Detecting Financial Data Dependence Structure by Averaging Mixture Copulas

Wei Long[a], Guannan Liu[b], Xinyu Zhang[c], and Qi Li[d,e*]

May, 2017

Abstract

A mixture copula is a linear combination of several individual copulas which can be used to generate dependence structures that do not belong to existing copula families. Thus it is useful in modeling the dependence in financial data, because different pairs of markets may exhibit quite different dependence structures in empirical studies. Therefore, rather than selecting one single copula through certain criteria, we propose using a model averaging approach to estimating financial data dependence structure in a mixture copula framework. We select weights (for averaging) through a J-fold Cross-Validation procedure. We prove that the model averaging estimator is asymptotically optimal in the sense that it minimizes a squared estimation loss. Our simulation results show that the model averaging approach outperforms some competing methods when the working mixture model is misspecified. Using 12 years of data on daily returns of four developed economies' stock indexes (United States, United Kingdom, Hong Kong and Japan), we show that the model averaging approach more accurately estimates their dependence structures than do some competing methods.

Keywords: Copula, mixture copula, model averaging, modeling dependence.

JEL Classification Numbers: C5; G15

*a: Department of Economics, Tulane University; b: School of Economics, Xiamen University, China; c: Academy of Mathematics and Systems Science, Chinese Academy of Sciences, China; d: ISEM, Capital University of Economics and Business, China; e: Department of Economics, Texas A&M University, U.S.A. The corresponding author: Qi Li, email: qi-li@tamu.edu (Guannan Liu, email: gliu@econmail.tamu.edu. Wei Long, email: weilong2@tulane.edu. Xinyu Zhang, email: xinyu@amss.ac.cn). Zhang and Li's researches are partially supported by National Natural Science Foundation of China (projects 71522004, 11471324 and 71631008 for Zhang; 71133001 and 71601130 for Li).

1 Introduction

Nelsen (2006) provides a thorough introduction about copula which is defined as "functions that join or couple multivariate distribution functions to their one-dimensional marginal distribution functions". Specifically, let $\boldsymbol{X} = (X_1, ..., X_p)^\top$ be a random vector and the respective marginal cumulative distribution functions defined as F_i for $i \in \{1, ..., p\}$. Then there exists a copula $C : [0,1]^p \to [0,1]$ such that $\forall\ \boldsymbol{x} = (x_1, ..., x_p) \in \mathbb{R}^p$, $F(\boldsymbol{x}) = C\{F_1(x_1), ..., F_p(x_p)\}$ (see Sklar, 1959). Therefore, copula is flexible; it does not constrain the selection of marginal distributions so that one can couple various margins together via a copula to obtain a flexible distribution function. In this paper we propose using a model averaging approach to estimate a mixture copula model. The mixture copula is a linear combination of multiple individual copulas.

Copula model primarily are used to study dependence patterns among variables, e.g., the co-movements among international equity markets. They have been extensively applied in a number of empirical studies such as: estimating default correlations (Li, 2000); examining the difference in dependence structures between developed and developing economies (Chollete, Peña and Lu, 2005, Chollete et al., 2009, and Aloui et al., 2011); structure breaks in exchange rates (Patton, 2006); financial contagions (Rodriguez, 2007); and the cross-state housing prices during the subprime mortgage crisis (Zimmer, 2012). One fundamental issue for such studies is how to select an appropriate copula to satisfactorily describe the dependence structure among the variables under study. Almost all of the works mentioned above presumptively build a candidate set and rely on certain statistical criteria to select one copula; they then use that copula to measure the degree of dependence among variables. For example, when studying housing crisis dependence among different states in the United States, Zimmer (2012) uses both Bayesian Information Criterion (BIC) and the Vuong test. He shows that a Clayton-Gumbel mixture copula provides a better estimate of dependence structures than a Gaussian, or Clayton or Gumbel copula.[1]

In practice, there are many types of copula. One might want to fit the data under analysis to each existing copula family and then select the most appropriate one. However, this strategy is not feasible because one can always create a new copula by making certain transformations on an existing copula.[2] Thus, most empirical researchers rely only on several common copulas, e.g., Gaussian, Clayton and Gumbel, to construct their candidate set. The argument for this 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 is relatively easy to implement, it has a cost: one needs to assume that the observations are generated from

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

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

one of the copula included in the candidate set. That is, one selects the most "appropriate" copula from the candidate set based on some criteria, and then uses it to describe the dependence pattern and to evaluate the degree of dependence of the data under analysis. If one's candidate set includes the true copula that generates the observations, then the estimating procedure just described should be effective and efficient. However, the true data generating copula model is unknown to researchers. In practice, it is highly likely that one's candidate set fails to include the true copula. Under such a circumstance, a selected candidate copula based on some criteria (say, by minimizing BIC) may fail to provide an adequate description of the true dependence structure.

To take advantage of different copula shapes, Chollete et al. (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 summing to 1. A mixture copula is more flexible than an individual copulas because it nests several individual copulas with quite different dependence structures. As we shall see, a mixture copula can generate dependence structures that do not belong to any existing individual copula. By combining several widely used individual copulas, one can build a parsimonious yet flexible mixture copula to capture various dependence patterns in financial data, such as zero and non-zero tail dependence and symmetric or asymmetric tail dependence. In their analysis, Chollete et al. (2005) and Hu (2006) consider a mixture model including Gaussian, Gumbel and rotated Gumbel copulas to evaluate the dependence structures among stock indexes in developed economies. They find strong left-tail dependence because the weight associated with the rotated Gumbel copula tends to be non-zero, while the Gumbel tends to be filtered out due to its small weight. Thus they conclude that stock markets in developed economies tend to decline simultaneously. This 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 method to estimate weights and copula parameters simultaneously using the Smoothly Clipped Absolute Deviation (SCAD) method proposed by Fan and Li (2001) as a penalizer. Cai and Wang (2014) further establish the asymptotic result for their proposed estimator and use simulations to demonstrate that their proposed method yields satisfactory estimation results on weights and copula parameters. That is, different dependence structures are captured well by their estimated mixture copulas via the penalized likelihood method.

This paper contributes to the existing literature by providing another method to estimate mixture copula models. Specifically, we use a model averaging approach: rather than choosing one appropriate copula model by comparing different criteria, such as AIC or BIC, we first fit observations to each individual copula in the candidate set separately. Sometimes all individual copulas lead to poor fit. Therefore, we also 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 the associated weights by minimizing a leave-one-group-out cross-validation criterion. This manner is similar to the Jackknife model average proposed by Hansen and Racine (2012). We obtain solutions through a standard application of quadratic programming technique; the leave-one-group-out cross-validation criterion is a quadratic function of weights. Under certain regularity conditions, we prove that our model averaging estimator is asymptotically optimal, in that it achieves the infeasible lowest possible squared estimation loss. The chosen weights help us to construct the optimal combination of each candidate copula and their composite that can satisfactorily describe the dependence structure among the variables under study. The distance (measured by the estimation squared loss) between the estimated mixture copula and the unknown true model is asymptotically minimized. This is extremely important when the working model is misspecified, i.e., when observations are generated from copulas that are not included in the model. Cai and Wang (2014) argue that when the working model is misspecified, their penalized likelihood method will select a (mixture) copula exhibiting a dependence pattern similar to the unknown true copula, for example, when observations are generated from a combination of Gaussian and Clayton, while Clayton is absent in the working model but there is a rotated Gumbel which also exhibits left-tail dependence. In such a case, Cai and Wang's (2014) method will assign a certain weight to the rotated Gumbel to guarantee that the left-tail dependence pattern is captured by the mixture copula model. Then they conclude that the "best approximated" copula is chosen.

The model averaging method provides a more solid criterion for the *best approximated* copula when the working model is misspecified: the optimal mixture model is constructed to minimize its distance (the estimation squared loss) to the true model, so that it can best describe the dependence pattern among variables. Because applied researchers often use working mixture models that tend to be quite parsimonious, the misspecification problem should be common. Therefore, model averaging is an alternative method to estimate mixture copula models which leads to the estimation squared loss being asymptotically minimized.

In the empirical part of the paper, we use our model average approach to study the dependence structures of the daily returns on equity indexes of four developed economies (United States, UK, Hong Kong and Japan). The estimation results support the superiority of the model averaging method over some existing methods in this analysis. Compared to the penalized likelihood method, the standard copula selection method (such as BIC) and the maximum likelihood estimation (MLE) method, our model averaging approach exhibits the smallest squared losses in out-of-sample predictions. The empirical application suggests that the model averaging method is a useful tool in analyzing the dependence structures of stock markets and in risk management.

The rest of this paper is organized as follows. In Section 2 we introduce a mixture copula model. Section 3 describes how to implement the model averaging approach in estimating a mixture copula model. In Section 4, we use simulations to compare estimation losses under the model averaging approach, Cai and Wang's (2014) penalized likelihood method, the BIC method and the MLE method. Section 5 presents a real data example. Section 6 concludes the paper. Regularity conditions and the proof of the optimality of the model averaging method are presented in Appendix A.

2 Mixture Copula Models: A Brief Introduction

Suppose we have a series of $p-$dimensional vectors of random variables $\{\mathbf{X}_t\}_{t=1}^T$, where $\mathbf{X}_t = (X_{t1}, \ldots, X_{tp})^\top$ and p is a finite positive integer. Let $F^0(\mathbf{x})$ and $f^0(\mathbf{x})$ be the joint distribution and the density function of $\mathbf{X}$ evaluated at $\mathbf{x} \in \mathcal{R}^p$, and $F_i^0(x_i)$ and $f_i^0(x_i)$ be the marginal distribution and the density function of X_i evaluated at $x_i \in \mathcal{R}$, respectively, where $1 \leq i \leq p$.

As in Hu (2006) and Cai and Wang (2014), a mixture copula model is a linear combination of several individual copulas. Specifically, a mixture copula model can be written as

$$C(\mathbf{u}; \boldsymbol{\theta}, \omega) = \sum_{k=1}^{L} \omega_k C_k(\mathbf{u}; \boldsymbol{\theta}_k) = \sum_{k=1}^{L} \omega_k C_k \left\{ F_1^0(x_1), \ldots, F_p^0(x_p); \boldsymbol{\theta}_k \right\}, \tag{1}$$

where $\{C_1(\cdot), ..., C_L(\cdot)\}$ is a set of candidate copulas with a vector of unknown parameters $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^\top, \ldots, \boldsymbol{\theta}_L^\top)^\top$ and a p-dimensional marginal distribution $\mathbf{u} = (F_1^0(\cdot), \ldots, F_p^0(\cdot))$. Let $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_L)^\top$ denote the *weight parameters* with $0 \leq \omega_l \leq 1$ and $\sum_{l=1}^L \omega_l = 1$. In equation (1), both the copula parameters $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^\top, \ldots, \boldsymbol{\theta}_L^\top)^\top$ and the weights $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_L)^\top$ control the shape of the mixture copula's dependence structure.

One may want to include many existing individual copulas into a mixture model to cover every possible dependence pattern. But in application this may make the mixture model too complicated, and the estimation burden of such a large mixture model can be very costly. In practice, one may only consider a few individual candidate copulas and hope that a combination of them can generate a flexible copula which can catch the dependence structure well for a given data. We present the flexibility of some mixture copulas through scatter plots. Each panel in Figure 1 is a scatter-plot of an i.i.d. sample of size 1000 generated from three types of widely-used copulas. Each margin has the standard normal distribution and the parameter for the corresponding copula is calibrated to imply a strong dependence with Kendall's $\tau = 0.5$. From Figure 1 (a) - (c), it can be seen that the 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, the Gaussian copula is symmetric and the stronger dependence appears in the center. Figure 1 (d) - (f) present scatter plots from three mixture copulas with equal weights on each component. Figure 1 clearly demonstrates that, after mixing with Clayton and Gumbel, a Gaussian copula begins to exhibit some asymmetric tail dependence. Therefore, the flexibility of a mixture copula stems from its ability of nesting various copula shapes. Each individual copula is nested as a special case.

[INSERT FIGURE 1 ABOUT HERE]

In our model averaging approach, besides estimating parameters associated with copulas (and marginal densities), we need to further estimate a vector of weight parameters ($\boldsymbol{\omega}$) introduced by mixture copula models. Chollete et al. (2005) and Hu (2006) independently propose a two-stage semiparametric method in estimating a mixture copula model. Specifically, in the first stage, the marginal distributions are estimated nonparametrically to avoid the misspecification of marginals. In the second stage, the estimated marginals or the empirical CDFs are plugged into the copula and then copula parameters are estimated by the maximum likelihood method. Finally, to facilitate the estimation of weight parameters for each nested copula, an iterative procedure, namely, the EM algorithm, is implemented. Cai and Wang (2014) provide the asymptotic results for mixture copula estimators. They propose a data-driven copula selection method via the penalized likelihood with a shrinkage operator so that parameter estimation and model selection are achieved simultaneously. In our framework, the model averaging approach estimates a mixture copula based on a criterion function that minimizes a squared loss function. In the next section, we discuss the estimation procedure and prove that our proposed model averaging estimator is asymptotically optimal in the sense of minimizing a squared estimation loss.

3 Theoretical Model

We consider a class of Semiparametric COpula-based Multivariate DYnamic (SCOMDY) models proposed in Chen and Fan (2006b). Define $\left\{\mathbf{Y}_t^\top, \mathbf{Z}_t^\top\right\}_{t=1}^T$ as a vector stochastic process where $\mathbf{Y}_t$ is of dimension p, and $\mathbf{Z}_t$ is a vector of predetermined or exogenous variables. Denote $\mathcal{F}_{t-1}$ as the information set at time t, i.e., $\mathcal{F}_{t-1}$ is the sigma-field generated by $\{\mathbf{Z}_t, \mathbf{Z}_{t-1}, ...; \mathbf{Y}_{t-1}, \mathbf{Y}_{t-2}, ...\}$. The class of SCOMDY models are specified as follows:

$$\mathbf{Y}_t = \mu_t(\beta_{01}) + \sqrt{H_t(\beta_0)}\varepsilon_t,$$

where

$$\mu_t(\beta_{01}) = E\left(\mathbf{Y}_t \mid \mathcal{F}_{t-1}\right)$$

and

$$H_t(\beta_0) = diag(h_{1,t}(\beta_0), ..., h_{p,t}(\beta_0)),$$

in which

$$h_{i,t}(\beta_0) = E[(Y_{it} - \mu_{it}(\beta_{01}))^2 \mid \digamma_{t-1}], \qquad i = 1, ..., p.$$

The unknown parameters β_{01} and β_0 are of fixed dimensions and $\beta_0 = (\beta_{01}^\top, \beta_{02}^\top)^\top$, where β_{01} and β_{02} do not have common elements. We further assume that the standardized innovations $\left\{\varepsilon_t \equiv (\varepsilon_{1t}, ..., \varepsilon_{pt})^\top\right\}$ are independent of $\digamma_{t-1}$. Also, for each $i \in \{1, ..., p\}$, $\{\varepsilon_{it}\}_{t=1}^T$ are independently and identically distributed (i.i.d.) over t with $E(\varepsilon_{it}) = 0$ and $E(\varepsilon_{it}^2) = 1$. Moreover, different component of $\left\{\varepsilon_t \equiv (\varepsilon_{1t}, ..., \varepsilon_{pt})^\top\right\}$ is allowed to be contemporaneously correlated. That is, $\{\varepsilon_{it},\ i = 1, ..., p;\ t = 1, ..., T\}$ are independent across the t-index, but we allow for contemporaneously dependence across the i-index. The SCOMDY models specified here can cover many commonly used specifications such as ARCH, GARCH, and vector autoregressions (VAR); see Chen and Fan (2006b) for a detailed discussion on this.

Our purpose is to estimate the joint distribution of ε_t based on a mixture copula model. Since $\{\varepsilon_t\}$ are unobservable, following Chen and Fan (2006b), we first estimate β_0 by a moment-based method, and obtain estimator $\widehat{\beta} = (\widehat{\beta}_1^\top, \widehat{\beta}_2^\top)^\top$ and residuals $\{\widehat{\varepsilon}_t\} = [\mathbf{Y}_t - \mu_t(\widehat{\beta}_1)]/\sqrt{H_t(\widehat{\beta})}$. Given the residuals $\{\widehat{\varepsilon}_t\}$, we estimate the marginal distributions of ε_{it}, $i = 1, ..., p$, through the rescaled empirical distributions of the residuals:

$$\widetilde{F}_i(x_i) = \frac{1}{T+1} \sum_{t=1}^{T} I\left\{\widehat{\varepsilon}_{it} \le x_i\right\}, \quad i = 1, ..., p.$$

Suppose that we have $K-1$ individual copulas

$$C_k(\mathbf{u}; \boldsymbol{\theta}_k) = C_k\left\{F_1^0(x_1), \ldots, F_p^0(x_p); \boldsymbol{\theta}_k\right\}, \qquad k = 1, \ldots, K-1, \tag{2}$$

where $F_i^0(\cdot)$ is the true (but unknown) marginal distribution of ε_{it}, $i = 1, ..., p$, $\mathbf{u} = (F_1^0(x_1), \ldots, F_p^0(x_p))$ is an arbitrary point in $[0,1]^p$, and $\boldsymbol{\theta}_k$ is a finite dimensional parameter associated with the k^{th} copula. In this paper we assume that K is fixed. In applications, K is often small, therefore, we maintain the fixed K assumption and leave the theoretical investigation of allowing for $K \to \infty$ as $T \to \infty$ to a future research topic.

