Construction of Leading Economic Index for Recession Prediction using Vine Copulas *

Kajal Lahiri[†1] and Liu Yang[‡2]

[1]Department of Economics, University at Albany, SUNY, NY 12222, USA

[2]School of Economics, Nanjing University, Jiangsu 210093, P. R. China

Abstract

This paper constructs a composite leading index for business cycle prediction based on vine copulas that capture the complex pattern of dependence among individual predictors. This approach is optimal in the sense that the resulting index possesses the highest discriminatory power as measured by the receiver operating characteristic (ROC) curve. The model specification is semi-parametric in nature, suggesting a two-step estimation procedure, with the second-step finite dimensional parameter being estimated by QMLE given the first-step non-parametric estimate. To illustrate its usefulness, we apply this methodology to optimally aggregate the ten leading indicators selected by The Conference Board (TCB) to predict economic recessions in the United States. In terms of the discriminatory power, our method is significantly better than the Index used by TCB.

JEL Classifications: C14, C15, C32, C43, C51, C53, E37

Key words: Leading Economic Index, Receiver operating characteristic curve, Vine copula, Block bootstrap.

*We thank Badi Baltagi, Søren Johansen, Ataman Ozylidirim, Andrew Patton, Jeffery Racine and Robin Sickles for many helpful comments when we presented earlier versions of this paper at the 2nd IAAE Conference, the 11th World Congress of the Econometric Society, the New York Camp Econometrics XI and the 33rd CIRET conference. Any remaining errors are solely ours.

[†]Tel.: +1 518 442 4758. E-mail address: klahiri@albany.edu.

[‡]Corresponding author. Tel.: +86 18710075220. E-mail address: lyang2@nju.edu.cn.

1 Introduction

Inspired by the seminal work of Burns and Mitchell (1946), a great body of literature is devoted to the prediction of the underlying evolving state of the economy characterized by business cycles. These efforts can essentially be classified into two groups. The first approach, with the time series model as a typical example, is to employ the historical information of the past cycles to predict future based on the persistence exhibited by most economic series. The second one focuses on the contemporaneous information embodied in some appropriately selected indicators. As characterized by Burns and Mitchell (1946), "business cycles ... [which] consist of expansions occurring at about the same time in many economic activities, followed by similarly general recessions, contractions and revivals which merge into the expansion phase of the next cycle". These economic activities, as measured by various macroeconomic indicators, would evolve over time along with the unobserved cycles. The fact that the statistical properties of these indicators are drastically different when the state of the economy switches provides the basis for the second approach. Unlike the first one, this approach does not assume the continuity of the underlying economic series, which makes it more appropriate for turning point prediction. In this paper, we will follow the second approach by constructing a composite index by extracting the predictive information contained in a number of representative leading indicators with correlation matrix changes between regimes.

One of the most prominent macro forecasting tool in the United States is the Leading Economic Index (LEI) developed by The Conference Board (TCB). Rather than being model-based, it simply summarizes ten leading indicators in a straightforward fashion. Since it is easy to calculate and interpret, the LEI enjoys great popularity in business cycle forecasting for decades. However, due to its lack of a statistically rigorous foundation, we do not know if it is optimal in some sense. To overcome this drawback, we create a new index that combines the same ten leading indicators within a probabilistic framework. Unlike the current LEI, our index has a recession probability interpretation. It ia shown to be optimal in maximizing the predictive accuracy as measured by the receiver operating characteristic (ROC) curve.

