (1996) and the Smoothly Clipped Absolute Deviation (SCAD) due to Fan and Li (2001) for variable selection in regression models. Cai and Wang (2014) further establish the asymptotic theory for their method, and simulation results demonstrate that their proposed method gives satisfactory estimations on both weighting and copula parameters. That is, different dependence structures are captured well by this penalized likelihood method.

In this article we contribute to the literature by providing another method to estimate mixture copula model. Specifically, we utilize a model average approach to estimate a mixture model which could be flexibly constructed by any existing copulas. Rather than pinning down one appropriate copula model by comparing different criteria such as AIC or BIC, under mixture copula framework, one firstly fits observations to each individual copula in the candidate set respectively. Sometimes fitting the data to an individual copula could be quite poor, so we also consider to fit the data to the composite of all the candidate copulas. We then average over the estimates of each individual copula and their composite and select their associated weights by minimizing a leave-one-group-out cross-validation criterion, a manner similar to the Jackknife model average (JMA) proposed by Hansen and Racine (2012). We obtain the solutions through a standard application of quadratic programming technique, as the leave-one-group-out cross-validation criterion is a quadratic function of weights. Under certain regularity conditions, we are able to prove that our model average estimator is asymptotically optimal in the sense of achieving the infeasible lowest possible squared estimation losses. The chosen weights help us to construct the optimal combination of each candidate copula and their composite that is able to satisfactorily describe the dependence structure among variables, as the distance between the estimated mixture copula and the unknown true model is asymptotically minimized. This is extremely important when one's working model is misspecified, i.e., when observations are generated from copulas that are not included in our working mixture model. Cai and Wang (2014) argue that when the working model is misspecified, their method will select copulas exhibiting the same dependence patterns. For example, when observations are generated from a combination of Gaussian and Clayton while in the working model Clayton is absent but a rotated Gumbel, which also exhibits the left tail dependence, is included. In this case, Cai and Wang (2014)'s method will assign certain weight on the rotated Gumbel to guarantee that the left tail dependence patterns is captured by the mixture copula model, and then they conclude that the "best" copula is chosen. However, including another copula which exhibits similar tail dependence structure to the true one does not necessarily guarantee that the distance between the estimated model and the true model is asymptotically minimized. For example, even though both Clayton and rotated Gumbel exhibit left tail dependence, the mathematical forms of their respective CDFs/PDFs are quite different. Thus, our model average method provides a more solid criterion for the *best* copula when working model is misspecified: The optimal mixture model is constructed to minimize its distance to the true model, or the estimation loss, so that it best describes the dependence pattern among variables. Considering that our working mixture model is usually quite parsimonious, misspecification problem should be common. Therefore, our model average method should be viewed as a more reasonable alternative to estimate mixture copula model so that the estimation loss is asymptotically minimized.

In the empirical part of the paper, we implement the model average approach on the daily returns of equity indexes in four developed economies (UK, Hong Kong, Japan and United States). Estimation results support the superiority of model average method in capturing the dependence structures among the international equity markets. Last but not least, model average approach exhibits the smallest errors in the out-sample predictions, comparing with the penalized likelihood method and the standard copula selection method which chooses one most "appropriate" copula by BIC. Since the financial press has rigorous requirement for models in prediction, the model average approach on mixture copula should be an useful tool in risk management.

The rest of this paper is organized as follows. In Section 2 we briefly introduce mixture copula model. Section 3 specifies steps and procedures about how to implement model average approach on a mixture copula model. In Section 4, we compare estimation losses under model average approach, Cai and Wang's (2014) penalized likelihood method, and the BIC method through Monte Carlo simulations. A real data example is presented in Section 5. We give some concluding remarks in Section 6. Regularity conditions and proof of optimality are included in the Appendix.

2.1 Mixture Copula Model: A Brief Introduction

Suppose we have a series of independent $p-$dimensional vectors of random variables $\{\mathbf{X}_t\}_{t=1}^T$, where $\mathbf{X}_t = (X_{t1}, \ldots, X_{tp})^{\mathrm{T}}$. Denote $F(\mathbf{x})$ and $f(\mathbf{x})$ to be the joint distribution and density function of $\mathbf{X} \in \mathcal{R}^p$, $F_i(x_i)$ and $f_i(x_i)$ be the marginal distribution and density function of X_i, respectively, where $1 \leq i \leq p$.

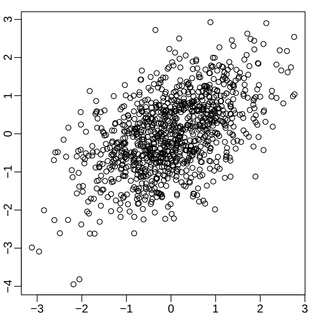
According to Hu (2006) and Cai and Wang (2014), a mixture copula model is a linear mixture of some copula families. Specifically, a mixture copula model can be written as

$$C(\mathbf{u}; \boldsymbol{\theta}, \boldsymbol{\omega}) = \sum_{l=1}^{L} \omega_l C_l(\mathbf{u}; \theta_l) = \sum_{l=1}^{L} \omega_l C_l(F_1(x_1; \alpha_1), ..., F_p(x_p; \alpha_p); \theta_l) \tag{2.1}$$

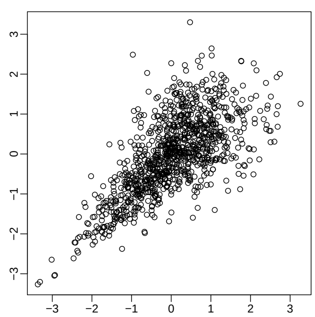
where $\{C_1(\cdot), ..., C_L(\cdot)\}$ is a set of candidate copulas with a vector of unknown associated parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_L)^{\mathrm{T}}$ and p-dimension marginal distribution $\boldsymbol{u} = (F_1(\cdot), \ldots, F_p(\cdot))$, $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_L)^{\mathrm{T}}$ denote the *weight parameters* with $0 \leq \omega_l \leq 1$ and $\sum_{l=1}^{L} \omega_l = 1$, and $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_p)^{\mathrm{T}}$ is the vector of parameters associated with each of the marginal distribution. In equation (1), $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_L)^{\mathrm{T}}$ control the dependence among the p-dimension variables and $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_L)^{\mathrm{T}}$ control the shape of the mixture copula's dependence.

One may want to include many existing individual copulas into the mixture model to cover every possible dependent pattern. But in application this would make the mixture model to be too complicated and the estimations to be less efficient, leaving mixture model less desirable. In practice, we only consider a few candidate individual copulas. We present the flexibility of mixture copula through scatter plots. Figure 2.1 (a) - (c) display scatter plots of 1000 i.i.d. samples generated from three types of widely-used copula model. For comparing purposes, each margin has the standard normal distribution and the parameter for the corresponding copula is calibrated to imply of Kendall's τ of one-half. It could be observed that Clayton copula displays strong dependence in the left tail while the Gumbel copula exhibits strong right tail dependence. Unlike Clayton and Gumbel which exhibit asymmetric dependence structure, Gaussian copula looks to be more symmetric and the stronger dependence appears in the center. Figure 2.1 (d) - (f) present scatter plots of 1000 i.i.d. samples generated respectively from three mixture copulas with equal weights on each component. Figure 2.1 clearly demonstrates that, after mixing with Clayton and Gumbel, Gaussian copula begins to exhibit some asymmetric tail dependence. Therefore, we can see that the flexibility of mixture copula stems from its ability of nesting various copula shapes. Each individual copula is nested as a special case.

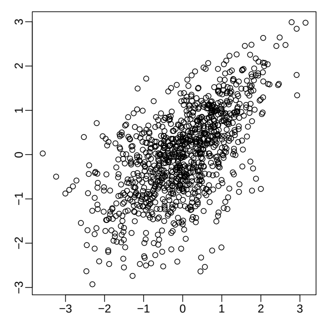
Fan and Patton (2014) summarize estimation methods for individual copula based models. Rather than simply estimating parameters for copulas and marginal densities, we need to further estimate a vector of weight parameter $\boldsymbol{\omega}$ introduced by mixture copula models. Chollete, Peña and Lu (2005) and Hu (2006) independently propose a two-stage semiparametric method in estimating a mixture copula model.



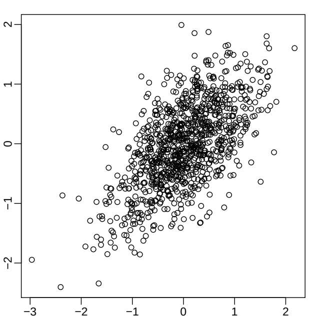
(a) Gaussian Copula



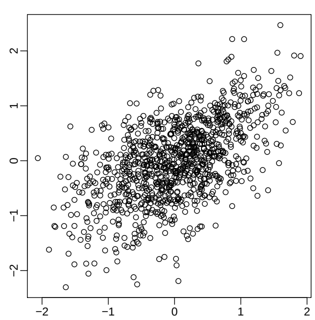
(b) Clayton Copula



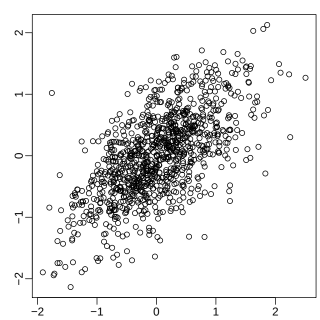
(c) Gumbel Copula



(d) Gaussian-Clayton Mixture



(e) Gaussian-Gumbel Mixture



(f) Clayton-Gumbel Mixture

Figure 2.1: Scatter plots for Gaussian, Clayton and Gumbel copula and their mixture.

Specifically, in the first stage, the marginal distributions are estimated nonparametrically to avoid misspecification of marginals. Then, in the second stage, the estimated marginals or the empirical CDFs are plugged into the copula so that copula parameters are estimated by Maximum Likelihood. Finally, to facilitate the estimation of weight parameters for each nested copula, iterative procedures, namely, the EM algorithm, are implemented. However, both works do not establish the asymptotic properties and lay theoretical foundations for the estimators. Cai and Wang (2014) provide a theoretical support about mixture copula estimation. Specifically, they discuss a data-driven copula selection method via penalized likelihood with a shrinkage operator so that all the related parameters estimation and model selection are achieved simultaneously. Their estimation procedure is similar to some variable selection strategies such as the Least Absolute Shrinkage and Selection Operator(LASSO) and Smoothly Clipped Absolute Deviation (SCAD) proposed by Tibshirani (1996) and Fan and Li (2001), respectively. In our work, the model average approach estimates a mixture copula based on the criterion function that minimizes the estimation losses. In the following section, we specify the estimating procedures and prove that the model average estimator is asymptotically optimal.

2.2 Theoretical Model

Consider $K-1$ candidate copulas

$$C_k(\mathbf{x}) = C_k\left\{F_1(x_1;\boldsymbol{\alpha}_1),\ldots,F_p(x_p;\boldsymbol{\alpha}_p);\boldsymbol{\theta}_k\right\}, \qquad k=1,\ldots,K-1.$$

When the data set $\{\mathbf{X}_t\}_{t=1}^{\mathrm{T}}$ is thought to be from the k^{th} copula, we can estimate $C_k(\mathbf{x})$ by maximizing the likelihood function. Denote the resulting estimator as

$$\widehat{C}_k(\mathbf{x}) = C_k\left\{F_1(x_1;\widehat{\boldsymbol{\alpha}}_{k,1}),\ldots,F_p(x_p;\widehat{\boldsymbol{\alpha}}_{k,p});\widehat{\boldsymbol{\theta}}_k\right\}, \qquad k=1,\ldots,K-1.$$

Noting that using a candidate copula to estimate mixture copula can be very poor, we also use maximum likelihood (ML) estimator of the mixture copula to construct our model average estimator of mixture copula. Let

$$\widehat{C}_K(\mathbf{x}) = \sum_{k=1}^{K-1} \widetilde{\omega}_k C_k \left\{ F_1(x_1; \widetilde{\boldsymbol{\alpha}}_1), ..., F_p(x_p; \widetilde{\boldsymbol{\alpha}}_p); \widetilde{\boldsymbol{\theta}}_k \right\},$$

where $\widetilde{\omega}_1, \ldots, \widetilde{\omega}_{K-1}$, $\widetilde{\boldsymbol{\alpha}}_1, \ldots, \widetilde{\boldsymbol{\alpha}}_p$ and $\widetilde{\boldsymbol{\theta}}_1, \ldots, \widetilde{\boldsymbol{\theta}}_{K-1}$ are estimators by ML estimation. Note that $\widetilde{\boldsymbol{\omega}}$ is constrained to be between 0 and 1 and summation equals to 1. For each copula k, $\widetilde{\theta}_k$ also has its own constraint. For example, the parameter for Gaussian copula should be between -1 and 1.

Write $\mathbf{w} = (w_1, \ldots, w_K)^{\mathrm{T}}$ as weight vector, belonging to the set

$$\mathcal{W} = \left\{ \mathbf{w} \in [0,1]^K : \sum\nolimits_{k=1}^{K} w_k = 1 \right\}.$$

Then, the model average estimator of mixture copula can be written as

$$\widehat{C}(\mathbf{x}, \mathbf{w}) = \sum_{k=1}^{K} w_k \widehat{C}_k(\mathbf{x}).$$

Let $C_0(\mathbf{x}) = C_0 \{G_1(x_1), \ldots, G_p(x_p); \boldsymbol{\theta}_0\}$ be the true copula. Note that $C_0(\mathbf{x})$ can be out of candidate set $\{C_1(\mathbf{x}), \ldots, C_{K-1}(\mathbf{x})\}$ and it may not be a mixture copula based on $\{C_1(\mathbf{x}), \ldots, C_{K-1}(\mathbf{x})\}$. The goal of this work is to estimate $C_0(\mathbf{x})$ by model average approach.

In this work we use the J-fold Cross-Validation (CV) to choose weights, which is similar to Jackknife model average (JMA) method (Hansen and Racine, 2012). Specifically, we divide the data set into J groups such that for each group, we have $M = T/J$ observations. In the j^{th} group, we have observations $\mathbf{X}_{(j-1)M+1}, \ldots, \mathbf{X}_{jM}$, where $j = 1, ..., J$. Write $\widetilde{C}_k^{(-j)}(\mathbf{x})$ as the estimator of $C_k(\mathbf{x})$ with the j^{th} group removed from the sample. So the corresponding average estimator is

$$\widetilde{C}^{(-j)}(\mathbf{x}, \mathbf{w}) = \sum_{k=1}^{K} w_k \widetilde{C}_k^{(-j)}(\mathbf{x}).$$

An empirical estimator of $C_0(\mathbf{x})$ is

$$\overline{C}_{(j)}(\mathbf{x}) = M^{-1} \sum_{m=1}^{M} I(\mathbf{X}_{(j-1)M+m} \leq \mathbf{x}),$$

where $I(\cdot)$ is an indicate function. We emphasize that the comparison between $\mathbf{X}_{(j-1)M+m}$ and $\mathbf{x}$ means the comparison of each component in these p-dimension vectors. Then, it is straightforward to show that

$$E\left\{\overline{C}_{(j)}(\mathbf{x})\right\} = C_0(\mathbf{x}). \tag{2.2}$$

So our J-fold CV criterion is formulated to be

$$CV_J(\mathbf{w}) = \sum_{j=1}^{J} \sum_{m=1}^{M} \left\{\widetilde{C}^{(-j)}(\mathbf{X}_{(j-1)M+m}, \mathbf{w}) - \overline{C}_{(j)}(\mathbf{X}_{(j-1)M+m})\right\}^2.$$

The resulting weight vector is

$$\widehat{\mathbf{w}} = \text{argmin}_{\mathbf{w} \in \mathcal{W}} CV_J(\mathbf{w}),$$

and the model average estimator of the copula is $\widehat{C}(\mathbf{x}, \widehat{\mathbf{w}})$.

To ease exposition, we introduce/summarize notations used in the paper.

The true copula

$$\mathbf{C}_0 = \{C_0(\mathbf{X}_1), \dots, C_0(\mathbf{X}_T)\}^{\mathrm{T}},$$

the copula estimated by the k^{th} candidate copula (when $k = K$, the candidate copula is a composite of the candidate copulas)

$$\widehat{\mathbf{C}}_k = \left\{\widehat{C}_k(\mathbf{X}_1), \dots, \widehat{C}_k(\mathbf{X}_T)\right\}^{\mathrm{T}},$$

the copula estimated by model averaging

$$\widehat{\mathbf{C}}(\mathbf{w}) = \left\{\widehat{C}(\mathbf{X}_1, \mathbf{w}), \dots, \widehat{C}(\mathbf{X}_T, \mathbf{w})\right\}^{\mathrm{T}},$$

the copula estimated by using the k^{th} candidate copula and J-fold CV

$$\widetilde{\mathbf{C}}_k = \left\{\widetilde{C}_k^{(-1)}(\mathbf{X}_1), \dots, \widetilde{C}_k^{(-1)}(\mathbf{X}_M), \widetilde{C}_k^{(-2)}(\mathbf{X}_{M+1}), \dots, \widetilde{C}_k^{(-J)}(\mathbf{X}_T)\right\}^{\mathrm{T}},$$

the copula estimated by using model averaging and J-fold CV

$$\widetilde{\mathbf{C}}(\mathbf{w}) = \left\{\widetilde{C}^{(-1)}(\mathbf{X}_1, \mathbf{w}), \dots, \widetilde{C}^{(-1)}(\mathbf{X}_M, \mathbf{w}), \widetilde{C}^{(-2)}(\mathbf{X}_{M+1}, \mathbf{w}), \dots, \widetilde{C}^{(-J)}(\mathbf{X}_T, \mathbf{w})\right\}^{\mathrm{T}},$$

and the empirical estimator of $\mathbf{C}_0$

$$\overline{\mathbf{C}} = \left\{\overline{C}_{(1)}(\mathbf{X}_1), \dots, \overline{C}_{(1)}(\mathbf{X}_M), \overline{C}_{(2)}(\mathbf{X}_{M+1}), \dots, \overline{C}_{(J)}(\mathbf{X}_T)\right\}^{\mathrm{T}}.$$

Let $\widetilde{\mathbf{h}}_k = \widetilde{\mathbf{C}}_k - \overline{\mathbf{C}}$ and $\widetilde{\mathbf{H}} = (\widetilde{\mathbf{h}}_1, \dots, \widetilde{\mathbf{h}}_K)$. Now, we can rewrite the J-fold CV criterion as

$$CV_J(\mathbf{w}) = \left\|\widetilde{\mathbf{C}}(\mathbf{w}) - \overline{\mathbf{C}}\right\|^2 = \mathbf{w}^{\mathrm{T}}\widetilde{\mathbf{H}}^{\mathrm{T}}\widetilde{\mathbf{H}}\mathbf{w},$$

which is a quadratic form of $\mathbf{w}$. So the minimization of $CV_J(\mathbf{w})$ with respect to $\mathbf{w}$ can be implemented easily.

Define a quadratic loss function of the model average estimator $\widehat{\mathbf{C}}(\mathbf{w})$ as $L_T(\mathbf{w}) = \|\widehat{\mathbf{C}}(\mathbf{w}) - \mathbf{C}_0\|^2$. Like literature on model selection and model averaging such as Shao (1997) and Hansen (2007), our goal is to reduce quadratic loss by using model averaging. The following theorem shows that our method minimizes the quadratic loss asymptotically.

<u>Theorem</u> 1. *Under Conditions (C.1) — (C.3) presented in Appendix A,*

$$\frac{L_T(\widehat{\mathbf{w}})}{\inf_{\mathbf{w}\in\mathcal{W}} L_T(\mathbf{w})} \to 1 \qquad \textit{in probability} \tag{2.3}$$

as $T \to \infty$*.*

The result (2.3) means that our model average estimator $\mathbf{C}(\widehat{\mathbf{w}})$ is asymptotically optimal in the sense that the squared loss of $\mathbf{C}(\widehat{\mathbf{w}})$ is asymptotically identical to that by the infeasible best possible model average estimator. The detailed proof for Theorem 1 is given in Appendix A.

2.3 Numerical Studies

We compare squared estimation losses of the proposed model average approach on mixture copula model with two other methods: Cai and Wang (2014)'s penalized likelihood method, and a BIC method which selects only one copula by comparing each candidate's BIC. Specifically, we consider two types of simulation. In Type I simulation, data are generated from copulas which are included in the mixture copula model. On the contrary, in Type II simulation, the working mixture copula model is misspecified. That is, data are generated from copulas which are not constituents of the working mixture model. We compare which method gives more reasonable description of dependence structure under the two types of settings.