When $\{\varepsilon_t\}_{t=1}^{\mathrm{T}}$ is thought to be generated from the k^{th} copula, we can estimate $C_k(\mathbf{u}; \boldsymbol{\theta}_k)$ by the Quasi Maximum likelihood estimation (MLE) method after replacing the unknown margins with the estimators, $\widetilde{F}_i(\cdot)$, $i = 1, ..., p$, as stated in Chen and Fan (2006b). Let $\widetilde{\mathbf{u}} = (\widetilde{F}_1(x_1), \ldots, \widetilde{F}_p(x_p))$ and denote the resulting estimator as

$$C_k(\widetilde{\mathbf{u}}; \widehat{\boldsymbol{\theta}}_k) = C_k\left\{\widetilde{F}_1(x_1), \ldots, \widetilde{F}_p(x_p); \widehat{\boldsymbol{\theta}}_k\right\}, \qquad k = 1, \ldots, K-1. \tag{3}$$

We note that if the candidate set only consists of single copulas while the true copula is a mixture of them and is not close to any of these individual copulas, an averaging estimator based on these single copulas can perform poorer than the ML estimator based on a mixture copula. Therefore, we also add the mixture copula into our model average copula candidate set, i.e., we add a mixture of the $K-1$ copulas into the candidate set. Consequently, our candidate set contains K copulas in total. The first $K-1$ are single copulas and the unknown parameters in each copula are estimated by the ML method, the last one is a mixture of the first $K-1$ copulas and the parameters as well as the weights in the mixture copula are also estimated by the ML method. We denote the mixture copula as the K^{th} copula. Let

$$C_K(\widetilde{\mathbf{u}}; \widehat{\boldsymbol{\theta}}_K) = C_K\left\{\widetilde{F}_1(x_1), \ldots, \widetilde{F}_p(x_p); \widehat{\boldsymbol{\theta}}_K\right\} \equiv \sum_{k=1}^{K-1} \breve{\omega}_k C_k\left\{\widetilde{F}_1(x_1), \ldots, \widetilde{F}_p(x_p); \breve{\boldsymbol{\theta}}_k\right\},$$

where $\breve{\omega}_1, \ldots, \breve{\omega}_{K-1}$ and $\breve{\boldsymbol{\theta}}_1, \ldots, \breve{\boldsymbol{\theta}}_{K-1}$ are the maximum likelihood estimates, and $\widehat{\boldsymbol{\theta}}_K \equiv (\breve{\omega}_1, \ldots, \breve{\omega}_{K-1}, \breve{\boldsymbol{\theta}}_1^\top, \ldots, \breve{\boldsymbol{\theta}}_{K-1}^\top)^\top$. Here $\breve{\omega}_k$ is constrained to be between 0 and 1 and the summation equals 1. For each single copula k, $\breve{\boldsymbol{\theta}}_k$ also has its own constraint. For example, the parameter (the correlation coefficient) for Gaussian copula should be between -1 and 1. We use the rescaled empirical distributions of the residuals to replace the unknown marginals, and use the ML method to estimate the finite dimensional parameters.

Let $C_0(\mathbf{u}; \boldsymbol{\theta}_0) = C_0\left\{F_1^0(x_1), \ldots, F_p^0(x_p); \boldsymbol{\theta}_0\right\}$ be the true copula. Note that $C_0(\mathbf{u}; \boldsymbol{\theta}_0)$ can be outside the set $\{C_1(\mathbf{u}; \boldsymbol{\theta}_1), \ldots, C_{K-1}(\mathbf{u}; \boldsymbol{\theta}_{K-1})\}$ and it may not be a mixture copula based on $\{C_1(\mathbf{u}; \boldsymbol{\theta}_1), \ldots, C_{K-1}(\mathbf{u}; \boldsymbol{\theta}_{K-1})\}$. The goal of this paper is to approximate $C_0(\mathbf{u}; \boldsymbol{\theta}_0)$ by the model averaging approach.

Write $\mathbf{w} = (w_1, \ldots, w_K)^\top$ as a weight vector belonging to the set[3]

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

Then, the model averaging method is to use the following weighted average of all candidate copulas to approximate the true unknown copula

$$C(\widetilde{\mathbf{u}}; \widehat{\boldsymbol{\theta}}, \mathbf{w}) = \sum_{k=1}^K w_k C_k(\widetilde{\mathbf{u}}; \widehat{\boldsymbol{\theta}}_k), \tag{4}$$

[3]Note that in Ando and Li (2014), $\sum_{k=1}^K w_k$ is not restricted to be one when selecting weights, but in other recent model averaging literature such as Cheng and Hansen (2015, JoE) and Zhang et al. (2016, JASA), the $\sum_{k=1}^K w_k = 1$ restriction is imposed. In this paper, we keep this restriction because less restriction on weights can make it harder to find the optimal weights. In Ando and Li (2014), the candidate models have no overlapping regressors, but in the current paper, the first $K-1$ copulas are nested in the K^{th} copula, using this restriction is appropriate for our method. In Appendix A, we compare model averaging methods imposing and without imposing this restriction. We find that the out-of-sample predicting error imposibg the restriction is smaller than that without imposing the restriction, but the difference is not statistically significant.

where $\widehat{\boldsymbol{\theta}} = \left(\widehat{\boldsymbol{\theta}}_1^\top, \ldots, \widehat{\boldsymbol{\theta}}_K^\top\right)^\top$.

We would like to make a comment on the above model averaging estimator. Here we will omit the $(\widetilde{F}_1(x_1), \ldots, \widetilde{F}_p(x_p))$ argument in copula functions to save space and simplify notations. The K^{th} copula is a 'mixture' of the first $(K-1)$ individual copulas which is estimated as $\sum_{k=1}^{K-1} \breve{\omega}_k C_k\left(\breve{\boldsymbol{\theta}}_k\right)$, where $\breve{\omega}_k$ and $\breve{\boldsymbol{\theta}}_k$, $k = 1, ..., K-1$, are the ML estimates of ω_k and $\boldsymbol{\theta}_k$ based on the mixture copula model. The model 'averaging' is a linear combination of the K copulas: $C_1\left(\widehat{\boldsymbol{\theta}}_1\right), \ldots, C_{K-1}\left(\widehat{\boldsymbol{\theta}}_{K-1}\right)$ and $\sum_{k=1}^{K-1} \breve{\omega}_k C_k\left(\breve{\boldsymbol{\theta}}_k\right)$, where $\widehat{\boldsymbol{\theta}}_k$ is the ML estimator of $\boldsymbol{\theta}_k$ based on the k^{th} copula, $k = 1, ..., K-1$. At a first glance, one may get an impression that the K^{th} copula is redundant as it looks like a linear combination of the first $K-1$ copulas. However, this is not the case because $\breve{\boldsymbol{\theta}}_k$ differs from $\widehat{\boldsymbol{\theta}}_k$ in general. Hence, the K^{th} mixture copula is not a linear combination of the first $K-1$ copulas. In fact, simulations (not reported here to save space) show that by adding the K^{th} mixture copula to the candidate set, the model averaging estimator performs much better than an estimator that only averages over the first $K-1$ copulas.

A crucial question in model average estimation is how to select the weight $\mathbf{w}$. In this paper we use a J-fold ($J > 1$ is a finite positive integer) Cross-Validation (CV) method to choose weights, which is similar to the Jackknife model averaging 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 except that the last group may have more data (if T/J is not an integer). In the j^{th} group, we have observations $\widehat{\varepsilon}_{(j-1)M+1}, \ldots, \widehat{\varepsilon}_{jM}$ for $j = 1, ..., J$. Write $C_k(\widetilde{\mathbf{u}}; \widehat{\boldsymbol{\theta}}_k^{(-j)})$ as the estimator of $C_k(\mathbf{u}; \boldsymbol{\theta}_k)$ with the j^{th} group removed from the sample, i.e., it is the same as $C_k(\widetilde{\mathbf{u}}; \widehat{\boldsymbol{\theta}}_k)$ except that one drops the j^{th} group in the MLE. Let $\widehat{\boldsymbol{\theta}}^{(-j)} = \left(\widehat{\boldsymbol{\theta}}_1^{(-j)\top}, \ldots, \widehat{\boldsymbol{\theta}}_K^{(-j)\top}\right)^\top$. Then similar to $C(\widetilde{\mathbf{u}}; \widehat{\boldsymbol{\theta}}, \mathbf{w})$ defined in (4), we define the leave j^{th} group out weighted averaging estimator $C(\widetilde{\mathbf{u}}; \widehat{\boldsymbol{\theta}}^{(-j)}, \mathbf{w})$ as

$$C(\widetilde{\mathbf{u}}; \widehat{\boldsymbol{\theta}}^{(-j)}, \mathbf{w}) = \sum_{k=1}^{K} w_k C_k(\widetilde{\mathbf{u}}; \widehat{\boldsymbol{\theta}}_k^{(-j)}), \qquad j = 1, \ldots, J.$$

An empirical estimator of $C_0(\mathbf{u}; \boldsymbol{\theta}_0)$ that only uses the j^{th} group data is denoted by

$$\widetilde{C}_{(j)}(\mathbf{x}) = \frac{1}{M} \sum_{m=1}^{M} I(\widehat{\varepsilon}_{(j-1)M+m} \leq \mathbf{x}), \tag{5}$$

where $\mathbf{x} = (x_1, ..., x_p)$ is an arbitrary point in $\mathcal{R}^p$, $I(\cdot)$ is an indicate function, and the comparison between $\widehat{\varepsilon}_{(j-1)M+m}$ and $\mathbf{x}$ means comparison componentwise in the p-dimensional vector. We emphasize that the superscript $(-j)$ denotes "leave the j^{th} group data out" and the subscript (j) means "only use the j^{th} group data".

Define $U_{0t} \equiv (F_1^0(\varepsilon_{1t}), \ldots, F_p^0(\varepsilon_{pt}))$ and $\widetilde{U}_t \equiv (\widetilde{F}_1(\widehat{\varepsilon}_{1t}), \ldots, \widetilde{F}_p(\widehat{\varepsilon}_{pt}))$, then our J-fold CV criterion is given by

$$CV_J(\mathbf{w}) = \sum_{j=1}^{J}\sum_{m=1}^{M}\left\{C(\widetilde{U}_{(j-1)M+m};\widehat{\boldsymbol{\theta}}^{(-j)},\mathbf{w}) - \widetilde{C}_{(j)}(\widehat{\varepsilon}_{(j-1)M+m})\right\}^2. \tag{6}$$

The weight $\mathbf{w}$ is selected via

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

and we estimate $C_0(\mathbf{u};\boldsymbol{\theta}_0)$ by the model averaging estimator $C(\widetilde{\mathbf{u}};\widehat{\boldsymbol{\theta}},\widehat{\mathbf{w}})$ as defined in (4) with $\widehat{\mathbf{w}}$ replacing $\mathbf{w}$.

To ease exposition, we introduce/summarize notations used in the paper. The $T\times 1$ vector of the true copula evaluated at $(\varepsilon_1,..., \varepsilon_T)$ is denoted by

$$\mathbf{C}_0 = \left\{C_0(U_{01};\boldsymbol{\theta}_0),\ldots,C_0(U_{0T};\boldsymbol{\theta}_0)\right\}^\top. \tag{7}$$

The vector of copula estimated by the k^{th} candidate copula using all observations (when $k=K$, the K^{th} copula is a composite of the candidate copulas) evaluated at $(\widehat{\varepsilon}_1,\ldots,\widehat{\varepsilon}_T)$ is

$$\widehat{\mathbf{C}}_k = \left\{C_k(\widetilde{U}_1;\widehat{\boldsymbol{\theta}}_k),\ldots,C_k(\widetilde{U}_T;\widehat{\boldsymbol{\theta}}_k)\right\}^\top. \tag{8}$$

The vector of the weighted average of the estimated candidate copulas $\left\{\widehat{\mathbf{C}}_1,\ldots,\widehat{\mathbf{C}}_K\right\}$ evaluated at $(\widehat{\varepsilon}_1,\ldots,\widehat{\varepsilon}_T)$ is denoted by

$$\widehat{\mathbf{C}}(\mathbf{w}) = \sum_{k=1}^{K} w_k\widehat{\mathbf{C}}_k = \left\{C(\widetilde{U}_1;\widehat{\boldsymbol{\theta}},\mathbf{w}),\ldots,C(\widetilde{U}_T;\widehat{\boldsymbol{\theta}},\mathbf{w})\right\}^\top. \tag{9}$$

The leave-M-out vector of copula estimated by using the k^{th} candidate copula evaluated at $(\widehat{\varepsilon}_1,\ldots\widehat{\varepsilon}_T)$ is

$$\overline{\mathbf{C}}_k = \left\{C_k(\widetilde{U}_1;\widehat{\boldsymbol{\theta}}_k^{(-1)}),\ldots,C_k(\widetilde{U}_M;\widehat{\boldsymbol{\theta}}_k^{(-1)}),...,C_k(\widetilde{U}_{(j-1)M+1};\widehat{\boldsymbol{\theta}}_k^{(-j)}),\ldots,C_k(\widetilde{U}_T;\widehat{\boldsymbol{\theta}}_k^{(-J)})\right\}^\top. \tag{10}$$

The vector of the weighted average of $\overline{\mathbf{C}}_1,\ldots,\overline{\mathbf{C}}_K$ evaluated at $(\widehat{\varepsilon}_1,\ldots,\widehat{\varepsilon}_T)$ is

$$\begin{aligned}\overline{\mathbf{C}}(\mathbf{w}) &= \sum_{k=1}^{K} w_k\overline{\mathbf{C}}_k \\ &= \left\{C(\widetilde{U}_1;\widehat{\boldsymbol{\theta}}^{(-1)},\mathbf{w}),\ldots,C(\widetilde{U}_M;\widehat{\boldsymbol{\theta}}^{(-1)},\mathbf{w}),\ldots,C(\widetilde{U}_T;\widehat{\boldsymbol{\theta}}^{(-J)},\mathbf{w})\right\}^\top. \end{aligned}\tag{11}$$

The vector of the empirical estimator of $\mathbf{C}_0$ using M observations evaluated at $(\widehat{\varepsilon}_1,\ldots,\widehat{\varepsilon}_T)$

is

$$\widetilde{\mathbf{C}} = \left\{\widetilde{C}_{(1)}(\widehat{\varepsilon}_1), \ldots, \widetilde{C}_{(1)}(\widehat{\varepsilon}_M), ..., \widetilde{C}_{(j)}(\widehat{\varepsilon}_{(j-1)M+1}), \ldots, \widetilde{C}_{(J)}(\widehat{\varepsilon}_T)\right\}^\top. \tag{12}$$

Finally, the vector of the model averaging estimator of $\mathbf{C}_0$ evaluated at $(\widehat{\varepsilon}_1, \ldots, \widehat{\varepsilon}_T)$ is

$$\widehat{\mathbf{C}}(\widehat{\mathbf{w}}) = \left\{C(\widetilde{U}_1; \widehat{\boldsymbol{\theta}}, \widehat{\mathbf{w}}), \ldots, C(\widetilde{U}_T; \widehat{\boldsymbol{\theta}}, \widehat{\mathbf{w}})\right\}^\top. \tag{13}$$

Note that in general we use the 'hat' notation to denote estimators based on semiparametric model estimation methods (because the margins are nonparametrically estimated), while we use the 'tilde' notation to denote nonparametric (empirical function based) estimators.

Proposition 1

(i). Under Conditions C.1 – C.4 presented in Appendix A.1, the bias of the J-fold CV criterion is a small term relative to the expected CV squared loss. That is,

$$\sup_{\mathbf{w}\in\mathcal{W}} \frac{E\left\|\overline{\mathbf{C}}(\mathbf{w}) - \mathbf{C}_0\right\|^2 - E\left\|\overline{\mathbf{C}}(\mathbf{w}) - \widetilde{\mathbf{C}}\right\|^2}{E\left\|\overline{\mathbf{C}}(\mathbf{w}) - \mathbf{C}_0\right\|^2} = o(1).$$

(ii). Under an i.i.d. situation, i.e., $\{\widehat{\varepsilon}_{it},\ i = 1, ..., p;\ t = 1, ..., T\}$ are independent across the t-index, the J-fold CV criterion is an unbiased estimator of the expected CV squared loss plus a term unrelated to $\mathbf{w}$. That is,

$$E\left\|\overline{\mathbf{C}}(\mathbf{w}) - \widetilde{\mathbf{C}}\right\|^2 = E\left\|\overline{\mathbf{C}}(\mathbf{w}) - \mathbf{C}_0\right\|^2 + \{\text{terms unrelated to } \mathbf{w}\}.$$

See Appendix A.3 for the proof of Proposition 1. Let $\breve{\mathbf{h}}_k = \overline{\mathbf{C}}_k - \widetilde{\mathbf{C}}$ and $\breve{\mathbf{H}} = (\breve{\mathbf{h}}_1, \ldots, \breve{\mathbf{h}}_K)$. We can rewrite the J-fold CV criterion as

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

which is quadratic in $\mathbf{w}$. Hence, the minimization of $CV_J(\mathbf{w})$ with respect to $\mathbf{w}$ can be implemented easily.

Define a quadratic loss function of the model averaging estimator by

$$L_T(\mathbf{w}) = \|\widehat{\mathbf{C}}(\mathbf{w}) - \mathbf{C}_0\|^2. \tag{14}$$

Like the 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.

THEOREM 1 *Under Conditions C.1 – C.4 presented in Appendix A.1,*

$$\frac{L_T(\widehat{\mathbf{w}})}{\inf_{\mathbf{w}\in\mathcal{W}} L_T(\mathbf{w})} \to 1 \quad \text{in probability} \ \ (\text{as } T \to \infty). \tag{15}$$

The detailed proof of Theorem 1 is given in Appendix A. Theorem 1 states that our model averaging estimator $\widehat{\mathbf{C}}(\widehat{\mathbf{w}})$ is asymptotically optimal in the sense that the squared loss of $\widehat{\mathbf{C}}(\widehat{\mathbf{w}})$ is asymptotically identical to that by the infeasible best possible model averaging estimator. So the squared loss is minimized approximately. In addition, it is obviously that $\inf_{\mathbf{w}\in\mathcal{W}} L_T(\mathbf{w}) \leq \inf_{k\in\{1,\ldots,K\}} \|\widehat{\mathbf{C}}_k - \mathbf{C}_0\|^2$, so it is expected that the model averaging can reduce estimation error relative to using a candidate copula.