The ROC curve is designed to evaluate the discriminatory capacity of a forecasting system to predict a binary event. Compared with the commonly used mean squared error, the ROC curve is not influenced by the marginal distribution of the target variable, making it more appropriate to use in forecasting relatively uncommon events, like an economic recession, see Lahiri and Yang (2013). This methodology was initially developed in the field of signal detection theory, where it was used to evaluate the discriminatory ability for a binary detection system to distinguish between two clearly-defined possibilities: signal plus noise and noise only. Thereafter, it has gained increasing popularity in many other related fields. A general introduction to ROC analysis can be found in Fawcett (2006), Pepe (2000), Swets et al. (2000), and Zhou et al. (2002). Until recently, ROC has not gained much attention from econometricians. Exceptions include Berge and Jordà (2011), Drehmann and Juselius (2014), Jordà and Taylor (2012), Lahiri and Wang (2013), Liu and Moench (2014), among a few others.

Our leading index depends on copula functions, a method dating back Sklar (1973). The optimal way to aggregate a set of indicators can be shown to be a non-linear functional of the conditional distribution of these predictors given the actual. One might use a fully specified parametric model, such as the multivariate normal distribution. However, it is subject to misspecification bias. Alternatively, we might not impose any constraint and use a non-parametric approach to estimate the conditional distribution. However, it suffers from the "curse of dimensionality" because the sample size is seldom large enough when compared with the number of predictors. The approach we take is semi-parametric in nature. The marginal distribution of each predictor is estimated non-parametrically, while the dependence parameters amongst predictors that are represented by copulas are estimated by quasi-maximum likelihood. This approach allows for a larger degree of flexibility in specifying the marginal distributions than the full parametric approach. In addition, the dimension of non-parametric estimation is such that the "curse of dimensionality" is resolved. The only difficulty comes from the specification of high-dimensional copulas. To capture the realistic yet complex dependence structure in a high-dimensional setting, we cannot use the direct extension of regular bivariate copulas since they require homogeneity in dependence. As an alternative, this paper develops multivariate distributions of the predictors based on vine

copulas, which rely on bivariate copulas as building blocks. It is well recognized that characterizing a high-dimensional dependence is harder than characterizing a bivariate dependence. Thus, our method is easy to use as long as a set of bivariate copulas are chosen appropriately. A two-step procedure is proposed to estimate the unknown quantities in our model, with each step being computationally affordable using a standard software package. In the empirical application, our composite index is shown to deliver a better classification of all time periods into the two regimes than the current LEI.

The paper is organized as follows. In Section 2, we construct the composite index by combining the multiple predictors based on the vine copulas and we show this combination rule to be optimal to maximize the ROC curve. A two-step estimation procedure is also described. Section 3 illustrates the usefulness of our method in predicting economic recessions in the United States using ten currently-used leading indicators of TCB. Finally, in Section 4, we offer concluding remarks and suggestions of further extensions.

2 A copula-based combination scheme

Throughout this section, we denote the individual predictor by X_i for $i = 1,2,...,I$, and the binary variable to be predicted is Z, which is one when the event of interest occurs and zero otherwise. We use upper case letters to denote cumulative distribution functions and corresponding lower case letters to denote the density functions. Our objective is to use the information content in each X_i to predict the occurrence of the binary event. For this purpose, forecast combination is a natural way to pursue. Let X be the vector of $(X_1, X_2, ..., X_I)'$. Given Z, we have two conditional distributions of X, namely, $F(\cdot|Z=1)$ and $F(\cdot|Z=0)$, both of which are $I-$dimensional cumulative distribution functions. Before outlining our combination scheme, it is necessary to discuss the modeling of $F(\cdot|Z=1)$ and $F(\cdot|Z=0)$, which are two building blocks of the scheme.