2.3.1 Simulation Type I

Type I simulation considers the scenario that data are generated from copulas which are constituents of our working model. The data generating process (DGP) is a process that the bivariate joint distribution has a form of copula function, and for simplicity, the two margins are normally distributed with marginal parameters $(\mu_1, \sigma_1) = (0, 1)$ and $(\mu_2, \sigma_2) = (0, 1)$. Our working mixture model includes three commonly used copulas: Gaussian, Clayton and Gumbel. One could always add in more candidate copulas, but we believe this parsimonious model has the ability to cover many possible dependence structure empirical researchers would encounter in the real world. Characteristics of the three copulas have been discussed in Section 2 and the simulated scatter plots have been displayed in Figure 2.1. Particularly, our presumed mixture copula could be formulated as:

$$C(u, v; \boldsymbol{\theta}, \boldsymbol{\omega}) = \omega_{Ga} C_{Ga}(u, v; \theta_1) + \omega_{Cl} C_{Cl}(u, v; \theta_2) + \omega_{Gu} C_{Gu}(u, v; \theta_3),$$

where C_{Ga}, C_{Cl} and C_{Gu} stand for Gaussian, Clayton and Gumbel copula, respectively, and u, v denote the two margins. By fitting the data into Gaussian, Clayton and Gumbel copula separately, one could obtain their ML estimates $\widehat{\theta}_1$, $\widehat{\theta}_2$ and $\widehat{\theta}_3$.

We have argued in Section 3 that it can be quite poor if we use any individual copula to estimate series of data generated by mixture copula, since for the mixture copula, any individual candidate copula is an "inappropriate" model. We thus include a maximum likelihood (ML) estimator of the mixture copula into our model average estimator. Specifically, let $C_{ML}(u, v; \widetilde{\boldsymbol{\theta}}, \widetilde{\boldsymbol{\omega}}) = \widetilde{\omega}_1 C_{Ga}(u, v; \widetilde{\theta}_1) + \widetilde{\omega}_2 C_{Cl}(u, v; \widetilde{\theta}_2) + \widetilde{\omega}_3 C_{Gu}(u, v; \widetilde{\theta}_3)$, where $\widetilde{\omega}_1, \widetilde{\omega}_2, \widetilde{\omega}_3$ and $\widetilde{\boldsymbol{\theta}} = (\widetilde{\theta}_1, \widetilde{\theta}_2, \widetilde{\theta}_3)$ are the ML estimates. Our method then averages over the the four components: three individual candidate copulas plus a ML estimator of their linear combination. We need to choose $w_{Ga}, w_{Cl}, w_{Gu}, w_{ML}$ in our working mixture copula model

$$C(u,v;\widehat{\boldsymbol{\theta}},\mathbf{w}) = w_{Ga}C_{Ga}(u,v;\widehat{\theta}_1)+w_{Cl}C_{Cl}(u,v;\widehat{\theta}_2)+w_{Gu}C_{Gu}(u,v;\widehat{\theta}_3)+w_{ML}C_{ML}(u,v;\widetilde{\boldsymbol{\theta}})$$

via model average method to minimize the estimation losses $L_T(\mathbf{w})$ defined in Section 3.

The simulation considers three sample sizes: $T = 100$, 200 and 500. Observations are simulated from different copulas by the following DGP and the model averaging estimators are computed. All simulations are repeated 500 times. As we concentrate on mixture copula situations, we simulate three mixture copulas with two components and one mixture copula with three components. Specifically, we have the following 4 cases for the setup of weights:

Case 1: $\omega_{Ga} = 1/2$, $\omega_{Cl} = 1/2$, $\omega_{Gu} = 0$;

Case 2: $\omega_{Ga} = 1/2$, $\omega_{Cl} = 0$, $\omega_{Gu} = 1/2$;

Case 3: $\omega_{Ga} = 0$, $\omega_{Cl} = 1/2$, $\omega_{Gu} = 1/2$;

Case 4: $\omega_{Ga} = 1/3$, $\omega_{Cl} = 1/3$, $\omega_{Gu} = 1/3$.

For each case of weight above, we consider three sets of copula parameters:

Parameter setting 1: $\theta_{Ga} = 0.5$, $\theta_{Cl} = 5.8$, $\theta_{Gu} = 5.1$;

Parameter setting 2: $\theta_{Ga} = 0.6$, $\theta_{Cl} = 6.8$, $\theta_{Gu} = 6.1$;

Parameter setting 3: $\theta_{Ga} = 0.7$, $\theta_{Cl} = 7.8$, $\theta_{Gu} = 7.1$.

Therefore, we will have $4 \times 3 = 12$ groups of DGPs in total.

Table 2.1: Mean of squared in-sample estimation losses for Type I simulation

	$\theta_{Ga}=0.5,\theta_{Cl}=5.8,\theta_{Gu}=5.1$			$\theta_{Ga}=0.6,\theta_{Cl}=6.8,\theta_{Gu}=6.1$			$\theta_{Ga}=0.7,\theta_{Cl}=7.8,\theta_{Gu}=7.1$		
Sample Size=100									
	MA	CW	BIC	MA	CW	BIC	MA	CW	BIC
	Mean	Mean	Mean	Mean	Mean	Mean	Mean	Mean	Mean
$\omega_{Ga}=0.5,\omega_{Cl}=0.5,\omega_{Gu}=0$	1.0909	1.0948	1.1808	1.0919	1.0932	1.1951	1.1052	1.0903	1.1870
$\omega_{Ga}=0.5,\omega_{Cl}=0,\omega_{Gu}=0.5$	1.0063	0.9837	1.0514	1.0135	0.9845	1.0568	1.0288	0.9892	1.0553
$\omega_{Ga}=0,\omega_{Cl}=0.5,\omega_{Gu}=0.5$	1.1372	1.1047	1.2147	1.1265	1.1041	1.1969	1.1225	1.0963	1.1867
$\omega_{Ga}=1/3,\omega_{Cl}=1/3,\omega_{Gu}=1/3$	1.1057	1.0722	1.2030	1.1097	1.0664	1.2067	1.1085	1.0621	1.1843
Sample Size=200									
	$\theta_{Ga}=0.5,\theta_{Cl}=5.8,\theta_{Gu}=5.1$			$\theta_{Ga}=0.6,\theta_{Cl}=6.8,\theta_{Gu}=6.1$			$\theta_{Ga}=0.7,\theta_{Cl}=7.8,\theta_{Gu}=7.1$		
	MA	CW	BIC	MA	CW	BIC	MA	CW	BIC
	Mean	Mean	Mean	Mean	Mean	Mean	Mean	Mean	Mean
$\omega_{Ga}=0.5,\omega_{Cl}=0.5,\omega_{Gu}=0$	0.5635	0.5651	0.6860	0.5618	0.5628	0.6866	0.5705	0.5598	0.6676
$\omega_{Ga}=0.5,\omega_{Cl}=0,\omega_{Gu}=0.5$	0.5137	0.5106	0.5675	0.5175	0.5168	0.5723	0.5242	0.5210	0.5714
$\omega_{Ga}=0,\omega_{Cl}=0.5,\omega_{Gu}=0.5$	0.5627	0.5263	0.6389	0.5581	0.5234	0.6368	0.5575	0.5241	0.6341
$\omega_{Ga}=1/3,\omega_{Cl}=1/3,\omega_{Gu}=1/3$	0.5721	0.5559	0.7076	0.5733	0.5567	0.6950	0.5835	0.5570	0.6665
Sample Size=500									
	$\theta_{Ga}=0.5,\theta_{Cl}=5.8,\theta_{Gu}=5.1$			$\theta_{Ga}=0.6,\theta_{Cl}=6.8,\theta_{Gu}=6.1$			$\theta_{Ga}=0.7,\theta_{Cl}=7.8,\theta_{Gu}=7.1$		
	MA	CW	BIC	MA	CW	BIC	MA	CW	BIC
	Mean	Mean	Mean	Mean	Mean	Mean	Mean	Mean	Mean
$\omega_{Ga}=0.5,\omega_{Cl}=0.5,\omega_{Gu}=0$	0.2245	0.2240	0.3499	0.2286	0.2251	0.3451	0.2270	0.2245	0.3315
$\omega_{Ga}=0.5,\omega_{Cl}=0,\omega_{Gu}=0.5$	0.2206	0.2133	0.2607	0.2218	0.2138	0.2623	0.2232	0.2141	0.2579
$\omega_{Ga}=0,\omega_{Cl}=0.5,\omega_{Gu}=0.5$	0.2198	0.2075	0.3476	0.2226	0.2102	0.3528	0.2222	0.2079	0.3408
$\omega_{Ga}=1/3,\omega_{Cl}=1/3,\omega_{Gu}=1/3$	0.2169	0.2137	0.3619	0.2162	0.2134	0.3477	0.2198	0.2129	0.3165

Table 2.1 displays how close the estimated copula is to the true copula in terms of mean squared estimation loss across the three methods we mentioned at the beginning: our proposed model average approach (MA), Cai and Wang (2014)'s penalized likelihood (CW), and the BIC method which selects one from a set of candidates based on BIC (BIC). For expositional ease, the estimation loss in each case is multiplied by 1000.

We make a few observations from Table 2.1. First, BIC gives the greatest mean squared estimation losses in all cases compared with the other two approaches. This implies that simply relying on one copula selected based on BIC comparison would give rather bad estimation results, even though the observations are mixture by copulas included in our candidate set and exhibit certain tail dependence structure. A visual comparison between the mixture copula and one copula model could be observed through boxplots in Figure 2.2: We calculate the ratio of estimation losses obtained from MA, CW and BIC, respectively, in each simulation. For example, MA/CW represents the ratio of estimation losses between MA and CW, and MA is superior to CW if the ratio is smaller than 1. To save space, we only present the 12 cases when sample size equals to 500. Results for 100 and 200 sample size are similar. In Figure 2.2, for all the four different setups of weight, on average about 75% of the ratios obtained from MA/BIC and CW/BIC are smaller than 1, indicating substantial superiority of mixture copula model estimated by MA and CW. As argued before, the superiority of the mixture copula model stems from its flexibility in nesting various dependence structures. Second, according to Table 2.1, the difference of mean squared estimation loss obtained by CW and MA is tiny. Figure 2.2 further supports this as the medians of the ratios of estimation losses between MA and CW are close to 1 for all 12 cases. That is to say, when our working mixture model is not misspecified, the two methods are not significantly superior to each other in terms of minimizing estimation losses. Simulation results in Cai and Wang (2014) demonstrate that, when data are generated from copulas included in the mixture model, their method is able to locate these data-generating copulas exactly and estimate their associated weights and parameters accurately. Thus, the estimated mixture copula is quite close to the true model as the estimation losses should be quite small. In this sense, the similarity between MA and CW in terms of estimation loss provides support for the optimality of model average approach.

Table 2.2 displays the out-sample predicting performance of the three competing methods. The number of out-sample observations is equal to the number of in-sample observations. Hence, for 100, 200 and 500 in-sample observations, the out-sample predictions include 100, 200 and 500 observations, respectively. The results in Table 2.2 are quite similar to those we found in Table 2.1: Mixture copula model still performs better than an individual copula. In the three competing methods, BIC gives the largest predicting errors for all cases in the three competing methods,

Table 2.2: Mean of squared out-sample predicting losses for Type I simulation

	In-sample = 100; Out-sample = 100								
	$\theta_{Ga}=0.5,\theta_{Cl}=5.8,\theta_{Gu}=5.1$			$\theta_{Ga}=0.6,\theta_{Cl}=6.8,\theta_{Gu}=6.1$			$\theta_{Ga}=0.7,\theta_{Cl}=7.8,\theta_{Gu}=7.1$		
	MA	CW	BIC	MA	CW	BIC	MA	CW	BIC
	Mean	Mean	Mean	Mean	Mean	Mean	Mean	Mean	Mean
$\omega_{Ga}=0.5,\omega_{Cl}=0.5,\omega_{Gu}=0$	1.0134	1.0277	1.0913	1.0232	1.0284	1.1074	1.0344	1.0304	1.1198
$\omega_{Ga}=0.5,\omega_{Cl}=0,\omega_{Gu}=0.5$	1.0117	0.9824	1.0861	1.0113	0.9825	1.0922	1.0276	0.9868	1.0813
$\omega_{Ga}=0,\omega_{Cl}=0.5,\omega_{Gu}=0.5$	1.1488	1.1096	1.2161	1.1390	1.1138	1.2126	1.1350	1.1048	1.1977
$\omega_{Ga}=1/3,\omega_{Cl}=1/3,\omega_{Gu}=1/3$	1.1108	1.0772	1.2490	1.1061	1.0685	1.2466	1.1055	1.0608	1.2208
	In-sample = 200; Out-sample = 200								
	$\theta_{Ga}=0.5,\theta_{Cl}=5.8,\theta_{Gu}=5.1$			$\theta_{Ga}=0.6,\theta_{Cl}=6.8,\theta_{Gu}=6.1$			$\theta_{Ga}=0.7,\theta_{Cl}=7.8,\theta_{Gu}=7.1$		
	MA	CW	BIC	MA	CW	BIC	MA	CW	BIC
	Mean	Mean	Mean	Mean	Mean	Mean	Mean	Mean	Mean
$\omega_{Ga}=0.5,\omega_{Cl}=0.5,\omega_{Gu}=0$	0.5523	0.5656	0.6763	0.5504	0.5643	0.6761	0.5597	0.5618	0.6623
$\omega_{Ga}=0.5,\omega_{Cl}=0,\omega_{Gu}=0.5$	0.5209	0.5155	0.5825	0.5244	0.5215	0.5891	0.5270	0.5250	0.5861
$\omega_{Ga}=0,\omega_{Cl}=0.5,\omega_{Gu}=0.5$	0.5598	0.5280	0.6396	0.5591	0.5280	0.6386	0.5580	0.5286	0.6318
$\omega_{Ga}=1/3,\omega_{Cl}=1/3,\omega_{Gu}=1/3$	0.5727	0.5612	0.7049	0.5720	0.5624	0.6926	0.5799	0.5623	0.6666
	In-sample = 500; Out-sample = 500								
	$\theta_{Ga}=0.5,\theta_{Cl}=5.8,\theta_{Gu}=5.1$			$\theta_{Ga}=0.6,\theta_{Cl}=6.8,\theta_{Gu}=6.1$			$\theta_{Ga}=0.7,\theta_{Cl}=7.8,\theta_{Gu}=7.1$		
	MA	CW	BIC	MA	CW	BIC	MA	CW	BIC
	Mean	Mean	Mean	Mean	Mean	Mean	Mean	Mean	Mean
$\omega_{Ga}=0.5,\omega_{Cl}=0.5,\omega_{Gu}=0$	0.2252	0.2252	0.3505	0.2284	0.2262	0.3465	0.2267	0.2255	0.3343
$\omega_{Ga}=0.5,\omega_{Cl}=0,\omega_{Gu}=0.5$	0.2212	0.2136	0.2660	0.2216	0.2141	0.2677	0.2220	0.2145	0.2634
$\omega_{Ga}=0,\omega_{Cl}=0.5,\omega_{Gu}=0.5$	0.2198	0.2076	0.3527	0.2229	0.2101	0.3562	0.2221	0.2080	0.3409
$\omega_{Ga}=1/3,\omega_{Cl}=1/3,\omega_{Gu}=1/3$	0.2158	0.2137	0.3672	0.2148	0.2138	0.3520	0.2188	0.2136	0.3197

while CW and MA are not superior to each other.

Above all, in terms of in-sample fitting losses and out-sample predicting errors, Type I simulation demonstrates the superiority of mixture copula model over one individual copula model. Furthermore, for the estimation of a mixture copula model, our proposed model average approach performs similarly to Cai and Wang (2014)'s method.

2.3.2 Simulation Type II

The working model is misspecified in Type II simulations. The purpose of Type II simulation is to see how the proposed model average approach performs when data are generated from copulas which are out of our candidate set, a situation which should be common in empirical studies as the true model is always unknown to econometricians.

Specifically, our working mixture model is still comprised by Gaussian, Clayton and Gumbel copulas but the true observations are generated separately from Frank, Survival Joe (SJ) and Joe copulas. These three copulas are also widely used in empirical studies. Frank copula is similar to Gaussian copula as it also does not exhibit tail dependence in both sides; but it has relatively stronger dependence in the center of the distribution. Joe copula, like Gumbel copula, exhibits right tail dependence. Survival Joe copula is a 180° rotation of Joe, so it exhibits left tail dependence as the Clayton copula does. We consider three cases that the true copula is generated from Frank, Joe and Survival Joe, respectively, plus their composite in which each component has equal weights. For Frank, Joe and Survival Joe, as we did in simulation Type I, we consider four weighting setups:

$$\text{Case 1: } \omega_{Frank} = 1/2,\, \omega_{SJ} = 1/2,\, \omega_{Joe} = 0;$$

$$\text{Case 2: } \omega_{Frank} = 1/2,\, \omega_{SJ} = 0,\, \omega_{Joe} = 1/2;$$

$$\text{Case 3: } \omega_{Frank} = 0,\, \omega_{SJ} = 1/2,\, \omega_{Joe} = 1/2;$$

$$\text{Case 4: } \omega_{Frank} = 1/3,\, \omega_{SJ} = 1/3,\, \omega_{Joe} = 1/3,$$

and three different copula parameter settings:

$$\text{Parameter setting 1: } \theta_{Frank} = 5.5,\, \theta_{SJ} = 4.8,\, \theta_{Joe} = 4.5;$$

$$\text{Parameter setting 2: } \theta_{Frank} = 6.5,\, \theta_{SJ} = 5.8,\, \theta_{Joe} = 5.5;$$

$$\text{Parameter setting 3: } \theta_{Frank} = 7.5,\, \theta_{SJ} = 6.8,\, \theta_{Joe} = 6.5.$$

Table 2.3 presents the mean of in-sample estimation losses of mixture copula model for MA, CW and BIC with 500 simulations when the working model is misspecified. The most significant difference between Table 2.1 and Table 2.3 is that, in Table 2.3, it exhibits that the estimation losses due to MA are uniformly smaller than the other two competing methods, while in Table 2.1, the estimation losses

Table 2.3: Mean of squared in-sample estimation losses for Type II simulation

	Sample Size=100								
	$\theta_{Frank}=5.5, \theta_{SJ}=4.8, \theta_{Joe}=4.5$			$\theta_{Frank}=6.5, \theta_{SJ}=5.8, \theta_{Joe}=5.5$			$\theta_{Frank}=7.5, \theta_{SJ}=6.8, \theta_{Joe}=6.5$		
	MA	CW	BIC	MA	CW	BIC	MA	CW	BIC
	Mean	Mean	Mean	Mean	Mean	Mean	Mean	Mean	Mean
$\omega_{Frank}=0.5, \omega_{SJ}=0.5, \omega_{Joe}=0$	1.1163	1.2195	1.2837	1.1218	1.1888	1.3198	1.0988	1.1463	1.3073
$\omega_{Frank}=0.5, \omega_{SJ}=0, \omega_{Joe}=0.5$	0.9837	1.0388	1.0012	0.9977	1.0629	1.0318	1.0137	1.0755	1.0571
$\omega_{Frank}=0, \omega_{SJ}=0.5, \omega_{Joe}=0.5$	1.1315	1.2573	1.2052	1.1371	1.1959	1.2314	1.1450	1.1810	1.2614
$\omega_{Frank}=1/3, \omega_{SJ}=1/3, \omega_{Joe}=1/3$	1.0716	1.1696	1.1343	1.0716	1.1404	1.1698	1.0766	1.0964	1.1943
	Sample Size=200								
	$\theta_{Frank}=5.5, \theta_{SJ}=4.8, \theta_{Joe}=4.5$			$\theta_{Frank}=6.5, \theta_{SJ}=5.8, \theta_{Joe}=5.5$			$\theta_{Frank}=7.5, \theta_{SJ}=6.8, \theta_{Joe}=6.5$		
	MA	CW	BIC	MA	CW	BIC	MA	CW	BIC
	Mean	Mean	Mean	Mean	Mean	Mean	Mean	Mean	Mean
$\omega_{Frank}=0.5, \omega_{SJ}=0.5, \omega_{Joe}=0$	0.5306	0.6352	0.7006	0.5425	0.6094	0.7220	0.5502	0.5917	0.7399
$\omega_{Frank}=0.5, \omega_{SJ}=0, \omega_{Joe}=0.5$	0.5300	0.5864	0.5539	0.5401	0.6135	0.5801	0.5505	0.6241	0.6000
$\omega_{Frank}=0, \omega_{SJ}=0.5, \omega_{Joe}=0.5$	0.5387	0.6324	0.5967	0.5498	0.6014	0.6248	0.5395	0.5636	0.6295
$\omega_{Frank}=1/3, \omega_{SJ}=1/3, \omega_{Joe}=1/3$	0.5280	0.6153	0.6038	0.5318	0.5819	0.6274	0.5325	0.5529	0.6456
	Sample Size=500								
	$\theta_{Frank}=5.5, \theta_{SJ}=4.8, \theta_{Joe}=4.5$			$\theta_{Frank}=6.5, \theta_{SJ}=5.8, \theta_{Joe}=5.5$			$\theta_{Frank}=7.5, \theta_{SJ}=6.8, \theta_{Joe}=6.5$		
	MA	CW	BIC	MA	CW	BIC	MA	CW	BIC
	Mean	Mean	Mean	Mean	Mean	Mean	Mean	Mean	Mean
$\omega_{Frank}=0.5, \omega_{SJ}=0.5, \omega_{Joe}=0$	0.2516	0.3366	0.4312	0.2550	0.3107	0.4534	0.2578	0.2920	0.4672
$\omega_{Frank}=0.5, \omega_{SJ}=0, \omega_{Joe}=0.5$	0.2699	0.3471	0.2796	0.2857	0.3684	0.3004	0.2929	0.3726	0.3177
$\omega_{Frank}=0, \omega_{SJ}=0.5, \omega_{Joe}=0.5$	0.2385	0.3466	0.3139	0.2354	0.3024	0.3378	0.2346	0.2780	0.3517
$\omega_{Frank}=1/3, \omega_{SJ}=1/3, \omega_{Joe}=1/3$	0.2396	0.3328	0.3265	0.2403	0.2976	0.3558	0.2425	0.2727	0.3766

between MA and CW are tiny and there is no significant superiority of one method over the other. But when the working mixture model is misspecified, MA yields better performance than CW.