Another important question is how to decide the appropriate values of K and J. For the choice of K, we propose using Cai and Wang (2015)'s method to filter out those candidates with zero weights. Then the proposed model averaging method could be implemented to those preserved candidates after the screening. For the choice of J, we follow a method proposed by Ma and Zhu (2012). Specifically, we choose J from a discrete set consisting of l candidates $J \in \{J_1, J_2, ..., J_l\}$. For each $s \in \{1, ..., l\}$, we use the block bootstrap method, with a block size m, to generate a bootstrap sample, and we repeat the process B times to obtain B bootstrap samples: $[\{\mathbf{Y}_t^\top, \mathbf{Z}_t^\top\}_{t=1}^T]^{[1]}, \ldots, [\{\mathbf{Y}_t^\top, \mathbf{Z}_t^\top\}_{t=1}^T]^{[B]}$. Using the model averaging procedure discussed above, we can calculate $\widehat{\mathbf{w}}_1^1, ..., \widehat{\mathbf{w}}_B^1, ..., \widehat{\mathbf{w}}_1^l, ..., \widehat{\mathbf{w}}_B^l$. For candidate J_s, we can calculate the sample variance matrix using $\widehat{\mathbf{w}}_1^s, ..., \widehat{\mathbf{w}}_B^s$ and denote it as V_{J_s} for $s = 1, ..., l$. Finally, we select J^* that gives the minimum value of $trace(V_J)$, i.e., $J^* = \arg\min_{J\in\{J_1,...,J_l\}} trace(V_J)$. In our simulation, we use $m = 10$ and $B = 100$.

4 Numerical Studies

We compare squared estimation losses using our proposed model average method with three other methods: Cai and Wang's (2014) penalized likelihood method, the BIC method which selects one copula by comparing each candidate's BIC, and the maximum likelihood method (Hu (2006)). We consider two types of simulations. In Type I simulation, data are generated from copulas which are included in the mixture copula model. In contrast, 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 accurate description of the data dependence structure under these two settings.

4.1 Simulation Type I

In type I simulation we consider the scenario that data are generated from copulas which are constituents of our working model. We first consider a bi-variate case that the data are generated by an AR(1)-GARCH(1,1) process:

$$y_{it} = \gamma_i y_{i,t-1} + e_{it}, \qquad\qquad i = 1, 2;\ t = 1, ..., T,$$

where $e_{it} = \sigma_{it}\epsilon_{it}$, ϵ_{it} has a standard normal marginal distribution with the dependence structure between ϵ_{1t} and ϵ_{2t} governed by a given form of copula function, and

$$\sigma_{it} = \sqrt{\alpha_{i0} + \alpha_{i1} e^2_{i,t-1} + \beta_{i1}\sigma^2_{i,t-1}},$$

where $\gamma_1 = 0.05$, $\alpha_{10} = 0.0001$, $\alpha_{11} = 0.95$, $\beta_{11} = 0.04$ for the first margin; and $\gamma_2 = 0.1, \alpha_{20} = 0.0001$, $\alpha_{21} = 0.90$, $\beta_{21} = 0.09$ for the second margin. Our working mixture model includes three commonly used copulas: Gaussian, Clayton and Gumbel. Characteristics of the three copulas have been discussed in Section 2 and the simulated scatter plots have been displayed in Figure 1. We first use Cai and Wang's (2014) method to screen the three candidate copulas and then decide which one should be kept in the candidate pool. Suppose all the three candidates have non-zero weights under Cai and Wang's (2014) method, our presumed mixture copula is formulated as:

$$C(\boldsymbol{u}; \boldsymbol{\theta}, \boldsymbol{\omega}) = \omega_{Ga} C_{Ga}(\boldsymbol{u}; \boldsymbol{\theta}_1) + \omega_{Cl} C_{Cl}(\boldsymbol{u}; \boldsymbol{\theta}_2) + \omega_{Gu} C_{Gu}(\boldsymbol{u}; \boldsymbol{\theta}_3),$$

where C_{Ga}, C_{Cl} and C_{Gu} stand for Gaussian, Clayton and Gumbel copulas, respectively, and $\boldsymbol{u} = (F_1(\cdot), F_2(\cdot))$ denote the two margins. By fitting the data into Gaussian, Clayton and Gumbel copulas separately, we obtain their maximum likelihood estimates $\widehat{\boldsymbol{\theta}}_1$, $\widehat{\boldsymbol{\theta}}_2$ and $\widehat{\boldsymbol{\theta}}_3$.

We argued in Section 3 that when the true copula is a mixture copula, an averaging estimator based on single copulas can lead to a poor fit. We thus include a maximum likelihood (ML) estimator of the mixture copula into our model average estimator. Specifically, let $C_{ML}(\tilde{\boldsymbol{u}}; \widehat{\boldsymbol{\theta}}_4) = \breve{\omega}_1 C_{Ga}(\tilde{\boldsymbol{u}}; \breve{\boldsymbol{\theta}}_1) + \breve{\omega}_2 C_{Cl}(\tilde{\boldsymbol{u}}; \breve{\boldsymbol{\theta}}_2) + \breve{\omega}_3 C_{Gu}(\tilde{\boldsymbol{u}}; \breve{\boldsymbol{\theta}}_3)$, where $\tilde{\boldsymbol{u}} = (\tilde{F}_1(\cdot), \tilde{F}_2(\cdot))$ are re-scaled empirical CDF estimates of $\boldsymbol{u}$, $\breve{\omega}_1, \breve{\omega}_2, \breve{\omega}_3$ and $\breve{\boldsymbol{\theta}}_1, \breve{\boldsymbol{\theta}}_2, \breve{\boldsymbol{\theta}}_3$ are the ML estimates of $\omega_1, \omega_2, \omega_3$ and $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3$, respectively. Let $\widehat{\boldsymbol{\theta}}_4 = (\breve{\omega}_1, \breve{\omega}_2, \breve{\omega}_3, \breve{\boldsymbol{\theta}}_1^\top, \breve{\boldsymbol{\theta}}_2^\top, \breve{\boldsymbol{\theta}}_3^\top)^\top$. Our method then averages over 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

$$C(\tilde{u}, \tilde{v}; \widehat{\boldsymbol{\theta}}, \mathbf{w}) = w_{Ga} C_{Ga}(\tilde{u}, \tilde{v}; \widehat{\boldsymbol{\theta}}_1) + w_{Cl} C_{Cl}(\tilde{u}, \tilde{v}; \widehat{\boldsymbol{\theta}}_2) + w_{Gu} C_{Gu}(\tilde{u}, \tilde{v}; \widehat{\boldsymbol{\theta}}_3) + w_{ML} C_{ML}(\tilde{u}, \tilde{v}; \widehat{\boldsymbol{\theta}}_4)$$

via the model averaging method by minimizing the estimation losses $L_T(\mathbf{w})$ defined in Section

3. For simplicity, the model averaging method with our proposed methods selecting K and J is denoted as MA^* in this section.

We consider two sample sizes: $T = 200$ and 500. We will focus on out-of-sample forecasting evaluations of different estimation methods. We choose in-sample size equals the out-of-sample size. For $T = 200$ (500), we use a sample of 200 (500) to estimate a model and then we evaluate the estimation squared loss using another 200 (500) out-of-sample data observations. We consider four different values for $J \in \{2, 4, 5, 10\}$. For each J, we use block bootstrap size $m = 10$ to generate $B = 100$ bootstrap samples, and use the method discussed at the end of Section 3 to select J^*. All simulations are repeated 2000 times. We simulate three mixture copulas with two components and one mixture copula with three components. Specifically, we have the following four cases for the setup of weights and the data are simulated from different copulas:

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

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

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

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

For each case of the above weights, we consider two sets of copula parameters:

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

$$\text{Parameter setting 2:} \quad \theta_{Ga} = 0.7,\ \theta_{Cl} = 7.8,\ \theta_{Gu} = 7.1.$$

Therefore, we have $4 \times 2 = 8$ DGPs in total.

Table 1 displays how close the estimated copula is to the true copula in terms of out-of-sample mean squared estimation loss across the four methods we mentioned earlier: our proposed model average approach (MA^*), Cai and Wang's (2014) penalized likelihood (CW), the BIC method which selects one from a set of candidates based on BIC (BIC), and the maximum likelihood method. To save space and for expositional ease, we only present the ratios of out-of-sample prediction errors of CW to MA^* (CW/MA^*), BIC to MA^* (BIC/MA^*) and MLE to MA^* (MLE/MA^*). Therefore, MA^* is superior to CW, or BIC, or MLE if the ratio is greater than one. To further check whether such a superiority is statistically significant, we implement the Diebold-Mariano (DM) test and report the $p-$value for each case.

[INSERT TABLE 1 ABOUT HERE]

As we stated above, Table 1 presents the out-of-sample prediction performance of the four competing methods. The number of out-of-sample observations is equal to the number of in-sample observations. Hence, for 200 and 500 in-sample observations, the out-of-sample predictions include 200 and 500 observations, respectively. The results in Table 1 show that mixture copula models perform better than using an individual copula (BIC). Among the four competing methods, BIC gives the largest prediction errors for all cases (BIC/MA^* is greater than one and larger than CW/MA^* and MLE/MA^*), while the performances of CW and MA^* are similar to each other because the estimation MSE ratios of CW/MA^* and MLE/MA^* are all close to one, and their $p-$values are quite large in almost all cases.

In terms of out-of-sample prediction errors, Type I simulation demonstrates the superiority of using a mixture copula model over a model that uses only one individual copula. Furthermore, for the estimation of a mixture copula model, the performance of our proposed model average approach is similar to those of Cai and Wang's (2014) method and the MLE method.

4.2 Simulation Type II

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

In the simulation setup, our working mixture model is still comprised by Gaussian, Clayton and Gumbel copulas but the true observations are generated from a linear combination of Frank, Survival Joe (SJ) and Joe copulas. These three copulas are also widely used in empirical studies. The Frank copula is similar to the Gaussian copula as it does not exhibit tail dependence, but has relatively stronger dependence in the center of the distribution. The Joe copula, like the Gumbel copula, exhibits right tail dependence. The Survival Joe copula is a 180° rotation of Joe, so it exhibits left tail dependence as the Clayton copula. We consider four cases: in the first three cases, the true copula is generated from a linear combination of two of the Frank, Joe and Survival Joe copulas. In the fourth case, each candidate copula has the equal weight, i.e., 1/3. Specifically, we consider the following four weighting setups:

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

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

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

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

and we use the following two sets of different copula parameters:

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

Parameter setting 2: $\theta_{Frank} = 7.5$, $\theta_{SJ} = 6.8$, $\theta_{Joe} = 6.5$.

We present the out-of-sample prediction results of the four competing methods under Type II simulation in Table 2. Same to the Type I case, the number of out-of-sample predictions is the same with the corresponding sample size of the in-sample data. The patterns in Table 2 show that MA^* exhibits more accurate out-of-sample predictions than CW, BIC and MLE because the estimation MSE ratios of CW to MA^*, BIC to MA^* and MLE to MA^* are all greater than one, and most p-values are quite small.

[INSERT TABLE 2 ABOUT HERE]

The simulation results show that our model averaging method outperforms CW, BIC and MLE methods when the working mixture copula is misspecified. This finding has important implications for empirical studies as the true copula is always unknown to researchers and the misspecification of one's working mixture model should be quite common in practice.

Next, we examine the performance of the four estimation methods with copula functions that have three or four components. We generate additional 3rd and 4th components via

$$y_{it} = \gamma_i y_{i,t-1} + e_{it}, \qquad i = 3, 4;\ t = 1, ..., T,$$

where $e_{it} = \sigma_{it}\epsilon_{it}$, ϵ_{it} has a standard normal marginal distribution with the dependence structure among ϵ_{1t}, ϵ_{2t}, ϵ_{3t} and ϵ_{4t} governed by a given form of copula function, and

$$\sigma_{it} = \sqrt{\alpha_{i0} + \alpha_{i1} e_{i,t-1}^2 + \beta_{i1}\sigma_{i,t-1}^2},$$

where $\gamma_3 = 0.09$, $\alpha_{30} = 0.0001$, $\alpha_{31} = 0.94$, $\beta_{31} = 0.05$ for the third margin; and $\gamma_4 = 0.15, \alpha_{40} = 0.0001$, $\alpha_{41} = 0.91$, $\beta_{41} = 0.08$ for the fourth margin. The other parts of the setup for simulations is similar to the two-component case discussed earlier. Both the in-sample (estimation) size and the out-of-sample (forecast) size are $T = 500$. To save space, we only conduct Type II simulations for the three and four components copula cases. Our working mixture model is still comprised by Gaussian, Clayton and Gumbel copulas but the true observations are generated from a linear combination of Frank, Survival Joe (SJ) and Joe copulas.

Table 3 displays the ratios of mean squared out-of-sample prediction losses of the model averaging method to the other three competing methods for the 3-component and 4-component copula cases. If the ratio is greater than 1, it implies that MA^* outperforms the corresponding competing method. As documented in Table 3, it can be seen that MA^* outperforms the other three competing methods in all cases.

[INSERT TABLE 3 ABOUT HERE]

In Appendix A, we run additionally simulations to compare the performance between: (1). the original model averaging method with fixed K and J (no selections of K and J; denoted as MA hereafter) and MA without $\sum_{k=1}^{K} w_k = 1$ constraint; (2). MA and MA with the screening step to select K (MAS); (3). MA with our proposed data-driven method selecting J and MA with $J \in \{2, 4, 5, 10\}$, respectively. All the simulation results are documented in Appendix A. In Table A1, the out-of-sample forecast performance between MA and MA without $\sum_{k=1}^{K} w_k = 1$ constraint does not exhibit significant difference because the ratio is close to 1. Table A2 demonstrates that, even though the ratio between MAS and MA is less than 1, the difference of the two methods is not statistically significant based on the DM test. This indicates that the additional screening step does not effectively improve the performance of MA. Table A3 shows that MA with selected J outperforms MA with fixed J in most cases and the difference is significant in most cases. This proves that the proposed J selection strategy is useful in improving the performance of MA.

5 An Empirical Study

In this section we use a real data to examine the performance of using the model average approach to estimate a mixture copula model. We consider daily returns of Morgan Stanley Capital International (MSCI) equity indexes for four developed economies: United States (US), United Kingdom (UK), Hong Kong (HK) and Japan (JP). The daily data span 12 years from August, 2002 to December, 2014, for a total of 3220 observations. We download these equity indexes from Datastream and calculate log returns of the four indexes. For comparing purposes, the currency for the daily indexes in United Kingdom, Hong Kong and Japan are converted into US dollars based on their respective contemporary exchange rates.

We split the data into two equal parts: the first 1610 observations (training set), ranging from August of 2002 to October of 2008, are used to fit a model, and the remaining 1610 observations (testing set) are used to examine the out-of-sample prediction accuracy across the competing methods. Table 4 displays the summary statistics for daily log-returns of MSCI indexes for the four markets. Over the 6 years between 2002 and 2008, HK market had the highest average daily return while the UK market had the highest median daily return. The skewness is negative for all markets, indicating higher probabilities in having extreme daily losses than having extreme daily gains in these markets. Kurtosis for US, UK and HK markets are greater than 3, while it is smaller than 3 for JP market. These statistics indicate that it can be difficult to correctly specify each marginal distribution in practice and nonparametric methods should be used to estimate the margins.

[INSERT TABLE 4 ABOUT HERE]

Table 5 demonstrates the linear correlation coefficients and Kendall's τs (in parentheses) for each pair. We observe that the HK-JP pair has the strongest correlation or degree of dependence based on both correlation coefficients and Kendall's τs. Figure 2 shows pairwise scatter plot for each pair. Daily returns in each pair appear to be positively correlated especially for US-UK, JP-HK and UK-HK pairs. Figure 2 also displays a violation of the elliptical multivariate distributions. Asymmetry and extreme data can be observed from each pair. Figure 2 further confirms the existence of large amount of extreme data 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 not be able to discover characteristics of the joint distribution. To take advantage of the flexibility of each individual copula, we consider a mixture copula model.

[INSERT TABLE 5 ABOUT HERE]

[INSERT FIGURE 2 ABOUT HERE]

Spurious regression results may be generated if a pair of time series data is processed inappropriately (see Granger and Newbold, 1974; Chen and Fan, 2006a). Hu (2006) also argues that, due to the clustering of large volatilities, data with conditional heteroscedasticity can lead to underestimation of the degree of dependence. Preliminary examination indicated the existence of both autocorrelation and conditional heteroscedasticity in the daily returns for the four economies. To filter both effects, we specify an AR(1)-GARCH(1,1) model. The filtered daily percentage changes (the residuals) are then substituted into the working mixture copula model. An AR(1)-GARCH(1,1) model is a special case of the class of SCOMDY models stated in Section 3. According to Theorem 1, the model averaging method should generate an asymptotically optimal estimator in the sense of minimizing the squared estimation loss.