The simplest case to deal with is when X follows a multivariate normal distribution given Z. In this scenario, the optimal combined forecast takes an interesting form, as shown later. Despite its parsimony and familiarity, the normality assumption has several key shortcom-

ings. First, it requires that each X_i be normally distributed. Thus, it rules out those predictors, whose distributions are skewed, fat-tailed, or multimodal. Furthermore, the discrete predictor is not allowed. Second, the dependence structure among predictors is also restricted. In the multivariate statistics literature, more than one dependence measures exist. The most commonly used is the Pearson correlation coefficient. The multivariate normal distribution is flexible in terms of this particular measure since it permits any type of correlation matrix as long as it is positive definite. However, the Pearson correlation is affected by the marginal distributions of the predictors, and thereby is often not regarded as the true dependence measure (Joe (1997)). Moreover, it merely measures the dependence in the center of the distributions. In practice, interest may center on the dependence behavior in the left and right tails of the distributions. As an illustration, we consider the first two predictors, i.e. X_1 and X_2. The upper tail dependence coefficient is defined as $\lim_{q\to 1^-} P(X_2 > F_2^{-1}(q)|X_1 > F_1^{-1}(q))$, and the lower tail dependence coefficient is $\lim_{q\to 0^+} P(X_2 \leq F_2^{-1}(q)|X_1 \leq F_1^{-1}(q))$, where $F_j^{-1}(\cdot)$ is the quantile function of X_j for $j = 1,2$. Roughly speaking, both coefficients are the probability that one predictor is very large (small) given that the other is also very large (small). As a result, they measure the dependence between X_1 and X_2 when both are extremely large (small). Many predictors in X may exhibit strong tail dependence in practice. Unfortunately, the normality assumption does not entail positive dependence in both tails so that extreme events appear to be uncorrelated, see Demarta and McNeil (2005) for more details. For this reason, we will not impose multivariate normality. The approach we are going to take is much more robust and it remedies the two aforementioned pitfalls.

The first pitfall is concerned with the non-normality of the marginal distributions. In our empirical application, each predictor in X reflects one aspect of the whole economy, which is related to future economic activities. The Conference Board uses different transformations to construct these predictors. For example, TCB computes the symmetric percentage change of the average weekly hours, but it sticks with the level form of the interest rate spread. These transformations are likely to change the marginal distributions of some predictors, even though the original series can be represented by a simple data generating process (DGP). It is hard, if not impossible, to figure out the implied distribution of the transformed predictor when the DGP of the original series is known. Following the same line of reasoning as

Harding and Pagan (2011), we will not impose any additional distributional restriction on each X_i beyond some smoothness conditions. Instead, the marginal distributions are estimated by non-parametric method. Specifically, let $k(\cdot)$ be a second-order symmetric kernel function. The marginal densities of X_i given $Z = 1$ and $Z = 0$ are estimated by

$$\hat{f}_i^1(x_i) \equiv \frac{1}{h_i^1 \sum_{t=1}^T I(Z_t = 1)} \sum_{t=1}^T I(Z_t = 1) k(\frac{x_i - X_{it}}{h_i^1}),$$

and

$$\hat{f}_i^0(x_i) \equiv \frac{1}{h_i^0 \sum_{t=1}^T I(Z_t = 0)} \sum_{t=1}^T I(Z_t = 0) k(\frac{x_i - X_{it}}{h_i^0}), \tag{1}$$

respectively, where $I(\cdot)$ is the indicator function which equals one when the condition in the parenthesis is true and zero otherwise. h_i^1 (when $Z = 1$) and h_i^0 (when $Z = 0$) are the bandwidths for X_i chosen by the researcher. The choice of h_i depends upon the tradeoff between the bias and the variance of the resulting estimator and the selected bandwidth may vary across i. A large body of literature discusses the optimal choice of h_i, see Li and Racine (2008) for a comprehensive review.