Like Figure 2.2, Figure 2.3 also gives a more straightforward way to compare the estimation performance among MA, CW and BIC in each of the 500 simulations when the sample size is 500. Comparing with the Simulation Type I, results presented in Figure 2.2 demonstrate the significant superiority of MA over CW, since of all the 12 cases displayed, on average about 75% of the ratios between MA and CW obtained in 500 simulations are smaller than 1. This implies the superiority of MA over CW when the working model is misspecified.

We also present the out-sample predicting performance of the three competing methods under Type II simulation in Table 2.4. Like Type I simulation, the number of out-sample predictions is the same with the corresponding size of the in-sample data. The patterns in Table 2.4 are similar to those in Table 2.3: MA exhibits more accurate out-sample predictions than CW and BIC do, and mixture copula provides

Table 2.4: Mean of squared out-sample predicting losses for Type II simulation

	In-sample = 100; Out-sample = 100								
	$\theta_{Frank}=5.5, \theta_{SJ}=4.8, \theta_{Joe}=4.5$			$\theta_{Frank}=6.5, \theta_{SJ}=5.8, \theta_{Joe}=5.5$			$\theta_{Frank}=7.5, \theta_{SJ}=6.8, \theta_{Joe}=6.5$		
	MA	CW	BIC	MA	CW	BIC	MA	CW	BIC
	Mean	Mean	Mean	Mean	Mean	Mean	Mean	Mean	Mean
$\omega_{Frank}=0.5, \omega_{SJ}=0.5, \omega_{Joe}=0$	1.1125	1.2418	1.2835	1.1196	1.2087	1.3225	1.0992	1.1659	1.3120
$\omega_{Frank}=0.5, \omega_{SJ}=0, \omega_{Joe}=0.5$	0.9642	1.0149	0.9920	0.9805	1.0399	1.0233	0.9958	1.0534	1.0488
$\omega_{Frank}=0, \omega_{SJ}=0.5, \omega_{Joe}=0.5$	1.1212	1.2154	1.2117	1.1239	1.1596	1.2390	1.1315	1.1485	1.2667
$\omega_{Frank}=1/3, \omega_{SJ}=1/3, \omega_{Joe}=1/3$	1.0564	1.1419	1.1292	1.0807	1.1144	1.1652	1.0620	1.0735	1.1887
	In-Sample = 200; Out-sample = 200								
	$\theta_{Frank}=5.5, \theta_{SJ}=4.8, \theta_{Joe}=4.5$			$\theta_{Frank}=6.5, \theta_{SJ}=5.8, \theta_{Joe}=5.5$			$\theta_{Frank}=7.5, \theta_{SJ}=6.8, \theta_{Joe}=6.5$		
	MA	CW	BIC	MA	CW	BIC	MA	CW	BIC
	Mean	Mean	Mean	Mean	Mean	Mean	Mean	Mean	Mean
$\omega_{Frank}=0.5, \omega_{SJ}=0.5, \omega_{Joe}=0$	0.5322	0.6472	0.6994	0.5422	0.6195	0.7209	0.5501	0.6009	0.7409
$\omega_{Frank}=0.5, \omega_{SJ}=0, \omega_{Joe}=0.5$	0.5275	0.5787	0.5566	0.5377	0.6071	0.5834	0.5491	0.6183	0.6034
$\omega_{Frank}=0, \omega_{SJ}=0.5, \omega_{Joe}=0.5$	0.5329	0.6135	0.6010	0.5443	0.5861	0.6291	0.5346	0.5505	0.6330
$\omega_{Frank}=1/3, \omega_{SJ}=1/3, \omega_{Joe}=1/3$	0.5264	0.6057	0.6083	0.5295	0.5734	0.6322	0.5301	0.5460	0.6502
	In-sample = 500; Out-sample = 500								
	$\theta_{Frank}=5.5, \theta_{SJ}=4.8, \theta_{Joe}=4.5$			$\theta_{Frank}=6.5, \theta_{SJ}=5.8, \theta_{Joe}=5.5$			$\theta_{Frank}=7.5, \theta_{SJ}=6.8, \theta_{Joe}=6.5$		
	MA	CW	BIC	MA	CW	BIC	MA	CW	BIC
	Mean	Mean	Mean	Mean	Mean	Mean	Mean	Mean	Mean
$\omega_{Frank}=0.5, \omega_{SJ}=0.5, \omega_{Joe}=0$	0.2509	0.3419	0.4316	0.2542	0.3153	0.4548	0.2569	0.2961	0.4695
$\omega_{Frank}=0.5, \omega_{SJ}=0, \omega_{Joe}=0.5$	0.2680	0.3431	0.2803	0.2841	0.3648	0.3013	0.2914	0.3693	0.3187
$\omega_{Frank}=0, \omega_{SJ}=0.5, \omega_{Joe}=0.5$	0.2359	0.3394	0.3140	0.2332	0.2964	0.3382	0.2329	0.2730	0.3521
$\omega_{Frank}=1/3, \omega_{SJ}=1/3, \omega_{Joe}=1/3$	0.2374	0.3262	0.3268	0.2382	0.2922	0.3563	0.2400	0.2680	0.3766

more robust estimates and relatively smaller predicting losses.

The simulation results show that our proposed model average method outperforms both CW and the BIC method when the working mixture copula is misspecified. This finding has important implications for empirical studies as the true copula is always unknown to econometricians and the misspecification of one's working mixture model should be quite common. Given the great flexibility of mixture copula model in nesting various dependence structures, the model average method is able to provide a more accurate estimate of the dependence structure in terms of estimation losses. We consider a real data example in the following section.

2.4 An Empirical Study

We consider a real data example to examine the performance of model average approach in mixture copula model. Specifically, we consider daily returns of Morgan Stanley Capital International (MSCI) equity indexes for four developed economies: United Kingdom (UK), Hong Kong (HK), Japan (JP) and United States (US). The

Table 2.5: Summary statistics for daily returns.

	UK	HK	JP	US
Mean	0.0091	0.0005	-0.0034	-0.0163
Median	0.0242	-0.0119	0.0180	0.0473
min	-4.8860	-4.1028	-4.7666	-6.6120
max	4.1248	3.3496	3.6153	2.8370
S.D.	0.9993	1.0110	1.0136	0.9979
Skewness	-0.2063	-0.2879	-0.2539	-0.4386
Kurtosis	3.9539	4.0051	3.8863	4.6069
J-B Stat	180.2803	195.9569	241.0026	359.7925

daily data span about 13 years from January 1, 2003 to July 31, 2014, for a total of 3020 observations. We download these equity indexes from Datastream and then calculate log returns of the four indexes. For comparing purposes, the currency for the daily index in Japan, United Kingdom and Hong Kong is converted into US dollar based on their respective contemporary exchange rates.

We split the data into two equal parts: The first 1510 observations (training set), ranging from January of 2003 to October of 2008, are used to fit the mixture model, and the remained 1510 observations (testing set) are used to examine the out-sample predicting accuracy across the competing models. Table 2.5 displays the summary statistics for daily log-returns of MSCI index for the four markets. Over the 5 years between 2003 and 2008, UK market gave the highest average daily return while the median of daily return was relative higher in the US market. The skewness is negative for all, indicating higher probability in having extreme daily losses. Kurtosis for all four cases is greater than 3, implying a deviation from Normality.

Table 2.6 demonstrates the linear correlation coefficients for each pair. The strongest linear correlation is in HK-JP pair while the weakest one is in US-JP pair. Figure 2.4 shows pairwise scatter plot for each pair. Daily returns in each pair

Table 2.6: Linear correlation coefficients across four markets.

	UK	HK	JP	US
UK	1.0000	0.3436	0.2593	0.4186
HK		1.0000	0.4592	0.1620
JP			1.0000	0.0741
US				1.0000

appear to be positively correlated especially for US-UK, JP-HK and UK-HK pairs. Figure 2.4 also displays a violation of the elliptical multivariate distributions: Different mass in joint tails of the distribution, asymmetry and outliers could be observed directly from each pair. Figure 2.4 further confirms the existence of large amount of outliers in the lower left corner for UK-HK, JP-HK and JP-UK pairs. Simply choosing one most "appropriate" copula from a candidate set may be only helpful in covering one characteristic of the joint distribution. To fully take advantage of the flexibility of copula theory, we consider mixture copula model.

Spurious regression results will be generated if a pair of time series data is processed inappropriately (see Granger and Newbold 1974; Chen and Fan 2006). Hu (2006) also argues that data with conditional heteroscedasticity lead to underestimation of the degree of dependence, due to the clustering of large volatilities. Preliminary examination has indicated the existence of both autocorrelation effects and conditional heteroscedasticity in the daily returns for the four economies. To filter both effect, we specify an AR(1)-GARCH(1,1) model. The filtered monthly percentage changes will be substituted into the working mixture copula model.

Statistics from Jarque-Bera test for each series list in Table 2.5 also confirm the deviation from normality. To better capture the heavy tails on both sides, we follow Cai and Wang (2014) and assume that marginal distributions follow the student t-distribution.

We then fit the filtered daily returns in the four economies into the mixture copula model which includes Gaussian, Clayton and Gumbel. We implement the model average method, Cai and Wang (2014)'s penalized likelihood method and the BIC method to estimate the mixture copula model respectively. To compare the estimating performance across the three methods, we follow procedures that are similar in manner to Genest and Rivest (1993). As the true model is unknown to us, the ultimate purpose of this comparison is to examine whether the mixture copula model estimated via model average approach has relatively smaller estimation losses and captures the dependence structure satisfactorily for the six pairs .

Specifically, following Genest and Rivest (1993), we construct four 5×5 cross-classifications which are presented in Table 2.7.[3] The cross-classifications in the first column in Table 2.7 is for the observations and the other three are for the competing methods. Genest and Rivest (1993) does not specify the dimension of a cross-classification. The choice of this number should be a trade-off: one needs both enough number of groups to test contingent dependence and sufficient observations for each cell. Let G represents the table and $G(i, j)$ be the cell in the ith row and jth column, where $i, j = 1, ..., 5$. For cell $G(i, j)$, let u_i and v_i be the lower bounds for the cell, where u_i and v_i are defined as the $i/5$ and $j/5$ percentiles for the two series of observations, respectively. The cell boundaries for the two variables were taken as the order statistics of rank $[1510 * j/5]$ in the respective economy, for $j = 1, ..., 4$ and 1510 is the number of observations. Then a pair of observations (u, v) belongs to the cell $G(i, j)$ if $u_i < u \leq u_{i+1}$ and $v_i < v \leq v_{i+1}$. Thus, the number in $G(i, j)$ implies the number of times the daily return of market 1 is between the $i/5$ and the $(i+1)/5$ percentile of its range, and that of market 2 is within the $j/5$ and $(j+1)/5$

[3] Genest and Rivest (1993) suggest to use 7×7 cross-classifications. We have tried this and the results are quite similar with the 5×5 ones. The results for 7×7 are available upon request.

percentile of its range. For example, the figure recorded in the cell $(3, 2)$ indicates the number of times that daily percentage changes of the first economy is between the 40th (2/5) and the 60th (3/5) percentile of its range, while that of the second economy is within 20th (1/5) and 40th (2/5) percentile of its range. Thus, if the two economies are perfectly positively correlated, we should see that most observations lie on the principal diagonal. If they are perfectly negatively correlated, then most observations should lie on the diagonal which is perpendicular to the principal one. If they are independent to each other, then the number of observations in each cell should similar to each other.

The observed and predicted frequencies are displayed in Table 2.7. We interpret Table 2.7 by taking UK-HK pair as an example. For observed frequencies, the cell at the top-left represents the number of times when indexes in UK and Hong Kong market are both below the 20th (1/5) percentile of their respective ranges; that is, the number of times when both markets face downturn risk simultaneously. Correspondingly, the cell at the bottom-right shows the frequency that both daily returns are between the 80th (4/5) and 100th (5/5) percentile. Of the total 1510 observations, there are 114 times that daily returns of MSCI indexes in UK and Hong Kong are both lower than their 20th percentile. Correspondingly, during the period between January 2003 and October 2008, there are 99 times that daily returns in UK and Hong Kong market are both higher than their 80th percentile. Thus, UK and Hong Kong stock markets exhibit higher probability to co-move downward than to move up simultaneously. However, such a difference is not significant in the other 5 pairs: The difference between the up-left and the bottom-right cell, which represents the difference between the number of extreme events in both tails, is no larger than 8, exhibiting similar dependence structures in both tails.

We then evaluate the fit of the estimated mixture copula models by comparing

Pair	Observed frequency					MA predicted frequency					CW predicted frequency					BIC predicted frequency				
UK-HK	114	69	52	38	29	134	72	47	31	18	155	68	39	24	15	136	77	49	29	11
	62	64	66	69	41	72	77	66	52	35	68	81	66	51	36	77	78	67	52	29
	45	57	71	74	55	47	66	71	67	51	39	66	72	68	57	49	67	70	67	49
	44	56	69	55	78	31	52	67	78	73	24	51	68	79	80	29	52	67	78	77
	37	56	44	66	99	18	35	51	73	125	15	36	57	80	114	11	29	49	77	136
	Est.error: $Q_{MA} = 148.24, Q_{CW} = 219.90, Q_{BIC} = 231.08$																			
JP-HK	135	66	53	36	12	142	78	47	26	10	173	68	33	18	11	134	81	50	27	10
	75	69	58	64	36	78	85	70	48	22	68	90	68	46	31	81	85	71	47	19
	43	64	77	66	52	47	70	78	69	38	33	68	80	70	51	50	71	79	70	33
	30	63	69	65	75	26	48	69	89	70	18	46	70	89	79	27	47	70	91	67
	19	40	45	71	127	10	22	38	70	161	11	31	51	79	130	10	19	33	67	174
	Est.error: $Q_{MA} = 156.32, Q_{CW} = 181.42, Q_{BIC} = 227.84$																			
JP-UK	96	74	50	46	36	122	69	49	37	25	137	65	42	32	26	141	65	42	30	23
	81	61	48	54	58	69	71	64	55	43	65	76	65	53	43	65	71	63	55	48
	45	57	78	74	48	49	64	67	65	57	42	65	69	67	59	42	63	67	66	64
	43	65	58	66	70	37	55	65	73	73	32	53	67	75	75	30	55	66	73	78
	37	45	68	62	90	25	43	57	73	103	26	43	59	75	99	23	48	64	78	89
	Est.error: $Q_{MA} = 110.15, Q_{CW} = 174.94, Q_{BIC} = 185.06$																			
US-HK	80	70	57	53	42	85	70	60	50	36	85	70	60	50	36	85	70	60	50	36
	80	75	54	50	43	70	68	64	57	44	70	68	64	57	43	70	68	64	57	43
	49	47	70	70	66	60	64	64	63	52	60	64	65	63	51	60	64	64	62	51
	35	55	65	68	79	50	57	63	67	65	50	57	63	68	64	50	57	62	68	64
	58	55	56	61	72	36	44	52	65	106	36	43	51	64	107	36	43	51	64	107
	Est.error: $Q_{MA} = 180.89, Q_{CW} = 185.82, Q_{BIC} = 187.29$																			
US-UK	120	89	46	28	19	140	77	47	27	12	135	80	49	27	10	132	81	51	28	10
	66	73	63	52	48	77	83	69	49	25	80	84	70	48	20	81	84	70	48	19
	49	55	83	72	43	47	69	76	69	42	49	70	78	69	35	51	70	78	69	34
	43	47	64	72	76	27	49	69	86	71	27	48	69	89	68	28	48	69	90	67
	24	38	46	78	116	12	25	42	71	152	10	20	35	68	169	10	19	34	67	171
	Est.error: $Q_{MA} = 155.10, Q_{CW} = 241.31, Q_{BIC} = 241.09$																			
US-JP	68	71	60	56	47	76	67	61	54	44	76	67	61	54	44	76	67	61	54	44
	76	65	54	58	49	67	65	62	58	49	67	65	62	58	49	67	65	62	58	49
	52	51	68	60	71	61	62	63	62	55	61	62	63	62	55	61	62	63	62	55
	48	58	60	68	68	54	58	62	64	63	54	58	62	65	63	54	58	62	64	63
	58	57	60	60	67	44	49	55	63	91	44	49	55	63	91	44	49	55	63	91
	Est.error: $Q_{MA} = 67.49, Q_{CW} = 67.60, Q_{BIC} = 67.49$																			

Table 2.7: Goodness-of-fit: comparison between observed and predicted frequencies.

the three competing methods' estimated frequencies with the observed frequencies of all the cells. To compute the estimated frequencies, let $\widehat{C}$ denote our estimated copula. Then the estimated count in cell (u_2, v_2), for example, could be obtained by multiplying the probability $\widehat{C}(u_2, v_2) - \widehat{C}(u_2, v_1) - \widehat{C}(u_1, v_2) + \widehat{C}(u_1, v_1)$ to the sample size 1510. Let $G_{i,j}$ and $G_{i,j}^{MA}$ denote the frequency observed and the frequency estimated by model average in cell (i, j) respectively. Then, we define the estimation error as:

$$Q_{MA} = \frac{1}{k^2} \sum_{i=1}^{k} \sum_{j=1}^{k} (G_{i,j} - G_{i,j}^{MA})^2. \tag{2.4}$$

In our case, $k = 5$. By the same manner, we could calculate estimating errors for CW and the BIC method, which are respectively denoted as Q_{CW} and Q_{BIC}. The estimation errors are displayed at the bottom of each pair in Table 2.7. For the first five pairs, the model average method exhibits the smallest estimation errors, while the BIC method which relies on the comparison of BIC among Gaussian, Clayton and Gumbel copulas gives the largest estimation errors. In US-JP pair, the model average and BIC methods exhibit the same estimation loss and the superiority of CW is tiny. The empirical studies again show that the model average approach gives satisfactory estimates of dependence structures comparing with other competing methods.

We further examine the predicting performance of the three competing methods based on the daily observations between October 2008 and July 2014. We could obtain a similar table as Table 2.7, and the estimated frequencies are based on our estimated mixture copula model under MA, CW and BIC, respectively. The out-sample predicting errors are calculated in the same way we used to calculate the in-sample estimation losses. To save space, we only present the predicting errors of each competing method for each pair. The results are displayed in Table 2.8. It is obvious that model average method exhibits the smallest predicting errors compared

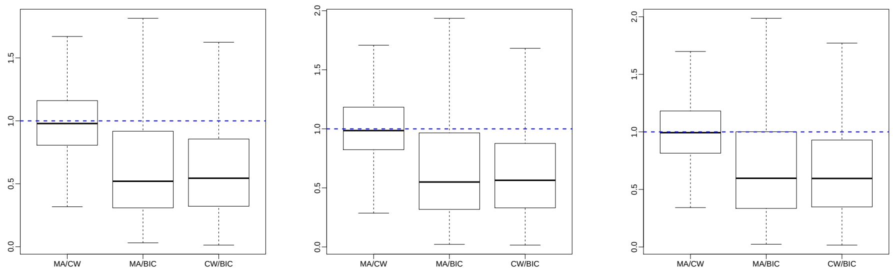
Table 2.8: Mean of out-sample predicting errors based on MA, CW and BIC.

	Q_{MA}	Q_{CW}	Q_{BIC}
UK-HK	70.8530	116.7931	130.1514
JP-HK	223.8020	246.5037	299.9155
JP-UK	171.5280	249.8285	247.5502
US-HK	80.7417	82.6520	83.6471
US-UK	61.0112	85.5270	102.5451
US-JP	96.1258	96.0680	96.1258

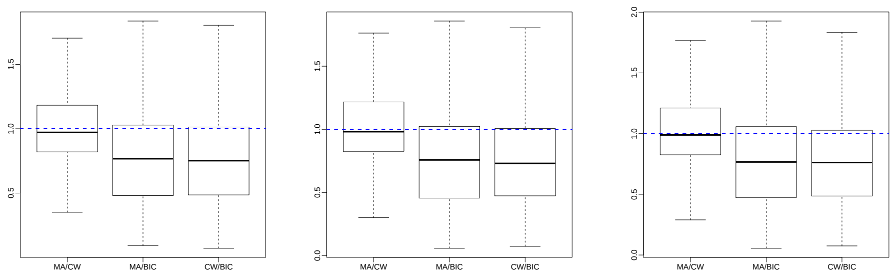
with CW and BIC methods, even though for the US-JP pair, these three methods have quite similar predicting performance. Thus, the model average method also provides relatively satisfactory predicting performance.