Next, we fit the filtered daily returns in the four economies into a mixture copula model which includes Student's t, Clayton and Gumbel.[4] We implement the model averaging

[4] We thank a referee for suggesting the use of a Student's t copula. Compared with a Gaussian copula, a Student's t copula is more suitable to financial data since it has the tail dependence property and is able to capture correlations in the extreme market movements.

method, Cai and Wang's (2014) penalized likelihood method, the BIC method and the MLE method to estimate the mixture copula model respectively. To compare the estimation performance across the four methods, we follow a procedure as suggested by Genest and Rivest (1993). Specifically, we construct five 7×7 cross-classification tables for each of the six pairs. For each pair, the numbers in the first cross-classification table is computed using the observed data and the other four tables are computed using the four competing methods, respectively. Let G represents a table and $G_{i,j}$ be the number appears at the cell (i, j) (i.e., in the ith row and jth column of table G) for a pair of markets (call them markets 1 and 2), where $i, j = 1, ..., 7$. Let u_i and v_j be the upper bounds for determining counts $G_{i,j}$ in cell (i, j), where u_i and v_j are defined as the $i/7$ and $j/7$ percentiles of the two return series, respectively $(i, j = 1, ..., 7)$. Then a pair of observations (u, v) belongs to the cell (i, j) if $u_{i-1} < u \leq u_i$ and $v_{j-1} < v \leq v_j$. Thus, $G_{i,j}$, the entry in cell (i, j), is the number of times the daily return of market 1 falls between the $(i-1)/7$ and the $i/7$ percentile of its data range, and that of market 2 is within the $(j-1)/7$ and the $j/7$ percentile of its data range. For example, the number recorded in the cell $(3, 2)$ indicates the number of times that daily return percentage changes of the first market is between the 29th (2/7) and the 43rd (3/7) percentile of its data range, while that of the second market falls within 14th (1/7) and 29th (2/7) percentile of its data range. Thus, if the two markets are strongly positively correlated, we should see that most observations lie on the principal diagonal. If they are strongly 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 be similar to each other.

We take 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 14th (1/7) percentile of their respective ranges; that is, the number of times when both markets face downturn risk simultaneously. Similarly, the cell at the bottom-right shows the frequency that both daily returns are above the 86th (6/7) percentile. During the period between August 2002 and October 2008 (a total of 1610 observations), there are 83 times that daily returns of MSCI indexes in UK and Hong Kong are both lower than their 14th percentile. In contrast, there are 64 times that daily returns in UK and Hong Kong market are both higher than their 86th percentile. Thus, UK and Hong Kong stock markets exhibit higher probability to co-move downward than to co-move upward. US-HK, UK-JP and JP-HK display similar left tail dependence structures. Such a pattern is not significant in the US-JP and US-UK markets.

We evaluate the fit of an estimation method by comparing the sum of squared differences between the estimated count and the observed count over all the $7 \times 7 = 49$ cells. We now describe how to compute the estimated count in each cell. Let $\widehat{C}$ denotes an estimated copula. The estimated count in cell (i, j) for example, can be obtained by multiplying the probability $\widehat{C}(u_i, v_j) - \widehat{C}(u_i, v_{j-1}) - \widehat{C}(u_{i-1}, v_j) + \widehat{C}(u_{i-1}, v_{j-1})$ to the sample size 1610. Let $G_{i,j}$ and $G_{i,j}^{MA^*}$ denote the frequency observed and the frequency estimated by the model averaging method in cell (i, j) respectively. Then, we define the estimation squared error as:

$$Q_{MA^*} = \frac{1}{k^2} \sum_{i=1}^{k} \sum_{j=1}^{k} (G_{i,j} - G_{i,j}^{MA^*})^2. \tag{16}$$

In our case, $k = 7$. By the same manner, we can calculate estimation squared errors for CW, BIC and MLE methods, which are respectively denoted as Q_{CW}, Q_{BIC} and Q_{MLE}. The estimation squared errors and the ratios are displayed in Table 6. Among all cases, the model averaging method exhibits the smallest estimation squared errors, while the BIC method which relies on the comparison of BIC among Student's t, Clayton and Gumbel copulas gives the largest estimation errors. The empirical study again shows that the model average approach gives more accurate estimates of dependence structures comparing with other competing methods. The in-sample fit results reported in Table 6 supports our model averaging method.

Next, we consider the out-of-sample prediction performance of the four competing methods based on the daily observations between October 2008 and December 2014. The estimated frequencies are based on our estimated mixture copula model under MA^*, CW, BIC and MLE, respectively. The out-of-sample predicting errors are calculated in a similar way as we did to calculate the in-sample estimation losses. We present the prediction squared errors of each competing method and the ratios of CW to MA^*, BIC to MA^* and MLE to MA^* for each pair. The results are displayed in Table 7. It is obvious that the model averaging method exhibits the smallest predicting errors, indicating a good out-of-sample prediction performance of the model averaging method.

We notice that the dependent structure among different stock markets may change significantly as time goes by, especially during the period that the financial markets fluctuate acutely. Ideally, one should allow for the parameters in a mixture model to change over time to capture the dynamics among the international financial markets. Although some approaches on time-varying copulas have been developed (see, for example, Patton (2006) and Manner and Reznikova (2012)), the extension of our semiparametric (with nonparametric marginals) model averaging method to the time varying copula framework is a challenge research problem (e.g., Fan and Patton (2014)) and we leave it as a future research topic. To partially deal with the time varying nature of the dependence structure among the returns from the four developed stock markets, we use a rolling window method to examine the performance among MA^*, CW, BIC and MLE. Specifically, we first estimate MA^*, CW, BIC and MLE based on six years' observations from 2002 to 2008 and then predict the dependence structure among US, UK, JP and HK in 2009. Next, we re-estimate MA^*, CW, BIC and MLE by using the observations from 2003 to 2009 and then predict the dependence in 2010. We repeat this procedure and make the final prediction on the dependence structure in 2014 based on the previous 6 years' observations. By doing this, we can update our mixture model by using the most recent 6 years' observations. We define the estimation squared error of MA^* based on the rolling window method (RW) as:

$$Q^{rw}_{MA^*} = \frac{1}{k^2} \sum_{i=1}^{k} \sum_{j=1}^{k} \sum_{l=1}^{6} (G_{l,i,j} - G^{MA^*}_{l,i,j})^2,$$

and in our case, $k = 7$. By the same manner, we calculate estimation squared errors for CW, BIC and MLE methods, which are respectively denoted as Q^{rw}_{CW}, Q^{rw}_{BIC} and Q^{rw}_{MLE}. The estimation squared errors and the ratios are displayed in Table 8.

[INSERT TABLE 7 & 8 ABOUT HERE]

We have two observations in Table 8. First, our model averaging estimator still outperforms CW, BIC and MLE in terms of squared estimation loss based on the rolling window method. Second, compared with the results in Table 7, the estimation losses in Table 8 are smaller in magnitude (except the HK-JP pair). This is expected as we can improve the prediction accuracy by updating the copula model via using the most recent data through the rolling window method.

Finally, we compare the performance of the four methods: MA^*, CW, BIC and MLE based on a four-component copula model by including all the four markets in a four-component copula. The calculating procedure is exactly the same as the two-component copula. However, in calculating the prediction squared errors under the four-component copula we do not split the data into the 7^4 cross-classification table because doing so will make observations in each cell too small. Instead, we respectively consider results from 2^4, 3^4 and 4^4 cross-classification tables. For example, for the 2^4 cross-classification case, we order the returns of each stock and divide them into two parts, one part below the median return (we use 1 to denote this) and the other part above the median return (we use 2 to denote this). Then we can label the $2^4 = 16$ cells as (i_1, i_2, i_3, i_4) for $i = 1, 2$ and $j = 1, 2, 3, 4$. The cell (1,1,1,1) denotes the case that all four indices are below their respective median returns; similarly, (1,1,2,2) denotes the case that the first two indexes (say, United States and United Kingdom) are below their medians while the last two indexes (say, Hong Kong and Japan) are above their medians. Let G_{i_1,i_2,i_3,i_4} and $G^{MA^*}_{i_1,i_2,i_3,i_4}$ denote the frequency observed and the frequency estimated by the model averaging method in cell (i_1, i_2, i_3, i_4), respectively. Then, we define the estimation squared error as:

$$Q_{2^4,MA^*} = \frac{1}{2^4}\sum_{i_1=1}^{2}\sum_{i_2=1}^{2}\sum_{i_3=1}^{2}\sum_{i_4=1}^{2}(G_{i_1,i_2,i_3,i_4} - G^{MA^*}_{i_1,i_2,i_3,i_4})^2. \tag{17}$$

The 3^4 and 4^4 cross-classification tables are defined similarly. The corresponding prediction squared errors of the four methods for each cross-classification table are reported in different rows of Table 9. From Table 9 one can see that our MA^* method outperforms the other three methods in terms of prediction squared errors using a four-component copula.

[INSERT TABLE 9 ABOUT HERE]

6 Concluding Remarks

In this paper we propose to use a model averaging method to estimate a mixture copula model. Unlike the BIC method which selects only one individual copula based on the comparison of BIC, the model averaging method estimates a mixture copula model by choosing weights (associated with components of individual copula) optimally in the sense of minimizing the estimation squared loss. Simulation studies show that the model averaging method performs similarly to the penalized likelihood method proposed by Cai and Wang (2014) and the MLE method when observations are generated from copulas included in the working mixture copula model. However, when the working mixture copula model is misspecified, that is, when observations are generated from copulas not included in the working mixture model, the model averaging method outperforms Cai and Wang's (2014) penalized likelihood method, the BIC method and the MLE method. An empirical example shows that the model averaging method provides satisfactory estimates of the dependence structures among four international stock markets. Thus, the model averaging method provides a useful tool to estimate mixture copula models and can be utilized by practioners in financial industry for portfolio diversification and risk management.

References

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

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

[3] Andrews, Donald W.K., 1991. Heteroskedasticity and autocorrelation consistent covariance matrix estimation. Econometrica 59, 817-858.

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

[5] Chen, X. and Fan, Y., 2006a. Estimation of Copula-based Semiparametric Time Series Models, Journal of Econometrics, 130, 307-335.

[6] Chen, X. and Fan, Y., 2006b. Estimation and model selection of semiparametric copula-based multivariate dynamic models under copula misspecification. Journal of Econometrics, 135, 125-154.

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

[8] Chollete, L., Peña, V. and Lu, C., 2005. Comovement of international financial markets. Unpublished manuscript.

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

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

[11] Fan, Y. and Patton, Andrew J., 2014. Copulas in econometrics. Annual Review of Economics 6, 179-200.

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

[13] Granger, Clive W. and Newbold, P., 1974. Spurious regressions in econometrics. Journal of Econometrics 2, 111-120.

[14] Hansen, Bruce E. and Racine, J.S., 2012. Jackknife model averaging. Journal of Econometrics 167, 38-46.

[15] Hansen, Bruce E., 2007. Least squares model averaging. Econometrica 75, 1175-1189.

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

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

[18] Li, David X., 2000. On default correlation: A copula function approach. Journal of Fixed Income 9, 43-54.

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

[20] Ma, Y. and Zhu, L., 2012. A semiparametric approach to dimension reduction. Journal of the American Statistical Association 497, 168-179.

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

[22] Nelsen, Roger B., 2006. An Introduction to Copulas. New York: Springer. 2nd ed.

[23] Patton, Andrew J., 2006. Modelling asymmetric exchange rate dependence. International Economic Review 47, 527-556.

[24] Patton, Andrew J., 2012. A review of copula models for economic time series. Journal of Multivariate Analysis 110, 4-18.

[25] Rodriguez, Juan C., 2007. Measuring financial contagion: A copula approach. Journal of Empirical Finance 14, 401-423.

[26] Shao, J., 1997. An asymptotic theory for linear model selection. Statistica Sinica 7, 221-264.

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

[28] Varian, Hal R., 1975. A Bayesian approach to real estate assessment. Studies in Bayesian Econometrics and Statistics in Honor of Leonard J. Savage, Amsterdam: North-Holland.

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

[30] Zellner, A., 1986. Bayesian estimation and prediction using asymmetric loss functions. Journal of American Statistical Association 81(394), 446-451.

[31] Zhang, X., Wan, A. T. K. and Zou, G., 2013. Model averaging by Jackknife criterion in models with dependent data. Journal of Econometrics 174, 82-94.

[32] Zhang, X., Yu, D., Zou, G. and Liang, H. 2016. Optimal model averaging estimation for generalized linear models and generalized linear mixed-effects models. Journal of American Statistical Association, 111, 1775-1790.

[33] Zimmer, David M., 2012. The role of copulas in the housing crisis. Review of Economics and Statistics 94, 607-620.

Type I simulation: Out-of-sample forecast (sample size=200)						
	$\theta_{Ga}=0.5,\theta_{Cl}=5.8,\theta_{Gu}=5.1$			$\theta_{Ga}=0.7,\theta_{Cl}=7.8,\theta_{Gu}=7.1$		
	CW/MA^*	MLE/MA^*	BIC/MA^*	CW/MA^*	MLE/MA^*	BIC/MA^*
$\omega_{Ga}=0.5,\omega_{Cl}=0.5,\omega_{Gu}=0$	0.9931	1.0821	1.2043	0.9943	1.0526	1.1832
	(0.61)	(0.08)	(0.00)	(0.23)	(0.11)	(0.01)
$\omega_{Ga}=0.5,\omega_{Cl}=0,\omega_{Gu}=0.5$	0.9831	1.0574	1.1937	0.9922	1.0169	1.1491
	(0.59)	(0.11)	(0.01)	(0.63)	(0.87)	(0.04)
$\omega_{Ga}=0,\omega_{Cl}=0.5,\omega_{Gu}=0.5$	0.9951	1.0643	1.1739	0.9816	1.0713	1.1756
	(0.49)	(0.13)	(0.02)	(0.41)	(0.07)	(0.01)
$\omega_{Ga}=1/3,\omega_{Cl}=1/3,\omega_{Gu}=1/3$	1.0022	1.0754	1.1928	0.9983	1.0983	1.1749
	(0.64)	(0.07)	(0.01)	(0.89)	(0.05)	(0.00)
Type I simulation: Out-of-sample forecast (sample size=500)						
	$\theta_{Ga}=0.5,\theta_{Cl}=5.8,\theta_{Gu}=5.1$			$\theta_{Ga}=0.7,\theta_{Cl}=7.8,\theta_{Gu}=7.1$		
	CW/MA^*	MLE/MA^*	BIC/MA^*	CW/MA^*	MLE/MA^*	BIC/MA^*
$\omega_{Ga}=0.5,\omega_{Cl}=0.5,\omega_{Gu}=0$	1.0006	1.0679	1.1923	0.9926	1.0538	1.1749
	(0.71)	(0.08)	(0.00)	(0.83)	(0.12)	(0.01)
$\omega_{Ga}=0.5,\omega_{Cl}=0,\omega_{Gu}=0.5$	1.0135	1.0365	1.2173	1.0219	1.0766	1.2207
	(0.19)	(0.14)	(0.00)	(0.14)	(0.09)	(0.00)
$\omega_{Ga}=0,\omega_{Cl}=0.5,\omega_{Gu}=0.5$	0.9908	1.0749	1.1697	0.9877	1.0621	1.2134
	(0.91)	(0.08)	(0.02)	(0.16)	(0.13)	(0.00)
$\omega_{Ga}=1/3,\omega_{Cl}=1/3,\omega_{Gu}=1/3$	0.9945	1.0693	1.2158	0.9833	1.0421	1.2036
	(0.54)	(0.09)	(0.00)	(0.67)	(0.12)	(0.00)

Table 1: Ratios of squared prediction losses for Type I simulation. The data are generated from linear combinations of Gaussian, Clayton and Gumbel copulas with the associated weights denoted by $\omega_{Ga},\omega_{Cl},\omega_{Gu}$. $\theta_{Ga},\theta_{Cl},\theta_{Gu}$ represent for copula parameters for Gaussian, Clayton and Gumbel. MA^* = Model Average with selected J and selected K; CW = Cai and Wang's (2014) penalized maximum likelihood; BIC = Baseline approach based on BIC comparison; MLE = maximum likelihood estimation method. Each simulation is repeated 2000 times. Values in the parenthesis indicate the $p-$value obtained through the Diebold-Mariano Test.

Type II simulation: Out-of-sample forecast (sample size=200)						
	$\theta_F = 5.5, \theta_{SJ} = 4.8, \theta_J = 4.5$			$\theta_F = 7.5, \theta_{SJ} = 6.8, \theta_J = 6.5$		
	CW/MA^*	MLE/MA^*	BIC/MA^*	CW/MA^*	MLE/MA^*	BIC/MA^*
$\omega_F = 0.5, \omega_{SJ} = 0.5, \omega_J = 0$	1.0801	1.1728	1.2159	1.0739	1.1642	1.2207
	(0.06)	(0.01)	(0.00)	(0.07)	(0.02)	(0.00)
$\omega_F = 0.5, \omega_{SJ} = 0, \omega_J = 0.5$	1.0933	1.1324	1.2283	1.0845	1.1175	1.1934
	(0.05)	(0.05)	(0.00)	(0.06)	(0.03)	(0.01)
$\omega_F = 0, \omega_{SJ} = 0.5, \omega_J = 0.5$	1.1291	1.1467	1.1908	1.1022	1.1236	1.1968
	(0.04)	(0.04)	(0.00)	(0.05)	(0.04)	(0.01)
$\omega_F = 1/3, \omega_{SJ} = 1/3, \omega_J = 1/3$	1.1327	1.1055	1.1997	1.1237	1.0978	1.2114
	(0.04)	(0.04)	(0.01)	(0.04)	(0.05)	(0.00)
Type II simulation: Out-of-sample forecast (sample size=500)						
	$\theta_F = 5.5, \theta_{SJ} = 4.8, \theta_J = 4.5$			$\theta_F = 7.5, \theta_{SJ} = 6.8, \theta_J = 6.5$		
	CW/MA^*	MLE/MA^*	BIC/MA^*	CW/MA^*	MLE/MA^*	BIC/MA^*
$\omega_F = 0.5, \omega_{SJ} = 0.5, \omega_J = 0$	1.1187	1.2001	1.2567	1.1304	1.2213	1.2265
	(0.04)	(0.01)	(0.01)	(0.05)	(0.01)	(0.00)
$\omega_F = 0.5, \omega_{SJ} = 0, \omega_J = 0.5$	1.1531	1.2093	1.2317	1.1613	1.2764	1.2618
	(0.03)	(0.00)	(0.00)	(0.02)	(0.00)	(0.00)
$\omega_F = 0, \omega_{SJ} = 0.5, \omega_J = 0.5$	1.1664	1.2231	1.2504	1.1562	1.2376	1.2703
	(0.02)	(0.00)	(0.00)	(0.02)	(0.00)	(0.00)
$\omega_F = 1/3, \omega_{SJ} = 1/3, \omega_J = 1/3$	1.1981	1.1944	1.2576	1.1683	1.2247	1.2629
	(0.00)	(0.00)	(0.00)	(0.02)	(0.00)	(0.00)

,

Table 2: Ratios of squared prediction losses for Type II simulation. The data are generated from linear combinations of Frank, Survival Joe and Joe copulas with the associated weights denoted by $\omega_F, \omega_{SJ}, \omega_J$. $\theta_F, \theta_{SJ}, \theta_J$ represent for copula parameters for Frank, Survival Joe and Joe. MA^* = Model Average with selected J and selected K; CW = Cai and Wang's (2014) penalized maximum likelihood; BIC = Baseline approach based on BIC comparison; MLE = maximum likelihood estimation method. Each simulation is repeated 2000 times. Values in the parenthesis indicate the $p-$value obtained through the Diebold-Mariano Test.