To address the second pitfall, we adopt copula functions. Since the seminal work of Sklar (1973), the multivariate modeling based on copulas has received growing attention in statistics. A copula is a multivariate distribution function whose univariate marginals are uniforms between zero and one. For any $I-$dimensional distribution function $F(x_1, x_2, ..., x_I)$ with $F_i(\cdot)$ as its marginal distribution, the usefulness of copula roots in the following decomposition theorem, i.e.,

$$F(x_1, x_2, ..., x_I) = C(F_1(x_1), F_2(x_2), ..., F_I(x_I)), \tag{2}$$

for all $(x_1, x_2, ..., x_I) \in R^I$, where C is the copula associated with $F(x_1, x_2, ..., x_I)$. If X is continuously distributed, C is uniquely determined. To uncover the corresponding copula, we can use the inverse of (2), i.e.,

$$C(v_1, v_2, ..., v_I) = F(F_1^{-1}(v_1), F_2^{-1}(v_2), ..., F_I^{-1}(v_I)), \tag{3}$$

where $F_i^{-1}(\cdot)$ is the inverse distribution function of X_i. Both (2) and (3) provide a way of isolating marginal distributions with their dependency structure. One can model $F_i(\cdot)$ individually, and then choose a reasonable copula C to form a joint distribution. This theorem assures that $F(x_1, x_2, ..., x_I)$ resulting from (2) is a valid multivariate distribution function. A general introduction to the modeling strategies based on copulas was given by Joe (1997), Nelsen (2006), Patton (2012), and Trivediand Zimmer (2005). Anatolyev (2009), Patton (2006) , Patton (2013) and Scotti (2011) applied this methodology to predict multiple economic events. Patton and Fan (2014) provided a recent survey on copula methods from the perspective of econometrics.

Like marginal distributions, the copula can be estimated non-parametrically. However, the "curse of dimensionality" argument implies that the estimate is not reliable if the number of predictors, I, exceeds two and the sample size T is not large. Unfortunately, this is the case in our empirical application, where we have a sample of moderate size with 10 predictors in X. Chen and Fan (2006) considered a semi-parametric copula-based model and developed a two-step estimation procedure. In the first step, the marginal distributions of standardized innovations are estimated non-parametrically. In the second step, the parameters in the copula are estimated by quasi-maximum likelihood when each marginal in the likelihood function is replaced by the first-step estimate. Like other semi-parametric models, Chen and Fan's approach avoids the "curse of dimensionality" in that the multivariate component, i.e. the copula, is parameterized. In order to apply this methodology in our context, a parametric family of copula must be specified. The analysts often know little about the dependence structure though they might be pretty sure of the marginal distributions. Therefore, selecting an appropriate copula seems to be quite challenging especially in a high-dimensional setting. When $I = 2$, there are a large number of bivariate candidates. These include, but are not limited to, elliptical class (Gaussian copula and t copula), Archimedean families (Clayton copula, Gumbel copula, Frank copula, etc.), and the finite mixture of bivariate copulas. Many graphic and analytical goodness-of-fit tests are proposed to provide guidelines in specifying an adequate bivariate copula, see Chen and Fan (2006, 2007), Fermanian (2005), Genest *et al.* (2009), Klugman and Parsa (1999), just to name a few.

When $I > 2$, the parametric families of copulas are very scare. The most obvious choice

is multivariate elliptical class. The $I-$dimensional Gaussian copula can be generated by (3) if F is the $I-$dimensional normal distribution function. Alternatively, we can replace F by the $I-$dimensional t distribution to obtain the multivariate t copula. However, these higher dimensional extensions cannot accommodate the complicated tail dependence structures among different pairs of predictors. For example, the multivariate t copula has only one degree of freedom, meaning that all pairs of predictors share the same tail dependence pattern, which is not realistic. In this paper, we exploit the recent advances on vine copulas (Aas *et al.* (2009), Berg and Aas (2009), Czado (2010), Kurowicka and Joe (2011)) since they are constructed from the bivariate copulas, which are much easier to specify and estimate.