2.5 Conclusion

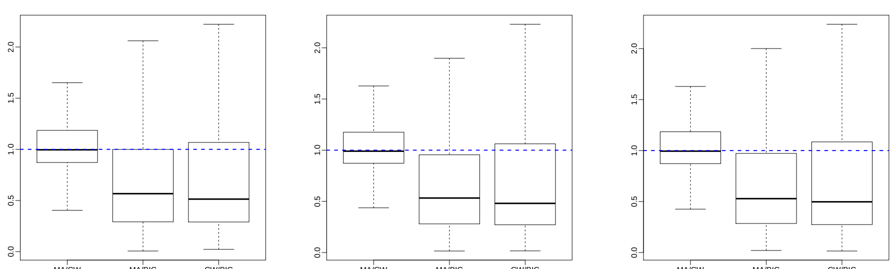
In this article we propose a model average method to estimate mixture copula models due to Hu (2006). Unlike the BIC method which selects only one individual copula based on the comparison of BIC, model average method estimates a mixture copula model based on choosing the optimal weights associated to the components in an averaging model. Simulation studies show that the model average method performs similarly to the penalized likelihood method proposed by Cai and Wang (2014) when observations are generated from copulas included in the working mixture copula model. However, when the working mixture copula model is misspecified, or observations are generated from copulas not included in the working mixture model, the model average method significantly outperforms Cai and Wang (2014)'s penalized likelihood method. In addition, mixture copula, no matter estimated by the model average method or the penalized likelihood method, exhibits superior estimating accuracy than the BIC method, indicating the flexibility of mixture copula model in nesting various dependence structures. An empirical example shows that the model average method provides satisfactory estimates of the dependence structures among four international stock markets. Thus, the model average method on mixture copula can be utilized by practioners in financial industry for portfolio diversification and risk management.



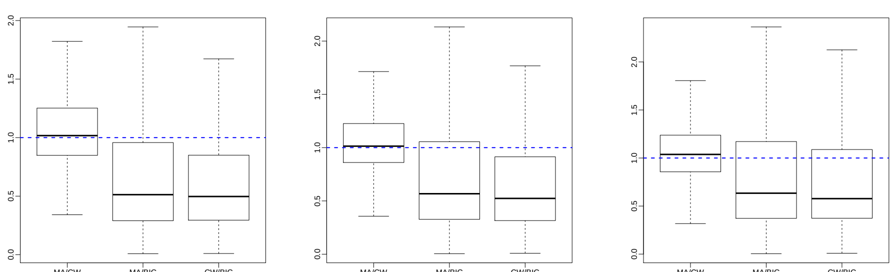
(a) $\omega_{Gaussian} = 0.5, \omega_{Clayton} = 0.5, \omega_{Gumbel} = 0$



(b) $\omega_{Gaussian} = 0.5, \omega_{Clayton} = 0, \omega_{Gumbel} = 0.5$

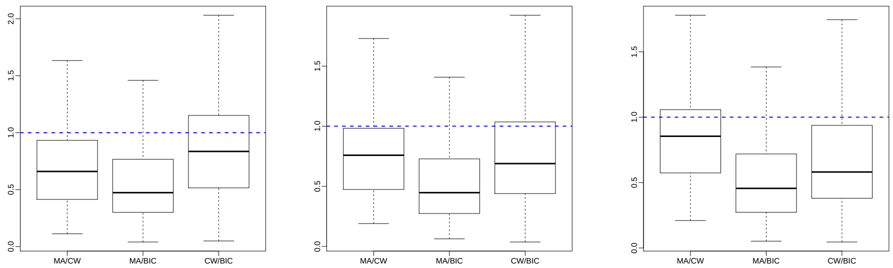


(c) $\omega_{Gaussian} = 0, \omega_{Clayton} = 0.5, \omega_{Gumbel} = 0.5$

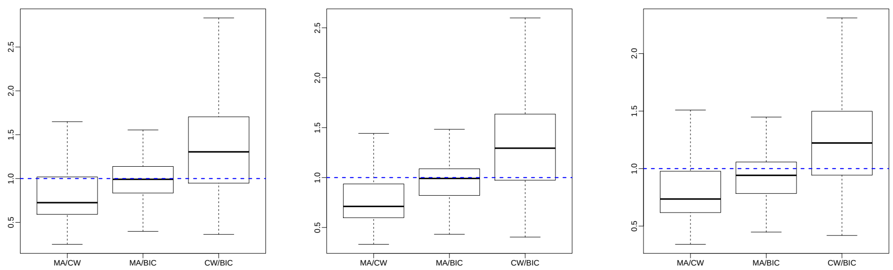


(d) $\omega_{Gaussian} = 1/3, \omega_{Clayton} = 1/3, \omega_{Gumbel} = 1/3$

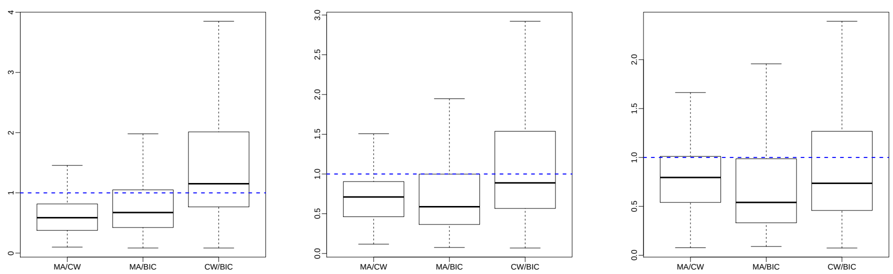
Figure 2.2: Boxplots for ratio of estimation losses in Type I simulation.



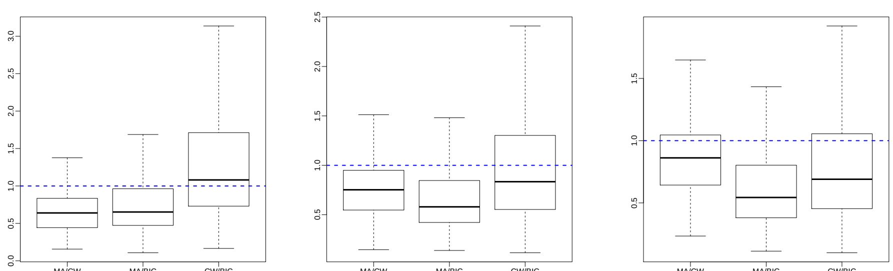
(a) $\omega_{Frank} = 0.5, \omega_{SJ} = 0.5, \omega_{Joe} = 0$



(b) $\omega_{Frank} = 0.5, \omega_{SJ} = 0, \omega_{Joe} = 0.5$



(c) $\omega_{Frank} = 0, \omega_{SJ} = 0.5, \omega_{Joe} = 0.5$



(d) $\omega_{Frank} = 1/3, \omega_{SJ} = 1/3, \omega_{Joe} = 1/3$

Figure 2.3: Boxplots for ratio of estimation losses in Type II simulation.

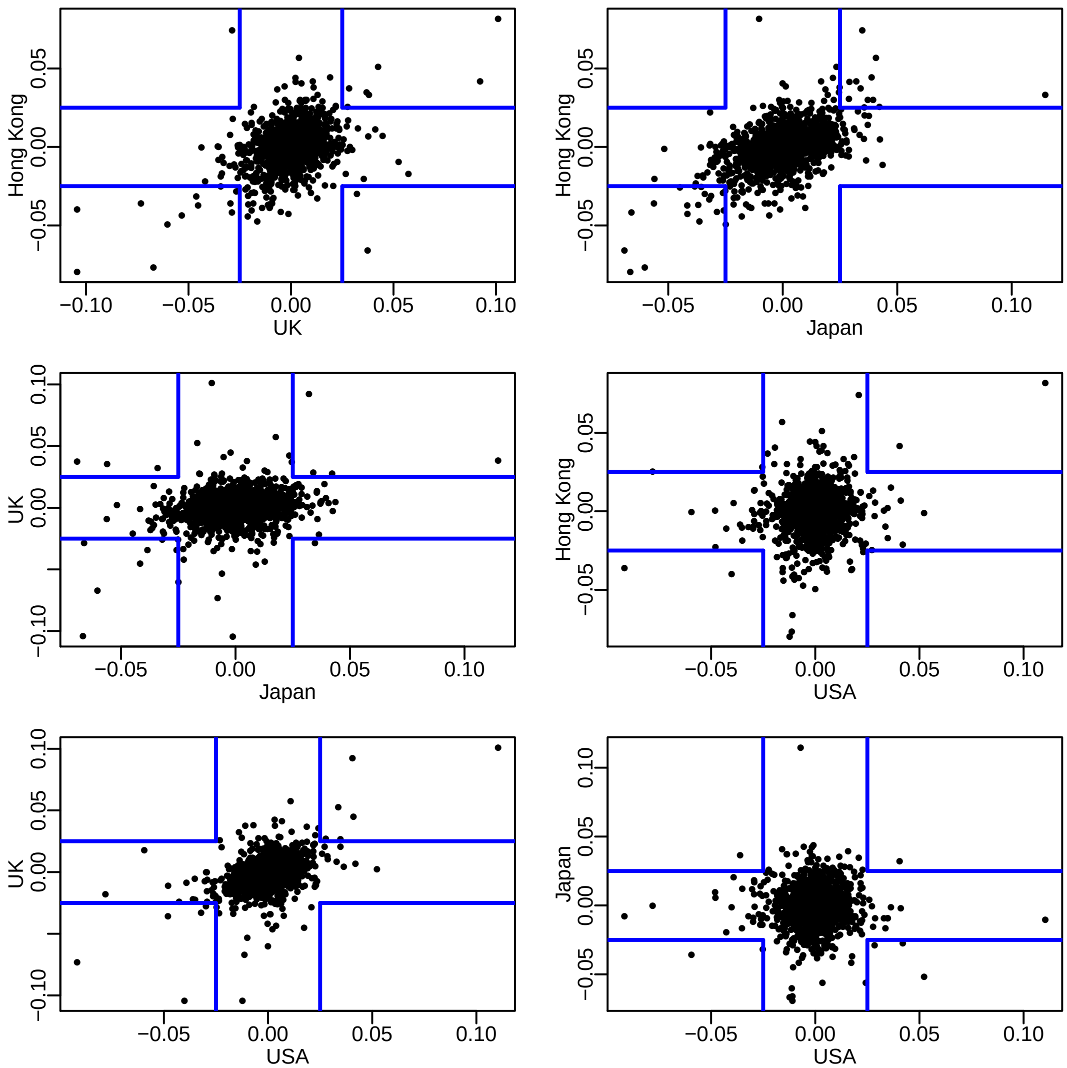


Figure 2.4: Scatter plots for daily return of MSCI Index.

3. ESTIMATING AVERAGE TREATMENT EFFECT OF THE JUSTICE REFORM IN VIRGINIA

3.1 Introduction

This paper evaluates the treatment effect of the justice reform in Virginia. From January 1, 1995, Virginia abolished discretionary parole for all violent crimes, reformed its sentencing systems by establishing the Truth-in-Sentencing (TIS) structure, and extensively enhanced the sentences on all violent offenders. Table 3.1 demonstrates the details: Punishment becomes tougher on repeat violent offenders with prior conviction greater than or equal to 40 years. For example, the median serving time for a first degree murder offender, who has prior conviction greater than or equal to 40 years, has increased to about 80 years during 1999 - 2001 compared to 14 years during 1988 - 1992. Even for offenders without prior crime records, the median serving years have almost doubled (for rape) or tripled (for first/second degree murder and robbery).

Table 3.1: Median Years Violent Offenders Served in Virginia

	FY 1988 - FY 1992			FY 1999 - FY 2001		
	No Prior	Prior<40	Prior>40	No Prior	Prior<40	Prior>40
1^{st} Degree Murder	12.4	14.1	14.7	35.3	51.5	80.3
2^{nd} Degree Murder	4.9	6.6	7.2	13.6	22.7	20.0
Rape	5.6	6.7	6.7	9.0	13.5	34.3
Robbery	1.4	2.2	2.3	3.7	6.2	7.3

Even though there are voluminous literatures about the relationship between incarcerating time and crime rate, the conclusions are far from consensus. For example, Myers (1980) argues that tougher punishment does not necessarily lead to substantial rehabilitative effect. Marvell and Moody (1996) also do not find convincing evidence to support the argument that the determinate sentencing laws (DSL) and abolition of parole significantly suppress the growth of prison population, because the estimated impacts on commitments vary state by state and there is little or no evidence that DSL affect crime rate. In a more related research, Sridharan, Vujic and Koopman (2003) use time series intervention analysis on the violent crime rates in Virginia. Using ARIMA and structural time series models and controlling the serial dependence between adjacent error terms, Sridharan, Vujic and Koopman find evidence that the parole abolition and TIS laws only had deterring effects on rape and murder, while the deterring effect for property crimes and aggravated assaults is not statistical significant.

In contrast to the findings above, McPheters et. al (1984) examine the deterrent response of robbery with a firearm in Arizona as penalties became tougher for using firearms. They conclude that offenders reduce the number of robberies with a firearm and the response is abrupt rather than gradual. Levitt (1996) argues that the elasticity is -0.4 for changes in violent crime with respect to changes in prison population, after controlling various covariates such as economics factors, percent changes in police staffing, racial composition and the age distribution. Kuziemko (2013) focuses on micro-data from Georgia and exploits the 1981 mass release in Georgia, a rare event, as a quasi-experiment to estimate how the elimination of discretionary prison release affects the social cost of crime. Similar to Levitt (1996), Kuziemko (2013) finds that "longer prison terms decrease recidivism. . . . the benefits of parole (the ability to ration prison resources based on recidivism risk and the creation of incentive) outweight the costs (lost incapacitation due to shorter prison terms)." She also argues that severely limiting the discretion of parole boards may leave some valuable information unused because parole boards "have access to information revealed after sentencing and therefore may be better than judges at forecasting inmates' expected recidivism risk." Shepherd (2002) checks the effect of TIS laws in deterring violent crimes using county level data. The empirical results demonstrate that TIS laws could deter violent crimes through increasing both the probability of arrest and the maximum imposed prison sentences. He specifies that TIS laws decrease rates of murder by 16%, rape by 12%, robbery by 24%, aggravated assault by 12%, and larceny by 3%. He also finds that under the TIS framework offenders tend to substitute to commit more property crimes such as burglaries and auto thefts for less severe punishments. Ross (2011) evaluates the impact of spatial variation in crime prevention policies on the migration of criminal activities into nearby locations. He uses the panel data approach and finds some deterring impact on crimes in states which adopt TIS laws and migration of crimes to neighboring states which do not establish TIS laws.

In this paper, we contribute to the literature by measuring how much the justice reform deters violent crimes in Virginia. We use a panel data approach proposed by Hsiao, Ching and Wan (2012, HCW hereafter) to conduct a counterfactual analysis. Unlike the popular Difference-in-Difference (DID) approach, HCW method does not suffer from sample selection bias problem since the method does not require the treatment unit and the control units follow parallel paths over time in the absence of treatment (Abadie, 2005, and Athey and Imbens, 2006). HCW suggest using outcomes of control units which do not receive treatment effects to predict the counterfactual path for the treated unit. The idea behind HCW method is that some common factors (they may be unobservable) drive both treatment units and control units over time so that cross-sectional units are correlated with each other.

The rest of the paper is organized as follows. In Section 2 we briefly review the history of parole system in the U.S. In Section 3 we describe HCW's method for estimating average treatment effects. Section 4 reports the empirical results obtained through HCW method. We conclude the paper in section 5. Formal definitions for violent/property crimes is given in Appendix A. We collect more detailed estimation results in Appendix B which is available from the authors upon request.

3.2 Parole and Truth-in-Sentencing in U.S.

Parole was introduced into U.S. in the 1800s and was mainly used to efficiently manage the population in prisons and prepare inmates for release.[1] By 1942, both the state and the federal governments have established their parole systems run by parole boards, whose discretionary power of releasing inmates was huge during the 1970s, a period in which judges only provided indeterminate sentences reflecting a range of minimum and maximum incarcerating years. Through structured decision-making process, a parole board might release an inmate as long as he/she served the minimum convicted sentence after evaluating his/her potential recidivism risk. For example, in Alaska, where the discretionary power of parole board still exists, the discretionary parole is defined as the following:

> *According to Alaska Stat.* **S***33.16.900, "discretionary parole" means the release of a prisoner by the board before the expiration of a term, subject to conditions imposed by the board and subject to its custody and jurisdiction; "discretionary parole" does not include "special medical parole".*

The percentage of U.S. prisoners released on parole reached a high level of 69% in 1977.[2] As shown in Figure 3.1, even in the early 1980s, the discretionary paroles still account for more than half of all the prisoners released in the United States. To

[1] For details, the first chapter of *Parole: Then & Now* by Texas Senate Research Center provides an excellent reference.

[2] Trends in State parole, 1990-2000, Bureau of Justice Statistics, 2001.

control the number of release on parole, in 1984, *The United States Federal Sentencing Guidelines* abolished parole for those committed federal crimes and limited early release from prison for good behavior on the federal level, as more and more states moved to determinate sentencing system and mandatory supervised release during the 1980s. This abolition narrowed the discrepancy between the sentenced years and the actual years served. Consequently, as shown in Figure 3.1, the percentage of discretionary parole continue to decline after 1980 while mandatory parole began to account for a larger fraction.[3] Correspondingly, there was an enhancement in punishment for offenders. In 1990, the mean value of the maximum sentenced years for the most serious offense was 99 months and the actual mean serving time is 43.8% of the sentenced terms. By 1999, this percentage has increased to 55%. Also, in 1990, the percentages of served sentence for violent crimes (murder, manslaughter, rape and robbery) are all less than 46%. But in 1999, all violent offenders have to serve more than 50% of their sentenced terms.[4]

To assure that felons serve a substantial portion of their sentences, the federal government launched Violent Offender Incarceration and Truth-in-Sentencing (VOI/TIS) Incentive Formula Grant Program. This program funds states to build or expand current correctional facilities for confinement of persons convicted of violent crimes, and the funded states should warrant violent felons to serve at least 85% of their sentenced terms. These measures led to a sharp decline in the percentage of discretionary paroles in the 1980s. The percentage of discretionary paroled prisoner declined to 38% in 1989 from 55% in 1980, while mandatory paroles increased from

[3]According to Virginia Department of Correction, mandatory parole is " the automatic release of an offender six months before completion of his or her sentence." Unlike discretionary parole, parole board members might impose some special conditions for this type of parole but will not make the parole decision through voting.

[4]Table 5 of Trends in State Parole, 1990-2000, Bureau of Justice Statistics, 2001.

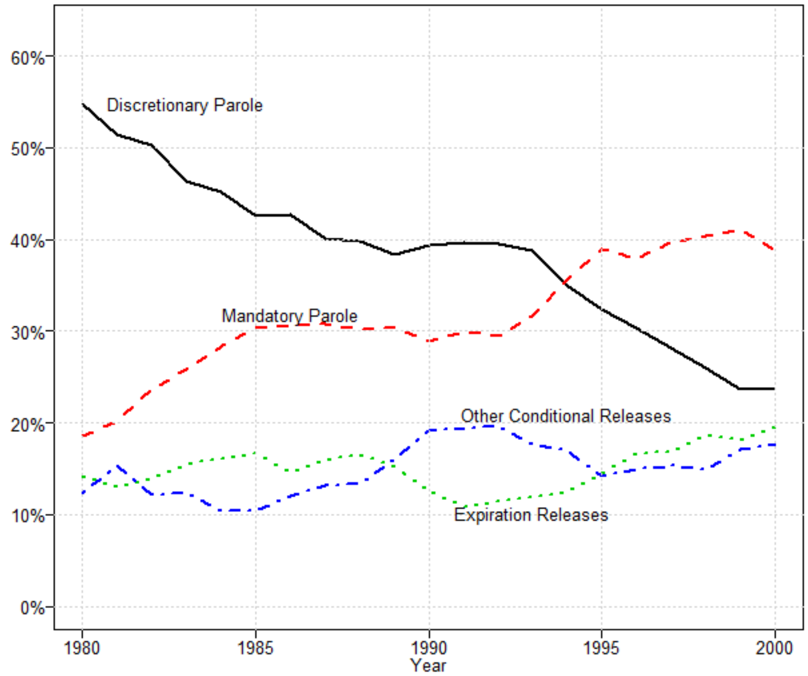


Figure 3.1: Percentage of Releases from State Prison by Forms: 1980 - 2000

19% to 30%.[5] The percentage of mandatory paroles surpassed that of discretionary paroles in 1994 and continued to increase. For Virginia, Figure 3.2 compares the actual serving percentages of sentenced terms by various types of offenders before and after the justice reform. Before 1995, except for those who convicted of rape or forcible sodomy, offenders on average served less than 40 percent of their sentences before being eligible for parole. After 1994, all offenders had to serve at least 85 percent of their sentences due to the TIS laws. We will then estimate whether enhancement in punishment and longer incarceration would effectively decrease violent crimes in Virginia.