3-component copula: Out-of-sample forecast (sample size=500)						
	$\theta_F = 5.5, \theta_{SJ} = 4.8, \theta_J = 4.5$			$\theta_F = 7.5, \theta_{SJ} = 6.8, \theta_J = 6.5$		
	CW/MA^*	MLE/MA^*	BIC/MA^*	CW/MA^*	MLE/MA^*	BIC/MA^*
$\omega_F = 0.5, \omega_{SJ} = 0.5, \omega_J = 0$	1.1631	1.3859	2.0968	1.1954	1.2207	2.5491
	(0.01)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
$\omega_F = 0.5, \omega_{SJ} = 0, \omega_J = 0.5$	1.1927	1.2157	1.9804	1.1368	1.2544	1.8093
	(0.01)	(0.00)	(0.00)	(0.01)	(0.00)	(0.00)
$\omega_F = 0 \omega_{SJ} = 0.5, \omega_J = 0.5$	1.1179	1.1038	1.4353	1.1277	1.1184	1.3965
	(0.05)	(0.04)	(0.00)	(0.03)	(0.04)	(0.00)
$\omega_F = 1/3, \omega_{SJ} = 1/3, \omega_J = 1/3$	1.0915	1.1179	1.4843	1.1302	1.1408	1.3951
	(0.04)	(0.01)	(0.00)	(0.01)	(0.00)	(0.00)
4-component copula: Out-of-sample forecast (sample size=500)						
	$\theta_F = 5.5, \theta_{SJ} = 4.8, \theta_J = 4.5$			$\theta_F = 7.5, \theta_{SJ} = 6.8, \theta_J = 6.5$		
	CW/MA^*	MLE/MA^*	BIC/MA^*	CW/MA^*	MLE/MA^*	BIC/MA^*
$\omega_F = 0.5, \omega_{SJ} = 0.5, \omega_J = 0$	1.1326	1.2605	2.9317	1.1429	1.2063	2.8970
	(0.07)	(0.01)	(0.00)	(0.01)	(0.00)	(0.00)
$\omega_F = 0.5, \omega_{SJ} = 0, \omega_J = 0.5$	1.1018	1.1916	1.4381	1.0988	1.1543	1.3806
	(0.05)	(0.00)	(0.00)	(0.04)	(0.00)	(0.00)
$\omega_F = 0, \omega_{SJ} = 0.5, \omega_J = 0.5$	1.1942	1.1139	1.4026	1.2237	1.1435	1.4109
	(0.00)	(0.01)	(0.00)	(0.00)	(0.01)	(0.00)
$\omega_F = 1/3, \omega_{SJ} = 1/3, \omega_J = 1/3$	1.1079	1.1834	1.3072	1.1406	1.2267	1.3869
	(0.03)	(0.01)	(0.00)	(0.01)	(0.00)	(0.00)

Table 3: Ratios of squared out-of-sample predicting losses for Type II simulation with 3-component and 4-component multivariate copulas. The data are generated from linear combinations of Frank, Survival Joe and Joe copula with the associated weights denoted by $\omega_F, \omega_{SJ}, \omega_J$. $\theta_F, \theta_{SJ}, \theta_J$ represent for copula parameters for Frank, Survival Joe and Joe. MA^* = Model Average with selected J and selected K; CW = Cai and Wang's (2014) penalized maximum likelihood; BIC = Baseline approach based on BIC comparison; MLE = Maximum likelihood estimation method. Each simulation is repeated 2000 times. Values in the parenthesis indicate the $p-$value obtained through the Diebold-Mariano Test.

	US	UK	HK	JP
Mean	0.0186	0.0084	0.0295	0.0070
Median	0.0307	0.0492	0.0117	0.0174
min	-9.2002	-10.494	-9.0137	-7.3300
max	5.2785	9.2065	10.192	4.2476
S.D.	1.0752	1.2243	1.3534	1.3763
Skewness	-0.7713	-0.4470	-0.1229	-0.4189
Kurtosis	7.7132	8.1975	7.3170	1.7011

Table 4: The summary statistics for daily log-returns of MSCI Indexes of United States, United Kingdom, Hong Kong and Japan.

	UK	JP	HK
US	0.4304 (0.271)	0.0274 (0.037)	0.1147 (0.072)
UK		0.1995 (0.140)	0.3753 (0.209)
JP			0.4880 (0.306)

Table 5: Linear correlation coefficients and Kendall's τs (Kendall's τs are in parentheses) across four markets.

	Q_{MA^*}	Q_{CW}	Q_{BIC}	Q_{MLE}	Q_{CW}/Q_{MA^*}	Q_{BIC}/Q_{MA^*}	Q_{MLE}/Q_{MA^*}
US-UK	17.59	22.09	28.32	21.13	1.256	1.610	1.201
US-HK	31.17	35.91	36.17	34.49	1.152	1.160	1.107
US-JP	27.12	29.24	29.61	28.74	1.078	1.092	1.060
UK-HK	30.19	33.77	35.43	33.16	1.119	1.174	1.098
UK-JP	20.04	25.21	25.58	24.41	1.258	1.276	1.218
HK-JP	29.17	31.41	35.04	33.11	1.077	1.201	1.135

Table 6: Mean of in-sample estimation errors based on MA^*, CW, BIC and MLE.

	Q_{MA^*}	Q_{CW}	Q_{BIC}	Q_{MLE}	Q_{CW}/Q_{MA^*}	Q_{BIC}/Q_{MA^*}	Q_{MLE}/Q_{MA^*}
US-UK	112.376	115.403	128.514	119.027	1.027	1.144	1.059
US-HK	42.3517	44.8726	44.1107	44.6314	1.060	1.042	1.054
US-JP	19.0015	20.6804	22.0029	22.6441	1.088	1.158	1.192
UK-HK	30.4812	32.9507	35.1903	34.6621	1.081	1.154	1.137
UK-JP	19.0735	22.2811	21.9532	23.2734	1.168	1.151	1.220
HK-JP	33.7658	35.1526	36.4483	36.0025	1.041	1.079	1.066

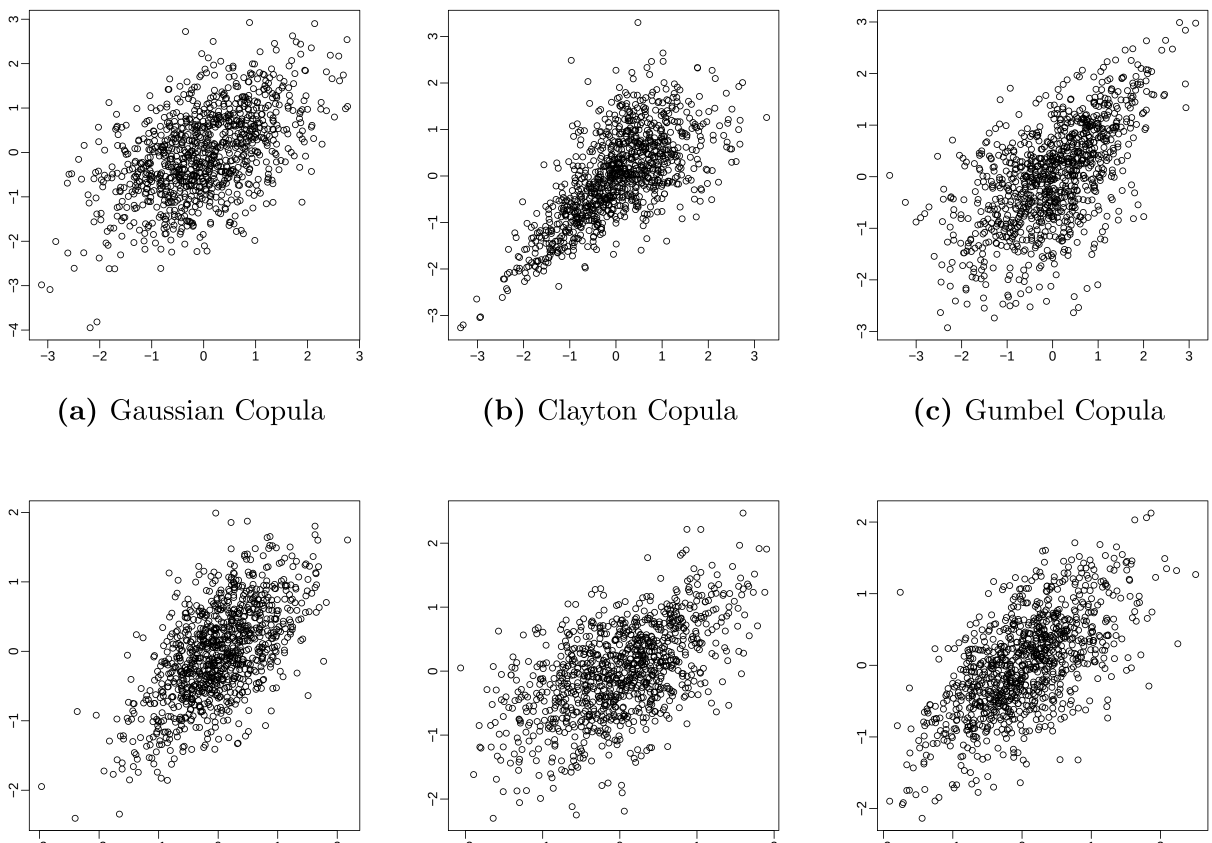
Table 7: Mean of out-of-sample predicting errors based on MA^*, CW, BIC and MLE.

	$Q^{rw}_{MA^*}$	Q^{rw}_{CW}	Q^{rw}_{BIC}	Q^{rw}_{MLE}	$Q^{rw}_{CW}/Q^{rw}_{MA^*}$	$Q^{rw}_{BIC}/Q^{rw}_{MA^*}$	$Q^{rw}_{MLE}/Q^{rw}_{MA^*}$
US-UK	32.2715	35.9321	39.2237	38.0735	1.113	1.215	1.180
US-HK	27.0634	28.9637	30.2680	30.1159	1.070	1.118	1.113
US-JP	14.5301	16.5926	17.1104	16.6603	1.142	1.178	1.147
UK-HK	17.1756	20.2027	21.3544	20.1138	1.176	1.243	1.171
UK-JP	17.0035	21.0038	21.1917	20.5104	1.235	1.246	1.206
HK-JP	34.1824	35.1027	39.0723	38.8361	1.027	1.143	1.136

Table 8: Mean of out-of-sample predicting errors of MA^*, CW, BIC and MLE based on the rolling window method (fixed window width).

	Q_{MA^*}	Q_{CW}	Q_{BIC}	Q_{MLE}	Q_{CW}/Q_{MA^*}	Q_{BIC}/Q_{MA^*}	Q_{MLE}/Q_{MA^*}
US-UK-HK-JP (2^4)	39.0027	47.2095	154.5832	45.3945	1.210	3.963	1.164
US-UK-HK-JP (3^4)	48.4601	55.1538	163.2781	61.1347	1.138	3.369	1.262
US-UK-HK-JP (4^4)	41.1866	44.7911	149.4092	47.0136	1.088	3.628	1.141

Table 9: Mean of out-of-sample predicting errors based on MA^*, CW, BIC and MLE with a 4-component copula. The mixture copula includes Student's t, Clayton and Gumbel copula. The first row considers $2^4 = 16$ cells, the second row considers $3^4 = 81$ cells and the third row considers $4^4 = 256$ cells.



(a) Gaussian Copula **(b)** Clayton Copula **(c)** Gumbel Copula

(d) Gaussian-Clayton Mixture **(e)** Gaussian-Gumbel Mixture **(f)** Clayton-Gumbel Mixture

Figure 1: Scatter plots for Gaussian, Clayton and Gumbel copula and their mixture. All have standard normal margins and Kendall's $\tau = 0.5$. For mixture copula, $\boldsymbol{\omega}_k = 0.5$.

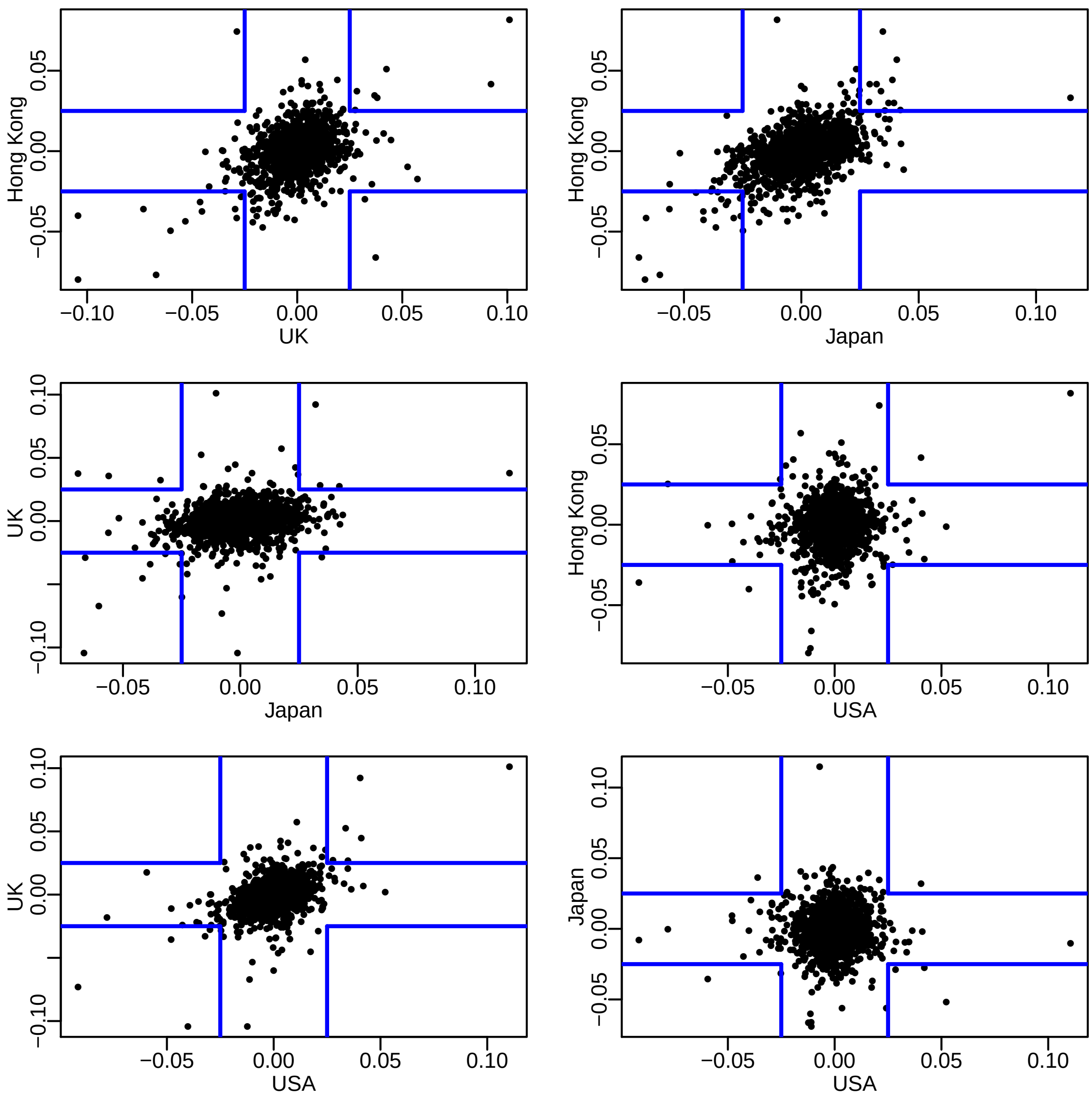


Figure 2: Scatter plots for daily return of MSCI Index.