[5]Trends in State parole, 1990-2000, Bureau of Justice Statistics, 2001.

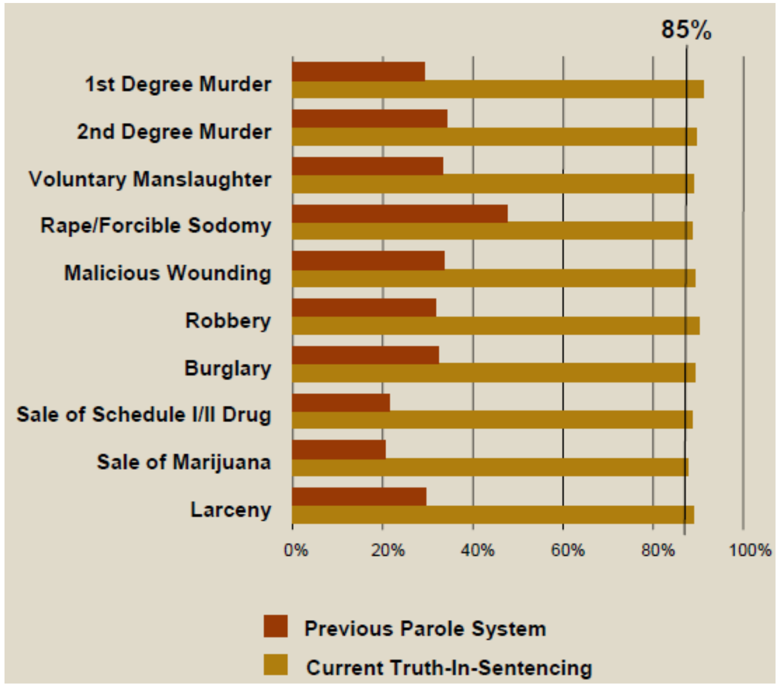


Figure 3.2: Percentage of Prison Sentence Served in Virginia

3.3 The Estimation Method

In this section we briefly review the estimation method proposed by Hsiao et al. (2012). Let y_{it} denotes the state i's violate crime rate at time t and the justice reform happens at time T_1. In the absence of any treatments, Hsiao et al. (2012) consider the case that

$$y_t = a + Bf_t + u_t, \tag{3.1}$$

for $t = 1, ..., T_1$, where $y_t = (y_{1t}, ..., y_{Nt})'$, $a = (a_1, ..., a_N)'$, f_t is a $K \times 1$ vector of common factors (they may be unobservable) that affect crime rates, B is a $N \times K$ matrix of factor loading, $u_t = (u_{1t}, ..., u_{Nt})'$ is a vector of idiosyncratic error. Following Hsiao et al. (2012), we assume that only the first unit, Virginia, receives the justice reform at time T_1, and all other units are not affected by the Virginia's policy intervention. Let y_{1t}^1 and y_{1t}^0 denote the violent crime rates of Virginia with and without the treatment, respectively. Given that there is an intervention at time T_1, we are interested in estimating the average treatment effects $\Delta_1 = E(y_{1t}^1 - y_{1t}^0)$. The difficulty is that we can not observe y_{1t}^0 for $t \geq T_1+1$. Hsiao et al. (2012) suggest using control group's y_{jt}, $j \geq 2$, to estimate y_{1t}^0. This can be done by replacing f_t by $\tilde{y}_t = (y_{2t}, ..., y_{Nt})'$ in the first unit's equation $y_{1t} = a_1 + b_1' f_t + u_{1t}$ to obtain

$$y_{1t} = \gamma_1 + \tilde{y}_t' \gamma + v_{1t} \tag{3.2}$$

for $t = 1, ..., T_1$, where v_{1t} satisfies $E(v_{1t}) = 0$ and $E(v_{1t}\tilde{y}_t) = 0$. Let $\widehat{\gamma}_1$ and $\widehat{\gamma}$ denote the least square estimators of γ_1 and γ based on (3.2) using the pre-treatment period data, then we estimate the counterfactual outcome of y_{1t}^0 by

$$\widehat{y}_{1t}^0 = \widehat{\gamma}_1 + \tilde{y}_t' \widehat{\gamma} \tag{3.3}$$

for $t = T_1 + 1, ..., T$. The average treatment effect is estimated by

$$\widehat{\Delta}_1 = \frac{1}{T_2} \sum_{t=T_1+1}^{T} \left(y_{1t} - \widehat{y}_{1t}^0 \right), \tag{3.4}$$

where $T_2 = T - T_1$.

Under quite mild conditions including that $rank(B) = K$, Li and Bell (2014) derive the asymptotic distribution of $\widehat{\Delta}_1$ as follows:

$$\sqrt{T_2}(\widehat{\Delta}_1 - \Delta_1) \xrightarrow{d} N(0, \Sigma),$$

where $\Sigma = \Sigma_1 + \Sigma_2 + \Sigma_3$. $\Sigma_1 = \eta E(x_t)' V E(x_t)$, $\eta = \lim_{T_1, T_2 \to \infty} T_2/T_1$, Σ_2 and Σ_3 are the asymptotic variances of $T_2^{-1/2} \sum_{s=T_1+1}^{T} v_{1s}$ and $T_2^{-1/2} \sum_{s=T_1+1}^{T} (\Delta_{1s} - E(\Delta_{1s}))$, respectively.

In our case, HCW method is more appropriate than the popular DID approach for two reasons. First, DID approach assumes that there is no sample selection bias issue while HCW method does not need to do so. Sample selection bias is a big concern in the program evaluation literature (e.g., Heckman and Hotz, 1989). As mentioned at the beginning of the paper, in Virginia the public safety deteriorated after mid-1980s and the violent crime rate reached its historical high in 1993. The severe situation called for tougher punishment to deter violent crimes. The justice reform in Virginia thus should not be regarded as a random arrangement among all the other states. Hence, HCW method is more appropriate for our analysis. Second, the fundamental assumption for DID approach is the common trend for both treatment and control groups. HCW method does not require such an assumption, while a violation of this assumption would make DID approach invalid. Simulation results from Li and Bell (2014) compare the MSEs of DID and HCW in predicting the average treatment effect and show that when the common trend assumption is violated, the ratio of MSE for DID and HCW is uniformly greater than 1 at various levels of T_1 and T_2, indicating that HCW outperforms DID in estimating the treatment effect.

One can also use the synthetic control method proposed by Abadie, Diamond and Hainmueller (2010, ADH hereafter) which usually give similar estimation result as that of HCW when the treatment and the control units are drawn from a random sample. However, when the sample are not randomly selected, ADH method can suffers sample selection bias problem. This is because the weights attached to the control units are restricted to adding up to one.

Nevertheless, we still would like to stress that the comparison between DID and HCW method discussed above should be interpreted with caution. Even though more flexible, HCW method also has limitation: it requires longer time series data to estimate the counterfactual outcomes. When there are many units in both treatment and control groups or the pre-treatment period is quite short, HCW method would not be applicable while DID is still likely to work well. Hence, HCW method should be viewed as a complement to DID.

3.4 Empirical Results

3.4.1 Data

Sabol et al. (2002) summarize different states' policies for parole and early release.[6] According to their discussions, 17 states — Alabama, Arkansas, Colorado, Hawaii, Maryland, Massachusetts, Nebraska, Nevada, New Hampshire, New Mexico, South Dakota, Rhode Island, Texas, Utah, Vermont, West Virginia, and Wyoming — still keep parole system for certain offenders and have not established stringent TIS laws.[7] Some of them require certain minimum incarcerating periods. For example, Texas and Maryland demand all felons to serve at least 50% of their sentences while Arkansas requires certain offenders to serve 70%. Colorado separates violent offenders by the number of time for prior violent convictions: felons with two prior violent convictions to serve 75% and with one prior violent conviction, 56%. Some even still keep discretionary power of correctional committee for parole. For example, Rhode Island still grants discretionary parole on inmates who have been imprisoned for more than six months and who have served no less than one-third of sentenced terms.

We collect both violent crime rates and property crime rates per 100,000 population from FBI's Uniform Crime Report. The data, collected by the Uniform Crime Report, cover the period between 1960 and 2010. The violent crime includes murder,

[6] For details, see page 20, Chapter 2 of *The Influences of Truth-in-Sentencing: Reforms on Changes in States' Sentencing Practices and Prison Populations.*

[7] Utah does not have Truth-in-Sentencing statutes but received federal grant funding on the basis of its Truth-in-Sentencing practices.

Table 3.2: Descriptive Statistics: Violent/Property Crimes Rates (1960 - 2010)

	Aggregated Violent Crime Rates				Aggregated Property Crime Rates			
	min	max	mean	S.D	min	max	mean	S.D
U.S.	72.48	324.81	209.60	70.08	1726.3	5353.3	3896.61	1028.98
Virginia	42.47	182.38	126.66	37.70	1469.20	4349.10	3144.75	792.60
Alabama	38.36	228.66	141.82	55.23	985.50	4521.40	3276.42	1117.66
Arkansas	36.64	191.36	114.34	42.37	926.40	4581.70	3113.49	1079.20
Colorado	74.15	223.01	149.06	41.47	2035.10	6821.40	4601.98	1438.82
Hawaii	15.85	233.65	117.77	55.32	2276.50	7182.80	5010.10	1282.62
Maryland	49.99	491.42	316.29	117.07	1518.80	5777.70	4143.38	1085.49
Massachusetts	26.71	301.96	161.75	70.42	1170.30	5635.30	3460.32	1217.11
Nebraska	22.01	113.91	80.24	26.21	1177.80	4162.50	3151.82	882.39
Nevada	95.35	547.79	305.23	111.14	2774.70	7996.00	5142.41	1419.44
New Hampshire	7.09	80.68	46.42	22.15	676.40	4499.80	2452.77	986.28
New Mexico	55.60	237.42	157.98	51.66	2062.40	6053.20	4593.11	1174.94
Rhode Island	15.38	157.47	103.74	40.59	1833.30	5524.10	3778.88	1091.71
South Dakota	14.97	91.11	45.63	19.78	1065.20	3116.30	2179.93	572.42
Texas	48.03	355.17	199.60	77.65	1970.50	7365.10	4596.80	1426.20
Utah	28.63	118.26	84.27	25.07	2047.90	5762.00	4297.66	1013.13
Vermont	4.87	72.19	35.60	15.50	796.70	5115.00	2793.35	1169.34
West Virginia	20.49	80.67	55.95	17.02	599.20	2639.90	1914.02	642.95
Wyoming	24.93	82.72	50.54	13.80	1564.70	4701.80	3285.44	822.14

rape and robbery. The property crime includes burglary, larceny and motor vehicle theft. Table 3.2 displays descriptive statistics for both national level and state level violent crime rate and property crime rate. We find that, even though lower than the national average level, the average violent crime rate in Virginia is only lower than that of seven states — Alabama, Colorado, Maryland, Massachusetts, Nevada, New Mexico and Texas, but significantly higher than that of other ten states in the control group. For property crime, Virginia has lower average property crime rate during 1960 - 2010: Only Arkansas, New Hampshire, South Dakota, Vermont and West Virginia have lower average property crime rates.

3.4.2 Aggregated Violent Crime

We check the treatment effect on the aggregated violent crime rates first. Implementing the method by Hsiao et al. (2012), we regress the violent crime rates in Virginia before 1995 on violent crime rates of various combinations of the 17 states in the control group. For the 17 states, there are $2^{17} - 1 = 131071$ different combinations. We categorize these combinations into 17 groups by the number of the control states selected, and then pick up one "optimal" model from each of the 17 groups by comparing their adjusted R^2 coefficient. For the 17 selected models, we finally choose the one gives the smallest AICC value.[8]

Based on HCW approach, the AICC method selects 8 states (as the control group): Alabama, Arkansas, Colorado, Maryland, Nevada, Rhode Island, Utah and West Virginia. These states are geographically scattered. The detailed estimation result is given in Table 3.3. The OLS coefficients for all selected states are significant at 5% level. The adjusted R^2 under this specification is about 0.98, indicating that the unobserved common factors which impact all the 8 selected states as well as Virginia could explain about 98 percent of the variation in violent crime rates in Virginia between 1960 and 1994. In fact, the 8 selected states, as well as Virginia, have similar evolving patterns of violent crime rates to the nation's. In the book "*The Crime Drop In America*", Blumstein gives some specifications about these patterns: (i) The remarkable increase in violent crime from 1960 to the mid-1970s is "a result of the decline in perceived legitimacy of America social and governmental authority during this turbulent period, which contained the civil rights movement and the strident opposition to the war in Vietnam." (ii) The uptrend during late 1970s to early 1980s

[8]We have also used the AIC standard for state selections and actually get quite similar results. AICC is a more conservative model selection standard and prefers more parsimonious model. Thus, in the remaining part of the paper, the optimal models are all selected under AICC.

and the small decline to 1985 are due to “the movement of the baby-boom generation into and then out-of the high-crime ages of the late teens and early twenties.” (iii) Violent crime rate became stabilized during the mid-1980s, but increased again due to drug-dealings. Based on these argument, Blumstein compiles some factors such as changing demographics, policing and community policing, growth in prison expansion and expanding economy as factors that impact every state. Based on Hsiao et al. (2012), we assume that the impact of the factors has been mirrored by the variations of violent crime rates in these states.

Table 3.3: Aggregated Violent Crime Rates: Weights for control group

	Panel A (no time trend)				Panel B (with time trend)			
	Estimate	Std. Error	t value	Pr(>|t|)	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	12.9183	4.8417	2.67	0.0130	13.2864	5.0506	2.63	0.0144
AL	0.6232	0.1440	4.33	0.0002	0.6402	0.1553	4.12	0.0004
AR	-0.2975	0.1364	-2.18	0.0384	-0.2898	0.1407	-2.06	0.0500
CO	0.1437	0.0514	2.79	0.0096	0.1327	0.0619	2.14	0.0420
MD	0.1058	0.0257	4.11	0.0004	0.1076	0.0267	4.02	0.0005
NV	-0.1497	0.0323	-4.63	0.0001	-0.1487	0.0330	-4.50	0.0001
RI	0.2094	0.0684	3.06	0.0050	0.2294	0.0920	2.49	0.0196
UT	0.8250	0.1931	4.27	0.0002	0.8008	0.2097	3.82	0.0008
WV	-0.6003	0.2879	-2.08	0.0471	-0.5789	0.3000	-1.93	0.0651
Trend					-0.1967	0.5938	-0.33	0.7433

Figure 3.3 gives a visual demonstration about how good the selected 8 states simulate the actual path of the violent crime rate in Virginia between 1960 and 1994. The black line represents the actual path of the violent crime rate per 100,000 population in Virginia. Comparing with the actual path, we find that our synthetic path (in red line) almost perfectly matches the actual one especially for periods during 1960 – 1975 and 1984 – 1994. For the period between 1976 and 1983, our synthetic path also captures the general patterns of the actual one.

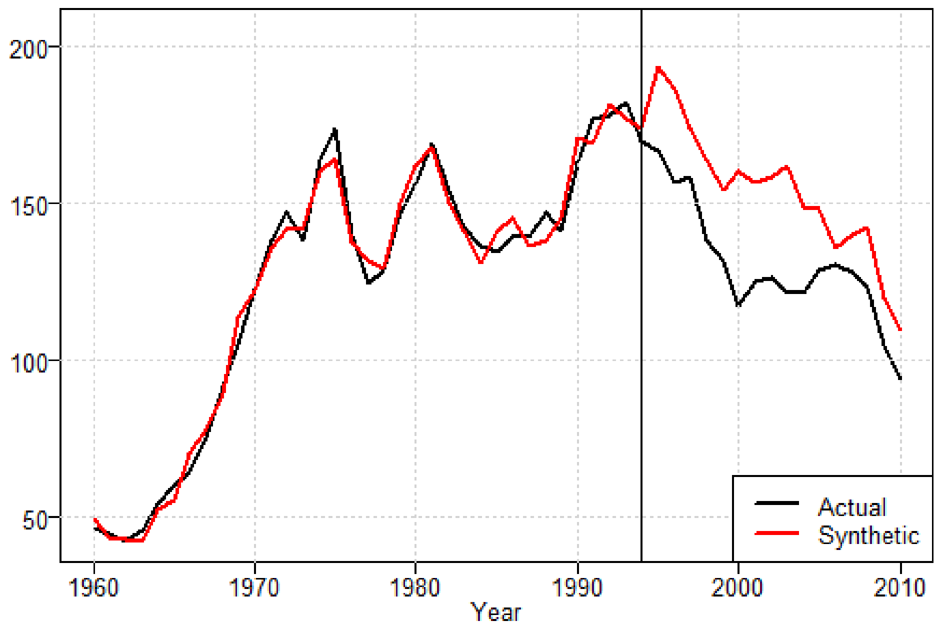


Figure 3.3: Violent Crime Rates in Virginia: Actual Path and Predicting Path

The actual path and the synthetic path begin to diverge abruptly after 1994. Figure 3.3 shows that synthetic path jumps at 1995 and is above the actual path uniformly from then on. Overall, the average treatment effect over the 16 years from 1995 to 2010 is -23.83 (a 16% drop), indicating that the violent crime rate would have been on average 23.83 per 100,000 population higher in the absence of the reform. The calculated t statistic is -9.27, which is significant at any conventional level. The estimated yearly treatment effects are documented at Table 3.4.

3.5 Robustness Test

To check the robustness of the treatment effect, we conduct several robustness tests. Similar to Abadie and Gardeazabal (2003) and Abadie, Diamond and Hainmueller (2010), we firstly conduct a placebo test by choosing a state randomly from

Table 3.4: Aggregated Violent Crime Rates: Treatment effect between 1995 - 2010

	Panel A (no time trend)			Panel B (with time trend)		
Year	Actual	Prediction	Trt.Effect	Actual	Prediction	Trt.Effect
1995	166.52	193.51	-26.99	166.52	192.89	-26.37
1996	156.76	186.51	-29.74	156.76	184.86	-28.10
1997	158.76	174.15	-15.39	158.76	172.17	-13.41
1998	138.46	163.79	-25.33	138.46	160.65	-22.19
1999	131.81	154.55	-22.74	131.81	151.65	-19.84
2000	117.43	160.25	-42.82	117.43	157.47	-40.04
2001	124.97	156.53	-31.56	124.97	153.44	-28.46
2002	126.07	158.56	-32.49	126.07	155.37	-29.29
2003	121.33	161.88	-40.55	121.33	158.42	-37.09
2004	121.70	148.91	-27.21	121.70	144.96	-23.26
2005	128.45	148.50	-20.06	128.45	144.91	-16.47
2006	130.50	135.84	-5.34	130.50	132.42	-1.92
2007	128.21	140.19	-11.98	128.21	136.60	-8.39
2008	123.59	142.49	-18.90	123.59	139.04	-15.46
2009	105.09	119.84	-14.74	105.09	115.68	-10.59
2010	94.46	109.83	-15.37	94.46	104.44	-9.98
Mean	129.63	153.46	-23.83	129.63	150.31	-20.68
S.D	18.66	21.71	10.27	18.66	22.72	10.62

the control group instead of using Virginia. We also consider the exogeneity of the states in the control group by removing Virginia's neighboring states. Finally, we explicitly introduce a time trend variable on the right hand side of equation (2) to reduce the near multicollinearity concern and examine whether a time trend would affect our conclusion.

3.5.1 Placebo Tests

We firstly carry out a series of placebo studies by conducting the HCW approach iteratively. In each iteration, we take Virginia as a control unit and assume that one of the 17 states in the control group introduces the justice reform on January 1, 1995. We then calculate the "treatment effects" in each placebo run for those states where no treatment actually takes place.

Figure 3.4(a) gives the results of the placebo tests. Instead of plotting both the actual and the synthetic paths respectively, we draw the differences (gap lines) of violent crime rates between the actual paths and the synthetic counterparts. Figure 3.4(a) suggests that the violent crime rate drops immediately after 1994 and the justice reform had a significant and uniform treatment effect on the violent crime rates in Virginia. The fit for the period before 1995 is also very good with mean squared prediction error (MSPE) approximately equals to 25, while the median of MSPE of all the 18 iterations is about 33. This indicates that Virginia has better pre-treatment fit than half of the control units. However, Figure 4(a) also suggests that some states have bad pre-treatment fit. Of all the 17 states in the control unit, Maryland has the worst fit for the pre-treatment period, with a MSPE of 788. Since 1964, Maryland's violent crime rate has been uniformly higher than the national rate and "a very likely explanation for Maryland's high violent crime rate may be its sentencing system that is too lenient, especially for violent crimes."[9]

If the pre-treatment synthetic path deviates significantly from the actual one, the gap between the two paths after 1995 should be interpreted as the lack of fit and would not give much valuable information about the robustness of the results. For this reason, as Abadie et al. (2010) do, in Figure 3.4(b) we exclude those states which have pre-treatment MSPE greater than 5 times of Virginia's MSPE. It is clear that the Virginia gap line is almost the lowest for the entire post-treatment period compared with the 11 remained states. In Figure 3.4(c) we further lower the threshold to exclude states with pre-treatment MSPE that is 1.5 times greater than Virginia's MSPE. Figure 3.4(c) displays that Virginia's gap line is obviously the

[9] *Why Maryland Needs Truth-in-Sentencing*, Statement before the Judiciary Committee of the Maryland House of Delegates Senate, David B. Muhlhausen.

lowest compared with the remained 7 states. This unusual gap line provides solid evidence that the drop in violent crime rates in Virginia after 1995 is not driven by statistical coincidence.

3.5.2 Out-of-Sample Prediction

We assume that the justice reform happened 10 years earlier and re-run the regression model specified in equation (2) using the new pre-treatment data. If the results are robust, we should not observe significant treatment effect between 1985 - 1994. Figure 3.5 indicates that a remarkable deviation between the actual and the synthetic paths still emerges only after 1994, even though we assume that the treatment happens 10 years earlier. According to Figure 3.5, we still obtain a good fit during the pre-1985 period. During the 1985 - 1994 period, the synthetic path follows the pattern of the actual path, and the mean value of the estimating errors during this period is -0.29. After 1994, there is an abrupt jump in the synthetic path and it is uniformly above the actual path during the post-1994 period, with the mean of the estimated treatment effect equals to -32.92. Again this result supports that the decline of violent crime in Virginia is not a statistical coincidence. The detailed regression results and the specific treatment effect information are demonstrated at Table 3.5 and Table 3.6.

3.5.3 Exogeneity of the States in Control Group

Controlling for urban area-year fixed effects, Ross (2011) finds some migrating effects of violent/property crimes from the state which adopts TIS laws to its neighboring non-adopting states. This is intuitive since potential offenders would tend to commit crimes at the lowest “expenses”. Under HCW’s settings, all states included in the control group are assumed to be exogenous to the treatment. For a robustness check, we remove two neighboring states of Virginia — Maryland and West Virginia

Table 3.5: Robustness Test: Weights for Control Group Based on Truncated Data

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	7.3449	3.3928	2.16	0.0449
AL	0.6774	0.1499	4.52	0.0003
NM	0.2095	0.0649	3.23	0.0049
AR	-0.4823	0.1546	-3.12	0.0062
CO	0.1132	0.0518	2.19	0.0431
MD	0.1332	0.0181	7.36	0.0000
NV	-0.1883	0.0255	-7.37	0.0000
UT	0.6423	0.1887	3.40	0.0034

— from the control group and re-run the model. This time, the selected optimal model includes 5 states: Alabama, Colorado, Nevada, Rhode Island and Utah. The estimated treatment effect now shows that, on average, the violent crimes would decrease by about 32 per 100,000 population, or 20% drop. This result is in line with the previously obtained 16% drop. Thus, our estimated mean of the treatment effect of the reform in Virginia is between 16 and 20%. See Table 3.7 and Table 3.8 for details.

3.5.4 Time Trend

Bai, Li and Ouyang (2013) extend Hsiao et al. (2012) to I(1) process and prove that HCW's approach still gives consistent estimates for weights. If both y_{1t} and $\tilde{y}_t$ in equation (2) are I(1) process and ν_{1t} is I(0), y_{1t} and $\tilde{y}_t$ are cointegrated. However, since some or all components in $\tilde{y}_t$ may contain drift terms, these series will be dominated by their non-zero drift terms. One way to estimate the cointegrated model is to add a time trend regressor to capture the time trend components of the I(1) regressors. We show the regression results with an explicitly time trend term in Panel B of Table 3.3. Comparing with the results in Panel A, the OLS estimated weights are almost the same. The coefficient for time trend is not significantly different from

Table 3.6: Treatment Effect of Violent Crime Rates Based on Truncated Data

Year	Actual	Prediction	Trt.Effect
1985	134.53	147.39	-12.87
1986	139.31	148.92	-9.61
1987	139.19	134.17	5.02
1988	147.40	129.29	18.11
1989	141.23	143.12	-1.89
1990	163.01	167.00	-3.99
1991	176.79	161.97	14.82
1992	178.12	188.29	-10.17
1993	182.38	178.34	4.03
1994	170.07	176.43	-6.36
1995	166.52	202.71	-36.19
1996	156.76	198.23	-41.46
1997	158.76	184.51	-25.75
1998	138.46	177.06	-38.60
1999	131.81	167.63	-35.82
2000	117.43	170.05	-52.63
2001	124.97	162.17	-37.20
2002	126.07	161.57	-35.49
2003	121.33	161.76	-40.43
2004	121.70	158.71	-37.01
2005	128.45	158.25	-29.81
2006	130.50	147.80	-17.30
2007	128.21	150.97	-22.76
2008	123.59	153.30	-29.72
2009	105.09	134.31	-29.22
2010	94.46	111.77	-17.31
Mean	140.24	160.61	-20.37
S.D	22.86	21.20	18.81

zero. The adjusted R^2 is still very high. Similarly, in Panel B of Table 3.4, the estimated treatment effects are documented and the estimated average treatment effect is around -20, which is also very close to -23 of the result in Panel A. Thus, our results are robust even after considering the time trend explicitly.

Table 3.7: Robustness Test: Weights for Control Group (excluding WV, MD)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.5100	5.1831	0.29	0.7729
AL	0.2814	0.0746	3.77	0.0007
CO	0.1521	0.0676	2.25	0.0322
NV	-0.1893	0.0300	-6.32	0.0000
RI	0.2603	0.0705	3.69	0.0009
UT	1.2292	0.2059	5.97	0.0000

Table 3.8: Robustness Test: Treatment effect between 1995-2010 (excluding WV, MD)

Year	Actual	Prediction	Trt.Effect
1995	166.52	184.05	-17.53
1996	156.76	182.05	-25.28
1997	158.76	181.92	-23.16
1998	138.46	173.66	-35.20
1999	131.81	158.21	-26.41
2000	117.43	166.60	-49.17
2001	124.97	164.87	-39.89
2002	126.07	162.48	-36.41
2003	121.33	167.65	-46.32
2004	121.70	162.17	-40.47
2005	128.45	156.65	-28.20
2006	130.50	144.27	-13.77
2007	128.21	153.38	-25.16
2008	123.59	155.51	-31.92
2009	105.09	144.13	-39.03
2010	94.46	138.80	-44.34
Mean	129.63	162.27	-32.64
S.D	18.66	13.69	10.38

3.5.5 Violent Crime In Details

We check the treatment effects on murder, rape and robbery separately. The formal definitions of each kind are given in the Appendix B. Figure 3.6 implies that there are clear treatment effects on the three types of violent crime. It is notable that the projected paths closely follow the actual ones for rape and robbery (with adjusted $R^2 = 0.97$ and 0.98 respectively) before the justice reform in 1995. For rape, Table 3.11 shows that five states — Alabama, Hawaii, Maryland, Nebraska and West Virginia — are selected and except for the coefficient of Alabama, which is significant at 10% level, all the others are significant at 5% level. The average of the actual rape rate per 100,000 population from 1995 to 2010, shown in Table 3.12, is 24.14 while the estimated average of rape rate per 100,000 population without the 1995 justice reform is 25.88, which is a 7% decline.

As for robbery, Table 3.13 shows that six states are included: Alabama, New Mexico, Maryland, Nevada, Utah and Vermont. The adjusted R^2 coefficient is 0.98, implying a nearly perfect fit to the actual path. Table 3.14 demonstrates that the average treatment effects are -13.01, or a 12% decline.

Gap between the actual and the synthetic paths for murder from 1995 to 2010 could also be observed in Figure 3.6. Table 3.9 documents the estimated coefficients. Five states are selected: Alabama, Maryland, Rhode Island, South Dakota and Utah. The mean value of estimated treatment effect, shown in Table 3.10, is -1.69, or a 23% decline. However, the adjusted R^2 is only 0.59, which is significantly smaller than that in rape and robbery. Figure 3.6(a) displays that the main variation is from the first 10 years. During this period, the unobserved common factors should impact these 5 states and Virginia in quite different manners. But after 1970, the fitted path generally follows the true path of murder rate and obvious deviations become fewer.

Table 3.9: Murder: Weights for Control Group

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.1411	1.1171	2.81	0.0087
AL	0.4536	0.0875	5.18	0.0000
MD	0.2609	0.1020	2.56	0.0160
RI	-0.4735	0.2029	-2.33	0.0268
SD	0.2867	0.1587	1.81	0.0812
UT	-0.4450	0.2486	-1.79	0.0839

Table 3.10: Murder: Treatment Effect between 1995 - 2010

Year	Actual	Prediction	Trt.Effect
1995	7.60	8.52	-0.92
1996	7.50	8.44	-0.94
1997	7.20	8.36	-1.16
1998	6.20	7.75	-1.55
1999	5.70	7.15	-1.45
2000	5.70	5.99	-0.29
2001	5.10	6.38	-1.28
2002	5.30	6.39	-1.09
2003	5.60	6.74	-1.14
2004	5.20	6.78	-1.58
2005	6.10	7.61	-1.51
2006	5.30	8.29	-2.99
2007	5.40	8.87	-3.47
2008	4.70	8.14	-3.44
2009	4.70	7.25	-2.55
2010	4.60	6.29	-1.69
Mean	5.74	7.43	-1.69
S.D	0.96	0.93	0.93

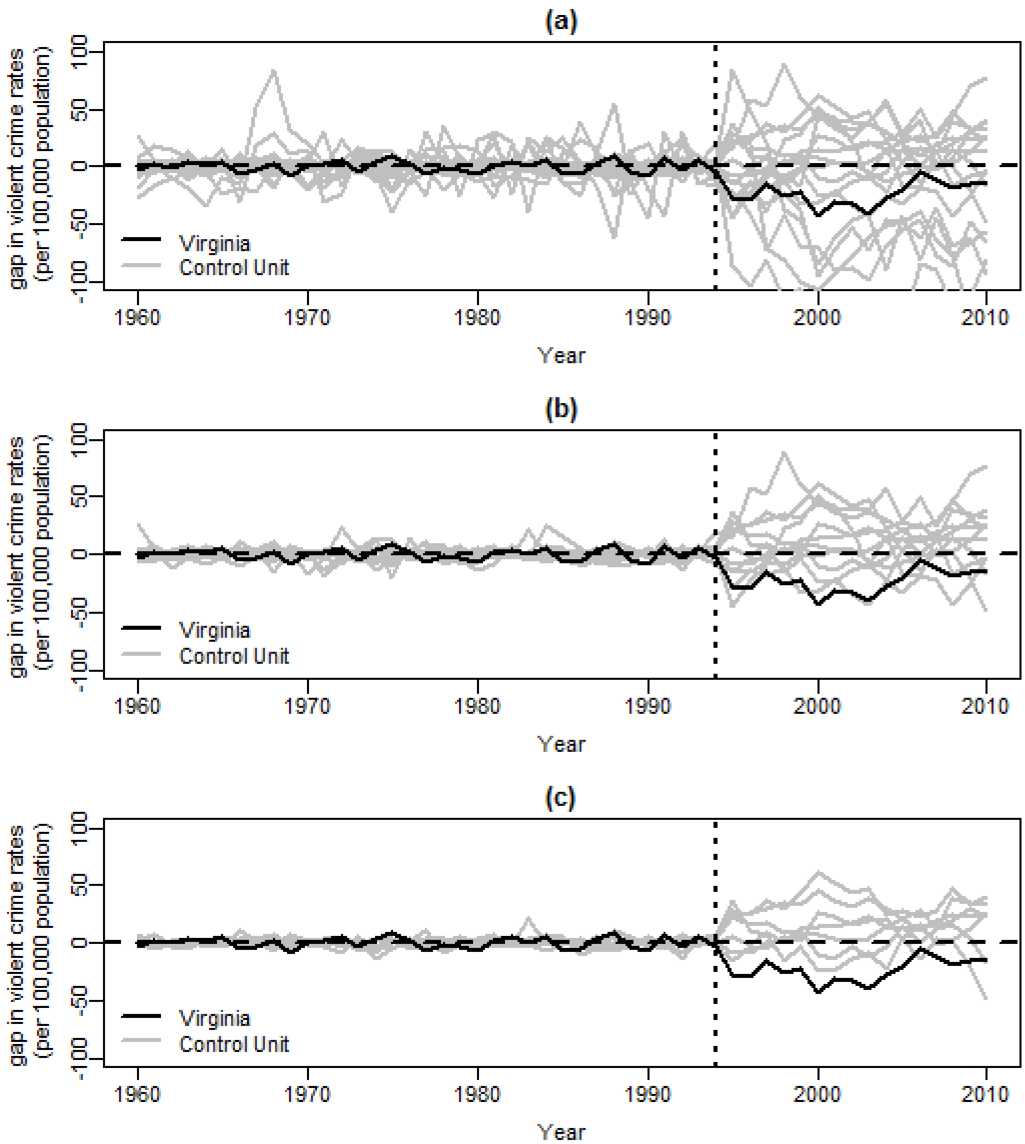


Figure 3.4: Placebo Test

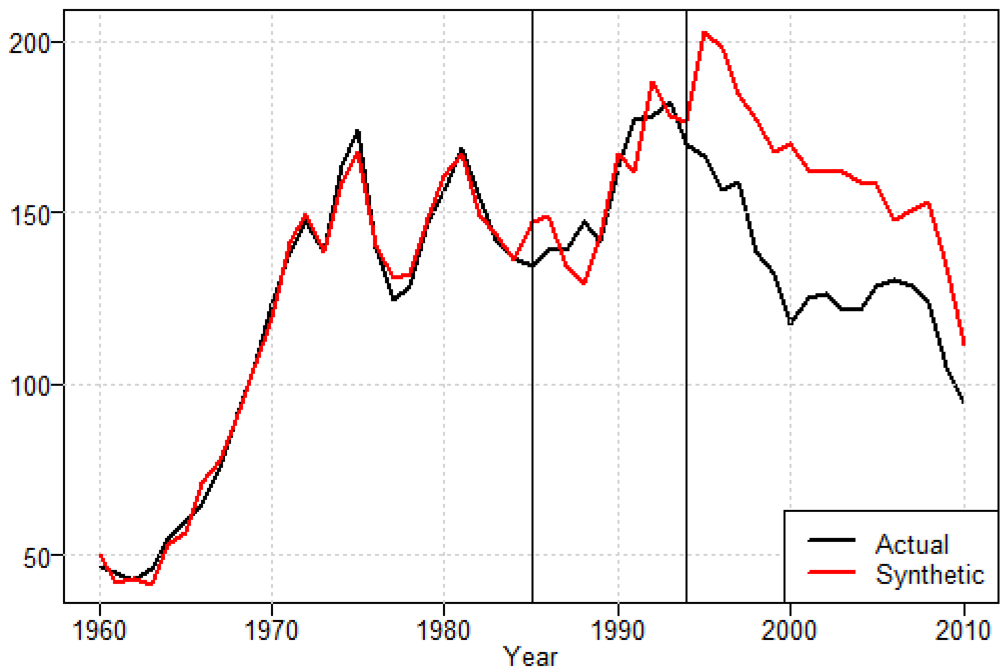
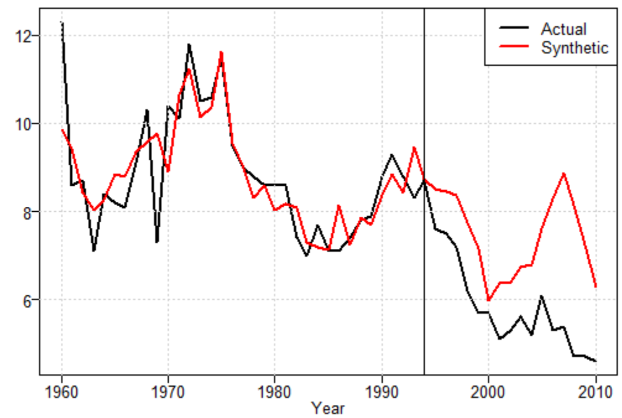
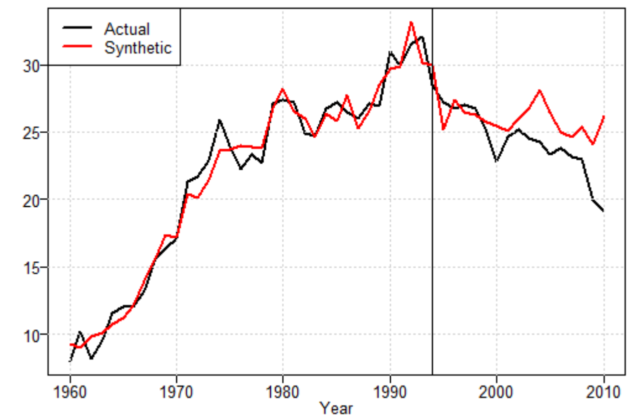


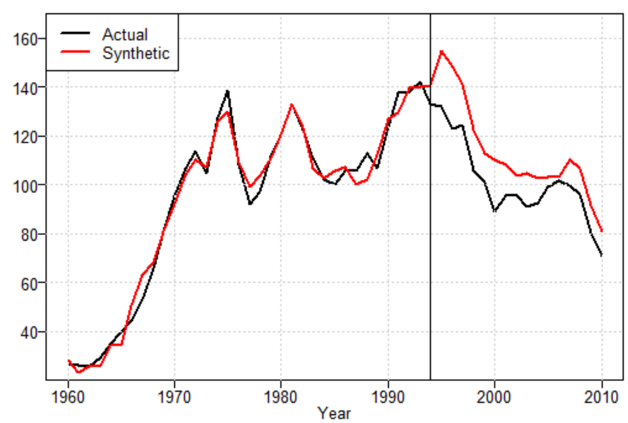
Figure 3.5: Robustness Test: Actual Path and Projected Path Based on the Truncated Data



(a) Murder



(b) Rape



(c) Robbery

Figure 3.6: Murder, Rape and Robbery Rates in Virginia: Actual Path and Synthetic Path

Table 3.11: Rape: Weights for Control Group

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	5.4284	0.7267	7.47	0.0000
AL	0.1839	0.1044	1.76	0.0888
HI	0.1382	0.0516	2.68	0.0121
MD	0.2305	0.0595	3.87	0.0006
NE	0.3473	0.1076	3.23	0.0031
WV	-0.3063	0.0997	-3.07	0.0046

Table 3.12: Rape: Treatment Effect between 1995 - 2010

Year	Actual	Prediction	Trt.Effect
1995	27.20	25.14	2.06
1996	26.70	27.37	-0.67
1997	27.00	26.44	0.56
1998	26.70	26.30	0.40
1999	25.00	25.72	-0.72
2000	22.80	25.46	-2.66
2001	24.60	25.12	-0.52
2002	25.20	25.97	-0.77
2003	24.50	26.73	-2.23
2004	24.30	28.10	-3.80
2005	23.30	26.48	-3.18
2006	23.80	25.04	-1.24
2007	23.20	24.62	-1.42
2008	23.00	25.36	-2.36
2009	19.90	24.09	-4.19
2010	19.10	26.16	-7.06
Mean	24.14	25.88	-1.74
S.D	2.32	1.03	2.18

3.6 Property Crimes

Even though the justice reform in Virginia targets on violent offenders, it is also meaningful to examine the response of non-violent or property crime offenders. Levitt

Table 3.13: Robbery: Weights for Control Group

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.1770	2.9759	0.06	0.9530
AL	0.2701	0.0613	4.41	0.0001
NM	0.1915	0.0748	2.56	0.0161
MD	0.1336	0.0200	6.67	0.0000
NV	-0.0727	0.0295	-2.47	0.0200
UT	0.6831	0.1617	4.22	0.0002
VT	-0.3560	0.1892	-1.88	0.0704

Table 3.14: Robbery: Treatment Effect between 1995 - 2010

Year	Actual	Prediction	Trt.Effect
1995	131.70	154.82	-23.12
1996	122.60	148.12	-25.52
1997	124.50	141.30	-16.80
1998	105.60	122.44	-16.84
1999	101.10	112.88	-11.78
2000	88.90	109.93	-21.03
2001	95.30	108.21	-12.91
2002	95.50	103.81	-8.31
2003	91.10	104.42	-13.32
2004	92.20	102.80	-10.60
2005	99.10	103.17	-4.07
2006	101.50	102.84	-1.34
2007	99.60	110.07	-10.47
2008	95.80	106.74	-10.94
2009	80.50	91.69	-11.19
2010	70.70	80.56	-9.86
Mean	99.73	112.74	-13.01
S.D	15.73	19.87	6.44

(1998) decomposes the reduction in crime rate into two channels. The first one is incapacitation: Tougher punishment leads to fewer crimes due to longer imprisonment. We have observed such an incapacitation effect in violent crimes as all murder, rape and robbery rate declined abruptly after 1994. The second one is deterrence: Severe punishment on one kind of crime will lead to a rise in another crime as offenders substitute away from the former. Table 3.15 compares the medians of actual serving time for burglary, larceny and motor vehicle theft in the year 1993 and in the period of 1999 - 2001. Changes in punishment for the three kinds of property crimes are relatively small compared with that for violent crimes.