Appendix A

A.1 Notations and conditions

Let $\widehat{\boldsymbol{\theta}} = \left(\widehat{\boldsymbol{\theta}}_1^\top, \ldots, \widehat{\boldsymbol{\theta}}_K^\top\right)^\top$, where K is a fixed positive integer. Denote the fixed dimension of $\widehat{\boldsymbol{\theta}}$ by $\varkappa$. Define $\boldsymbol{\theta}_k^*$ as the pseudo true value defined by

$$\boldsymbol{\theta}_k^* = \arg\max_{\boldsymbol{\theta}_k} E\left[\log c_k(F_1^0(\varepsilon_{1t}), \ldots, F_p^0(\varepsilon_{pt}); \boldsymbol{\theta}_k)\right], \qquad k = 1, ..., K,$$

where $c_k(F_1^0, \ldots, F_p^0; \boldsymbol{\theta}_k) \equiv \partial^p C_k(F_1^0, \ldots, F_p^0; \boldsymbol{\theta}_k)/\partial F_1^0 ... \partial F_p^0$ is the copula density. Denote $\boldsymbol{\theta}^* = \left(\boldsymbol{\theta}_1^{*\top}, \ldots, \boldsymbol{\theta}_K^{*\top}\right)^\top$. Let $\mathbf{C}^*(\mathbf{w}) = \widehat{\mathbf{C}}(\mathbf{w}) \mid_{\widehat{\boldsymbol{\theta}} = \boldsymbol{\theta}^*}$, $\nu_t(\mathbf{w}) = \partial C(\widetilde{U}_t; \widehat{\boldsymbol{\theta}}_k, \mathbf{w})/\partial \widehat{\boldsymbol{\theta}}|_{\widehat{\boldsymbol{\theta}} = \bar{\boldsymbol{\theta}}_t}$, for $t = 1, \ldots, T$, where $\bar{\boldsymbol{\theta}}_t$ is between $\widehat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}^*$, $\mathbf{Q}(\mathbf{w}) = \{\nu_1(\mathbf{w}), \ldots, \nu_T(\mathbf{w})\}^\top$, $L_T^*(\mathbf{w}) = \|\mathbf{C}^*(\mathbf{w}) - \mathbf{C}_0\|^2$, and $\xi_T = \inf_{\mathbf{w} \in \mathcal{W}} L_T^*(\mathbf{w})$. We assume that J is fixed and $M \to \infty$ as $T \to \infty$.

In addition, for the k^{th} copula, we denote $C_k\{v_1, \ldots, v_p; \boldsymbol{\theta}_k\} \equiv C_k(\mathbf{v}; \boldsymbol{\theta}_k)$ and $c_k\{v_1, \ldots, v_p; \boldsymbol{\theta}_k\} \equiv c_k(\mathbf{v}; \boldsymbol{\theta}_k)$. Also, we denote $l_k(\mathbf{v}; \boldsymbol{\theta}_k) \equiv \log c_k(\mathbf{v}; \boldsymbol{\theta}_k)$, $l_{\boldsymbol{\theta},k}(\mathbf{v}; \boldsymbol{\theta}_k) \equiv \partial l_k(\mathbf{v}; \boldsymbol{\theta}_k)/\partial\, \boldsymbol{\theta}_k$, $l_{\boldsymbol{\theta\theta},k}(\mathbf{v}; \boldsymbol{\theta}_k) \equiv \partial^2 l_k(\mathbf{v}; \boldsymbol{\theta}_k)/\partial \boldsymbol{\theta}_k \partial \boldsymbol{\theta}_k^\top$, $l_{\boldsymbol{\theta j},k}(\mathbf{v}; \boldsymbol{\theta}_k) \equiv \partial^2 l_k(\mathbf{v}; \boldsymbol{\theta}_k)/\partial u_j \partial \boldsymbol{\theta}_k^\top$ and $U_t \equiv (F_1(\varepsilon_{1t}), \ldots, F_p(\varepsilon_{pt}))$.

To prove the asymptotic optimality as stated in Theorem 1, we need the following regularity conditions.

Condition C.1: $\boldsymbol{\theta}^*$ is a finite dimensional vector with constant components and it takes value in a compact subset of $\mathcal{R}^\varkappa$, $\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^* = O_p(T^{-1/2})$.

Condition C.2: Uniformly for $\mathbf{w} \in \mathcal{W}$, the elements of $T \times \varkappa$ matrix $\mathbf{Q}(\mathbf{w})$ are uniformly bounded.

Condition C.3: Uniformly for $\mathbf{w} \in \mathcal{W}$, $T^{-1/2}\|\widehat{\mathbf{C}}(\mathbf{w}) - \overline{\mathbf{C}}(\mathbf{w})\|^2 = O_p(1)$, $T^{-1/2}\{\widehat{\mathbf{C}}(\mathbf{w}) - \overline{\mathbf{C}}(\mathbf{w})\}^\top\{\widehat{\mathbf{C}}(\mathbf{w}) - \widetilde{\mathbf{C}}\} = O_p(1)$, and $T^{-1/2}\{\widehat{\mathbf{C}}(\mathbf{w}) - \overline{\mathbf{C}}(\mathbf{w})\}^\top(\mathbf{C}_0 - \widetilde{\mathbf{C}}) = O_p(1)$.

Condition C.4: There exists a sequence $c_T \to 0$ such that $T\xi_T^{-2} \leq c_T$ almost surely.

Remark 2: Condition C.1 requires that $\boldsymbol{\theta}^*$ takes values in a compact set. It rules out some cases such as the true distribution is normal, while one fits a t_ν-distribution model and estimates the degree of freedom ν (as $\nu^* = \infty$). Condition C.1 also requires that the convergence rate of $\widehat{\boldsymbol{\theta}}$ to the pseudo true value $\boldsymbol{\theta}^*$ is $O_p(T^{-1/2})$. Chen and Fan (2006b) show that Condition C.1 holds true under quite general regularity conditions including: (i) $\boldsymbol{\theta}_k^*$ are in the interior of the parameter space for $k = 1, ..., K$, (ii) $\{\mathbf{Y}_t', \mathbf{Z}_t'\}_{t=1}^{\mathrm{T}}$ is stationary β mixing with the appropriate decay rate, (iii) $l_{\boldsymbol{\theta},k}(\mathbf{u}; \boldsymbol{\theta}_k)$, $l_{\boldsymbol{\theta\theta},k}(\mathbf{u}; \boldsymbol{\theta}_k)$ and $l_{\boldsymbol{\theta j},k}(\mathbf{u}; \boldsymbol{\theta}_k)$ satisfy some standard smoothness conditions for $k = 1, ..., K$, and (iv) $l_{\boldsymbol{\theta},k}(U_t; \boldsymbol{\theta}_k)$, $l_{\boldsymbol{\theta\theta},k}(U_t; \boldsymbol{\theta}_k)$ and $l_{\boldsymbol{\theta j},k}(U_t; \boldsymbol{\theta}_k)$ satisfy some appropriate moment conditions for $k = 1, ..., K$. Condition C.2 puts an restriction of the derivative of copulas and requires the derivatives are bounded uniformly.

With Conditions C.1 and C.2, we obtain that uniformly for $w \in \mathcal{W}$,

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

where $\lambda_{\max}(\cdot)$ denotes the maximum eigenvalue of a matrix (since C.2 implies that $\lambda_{\max}\left\{\mathbf{Q}^{\mathrm{T}}(\mathbf{w})\mathbf{Q}(\mathbf{w})\right\} = O_p(T)$), and

$$T^{-1/2}\left\{\mathbf{C}^*(\mathbf{w})-\mathbf{C}_0\right\}^{\top}\mathbf{Q}(\mathbf{w})(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*) = T^{-1/2}O_p(T^{1/2}) = O_p(1), \qquad (A.2)$$

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

Condition C.3 requires that as $T\to\infty$, the difference between the loss of the regular and leave-M-out estimators decreases at some rate. This is similar to condition (A.10) of Andrews (1991) and condition (A.5) of Hansen and Racine (2012). Let $\widehat{\boldsymbol{\theta}}^{(-j)}$ be the estimator of $\boldsymbol{\theta}$ with the j^{th} group removed from the sample and $\boldsymbol{\nu}_t^{(-j)}(\mathbf{w}) = \partial\widehat{C}^{(-j)}(\widehat{\varepsilon}_t,\mathbf{w})/\partial\widehat{\boldsymbol{\theta}}^{(-j)}|_{\widehat{\boldsymbol{\theta}}^{(-j)}=\bar{\boldsymbol{\theta}}_t^{(-j)}}$ for $t=1,\dots,T$, where $\bar{\boldsymbol{\theta}}_t^{(-j)}$ is between $\widehat{\boldsymbol{\theta}}^{(-j)}$ and $\boldsymbol{\theta}^*$, and $\overline{\mathbf{Q}}(\mathbf{w}) = \{\widetilde{\boldsymbol{\nu}}_1^{(-1)}(\mathbf{w}),\dots,\boldsymbol{\nu}_T^{(-J)}(\mathbf{w})\}^{\mathrm{T}}$. From Proposition 3.2 of Chen and Fan (2006b) and Assumptions D and N in that paper, the estimator $\widehat{\boldsymbol{\theta}}$ converges to $\boldsymbol{\theta}^*$ with a root-T rate and the estimators $\widehat{\boldsymbol{\theta}}^{(-j)}$ converges to $\boldsymbol{\theta}^*$ with rate root-$(T-M)$, which is uniformly over $j=1,\dots,J$ because J is a finite constant. When the elements of $T\times\kappa$ matrix $\overline{\mathbf{Q}}(\mathbf{w})$ are uniformly bounded, similar to the derivations in (A.1) and (A.2), we obtain that uniformly over $\mathbf{w}\in\mathcal{W}$, $T^{-1/2}\|\widehat{\mathbf{C}}(\mathbf{w})-\overline{\mathbf{C}}(\mathbf{w})\|^2 = O_p(T^{-1/2})$, $T^{-1/2}\{\widehat{\mathbf{C}}(\mathbf{w})-\overline{\mathbf{C}}(\mathbf{w})\}^{\top}\{\widehat{\mathbf{C}}(\mathbf{w})-\widetilde{\mathbf{C}}\} = O_p(1)$, and $T^{-1/2}\{\widehat{\mathbf{C}}(\mathbf{w})-\overline{\mathbf{C}}(\mathbf{w})\}^{\top}(\mathbf{C}_0-\widetilde{\mathbf{C}}) = O_p(1)$.

Condition C.4 imposes a limitation on situations to apply our asymptotic results. It requires that ξ_T grow at a rate faster than $T^{1/2}$, which implies all candidate copulas are misspecified. This is similar to the third part of condition (A7) in Zhang et al. (2013) and condition 7 in Ando and Li (2014). The assumption that all candidate models are misspecified is a common condition used in proving optimality properties of model averaging estimators.

Here we would like to comment on the case when the candidate mixture copula is not misspecified, i.e., the true copula is the mixture copula $C_K(\mathbf{u};\boldsymbol{\theta}_K) = \sum_{k=1}^{K-1}\omega_k C_k\left\{F_1(x_1),\dots,F_p(x_p);\boldsymbol{\theta}_k\right\}$ with $\sum_{k=1}^{K-1}\omega_k = 1$ and $\omega_k>0$ for $k=1,\dots,K-1$, we can show that under some regular conditions, the estimated weight of $C_K(\widetilde{\mathbf{u}};\widehat{\boldsymbol{\theta}}_K)$

$$\widehat{w}_K \to 1, \qquad (A.3)$$

in probability, as $T\to\infty$ (see Appendix A.4 for the regular conditions and the proof), thus our method will have the same large sample property as that of MLE. This result is expected and reasonable, and means our method is adaptable to the "lucky" case that the true copula is the mixture copula. The simulation results in Table 1 also indicate that under this case, our MA method performs similarly to MLE. Let $\mathbf{w}^* = (0, \ldots, 0, 1)^\top$. When the candidate mixture copula is not misspecified, we have $L_T^*(\mathbf{w}^\star) = \|\mathbf{C}^*(\mathbf{w}^\star) - \mathbf{C}_0\|^2 = O_p(1)$ because of the convergence of MLE to the true value. Therefore, Condition 4 is violated when the mixture copula contains the true copula.

A.2 Proof of Theorem 1

The result below follows from various definitions given in the paper (and some adding/subtracting terms manipulations)

$$
\begin{aligned}
CV_J(\mathbf{w}) &= \left\|\overline{\mathbf{C}}(\mathbf{w}) - \widetilde{\mathbf{C}}\right\|^2 \\
&= \left\|\left\{\widehat{\mathbf{C}}(\mathbf{w}) - \mathbf{C}_0\right\} - \left\{\widehat{\mathbf{C}}(\mathbf{w}) - \overline{\mathbf{C}}(\mathbf{w})\right\} + \left(\mathbf{C}_0 - \widetilde{\mathbf{C}}\right)\right\|^2 \\
&= L_T(\mathbf{w}) + \left\|\widehat{\mathbf{C}}(\mathbf{w}) - \overline{\mathbf{C}}(\mathbf{w})\right\|^2 + \left\|\mathbf{C}_0 - \widetilde{\mathbf{C}}\right\|^2 - 2\left\{\widehat{\mathbf{C}}(\mathbf{w}) - \overline{\mathbf{C}}(\mathbf{w})\right\}^\top \left(\mathbf{C}_0 - \widetilde{\mathbf{C}}\right) \\
&\quad -2\left\{\widehat{\mathbf{C}}(\mathbf{w}) - \overline{\mathbf{C}}(\mathbf{w})\right\}^\top \left\{\widehat{\mathbf{C}}(\mathbf{w}) - \mathbf{C}_0\right\} + 2\left\{\widehat{\mathbf{C}}(\mathbf{w}) - \mathbf{C}_0\right\}^\top \left(\mathbf{C}_0 - \widetilde{\mathbf{C}}\right) \\
&= L_T(\mathbf{w}) + \left\|\widehat{\mathbf{C}}(\mathbf{w}) - \overline{\mathbf{C}}(\mathbf{w})\right\|^2 - 2\left\{\widehat{\mathbf{C}}(\mathbf{w}) - \overline{\mathbf{C}}(\mathbf{w})\right\}^\top \left\{\widehat{\mathbf{C}}(\mathbf{w}) - \widetilde{\mathbf{C}}\right\} \\
&\quad +2\left\{\widehat{\mathbf{C}}(\mathbf{w}) - \overline{\mathbf{C}}(\mathbf{w})\right\}^\top (\mathbf{C}_0 - \widetilde{\mathbf{C}}) + 2\overline{\mathbf{C}}(\mathbf{w})^\top(\mathbf{C}_0 - \widetilde{\mathbf{C}}) - (\mathbf{C}_0 + \widetilde{\mathbf{C}})^\top(\mathbf{C}_0 - \widetilde{\mathbf{C}}) \\
&\equiv L_T(\mathbf{w}) + \Xi_T(\mathbf{w}) - (\mathbf{C}_0 + \widetilde{\mathbf{C}})^\top(\mathbf{C}_0 - \widetilde{\mathbf{C}}),
\end{aligned}
\tag{A.4}
$$

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

$$
\begin{aligned}
L_T(\mathbf{w}) &= \left\|\widehat{\mathbf{C}}(\mathbf{w}) - \mathbf{C}_0\right\|^2 \\
&= \left\|\left\{\widehat{\mathbf{C}}(\mathbf{w}) - \mathbf{C}^*(\mathbf{w})\right\} + \left\{\mathbf{C}^*(\mathbf{w}) - \mathbf{C}_0\right\}\right\|^2 \\
&= \left\|\widehat{\mathbf{C}}(\mathbf{w}) - \mathbf{C}^*(\mathbf{w})\right\|^2 + \|\mathbf{C}^*(\mathbf{w}) - \mathbf{C}_0\|^2 + 2\left\{\mathbf{C}^*(\mathbf{w}) - \mathbf{C}_0\right\}^\top \left\{\widehat{\mathbf{C}}(\mathbf{w}) - \mathbf{C}^*(\mathbf{w})\right\} \\
&= L_T^*(\mathbf{w}) + \left\|\widehat{\mathbf{C}}(\mathbf{w}) - \mathbf{C}^*(\mathbf{w})\right\|^2 + 2\left\{\mathbf{C}^*(\mathbf{w}) - \mathbf{C}_0\right\}^\top \left\{\widehat{\mathbf{C}}(\mathbf{w}) - \mathbf{C}^*(\mathbf{w})\right\} \\
&\equiv L_T^*(\mathbf{w}) + \Pi_T(\mathbf{w}).
\end{aligned}
\tag{A.5}
$$

We first prove that if

$$\sup_{\mathbf{w}\in\mathcal{W}} \left|\frac{L_T(\mathbf{w})}{L_T^*(\mathbf{w})} - 1\right| = o_p(1), \tag{A.6}$$

and $CV_J(\mathbf{w})$ can be written as

$$CV_J(\mathbf{w}) = L_T(\mathbf{w}) + a_T(\mathbf{w}) + b_T \tag{A.7}$$

with the term $a_T(\mathbf{w})$ satisfies

$$\sup_{\mathbf{w}\in\mathcal{W}} \frac{|a_T(\mathbf{w})|}{L_T^*(\mathbf{w})} = o_p(1) \tag{A.8}$$

and the term b_T is unrelated to $\mathbf{w}$, then Theorem 1 holds true. From (A.6), we know that

$$\inf_{\mathbf{w}\in\mathcal{W}} \frac{L_T(\mathbf{w})}{L_T^*(\mathbf{w})} \geq -\sup_{\mathbf{w}\in\mathcal{W}} \left|\frac{L_T(\mathbf{w})}{L_T^*(\mathbf{w})} - 1\right| + 1 \to 1, \tag{A.9}$$

in probability as $T \to \infty$. By the definition of $\widehat{\mathbf{w}}$ and (A.7), we have

$$\inf_{\mathbf{w}\in\mathcal{W}} \{L_T(\mathbf{w}) + a_T(\mathbf{w})\} = L_T(\widehat{\mathbf{w}}) + a_T(\widehat{\mathbf{w}}). \tag{A.10}$$

By Condition C.4, we know that there exists a non-negative sequence ν_T and a sequence of vectors $\mathbf{w}(T) \in \mathcal{W}$ such that as $T \to \infty$,

$$\nu_T \xi_T^{-1} \to 0 \tag{A.11}$$

and

$$\inf_{\mathbf{w}\in\mathcal{W}} L_T(\mathbf{w}) = L_T\left\{\mathbf{w}(T)\right\} - \nu_T. \tag{A.12}$$

From (A.9) and (A.11), we have

$$\inf_{\mathbf{w}\in\mathcal{W}} \frac{|L_T(\mathbf{w}) - \nu_T|}{L_T^*(\mathbf{w})} \geq \inf_{\mathbf{w}\in\mathcal{W}} \frac{L_T(\mathbf{w}) - \nu_T}{L_T^*(\mathbf{w})} \geq \inf_{\mathbf{w}\in\mathcal{W}} \frac{L_T(\mathbf{w})}{L_T^*(\mathbf{w})} - \frac{\nu_T}{\xi_T} \to 1, \tag{A.13}$$

in probability as $T \to \infty$. From (A.8)-(A.13), we obtain that, for any $\delta > 0$,

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

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

which implies that Theorem 1 holds true.

Hence from (A.4)-(A.8), Theorem 1 is valid if the following hold:

$$\sup_{\mathbf{w}\in\mathcal{W}}\frac{|\Xi_T(\mathbf{w})|}{L_T^*(\mathbf{w})}=o_p(1) \tag{A.14}$$

and

$$\sup_{\mathbf{w}\in\mathcal{W}}\frac{|\Pi_T(\mathbf{w})|}{L_T^*(\mathbf{w})}=o_p(1). \tag{A.15}$$

Using Taylor expansion,

$$\widehat{\mathbf{C}}(\mathbf{w})-\mathbf{C}^*(\mathbf{w})=\mathbf{Q}(\mathbf{w})(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*). \tag{A.16}$$

where $\mathbf{Q}(\mathbf{w})=\mathbf{Q}(\mathbf{w};\overline{\boldsymbol{\theta}}_1,\ldots,\overline{\boldsymbol{\theta}}_T)$ with $\overline{\boldsymbol{\theta}}_t$'s being between the line segment of $\widehat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}^*$ (see the first paragraph of Appendix A.1 for the definition of $\mathbf{Q}(\mathbf{w})$). From (A.16), (A.1), (A.2), and Condition C.4 we have

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

which is (A.15).

Similarly, from Conditions C.3 and C.4, we have

$$
\begin{aligned}
&\sup_{\mathbf{w}\in\mathcal{W}}\frac{\left|\left\|\widehat{\mathbf{C}}(\mathbf{w})-\overline{\mathbf{C}}(\mathbf{w})\right\|^2-2\left\{\widehat{\mathbf{C}}(\mathbf{w})-\overline{\mathbf{C}}(\mathbf{w})\right\}^\top\left\{\widehat{\mathbf{C}}(\mathbf{w})-\widetilde{\mathbf{C}}\right\}+2\left\{\widehat{\mathbf{C}}(\mathbf{w})-\overline{\mathbf{C}}(\mathbf{w})\right\}^\top(\mathbf{C}_0-\widetilde{\mathbf{C}})\right|}{L_T^*(\mathbf{w})}\\
&=o_p(1).
\end{aligned} \tag{A.17}
$$

Denote $F_0(\cdot)$ as the true distribution function of $\varepsilon_t \equiv (\varepsilon_{1t}, ..., \varepsilon_{pt})'$, $\widetilde{C}_T(\mathbf{x}) = T^{-1}\sum_{s=1}^T I\{\widehat{\varepsilon}_s \leq \mathbf{x}\}$ and $C_T(\mathbf{x}) = \frac{1}{T}\sum_{s=1}^T I\{\varepsilon_s \leq \mathbf{x}\}$.

Next, we want to show

$$T^{-1}\sum_{t=1}^T \left|F_0(\varepsilon_t) - \widetilde{C}_T(\widehat{\varepsilon}_t)\right| = O_p(T^{-1/2}). \tag{A.18}$$

We have

$$\begin{aligned}
& \frac{1}{T}\sum_{t=1}^T \left|F_0(\varepsilon_t) - \widetilde{C}_T(\widehat{\varepsilon}_t)\right| \\
= & \frac{1}{T}\sum_{t=1}^T \left|F_0(\varepsilon_t) - C_T(\varepsilon_t) + C_T(\varepsilon_t) - \widetilde{C}_T(\varepsilon_t) + \widetilde{C}_T(\varepsilon_t) - \widetilde{C}_T(\widehat{\varepsilon}_t)\right| \\
\leq & \frac{1}{T}\sum_{t=1}^T |F_0(\varepsilon_t) - C_T(\varepsilon_t)| + \frac{1}{T}\sum_{t=1}^T \left|C_T(\varepsilon_t) - \widetilde{C}_T(\varepsilon_t)\right| + \frac{1}{T}\sum_{t=1}^T \left|\widetilde{C}_T(\varepsilon_t) - \widetilde{C}_T(\widehat{\varepsilon}_t)\right|
\end{aligned} \tag{A.19}$$

Since $\{\varepsilon_t\}$ are i.i.d., it is a standard result that

$$\sqrt{T}\sup_{\mathbf{x}\in\mathcal{R}^p} | F_0(\mathbf{x}) - C_T(\mathbf{x}) |= O_p(1). \tag{A.20}$$

Therefore, the first term of (A.19) is $O_p(T^{-1/2})$. By the same proof method as used in Chen and Fan (2006b, Lemma A.1 (3)), one can show that

$$\sqrt{T}\sup_{\mathbf{x}\in\mathcal{R}^p} | C_T(\mathbf{x}) - \widetilde{C}_T(\mathbf{x}) |= O_p(1), \tag{A.21}$$

and the second term of (A.19) is also $O_p(T^{-1/2})$. Finally the third term of (A.19) can be written as

$$\begin{aligned}
& T^{-1}\sum_{t=1}^T \left|\widetilde{C}_T(\varepsilon_t) - \widetilde{C}_T(\widehat{\varepsilon}_t)\right| \\
= & T^{-2}\sum_{t=1}^T\sum_{s=1}^T |I\{\widehat{\varepsilon}_s \leq \varepsilon_t\} - I\{\widehat{\varepsilon}_s \leq \widehat{\varepsilon}_t\}| \\
= & T^{-2}\sum_{t=1}^T\sum_{s=1}^T |1 - I\{\widehat{\varepsilon}_s > \varepsilon_t\} - 1 + I\{\widehat{\varepsilon}_s > \widehat{\varepsilon}_t\}| \\
= & T^{-2}\sum_{t=1}^T\sum_{s=1}^T |I\{\widehat{\varepsilon}_t < \widehat{\varepsilon}_s\} - I\{\varepsilon_t < \widehat{\varepsilon}_s\}|
\end{aligned}$$

$$= \quad T^{-1}\sum_{s=1}^{T}\left|\widetilde{C}_T(\widehat{\varepsilon}_s)-C_T(\widehat{\varepsilon}_s)\right| = O_p(T^{-1/2}) \tag{A.22}$$

by (A.21). Hence, we have the third term of (A.19) is $O_p(T^{-1/2})$.

From $c_T \to 0$, (A.18), and the fact that all elements of the vector $|\widehat{\mathbf{C}}_k|$ are bounded by 1, we obtain that for any $j \in \{1,\ldots,J\}$,

$$\begin{aligned}
& c_T^{1/2}M^{-1/2}\left|\sum_{m=1}^{M} C_k(\widetilde{U}_{(j-1)M+m};\widehat{\boldsymbol{\theta}}_k^{(-j)})\left\{C_0(U_{0,(j-1)M+m};\boldsymbol{\theta}_0)-\widetilde{C}_{(j)}(\widehat{\varepsilon}_{(j-1)M+m})\right\}\right| \\
\le\; & c_T^{1/2}M^{1/2}M^{-1}\sum_{m=1}^{M}\left|C_0(U_{0,(j-1)M+m};\boldsymbol{\theta}_0)-\widetilde{C}_{(j)}(\widehat{\varepsilon}_{(j-1)M+m})\right| \\
=\; & c_T^{1/2}M^{1/2}M^{-1}\sum_{m=1}^{M}\left|F_0(\varepsilon_{(j-1)M+m})-\widetilde{C}_{(j)}(\widehat{\varepsilon}_{(j-1)M+m})\right| \\
=\; & o_p(1),
\end{aligned} \tag{A.23}$$

which, along with Condition C.4 and the assumption that K and J are fixed, implies that

$$\begin{aligned}
& \xi_T^{-1}\sup_{\mathbf{w}\in\mathcal{W}}\left|\overline{\mathbf{C}}(\mathbf{w})^\top(\mathbf{C}_0-\widetilde{\mathbf{C}})\right| \\
=\; & \xi_T^{-1}\sup_{\mathbf{w}\in\mathcal{W}}\left|\sum_{k=1}^{K} w_k\overline{\mathbf{C}}_k^\top(\mathbf{C}_0-\widetilde{\mathbf{C}})\right| \\
\le\; & \sum_{k=1}^{K}\frac{T^{1/2}}{\xi_T}T^{-1/2}\left|\overline{\mathbf{C}}_k^\top(\mathbf{C}_0-\widetilde{\mathbf{C}})\right| \\
=\; & \sum_{k=1}^{K} J^{-1/2}\frac{T^{1/2}}{\xi_T}\left|\sum_{j=1}^{J} M^{-1/2}\sum_{m=1}^{M} C_k(\widetilde{U}_{(j-1)M+m};\widehat{\boldsymbol{\theta}}_k^{(-j)})\left\{C_0(U_{0,(j-1)M+m};\boldsymbol{\theta}_0)-\widetilde{C}_{(j)}(\widehat{\varepsilon}_{(j-1)M+m})\right\}\right| \\
\le\; & \sum_{k=1}^{K} J^{-1/2}c_T^{1/2}\left|\sum_{j=1}^{J} M^{-1/2}\sum_{m=1}^{M} C_k(\widetilde{U}_{(j-1)M+m};\widehat{\boldsymbol{\theta}}_k^{(-j)})\left\{C_0(U_{0,(j-1)M+m};\boldsymbol{\theta}_0)-\widetilde{C}_{(j)}(\widehat{\varepsilon}_{(j-1)M+m})\right\}\right| \\
=\; & \sum_{k=1}^{K} J^{-1/2}\left|\sum_{j=1}^{J} c_T^{1/2}M^{-1/2}\sum_{m=1}^{M} C_k(\widetilde{U}_{(j-1)M+m};\widehat{\boldsymbol{\theta}}_k^{(-j)})\left\{C_0(U_{0,(j-1)M+m};\boldsymbol{\theta}_0)-\widetilde{C}_{(j)}(\widehat{\varepsilon}_{(j-1)M+m})\right\}\right| \\
=\; & o_p(1),
\end{aligned} \tag{A.24}$$

where the second '$\le$' holds almost surely. From (A.17) and (A.24), we obtain (A.14). This completes the proof for Theorem 1.

A.3 Proof of Proposition 1

First, we show the first part of Proposition 1 without imposing the i.i.d assumption. From the steps in deriving (A.24), we know that uniformly for $\mathbf{w} \in \mathcal{W}$,

$$\xi_T^{-1}\left\{\left\|\overline{\mathbf{C}}(\mathbf{w})-\mathbf{C}_0\right\|^2-\left\|\overline{\mathbf{C}}(\mathbf{w})-\widetilde{\mathbf{C}}\right\|^2\right\} \;=\; \xi_T^{-1}\left\{2\overline{\mathbf{C}}(\mathbf{w})-\widetilde{\mathbf{C}}-\mathbf{C}_0\right\}^{\mathrm{T}}(\widetilde{\mathbf{C}}-\mathbf{C}_0)=o_p(1).$$

From Conditions C.1-C.3, (A.6), and the steps used in the proof of (A.15), we have that uniformly for $\mathbf{w} \in \mathcal{W}$,

$$\begin{aligned}\left\|\overline{\mathbf{C}}(\mathbf{w})-\mathbf{C}_0\right\|^2 &= L_T(\mathbf{w})+\left\|\widehat{\mathbf{C}}(\mathbf{w})-\overline{\mathbf{C}}(\mathbf{w})\right\|^2-2\left\{\widehat{\mathbf{C}}(\mathbf{w})-\overline{\mathbf{C}}(\mathbf{w})\right\}^{\top}\left\{\widehat{\mathbf{C}}(\mathbf{w})-\mathbf{C}_0\right\}\\ &= L_T^*(\mathbf{w})+o_p(\xi_T).\end{aligned}$$

Thus, uniformly for $\mathbf{w} \in \mathcal{W}$,

$$\begin{aligned}\frac{E\left\|\overline{\mathbf{C}}(\mathbf{w})-\mathbf{C}_0\right\|^2-E\left\|\overline{\mathbf{C}}(\mathbf{w})-\widetilde{\mathbf{C}}\right\|^2}{E\left\|\overline{\mathbf{C}}(\mathbf{w})-\mathbf{C}_0\right\|^2} &= \frac{E\left[\xi_T^{-1}\left\{\left\|\overline{\mathbf{C}}(\mathbf{w})-\mathbf{C}_0\right\|^2-\left\|\overline{\mathbf{C}}(\mathbf{w})-\widetilde{\mathbf{C}}\right\|^2\right\}\right]}{E\left\{\xi_T^{-1}L_T^*(\mathbf{w})+o_p(1)\right\}}\\ &= \frac{E\left[o_p(1)\right]}{E\xi_T^{-1}L_T^*(\mathbf{w})+E[o_p(1)]}. \qquad (A.25)\end{aligned}$$

Suppose the above $o_p(1)$ terms are uniformly integrable, then (A.25) is $o(1)$ uniformly for $\mathbf{w} \in \mathcal{W}$, which is the first part of Proposition 1.

Next, we consider the i.i.d. data case, i.e., the second part of Proposition 1. Denoting $V_t = (F_1^0(X_{1t}), \ldots, F_p^0(X_{pt}))$ and $\widetilde{V}_t = (\widetilde{F}_1(X_{1t}), \ldots, \widetilde{F}_p(X_{pt}))$, it is easy to see that

$$\begin{aligned}& E\left\|\overline{\mathbf{C}}(\mathbf{w})-\mathbf{C}_0\right\|^2-E\left\|\overline{\mathbf{C}}(\mathbf{w})-\widetilde{\mathbf{C}}\right\|^2\\ =\;& 2E\left\{\overline{\mathbf{C}}^{\mathrm{T}}(\mathbf{w})(\widetilde{\mathbf{C}}-\mathbf{C}_0)\right\}-E\left\{(\widetilde{\mathbf{C}}+\mathbf{C}_0)^{\mathrm{T}}(\widetilde{\mathbf{C}}-\mathbf{C}_0)\right\}\\ =\;& 2\sum_{k=1}^{K}w_k\sum_{j=1}^{J}\sum_{m=1}^{M}E\left[C_k(\widetilde{V}_{(j-1)M+m};\widehat{\boldsymbol{\theta}}_k^{(-j)})\left\{\widetilde{C}_{(j)}(X_{(j-1)M+m})-C_0(V_{0,(j-1)M+m};\boldsymbol{\theta}_0)\right\}\right]\\ & -E\left\{(\widetilde{\mathbf{C}}+\mathbf{C}_0)^{\mathrm{T}}(\widetilde{\mathbf{C}}-\mathbf{C}_0)\right\}. \qquad (A.26)\end{aligned}$$

Let $X_t = (X_{1t}, \ldots, X_{pt})^{\top}$ and $\mathcal{X}^{(-j)} = \left\{X_1, \ldots, X_{(j-1)M}, X_{jM+1}, \ldots, X_T\right\}$. So from

$$E\left\{\widetilde{C}_{(j)}(\mathbf{x})\right\} = C_0(\mathbf{u};\boldsymbol{\theta}_0),$$

we know that for any $m \in \{1, \ldots, M\}$ and $j \in \{1, \ldots, J\}$,

$$E\left[C_k(\widetilde{V}_{(j-1)M+m};\widehat{\boldsymbol{\theta}}_k^{(-j)})\left\{\widetilde{C}_{(j)}(X_{(j-1)M+m})-C_0(V_{0,(j-1)M+m};\boldsymbol{\theta}_0)\right\}\mid \mathcal{X}^{(-j)}\right]$$

$$
\begin{aligned}
&= E_{X_{(j-1)M+m}}\Big(E\Big[C_k(\widetilde{V}_{(j-1)M+m};\widehat{\boldsymbol{\theta}}_k^{(-j)}) \\
&\qquad \times \Big\{\widetilde{C}_{(j)}(X_{(j-1)M+m}) - C_0(V_{0,(j-1)M+m};\boldsymbol{\theta}_0)\Big\}\,|X_{(j-1)M+m},\ \mathcal{X}^{(-j)}\Big]\Big) \\
&= E_{X_{(j-1)M+m}}\Big[C_k(\widetilde{V}_{(j-1)M+m};\widehat{\boldsymbol{\theta}}_k^{(-j)}) \\
&\qquad \times E\Big\{\widetilde{C}_{(j)}(X_{(j-1)M+m}) - C_0(V_{0,(j-1)M+m};\boldsymbol{\theta}_0)|X_{(j-1)M+m},\ \mathcal{X}^{(-j)}\Big\}\Big] \\
&= 0,
\end{aligned} \tag{A.27}
$$

where the expectation $E_{X_{(j-1)M+m}}$ is taken with respect to the randomness of $X_{(j-1)M+m}$. The above two formulas (A.26) and (A.27) imply that

$$
E\left\|\overline{\mathbf{C}}(\mathbf{w}) - \mathbf{C}_0\right\|^2 - E\left\|\overline{\mathbf{C}}(\mathbf{w}) - \widetilde{\mathbf{C}}\right\|^2 = -E\left\{(\widetilde{\mathbf{C}} + \mathbf{C}_0)^{\mathrm{T}}(\widetilde{\mathbf{C}} - \mathbf{C}_0)\right\},
$$

where the right-hand-side quantity does not depend on $\mathbf{w}$, completing the proof for the claim that the J-fold CV criterion is an unbiased estimator of the expected CV squared loss plus a term unrelated to $\mathbf{w}$.