Table 3.15: Median Years Property Offenders Served in Virginia

	FY 1993	FY 1999 - FY 2001		
	Prior $>$ 40	No Prior	Prior $<$ 40	Prior $>$ 40
Burglary	2.2	1.8	3.6	5.4
Larceny	1.3	1.1	1.8	2.3
Motor Vehicle Theft	1.3	1.3	1.8	2.7

Literatures have some evidence about the substitution from severely punished violent crimes to less stringently penalized property crimes. Using data on all counties in the United States, Shepherd (2002) finds that burglaries and auto thefts increase by 20% and 15% respectively after the enactment of TIS laws. As a matter of fact, sociological evidence (see Shafer, 1999) shows that for those who have experienced the sentencing system, many have learned from the path and thus realized that the less severe punishments on property crimes would not deter them from committing the same crimes in the future. However, when we impose more severe punishments on felons who also plan to engage in property crimes, we might also see a decline in property crimes as a by-product. Thus, the overall anticipated effect of the justice reform on property crimes is not clear in literatures.

Table 3.16 displays the weights for the five selected states, and Figure 3.7 demonstrates that property crime rate declines after the introduction of the justice reform.

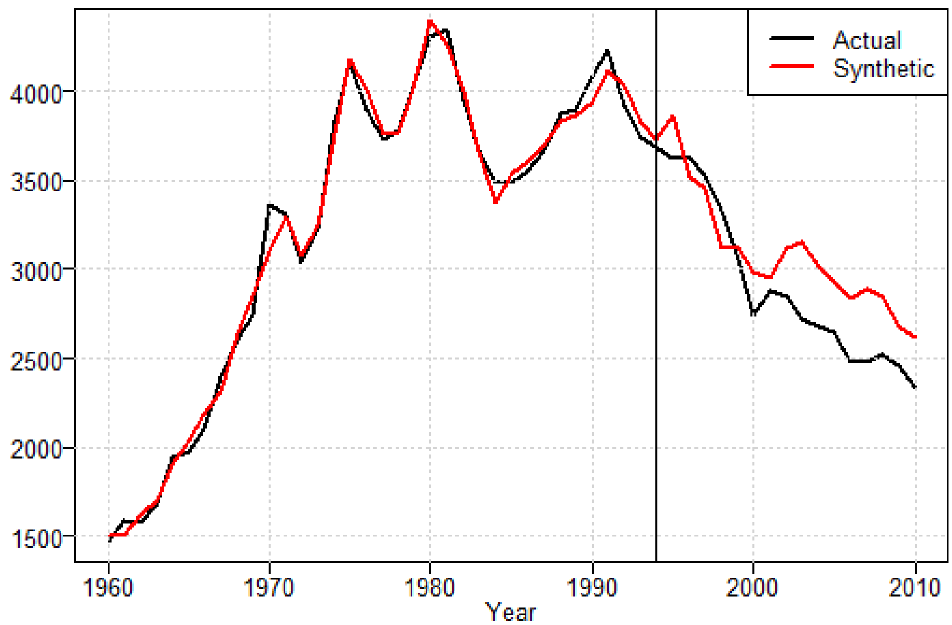


Figure 3.7: Property Crime Rates in Virginia: Actual Path and Synthetic Path

However, the story is not that straightforward: At least during the four years between 1996 and 1999 after the enactment of the justice reform, the actual property crime rates in Virginia are not significantly different from their predicting counterfactual counterparts. Specifically, according to Table 3.17, in the first 5 years after 1994, the property crime rates per 100,000 population have increased by about 19.43 on average, which is consistent with the theory of substitution to less severely penalized property crimes proposed by Levitt (1998) and Shepherd (2002). But we further find that even though violent offenders do not change their behaviors drastically after the reform in 1995, as time goes by, felons who plan to commit property crimes are incapacitated to do so due to the longer serving time in jails. This indicates some lagged treatment effect of the justice reform on property crimes, so that the actual property crime rate becomes uniformly lower than it should have been in the absence of the justice reform.

Table 3.16: Aggregated Property Crime Rates: weights of control group

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	339.0815	102.9194	3.29	0.0026
AL	0.3524	0.0598	5.89	0.0000
MA	0.5650	0.0505	11.20	0.0000
NH	-0.2927	0.0592	-4.94	0.0000
NM	-0.1466	0.0546	-2.68	0.0119
UT	0.2672	0.0872	3.06	0.0047

Table 3.17: Aggregated Property Crime Rates: Treatment effect between 1995 - 2010

Year	Actual	Prediction	Trt.Effect
1995	3627.70	3863.08	-235.38
1996	3627.00	3517.84	109.16
1997	3530.90	3461.72	69.18
1998	3334.70	3121.97	212.73
1999	3059.30	3118.16	-58.86
2000	2746.40	2979.39	-232.99
2001	2883.30	2952.10	-68.80
2002	2851.10	3119.66	-268.56
2003	2721.50	3149.70	-428.20
2004	2678.20	3016.65	-338.45
2005	2649.00	2927.00	-278.00
2006	2479.60	2834.74	-355.14
2007	2480.00	2893.24	-413.24
2008	2523.30	2844.94	-321.64
2009	2461.40	2683.71	-222.31
2010	2327.20	2612.55	-285.35
Mean	2873.79	3068.53	-194.74
S.D	435.53	320.20	191.76

3.7 Conclusion

In this paper we employ a panel data approach to evaluate the average treatment effect of Virginia's 1995 justice reform on violent/porperty crimes. Empirical results show that the average treatment effect of the justice reform on violent crimes is abrupt and significant. The robustness tests further confirm our findings. A closer examination on three types of violent crime — murder, rape and robbery — shows that the incapacitation effect is especially significant on murder and robbery. Treatment effect on property crime becomes significant four years later after 1994. This delay is consistent with criminological theories, which indicate that some violent offenders substitute to less severely penalized property crimes.

4. CONCLUSION

Model selection becomes a popular topic in econometrics and this dissertation contributes to the literature by proposing a model averaging approach to estimate the dependence structures among international stock markets. By averaging some well-known individual copulas, we prove that our model average estimates provide more realistic estimation about the dependence structures among the financial market in U.S., United Kingdom, Japan and Hong Kong. Model selection technique could also be employed to estimate the treatment effect of the justice reform in Virginia in 1995. We show that both the violent and property crime rates declines after the justice reform, based on the optimal model selected by some criteria.

REFERENCES

Abadie, A. Semiparametric difference-in-differences estimators. Review of Economic Studies 72(1), 2005, 1-19.

Abadie, A. and Gardeazabal, J. The economic costs of conflict: a case study of the Basque Country. American Economics Review 93(1), 2003, 113-132.

Abadie, A., Diamond, A. and Hainmueller, J. Synthetic control methods for comparative case studies: estimating the effects California tobacco control program. Journal of the American Statistical Association 105(490), 2010, 493-505.

Akaike, H. Statistical predictor identification. Annals of the Institute of Statistical Mathematics 22(1), 1970, 203-217

Aloui, R., Aïssa, M. and Nguyen, D. Global financial crisis, extreme interdependences, and contagion effects: The role of economic structure? Journal of Banking & Finance 35(1), 2011, 130-141.

Ando, T. and Li, K. A model-averaging approach for high-dimensional regression. Journal of the American Statistical Association 109(505), 2014, 254-265.

Andrews, D. Heteroskedasticity and autocorrelation consistent covariance matrix estimation. Econometrica 59(3), 1991, 817-858.

Athey, S. and Imbens, G. Identification and inference in nonlinear difference-in-differences models. Econometrica 74(2), 2006, 431-497.

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

Bai, J. Inferential theory for factor models of large dimensions. Econometrica 71(1), 2003, 135-171.

Becker, G. Crime and punishment: An economic approach. Journal of Political

Economics 76(22), 1968, 169-217.

Blumstein, A. and Wallman, J. The crime drop in America. Cambridge University Press, 2000.

Buckland, S., Burnham, K. and Augustin, N. Model selection: an integral part of inference. Biometrics 53(2), 1997, 603-618.

Cai, Z. and Wang, X. Selection of mixed copula model via penalized likelihood. Journal of the American Statistical Association 109(506), 2014, 788-801.

Chen, X. and Fan, Y. Estimation of copula-based semiparametric time series models. Journal of Econometrics 130(2), 2006, 307-335.

Chollete, L., Peña, V. and Lu, C. Comovement of international financial markets, 2005, unpublished manuscript.

Chollete, L., Heinen, A. and Valdesogo, A. Modeling international financial returns with a multivariate regime-switching copula. Journal of Financial Econometrics 7(4), 2009, 437-480.

DiIulio, J. and Piehl, A. Does prison pay? The stormy national debate over the cost-effectiveness of imprisonment. The Brookings Review, 1991, 28-35.

Fan, J. and Li, R. Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American Statistical Association 96(568), 2001, 1348-1360.

Fan, Y. and Patton, A. Copulas in econometrics. Annual Review of Economics 6, 2014, 179-200.

Genest, C. and Rivest, L. Statistical inference procedures for bivariate Archimedean copulas. Journal of the American Statistical Association 88(423), 1993, 1034-1043.

Granger, C. and Newbold, P. Spurious regressions in econometrics. Journal of Econometrics 2(2), 1974, 111-120.

Hansen, B. and Racine, J. Jackknife model averaging. Journal of Econometrics

167(1), 2012, 38-46.

Hansen, B. Least squares model averaging. Econometrica 75(4), 2007, 1175-1189.

Heckman, J. and Hotz, V. Choosing among alternative nonexperimental methods for estimating the impact if social programs: The case of manpower training. Journal of the American Statistical Association 84(408), 1989, 862-874.

Hoeting, J., Madigan, D., Raftery, A. and Volinsky, C. Bayesian model averaging: a tutorial. Statistical Science 14(4), 1999, 382-417.

Hsiao, C., Steve, H. and Wan, S. 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(5), 2012, 705-740.

Hu, L. Dependence patterns across financial markets: A mixed copula approach. Applied Financial Economics 16(2006), 2006, 717-729.

Hurvich, C. and Tsai, C., 1989. Regression and time series model selection in small samples. Biometrika 76(2), 297-307.

Joe, H. Multivariate Models and Dependence Concepts. London: Chapman & Hall, 1997.

Keegan, S. and Solomon, A. Prisoner reentry in Virginia. Justice Policy Center, Urban Institute, 2004.

Kuziemko, I. How should inmates be released from prison? An assessment of parole versus fixed sentence regimes. The Quarterly Journal of Economics 128(1), 2013, 371-424.

Levitt, S. The effect of prison population size on crime rates: Evidence from prison overcorwding litigation. The Quarterly Journal of Economics 111(2), 1996, 319-351.

Levitt, S. Why do increased arrest rates appear to reduce crime: Detterence, incapacitation, or measurement error? Economic Inquiry 36(3), 1998, 353-372.

Levitt, S. Understanding why crime fell in the 1990s: Four factors that explain the decline and six that do not. Journal of Economic Perspectives 18(1), 2004, 163-190.

Li, David X. On default correlation: A copula function approach. Journal of Fixed Income 9(4), 2000, 43-54.

Li, K. and Bell, D. Estimation of average treatment effect: What is the effect of opening a showroom by an online retailer? Research manuscript, The Wharton School, University of Pennsylvania, 2014.

Longin, F. and Solnik, B. Extreme correlation of international equity markets. The Journal of Finance 56, 2001, 649-676.

Manner, H. and Reznikova, O. A survey on time-varying copulas: Specification, simulations and application. Econometric Reviews 31(6), 2012, 654-687.

Marvell, T. and Moody, C. Specification problems, police levels and crime rates. Criminology 34(4), 1996, 609-646.

McPheters, L., Mann, R. and Schlagenhauf, D. Economic response to a crime deterrence program: Mandatory sentencing for robbery with a firearm. Economic Inquiry 22(4), 1984, 550-570.

Myers, S. The rehabilitation effect on punishment. Economic Inquiry 18(3), 1980, 353-366.

Nelsen, R. An Introduction to Copulas. New York: Springer. 2nd ed, 2006.

Patton, A. Modelling asymmetric exchange rate dependence. International Economic Review 47(2), 2006, 527-556.

Patton, A. A review of copula models for economic time series. Journal of Multivariate Analysis 110(2012), 2012, 4-18.

Petersilia, J. Parole and prisoner reentry in the United States. Crime and Justice 26, 1999, 479-529.

Rodriguez, J. Measuring financial contagion: A copula approach. Journal of

Empirical Finance 14(3), 2007, 401-423.

Ross, A. Crime, police, and truth-in-sentencing: The impact of state sentencing policy on local communities. Regional Science and Urban Economics 42(1), 2012, 144-152.

Sabol, W., Rosich, K., Kane, K., Kirk, D. and Dubin, G. The influence of truth-in-sentencing reforms on changes in states' sentencing practices and prison populations. Justice Policy Center, 2002, Urban Institute.

Schafer, J. The deterrent effect of three strikes law. FBI Law Enforcement Bulletin 68(4), 1999, 6-9.

Schwarz, G. Estimating the dimension of a model. Annals of Statistics 6(2), 1978, 461-464.

Shao, J. An asymptotic theory for linear model selection. Statistica Sinica 7(2), 1997, 221-264.

Shepherd, J. Police, prosecutors, criminals, and determinate sentencing: The truth about truth-in-sentencing laws. Journal of Law and Economics 45(2), 2002, 509-533.

Sklar, A. Fonctions de répartition à n dimensions et leurs marges. Publication de l'Institut de Statistique de l'Universite de Paris 8, 1959, 229-231.

Sridharan, S., Vujic, S. and Koopman, S. Intervention time series analysis of crime rates: The impact of sentence reforms in Virginia. Tinbergen Institute Discussion Paper, 2003.

Tibshirani, R. Regresion shrinkage and selection via the Lasso. Journal of the Royal Statistical Society, Series B 58(1), 1996, 267-288.

Wan, A., Zhang, X. and Zou, G. Least squares model average by Mallows criterion. Journal of Econometrics 156(2), 2010, 277-283.

Zimmer, D. The role of copulas in the housing crisis. Review of Economics and

Statistics 94(2), 2012, 607-620.

APPENDIX A

First, we define that

$$\widehat{\boldsymbol{\beta}} = \left(\widehat{\boldsymbol{\alpha}}_{1,1}^{\mathrm{T}}, \ldots, \widehat{\boldsymbol{\alpha}}_{1,p}^{\mathrm{T}}, \widehat{\boldsymbol{\theta}}_{1}^{\mathrm{T}}, \ldots, \widehat{\boldsymbol{\theta}}_{K-1}^{\mathrm{T}}, \widetilde{\omega}_1, \ldots, \widetilde{\omega}_{K-1}, \widetilde{\boldsymbol{\alpha}}_{1}^{\mathrm{T}}, \ldots, \widetilde{\boldsymbol{\alpha}}_{p}^{\mathrm{T}}, \widetilde{\boldsymbol{\theta}}_{1}^{\mathrm{T}}, \ldots, \widetilde{\boldsymbol{\theta}}_{K-1}^{\mathrm{T}}\right)^{\mathrm{T}}.$$

We assume K to be fixed and denote the fixed dimension of $\widehat{\boldsymbol{\beta}}$ by κ. Assume that there exists a vector $\boldsymbol{\beta}^*$ such that $\widehat{\boldsymbol{\beta}} \to \boldsymbol{\beta}^*$ in probability as $T \to \infty$. Let $\mathbf{C}^*(\mathbf{w}) = \widehat{\mathbf{C}}(\mathbf{w}) \mid_{\widehat{\boldsymbol{\beta}}=\boldsymbol{\beta}^*}$, $\boldsymbol{\nu}_t(\mathbf{w}) = \partial \widehat{C}(\mathbf{X}_t, \mathbf{w})/\partial \widehat{\boldsymbol{\beta}}|_{\widehat{\boldsymbol{\beta}}=\widetilde{\boldsymbol{\beta}}_t}$ for $t = 1, \ldots, T$ where $\widetilde{\boldsymbol{\beta}}_t$ is between $\widehat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}^*$, $\mathbf{Q}(\mathbf{w}) = \{\boldsymbol{\nu}_1(\mathbf{w}), \ldots, \boldsymbol{\nu}_T(\mathbf{w})\}^{\mathrm{T}}$, $L_T^*(\mathbf{w}) = \|\mathbf{C}^*(\mathbf{w}) - \mathbf{C}_0\|^2$, and $\xi_T = \inf_{\mathbf{w}\in\mathcal{W}} L_T^*(\mathbf{w})$. We assume J to be fixed and $M \to \infty$ as $T \to \infty$.

For proving the asymptotic optimality, we need the regularity conditions as the following:

Condition (C.1). *Uniformly for* $\mathbf{w} \in \mathcal{W}$, $T^{-1/2}\|\mathbf{Q}(\mathbf{w})(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|^2 = O_p(1)$ *and* $T^{-1/2}\{\mathbf{C}^*(\mathbf{w}) - \mathbf{C}_0\}^{\mathrm{T}}\mathbf{Q}(\mathbf{w})(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) = O_p(1)$.

Condition (C.2). *Uniformly,* $\mathbf{w} \in \mathcal{W}$, $T^{-1/2}\|\widehat{\mathbf{C}}(\mathbf{w}) - \widetilde{\mathbf{C}}(\mathbf{w})\|^2 = O_p(1)$, $T^{-1/2}\{\widehat{\mathbf{C}}(\mathbf{w}) - \widetilde{\mathbf{C}}(\mathbf{w})\}^{\mathrm{T}}\{\widehat{\mathbf{C}}(\mathbf{w}) - \overline{\mathbf{C}}\} = O_p(1)$, *and* $T^{-1/2}\{\widehat{\mathbf{C}}(\mathbf{w}) - \widetilde{\mathbf{C}}(\mathbf{w})\}^{\mathrm{T}}(\mathbf{C}_0 - \overline{\mathbf{C}}) = O_p(1)$.

Condition (C.3). *There exists a sequence* $c_T \to 0$ *such that* $T\xi_T^{-2} \le c_T$ *almost surely.*

Condition (C.1) constrains the convergence rate of $\widehat{\boldsymbol{\beta}}$ to its limit $\boldsymbol{\beta}^*$. Consider a general case with $\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^* = O_p(T^{-1/2})$. When the elements of $T \times \kappa$ matrix $\mathbf{Q}(\mathbf{w})$ are uniformly up-bounded, we obtain that uniformly for $\mathbf{w} \in \mathcal{W}$,

$$\begin{aligned}
& T^{-1/2}\left\|\mathbf{Q}(\mathbf{w})(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\right\|^2 \\
\le\ & T^{-1/2}\left\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right\|^2 \lambda_{\max}\left\{\mathbf{Q}^{\mathrm{T}}(\mathbf{w})\mathbf{Q}(\mathbf{w})\right\} \\
=\ & T^{-1/2}O_p(T^{-1})O_p(T) \\
=\ & O_p(T^{-1/2}),
\end{aligned}$$

where $\lambda_{\max}(\cdot)$ denotes the maximum eigenvalue of a matrix, and

$$T^{-1/2}\left\{\mathbf{C}^*(\mathbf{w})-\mathbf{C}_0\right\}^{\mathrm{T}}\mathbf{Q}(\mathbf{w})(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)=T^{-1/2}O_p(T^{1/2})=O_p(1),$$

where we also used the elements of vector $|\mathbf{C}^*(\mathbf{w})-\mathbf{C}_0|$ are uniformly up-bounded by 2.

Condition (C.2) needs that as $T\to\infty$, the difference between the loss of the regular and leave-M out estimators decrease at some rate, is similar to the condition (A.10) of Andrews (1991) and the condition (A.5) of Hansen and Racine (2012). When the elements of vector $\widehat{\mathbf{C}}(\mathbf{w})-\widetilde{\mathbf{C}}(\mathbf{w})$ are $O(T^{-1/2})$ uniformly, by the truth that the elements of vectors $|\widehat{\mathbf{C}}(\mathbf{w})-\overline{\mathbf{C}}|$ and $|\mathbf{C}_0-\overline{\mathbf{C}}|$ are uniformly up-bounded by 2, we obtain that uniformly for $\mathbf{w}\in\mathcal{W}$, $T^{-1/2}\|\widehat{\mathbf{C}}(\mathbf{w})-\widetilde{\mathbf{C}}(\mathbf{w})\|^2=O_p(T^{-1/2})$, $T^{-1/2}\{\widehat{\mathbf{C}}(\mathbf{w})-\widetilde{\mathbf{C}}(\mathbf{w})\}^{\mathrm{T}}\{\widehat{\mathbf{C}}(\mathbf{w})-\overline{\mathbf{C}}\}=O_p(1)$, and $T^{-1/2}\{\widehat{\mathbf{C}}(\mathbf{w})-\widetilde{\mathbf{C}}(\mathbf{w})\}^{\mathrm{T}}(\mathbf{C}_0-\overline{\mathbf{C}})=O_p(1)$.