A.4 Conditions and Proof of (A.3)

Let $\boldsymbol{\psi} = (\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_{K-1})^\top$ belonging to $\mathcal{Q} = \left\{\boldsymbol{\psi} \in [0,1]^{K-1} : \sum_{k=1}^{K-1}\psi_k = 1\right\}$, $\mathbf{C}^*(\boldsymbol{\psi}) = \sum_{k=1}^{K-1}\psi_k\widehat{\mathbf{C}}_k\,|_{\widehat{\boldsymbol{\theta}}_k=\boldsymbol{\theta}_k^*}$, $L_T^*(\boldsymbol{\psi}) = \|\mathbf{C}^*(\boldsymbol{\psi}) - \mathbf{C}_0\|^2$, and $\widetilde{\xi}_T = \inf_{\boldsymbol{\psi}\in\mathcal{Q}} L_T^*(\boldsymbol{\psi})$.

Condition C.5: There exists a sequence $\widetilde{c}_T \to 0$ such that $T\widetilde{\xi}_T^{-2} \le \widetilde{c}_T$ almost surely.

Condition C.5 is a counterpart of Condition C.4 and requires that $\widetilde{\xi}_T$ grow at a rate faster than $T^{1/2}$, which implies the true copula cannot be well approximated by combining estimated single copulas. This is reasonable because the true copula is a mixture copula with $\omega_k > 0$.

Next, we prove (A.3) by using Conditions C.1, C.2, C.3 and C.5. Let $\mathbf{w}^\star = (0, \ldots, 0, 1)^\top$ and $\widehat{\boldsymbol{\psi}} = (\widehat{w}_1/(1-\widehat{w}_K), \ldots, \widehat{w}_{K-1}/(1-\widehat{w}_K))^\top$. From Conditions C.1-C.3, and the steps used in the proof of (A.18), we have

$$
\begin{aligned}
CV_J(\mathbf{w}^\star) &= \left\|\overline{\mathbf{C}}(\mathbf{w}^\star) - \widetilde{\mathbf{C}}\right\|^2 \\
&= \left\|\overline{\mathbf{C}}_K - \widetilde{\mathbf{C}}\right\|^2 \\
&= \left\|\overline{\mathbf{C}}_K - \widehat{\mathbf{C}}_K + \widehat{\mathbf{C}}_K - \mathbf{C}_0 + \mathbf{C}_0 - \widetilde{\mathbf{C}}\right\|^2 \\
&= O_p(1).
\end{aligned} \tag{A.28}
$$

From Condition C.3, and the steps used in the proof of (A.15), we have

$$
CV_J(\widehat{\mathbf{w}}) = \left\|\overline{\mathbf{C}}(\widehat{\mathbf{w}}) - \widetilde{\mathbf{C}}\right\|^2
$$

$$
\begin{aligned}
&= \left\|\left\{\widehat{\mathbf{C}}(\widehat{\mathbf{w}})-\mathbf{C}_0\right\}-\left\{\widehat{\mathbf{C}}(\widehat{\mathbf{w}})-\overline{\mathbf{C}}(\widehat{\mathbf{w}})\right\}+\left(\mathbf{C}_0-\widetilde{\mathbf{C}}\right)\right\|^2 \\
&= L_T(\widehat{\mathbf{w}})+\left\|\widehat{\mathbf{C}}(\widehat{\mathbf{w}})-\overline{\mathbf{C}}(\widehat{\mathbf{w}})\right\|^2+\left\|\mathbf{C}_0-\widetilde{\mathbf{C}}\right\|^2-2\left\{\widehat{\mathbf{C}}(\widehat{\mathbf{w}})-\overline{\mathbf{C}}(\widehat{\mathbf{w}})\right\}^\top\left(\mathbf{C}_0-\widetilde{\mathbf{C}}\right) \\
&\quad -2\left\{\widehat{\mathbf{C}}(\widehat{\mathbf{w}})-\overline{\mathbf{C}}(\widehat{\mathbf{w}})\right\}^\top\left\{\widehat{\mathbf{C}}(\widehat{\mathbf{w}})-\mathbf{C}_0\right\}+2\left\{\widehat{\mathbf{C}}(\widehat{\mathbf{w}})-\mathbf{C}_0\right\}^\top\left(\mathbf{C}_0-\widetilde{\mathbf{C}}\right) \\
&= L_T^*(\widehat{\mathbf{w}})+\Pi_T(\widehat{\mathbf{w}})+\left\|\widehat{\mathbf{C}}(\widehat{\mathbf{w}})-\overline{\mathbf{C}}(\widehat{\mathbf{w}})\right\|^2-2\left\{\widehat{\mathbf{C}}(\widehat{\mathbf{w}})-\overline{\mathbf{C}}(\widehat{\mathbf{w}})\right\}^\top\left(\mathbf{C}_0-\widetilde{\mathbf{C}}\right) \\
&\quad -2\left\{\widehat{\mathbf{C}}(\widehat{\mathbf{w}})-\overline{\mathbf{C}}(\widehat{\mathbf{w}})\right\}^\top\left\{\widehat{\mathbf{C}}(\widehat{\mathbf{w}})-\mathbf{C}_0\right\}+2\left\{\widehat{\mathbf{C}}(\widehat{\mathbf{w}})-\mathbf{C}_0\right\}^\top\left(\mathbf{C}_0-\widetilde{\mathbf{C}}\right)+\left\|\mathbf{C}_0-\widetilde{\mathbf{C}}\right\|^2 \\
&= L_T^*(\widehat{\mathbf{w}})+O_p(T^{1/2})=\|\mathbf{C}^*(\widehat{\mathbf{w}})-\mathbf{C}_0\|^2+O_p(T^{1/2}) \\
&= \left\|(1-\widehat{w}_K)[\mathbf{C}^*(\widehat{\boldsymbol{\psi}})-\mathbf{C}_0]+\widehat{w}_K(\widehat{\mathbf{C}}_K\mid_{\widehat{\boldsymbol{\theta}}=\boldsymbol{\theta}^*}-\mathbf{C}_0)\right\|^2+O_p(T^{1/2}) \\
&= (1-\widehat{w}_K)^2\left\|\mathbf{C}^*(\widehat{\boldsymbol{\psi}})-\mathbf{C}_0\right\|^2+2(1-\widehat{w}_K)\widehat{w}_K[\mathbf{C}^*(\widehat{\boldsymbol{\psi}})-\mathbf{C}_0](\widehat{\mathbf{C}}_K\mid_{\widehat{\boldsymbol{\theta}}=\boldsymbol{\theta}^*}-\mathbf{C}_0) \\
&\quad +\widehat{w}_K^2\left\|\widehat{\mathbf{C}}_K\mid_{\widehat{\boldsymbol{\theta}}=\boldsymbol{\theta}^*}-\mathbf{C}_0\right\|^2+O_p(T^{1/2}) \\
&= (1-\widehat{w}_K)^2\left\|\mathbf{C}^*(\widehat{\boldsymbol{\psi}})-\mathbf{C}_0\right\|^2+O_p(T^{1/2})+O_p(1)+O_p(T^{1/2}) \\
&= (1-\widehat{w}_K)^2\left\|\mathbf{C}^*(\widehat{\boldsymbol{\psi}})-\mathbf{C}_0\right\|^2+O_p(T^{1/2}), \qquad \text{(A.29)}
\end{aligned}
$$

as $T\to\infty$. Since $\widehat{\mathbf{w}}=\text{argmin}_{\mathbf{w}\in\mathcal{W}}CV_J(\mathbf{w})$, we have $CV_J(\mathbf{w}^\star)\geq CV_J(\widehat{\mathbf{w}})$, which along with the above two results, implies

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

which, along with Condition C.4, implies (A.3).

A.5 More Simulation Results

Here, we present some additional simulation results. First, we compare the out-of-sample forecast performance between MA and MA without $\sum_{k=1}^{K} w_k = 1$ constraint (denoted as MA_{un}). For simplicity and comparing purposes, we do not add the screening step and fix $J = 4$ in all the simulations. In Table A1, the out-of-sample forecast performance between MA and MA without imposing $\sum_{k=1}^{K} w_k = 1$ does not exhibit significant difference because the ratio is close to 1.

Table A2 documents the results of the comparison between MA and MA with the screening step to select K (MAS). For simplicity, we still fix $J = 4$ in this simulation. Table A2 demonstrates that even though the ratio between MAS and MA is less than 1, such a difference is not statistically significant according to the DM test, indicating that the additional screening step does not effectively improve the performance of MA. Intuitively, when the mixture model is correctly specified, both MA and CW will impose zero weight to the unrelated candidate so the performance between MA and MAS should be quite similar. When the model is misspecified, many candidate copulas are often pass the screening step and enter the following MA step. Therefore, one would observe similar performance between MA and MAS no matter whether the true copula is contained in the mixture copula or not.

In Table A3 we compare the performance between MA with selected J and MA with J respectively equals to 2, 4, 5 and 10. As one can see, MA with selected J outperforms MA with fixed J in most cases and the difference is often statistically significant. This shows that the proposed J selection strategy is useful in improving the performance of MA.

Type I simulation: Out-of-sample forecast (sample size=500)		
	$\theta_{Ga} = 0.5, \theta_{Cl} = 5.8, \theta_{Gu} = 5.1$	$\theta_{Ga} = 0.7, \theta_{Cl} = 7.8, \theta_{Gu} = 7.1$
	MA_{un}/MA	MA_{un}/MA
$\omega_{Ga} = 0.5, \omega_{Cl} = 0.5, \omega_{Gu} = 0$	1.0135	1.0397
	(0.79)	(0.77)
$\omega_{Ga} = 0.5, \omega_{Cl} = 0, \omega_{Gu} = 0.5$	1.0272	1.0389
	(0.27)	(0.24)
$\omega_{Ga} = 0, \omega_{Cl} = 0.5, \omega_{Gu} = 0.5$	1.0347	1.0273
	(0.33)	(0.52)
$\omega_{Ga} = 1/3, \omega_{Cl} = 1/3, \omega_{Gu} = 1/3$	0.9536	1.0141
	(0.83)	(0.70)
Type II simulation: Out-of-sample forecast (sample size=500)		
	$\theta_F = 5.5, \theta_{SJ} = 4.8, \theta_J = 4.5$	$\theta_F = 7.5, \theta_{SJ} = 6.8, \theta_J = 6.5$
	MA_{un}/MA	MA_{un}/MA
$\omega_F = 0.5, \omega_{SJ} = 0.5, \omega_J = 0$	1.0242	1.0367
	(0.23)	(0.68)
$\omega_F = 0.5, \omega_{SJ} = 0, \omega_J = 0.5$	1.0241	1.0135
	(0.47)	(0.52)
$\omega_F = 0, \omega_{SJ} = 0.5, \omega_J = 0.5$	1.0206	1.0109
	(0.84)	(0.95)
$\omega_F = 1/3, \omega_{SJ} = 1/3, \omega_J = 1/3$	1.0137	1.0175
	(0.73)	(0.87)

Table A1: Ratios of prediction losses of both MA with $\sum_{k=1}^{K} w_k = 1$ constraint and MA without $\sum_{k=1}^{K} w_k = 1$ constraint for both Type I and Type II simulations. Each simulation is repeated 300 times. J is fixed to 4 for all the simulations. Values in the parenthesis indicate the $p-$value obtained through the Diebold-Mariano Test.

Type I simulation: Out-of-sample forecast (sample size=500)		
	$\theta_{Ga} = 0.5, \theta_{Cl} = 5.8, \theta_{Gu} = 5.1$	$\theta_{Ga} = 0.7, \theta_{Cl} = 7.8, \theta_{Gu} = 7.1$
	MAS/MA	MAS/MA
$\omega_{Ga} = 0.5, \omega_{Cl} = 0.5, \omega_{Gu} = 0$	0.9754	0.9932
	(0.88)	(0.90)
$\omega_{Ga} = 0.5, \omega_{Cl} = 0, \omega_{Gu} = 0.5$	0.9745	0.9941
	(0.90)	(0.92)
$\omega_{Ga} = 0, \omega_{Cl} = 0.5, \omega_{Gu} = 0.5$	0.9875	0.9896
	(0.91)	(0.88)
$\omega_{Ga} = 1/3, \omega_{Cl} = 1/3, \omega_{Gu} = 1/3$	0.9846	0.9837
	(0.89)	(0.93)
Type II simulation: Out-of-sample forecast (sample size=500)		
	$\theta_F = 5.5, \theta_{SJ} = 4.8, \theta_J = 4.5$	$\theta_F = 7.5, \theta_{SJ} = 6.8, \theta_J = 6.5$
	MAS/MA	MAS/MA
$\omega_F = 0.5, \omega_{SJ} = 0.5, \omega_J = 0$	0.9871	0.9803
	(0.94)	(0.90)
$\omega_F = 0.5, \omega_{SJ} = 0, \omega_J = 0.5$	0.9974	0.9923
	(0.89)	(0.91)
$\omega_F = 0, \omega_{SJ} = 0.5, \omega_J = 0.5$	0.9857	0.9908
	(0.95)	(0.89)
$\omega_F = 1/3, \omega_{SJ} = 1/3, \omega_J = 1/3$	0.9942	0.9936
	(0.94)	(0.92)

Table A2: Ratios of prediction losses of both MA and MAS method for both Type I and Type II simulations. Each simulation is repeated 300 times. J is fixed at 4 for all simulations. Values in the parenthesis indicate the $p-$value obtained through the Diebold-Mariano Test.

Type I simulation: Out-of-sample forecast (sample size=500)								
	$\theta_{Ga} = 0.5, \theta_{Cl} = 5.8, \theta_{Gu} = 5.1$				$\theta_{Ga} = 0.7, \theta_{Cl} = 7.8, \theta_{Gu} = 7.1$			
	$MA_{J^*}/MA_{J=2}$	$MA_{J^*}/MA_{J=4}$	$MA_{J^*}/MA_{J=5}$	$MA_{J^*}/MA_{J=10}$	$MA_{J^*}/MA_{J=2}$	$MA_{J^*}/MA_{J=4}$	$MA_{J^*}/MA_{J=5}$	$MA_{J^*}/MA_{J=10}$
$\omega_{Ga} = 0.5, \omega_{Cl} = 0.5, \omega_{Gu} = 0$	0.9769	0.8841	1.0073	0.9765	0.9348	0.9339	0.8931	0.9107
	(0.12)	(0.04)	(0.84)	(0.12)	(0.05)	(0.05)	(0.04)	(0.08)
$\omega_{Ga} = 0.5, \omega_{Cl} = 0, \omega_{Gu} = 0.5$	0.9337	0.8124	0.8675	0.9526	1.0011	0.8357	0.9431	0.8972
	(0.07)	(0.03)	(0.04)	(0.06)	(0.31)	(0.02)	(0.05)	(0.03)
$\omega_{Ga} = 0, \omega_{Cl} = 0.5, \omega_{Gu} = 0.5$	0.9179	0.8325	1.0041	0.9704	0.9321	0.7159	0.9327	0.9172
	(0.06)	(0.03)	(0.61)	(0.11)	(0.08)	(0.01)	(0.07)	(0.06)
$\omega_{Ga} = 1/3, \omega_{Cl} = 1/3, \omega_{Gu} = 1/3$	0.9769	0.9528	0.8709	0.9132	0.9537	0.9921	0.9069	0.8257
	(0.13)	(0.10)	(0.05)	(0.07)	(0.10)	(0.31)	(0.08)	(0.02)
Type II simulation: Out-of-sample forecast (sample size=500)								
	$\theta_F = 5.5, \theta_{SJ} = 4.8, \theta_J = 4.5$				$\theta_F = 7.5, \theta_{SJ} = 6.8, \theta_J = 6.5$			
	$MA_{J^*}/MA_{J=2}$	$MA_{J^*}/MA_{J=4}$	$MA_{J^*}/MA_{J=5}$	$MA_{J^*}/MA_{J=10}$	$MA_{J^*}/MA_{J=2}$	$MA_{J^*}/MA_{J=4}$	$MA_{J^*}/MA_{J=5}$	$MA_{J^*}/MA_{J=10}$
$\omega_F = 0.5, \omega_{SJ} = 0.5, \omega_J = 0$	0.8851	0.9501	0.8912	0.8537	0.8731	0.9269	0.9104	0.8922
	(0.03)	(0.08)	(0.04)	(0.03)	(0.02)	(0.06)	(0.05)	(0.04)
$\omega_F = 0.5, \omega_{SJ} = 0, \omega_J = 0.5$	0.9177	0.9025	0.8241	0.9030	0.9511	0.8546	0.9279	0.8732
	(0.08)	(0.07)	(0.02)	(0.05)	(0.08)	(0.04)	(0.06)	(0.03)
$\omega_F = 0, \omega_{SJ} = 0.5, \omega_J = 0.5$	1.0091	0.9279	0.8397	0.9012	1.0017	0.9346	0.8162	0.9341
	(0.35)	(0.06)	(0.03)	(0.05)	(0.81)	(0.06)	(0.01)	(0.07)
$\omega_F = 1/3, \omega_{SJ} = 1/3, \omega_J = 1/3$	0.8455	0.8703	0.8768	0.9347	0.9607	0.9225	0.8930	0.9134
	(0.03)	(0.04)	(0.05)	(0.07)	(0.06)	(0.06)	(0.04)	(0.05)

Table A3: Ratios of prediction losses of both MA with selected $J = J^*$ and MA under $J = 2, 4, 5, 10$ for both Type I and Type II simulations. Each simulation is repeated 300 times. Values in the parenthesis indicate the $p-$value obtained through the Diebold-Mariano Test.