Condition (C.3) imposes a limitation on the situation to apply our asymptotic results. It requires that ξ_T grow at a rate no slower than $T^{1/2}$, and implies all candidate copulas are misspecified. The condition (7) of Ando and Li (2014) is similar to our Condition (C.3).

We present a Lemma which will be useful in proving Theorem 1.

Lemma 1. *Write $CV_J(\mathbf{w})=L_T(\mathbf{w})+a_T(\mathbf{w})+b_T$, where the term b_T has nothing to do with $\mathbf{w}$. As $T\to\infty$, if*

$$\sup_{\mathbf{w}\in\mathcal{W}}\frac{|a_T(\mathbf{w})|}{L_T^*(\mathbf{w})}=o_p(1), \tag{A1}$$

$$\sup_{\mathbf{w}\in\mathcal{W}}\left|\frac{L_T(\mathbf{w})}{L_T^*(\mathbf{w})}-1\right|=o_p(1), \tag{A2}$$

and there exists a positive constant c such that

$$\xi_T\geq c \qquad \textit{almost surely}, \tag{A3}$$

then (2.3) holds.

Proof. From (A2), we know that

$$\begin{aligned}\inf_{\mathbf{w}\in\mathcal{W}} \frac{L_T(\mathbf{w})}{L_T^*(\mathbf{w})} &= \inf_{\mathbf{w}\in\mathcal{W}} \left(\frac{L_T(\mathbf{w})}{L_T^*(\mathbf{w})} - 1\right) + 1 \\ &\geq -\sup_{\mathbf{w}\in\mathcal{W}} \left|\frac{L_T(\mathbf{w})}{L_T^*(\mathbf{w})} - 1\right| + 1 \\ &\to 1, \end{aligned} \tag{A4}$$

in probability as $T \to \infty$. In addition, it is seen that there exists a non-negative sequence ν_T and a sequence of vectors $\mathbf{w}(T) \in \mathcal{W}$ such that as $T \to \infty$, $\nu_T \to 0$ and

$$\inf_{\mathbf{w}\in\mathcal{W}} L_T(\mathbf{w}) = L_T\{\mathbf{w}(T)\} - \nu_T. \tag{A5}$$

Thus, by (A2) and (A3), we have

$$\begin{aligned}\inf_{\mathbf{w}\in\mathcal{W}} \frac{|L_T(\mathbf{w}) - \nu_T|}{L_T^*(\mathbf{w})} &\geq \inf_{\mathbf{w}\in\mathcal{W}} \frac{L_T(\mathbf{w}) - \nu_T}{L_T^*(\mathbf{w})} \\ &\geq \inf_{\mathbf{w}\in\mathcal{W}} \frac{L_T(\mathbf{w})}{L_T^*(\mathbf{w})} - \frac{\nu_T}{\inf_{\mathbf{w}\in\mathcal{W}} L_T^*(\mathbf{w})} \\ &\geq -\sup_{\mathbf{w}\in\mathcal{W}} \left|\frac{L_T(\mathbf{w})}{L_T^*(\mathbf{w})} - 1\right| + 1 - \frac{\nu_T}{\inf_{\mathbf{w}\in\mathcal{W}} L_T^*(\mathbf{w})} \\ &\to 1, \end{aligned} \tag{A6}$$

in probability as $T \to \infty$. In addition, by the definition of $\widehat{\mathbf{w}}$, we have

$$\inf_{\mathbf{w}\in\mathcal{W}} \{L_T(\mathbf{w}) + a_T(\mathbf{w})\} = L_T(\widehat{\mathbf{w}}) + a_T(\widehat{\mathbf{w}}). \tag{A7}$$

Now, from (A1), (A3), (A4), (A5), (A6), (A7), and $\nu_T \to 0$, we obtain that, for any $\delta > 0$,

$$\begin{aligned} & \Pr\left\{\left|\frac{\inf_{\mathbf{w}\in\mathcal{W}} L_T(\mathbf{w})}{L_T(\widehat{\mathbf{w}})} - 1\right| > \delta\right\} \\ = \;& \Pr\left\{\frac{L_T(\widehat{\mathbf{w}}) - \inf_{\mathbf{w}\in\mathcal{W}} L_T(\mathbf{w})}{L_T(\widehat{\mathbf{w}})} > \delta\right\} \end{aligned}$$

$$
\begin{aligned}
&= \Pr\left\{\frac{\inf_{\mathbf{w}\in\mathcal{W}}\left(L_T(\mathbf{w})+a_T(\mathbf{w})\right)-a_T(\widehat{\mathbf{w}})-\inf_{\mathbf{w}\in\mathcal{W}}L_T(\mathbf{w})}{L_T(\widehat{\mathbf{w}})}>\delta\right\}\\
&\le \Pr\left\{\frac{L_T\left\{\mathbf{w}(T)\right\}+a_T\left\{\mathbf{w}(T)\right\}-a_T(\widehat{\mathbf{w}})-L_T\left\{\mathbf{w}(T)\right\}+\nu_T}{L_T(\widehat{\mathbf{w}})}>\delta\right\}\\
&\le \Pr\left\{\frac{\left|a_T\left\{\mathbf{w}(T)\right\}\right|}{L_T(\widehat{\mathbf{w}})}+\frac{|a_T(\widehat{\mathbf{w}})|}{L_T(\widehat{\mathbf{w}})}+\frac{\nu_T}{L_T(\widehat{\mathbf{w}})}>\delta\right\}\\
&\le \Pr\left\{\frac{\left|a_T\left\{\mathbf{w}(T)\right\}\right|}{\inf_{\mathbf{w}\in\mathcal{W}}L_T(\mathbf{w})}+\frac{|a_T(\widehat{\mathbf{w}})|}{L_T(\widehat{\mathbf{w}})}+\frac{\nu_T}{L_T(\widehat{\mathbf{w}})}>\delta\right\}\\
&= \Pr\left\{\frac{\left|a_T\left\{\mathbf{w}(T)\right\}\right|}{L_T\left\{\mathbf{w}(T)\right\}-\nu_T}+\frac{|a_T(\widehat{\mathbf{w}})|}{L_T(\widehat{\mathbf{w}})}+\frac{\nu_T}{L_T(\widehat{\mathbf{w}})}>\delta\right\}\\
&\le \Pr\left\{\sup_{\mathbf{w}\in\mathcal{W}}\frac{|a_T(\mathbf{w})|}{L_T(\mathbf{w})-\nu_T}+\sup_{\mathbf{w}\in\mathcal{W}}\frac{|a_T(\mathbf{w})|}{L_T(\mathbf{w})}+\sup_{\mathbf{w}\in\mathcal{W}}\frac{\nu_T}{L_T(\mathbf{w})}>\delta\right\}\\
&\le \Pr\left\{\sup_{\mathbf{w}\in\mathcal{W}}\frac{|a_T(\mathbf{w})|}{L_T^*(\mathbf{w})}\sup_{\mathbf{w}\in\mathcal{W}}\frac{L_T^*(\mathbf{w})}{|L_T(\mathbf{w})-\nu_T|}+\sup_{\mathbf{w}\in\mathcal{W}}\frac{|a_T(\mathbf{w})|}{L_T^*(\mathbf{w})}\sup_{\mathbf{w}\in\mathcal{W}}\frac{L_T^*(\mathbf{w})}{L_T(\mathbf{w})}\right.\\
&\qquad\left.+\sup_{\mathbf{w}\in\mathcal{W}}\frac{\nu_T}{L_T^*(\mathbf{w})}\sup_{\mathbf{w}\in\mathcal{W}}\frac{L_T^*(\mathbf{w})}{L_T(\mathbf{w})}>\delta\right\}\\
&= \Pr\left\{\sup_{\mathbf{w}\in\mathcal{W}}\frac{|a_T(\mathbf{w})|}{L_T^*(\mathbf{w})}\left[\inf_{\mathbf{w}\in\mathcal{W}}\frac{|L_T(\mathbf{w})-\nu_T|}{L_T^*(\mathbf{w})}\right]^{-1}+\sup_{\mathbf{w}\in\mathcal{W}}\frac{|a_T(\mathbf{w})|}{L_T^*(\mathbf{w})}\left[\inf_{\mathbf{w}\in\mathcal{W}}\frac{L_T(\mathbf{w})}{L_T^*(\mathbf{w})}\right]^{-1}\right.\\
&\qquad\left.+\frac{\nu_T}{\inf_{\mathbf{w}\in\mathcal{W}}L_T^*(\mathbf{w})}\left[\inf_{\mathbf{w}\in\mathcal{W}}\frac{L_T(\mathbf{w})}{L_T^*(\mathbf{w})}\right]^{-1}>\delta\right\}\\
&\to 0,
\end{aligned}
$$

which implies (2.3).

It could be seen that

$$
\begin{aligned}
CV_J(\mathbf{w}) &= \left\|\widetilde{\mathbf{C}}(\mathbf{w})-\overline{\mathbf{C}}\right\|^2\\
&= \left\|\left\{\widehat{\mathbf{C}}(\mathbf{w})-\mathbf{C}_0\right\}-\left\{\widehat{\mathbf{C}}(\mathbf{w})-\widetilde{\mathbf{C}}(\mathbf{w})\right\}+\left(\mathbf{C}_0-\overline{\mathbf{C}}\right)\right\|^2\\
&= L_T(\mathbf{w})+\left\|\widehat{\mathbf{C}}(\mathbf{w})-\widetilde{\mathbf{C}}(\mathbf{w})\right\|^2+\left\|\mathbf{C}_0-\overline{\mathbf{C}}\right\|^2\\
&\quad -2\left\{\widehat{\mathbf{C}}(\mathbf{w})-\widetilde{\mathbf{C}}(\mathbf{w})\right\}^{\mathrm{T}}(\mathbf{C}_0-\overline{\mathbf{C}})\\
&\quad -2\left\{\widehat{\mathbf{C}}(\mathbf{w})-\widetilde{\mathbf{C}}(\mathbf{w})\right\}^{\mathrm{T}}\left\{\widehat{\mathbf{C}}(\mathbf{w})-\mathbf{C}_0\right\}+2\left\{\widehat{\mathbf{C}}(\mathbf{w})-\mathbf{C}_0\right\}^{\mathrm{T}}(\mathbf{C}_0-\overline{\mathbf{C}})\\
&= L_T(\mathbf{w})+\left\|\widehat{\mathbf{C}}(\mathbf{w})-\widetilde{\mathbf{C}}(\mathbf{w})\right\|^2-2\left\{\widehat{\mathbf{C}}(\mathbf{w})-\widetilde{\mathbf{C}}(\mathbf{w})\right\}^{\mathrm{T}}\left\{\widehat{\mathbf{C}}(\mathbf{w})-\overline{\mathbf{C}}\right\}\\
&\quad +2\left\{\widehat{\mathbf{C}}(\mathbf{w})-\widetilde{\mathbf{C}}(\mathbf{w})\right\}^{\mathrm{T}}(\mathbf{C}_0-\overline{\mathbf{C}})+2\widetilde{\mathbf{C}}(\mathbf{w})^{\mathrm{T}}(\mathbf{C}_0-\overline{\mathbf{C}})
\end{aligned}
$$

$$
\begin{aligned}
&-2(\mathbf{C}_0+\overline{\mathbf{C}})^{\mathrm{T}}(\mathbf{C}_0-\overline{\mathbf{C}}) \\
\equiv\ & L_T(\mathbf{w})+\Xi_T(\mathbf{w})-2(\mathbf{C}_0+\overline{\mathbf{C}})^{\mathrm{T}}(\mathbf{C}_0-\overline{\mathbf{C}}),
\end{aligned}
$$

where the last term has nothing to do with the weight vector $\mathbf{w}$, and

$$
\begin{aligned}
L_T(\mathbf{w}) &= \left\|\widehat{\mathbf{C}}(\mathbf{w})-\mathbf{C}_0\right\|^2 \\
&= \left\|\left\{\widehat{\mathbf{C}}(\mathbf{w})-\mathbf{C}^*(\mathbf{w})\right\}+\left\{\mathbf{C}^*(\mathbf{w})-\mathbf{C}_0\right\}\right\|^2 \\
&= \left\|\widehat{\mathbf{C}}(\mathbf{w})-\mathbf{C}^*(\mathbf{w})\right\|^2+\left\|\mathbf{C}^*(\mathbf{w})-\mathbf{C}_0\right\|^2 \\
&\quad +2\left\{\mathbf{C}^*(\mathbf{w})-\mathbf{C}_0\right\}^{\mathrm{T}}\left\{\widehat{\mathbf{C}}(\mathbf{w})-\mathbf{C}^*(\mathbf{w})\right\} \\
&= L_T^*(\mathbf{w})+\left\|\widehat{\mathbf{C}}(\mathbf{w})-\mathbf{C}^*(\mathbf{w})\right\|^2 \\
&\quad +2\left\{\mathbf{C}^*(\mathbf{w})-\mathbf{C}_0\right\}^{\mathrm{T}}\left\{\widehat{\mathbf{C}}(\mathbf{w})-\mathbf{C}^*(\mathbf{w})\right\} \\
&\equiv L_T^*(\mathbf{w})+\Pi_T(\mathbf{w}).
\end{aligned}
$$

From Condition (C.3), we can obtain (A3). Hence, from Lemma 1, Theorem 1 is valid if the following hold:

$$
\sup_{\mathbf{w}\in\mathcal{W}}\frac{|\Xi_T(\mathbf{w})|}{L_T^*(\mathbf{w})}=o_p(1) \tag{A8}
$$

and

$$
\sup_{\mathbf{w}\in\mathcal{W}}\frac{|\Pi_T(\mathbf{w})|}{L_T^*(\mathbf{w})}=o_p(1). \tag{A9}
$$

Using Taylor expansion,

$$
\widehat{\mathbf{C}}(\mathbf{w})-\mathbf{C}^*(\mathbf{w})=\mathbf{Q}(\mathbf{w})(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*). \tag{A10}
$$

From (A10), Conditions (C.1) and (C.3), and the truth that any element of vector

$|\mathbf{C}^*(\mathbf{w}) - \mathbf{C}_0|$ is up-bounded by 2, we have

$$\begin{aligned}
& \sup_{\mathbf{w}\in\mathcal{W}} \frac{|\Pi_T(\mathbf{w})|}{L_T^*(\mathbf{w})} \\
\leq\ & \xi_T^{-1} \sup_{\mathbf{w}\in\mathcal{W}} \left\|\widehat{\mathbf{C}}(\mathbf{w}) - \mathbf{C}^*(\mathbf{w})\right\|^2 + 2\xi_T^{-1} \sup_{\mathbf{w}\in\mathcal{W}} \left|\{\mathbf{C}^*(\mathbf{w}) - \mathbf{C}_0\}^{\mathrm{T}} \left\{\widehat{\mathbf{C}}(\mathbf{w}) - \mathbf{C}^*(\mathbf{w})\right\}\right| \\
=\ & \frac{T^{1/2}}{\xi_T} T^{-1/2} \sup_{\mathbf{w}\in\mathcal{W}} \left\|\mathbf{Q}(\mathbf{w})(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\right\|^2 \\
& + 2\frac{T^{1/2}}{\xi_T} T^{-1/2} \sup_{\mathbf{w}\in\mathcal{W}} \left|\{\mathbf{C}^*(\mathbf{w}) - \mathbf{C}_0\}^{\mathrm{T}} \mathbf{Q}(\mathbf{w})(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\right| \\
=\ & o_p(1),
\end{aligned}$$

which is (A9). Similarly, from Conditions (C.2) and (C.3), we have

$$\begin{aligned}
& \sup_{\mathbf{w}\in\mathcal{W}} \frac{\left|\left\|\widehat{\mathbf{C}}(\mathbf{w}) - \widetilde{\mathbf{C}}(\mathbf{w})\right\|^2 - 2\left\{\widehat{\mathbf{C}}(\mathbf{w}) - \widetilde{\mathbf{C}}(\mathbf{w})\right\}^{\mathrm{T}} \left\{(\widehat{\mathbf{C}}(\mathbf{w}) - \overline{\mathbf{C}}) + (\mathbf{C}_0 - \overline{\mathbf{C}})\right\}\right|}{L_T^*(\mathbf{w})} \\
& = o_p(1). \qquad \text{(A11)}
\end{aligned}$$

For any $\mathbf{x}$, from Central Limit Theory and (2.2), we have

$$C_0(\mathbf{x}) - \overline{C}_{(j)}(\mathbf{x}) = C_0(\mathbf{x}) - M^{-1} \sum_{m=1}^{M} I(\mathbf{X}_{(j-1)M+m} \leq \mathbf{x}) = O_p(M^{-1/2}),$$

which, along with the fact that $\mathbf{X}_1, \ldots, \mathbf{X}_T$ are i.i.d., implies that, uniformly for $t \in \{1, \ldots, T\}$,

$$C_0(\mathbf{X}_t) - \overline{C}_{(j)}(\mathbf{X}_t) = O_p(M^{-1/2}). \qquad \text{(A12)}$$

From $c_T \to 0$, (A12), and the truth that any element of vectors $|\widehat{\mathbf{C}}_k|$ are up-bounded by 1, we obtain that for any $j \in \{1, \ldots, J\}$,

$$c_T^{1/2} M^{-1/2} \left|\sum_{m=1}^{M} \widetilde{C}_k^{(-j)}(\mathbf{X}_{(j-1)M+m}) \left\{C_0(\mathbf{X}_{(j-1)M+m}) - \overline{C}_{(j)}(\mathbf{X}_{(j-1)M+m})\right\}\right|$$

$$
\begin{aligned}
&\leq c_T^{1/2} M^{-1/2} \sum_{m=1}^{M} \left| \widetilde{C}_k^{(-j)}(\mathbf{X}_{(j-1)M+m}) \left\{ C_0(\mathbf{X}_{(j-1)M+m}) - \overline{C}_{(j)}(\mathbf{X}_{(j-1)M+m}) \right\} \right| \\
&\leq c_T^{1/2} M^{-1/2} \sum_{m=1}^{M} \left| C_0(\mathbf{X}_{(j-1)M+m}) - \overline{C}_{(j)}(\mathbf{X}_{(j-1)M+m}) \right| \\
&= o_p(1),
\end{aligned}
$$

which, along with Condition (C.3) and the assumption that K and J are fixed, implies that

$$
\begin{aligned}
& \xi_T^{-1} \sup_{\mathbf{w} \in \mathcal{W}} \left| \widetilde{\mathbf{C}}(\mathbf{w})^{\mathrm{T}} (\mathbf{C}_0 - \overline{\mathbf{C}}) \right| \\
= \ & \xi_T^{-1} \sup_{\mathbf{w} \in \mathcal{W}} \left| \sum_{k=1}^{K} w_k \widetilde{\mathbf{C}}_k^{\mathrm{T}} (\mathbf{C}_0 - \overline{\mathbf{C}}) \right| \\
\leq \ & \sum_{k=1}^{K} \frac{T^{1/2}}{\xi_T} T^{-1/2} \left| \widetilde{\mathbf{C}}_k^{\mathrm{T}} (\mathbf{C}_0 - \overline{\mathbf{C}}) \right| \\
= \ & o_p(1),
\end{aligned}
\tag{A13}
$$

where the second '$\leq$' holds almost surely. From (A11) and (A13), we obtain (A8). This completes the proof.

APPENDIX B

Criminal homicide: a.) Murder and nonnegligent manslaughter: the willful (nonnegligent) killing of one human being by another. Deaths caused by negligence, attempts to kill, assaults to kill, suicides, and accidental deaths are excluded. The program classifies justifiable homicides separately and limits the definition to: (1) the killing of a felon by a law enforcement officer in the line of duty; or (2) the killing of a felon, during the commission of a felony, by a private citizen. b.) Manslaughter by negligence: the killing of another person through gross negligence. Deaths of persons due to their own negligence, accidental deaths not resulting from gross negligence, and traffic fatalities are not included in the category Manslaughter by Negligence.

Forcible rape: The carnal knowledge of a female forcibly and against her will. Rapes by force and attempts or assaults to rape, regardless of the age of the victim, are included. Statutory offenses (no force used — victim under age of consent) are excluded.

Robbery: The taking or attempting to take anything of value from the care, custody, or control of a person or persons by force or threat of force or violence and/or by putting the victim in fear.

Burglary (breaking or entering): The unlawful entry of a structure to commit a felony or a theft. Attempted forcible entry is included.

Larceny-theft (except motor vehicle theft): The unlawful taking, carrying, leading, or riding away of property from the possession or constructive possession of another. Examples are thefts of bicycles, motor vehicle parts and accessories, shoplifting, pocketpicking, or the stealing of any property or article that is not taken by force and violence or by fraud. Attempted larcenies are included. Embezzlement, confidence games, forgery, check fraud, etc., are excluded.

Motor vehicle theft: The theft or attempted theft of a motor vehicle. A motor vehicle is self-propelled and runs on land surface and not on rails. Motorboats, construction equipment, airplanes, and farming equipment are specifically excluded from this category.