To fix the idea, we consider a trivariate density, namely, $f_{123}(x_1,x_2,x_3)$. First, $f_{123}(x_1,x_2,x_3)$ can be decomposed into the product of one marginal density and two conditional densities,

$$f_{123}(x_1,x_2,x_3) = f_1(x_1)f_{2|1}(x_2|x_1)f_{3|12}(x_3|x_1,x_2). \tag{4}$$

It follows by twice differentiating (2) for $I = 2$ that

$$f_{12}(x_1,x_2) = f_1(x_1)f_2(x_2)c_{12}(F_1(x_1),F_2(x_2)), \tag{5}$$

where $c_{12}(v_1,v_2) \equiv \partial^2/\partial v_1 \partial v_2 C_{12}(v_1,v_2)$ is the copula density between X_1 and X_2. This implies that

$$f_{2|1}(x_2|x_1) = \frac{f_{12}(x_1,x_2)}{f_1(x_1)} = f_2(x_2)c_{12}(F_1(x_1),F_2(x_2)). \tag{6}$$

By the same reasoning, we have

$$f_{3|12}(x_3|x_1,x_2) = \frac{f_{3|2}(x_3|x_2)f_{1|23}(x_1|x_2,x_3)}{f_{1|2}(x_1|x_2)} = f_{3|2}(x_3|x_2)c_{13|2}(F_{1|2}(x_1|x_2),F_{3|2}(x_3|x_2)), \tag{7}$$

where $c_{13|2}$ is the copula between X_1 and X_3 given X_2. Substituting (6) and (7) into (4), $f_{123}(x_1,x_2,x_3)$ can be expressed as

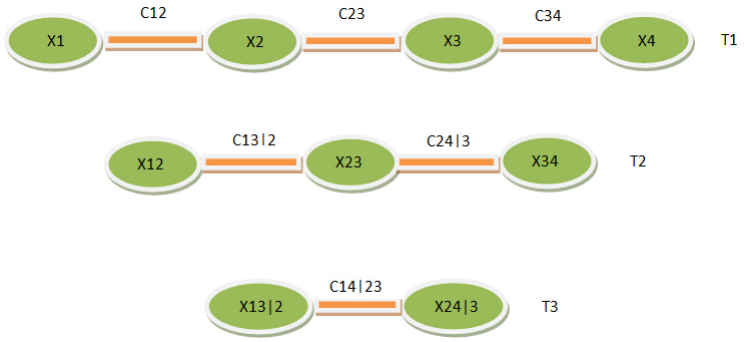
$$f_1(x_1)f_2(x_2)f_3(x_3)c_{12}(F_1(x_1),F_2(x_2))c_{23}(F_2(x_2),F_3(x_3))c_{13|2}(F_{1|2}(x_1|x_2),F_{3|2}(x_3|x_2)), \tag{8}$$

where we make use of the fact that $f_{3|2}(x_3|x_2) = f_3(x_3)c_{23}(F_2(x_2), F_3(x_3))$. Consequently, the trivariate density can be reformulated as the product of three marginal densities $(f_1(x_1), f_2(x_2), f_3(x_3))$ and three bivariate copula densities $(c_{12}, c_{23}, c_{13|2})$. Two copulas are unconditional (c_{12}, c_{23}) and one is conditional $(c_{13|2})$. In general, any $I-$dimensional density can be decomposed in this way. There are I marginal densities, $I-1$ unconditional copulas and $(I-2)(I-1)/2$ conditional copulas in the decomposition. The readers are referred to Aas *et al.* (2009) for the general formula when $I > 3$.

The decomposition in (8) is called pair-copula construction (PCC) of multivariate distribution. For high-dimensional densities, we can generate different decompositions by changing the ordering of the predictors. For example, there are totally 240 different constructions of a $5-$dimensional density. Bedford and Cooke (2001) suggested the use of a nested tree structure to visualize a specific pair-copula construction and called it *the regular vine*. Within the class of regular vine, we only focus on the so called *D-vine* (Kurowicka and Cooke (2004)), where all predictors are treated equally and none of them plays a central role. A specific example of the D-vine with four predictors is illustrated in Figures 1. The first tree T_1 has four nodes, each corresponding to an individual predictor. The edge connecting the adjacent nodes represents the bivariate unconditional copula describing the dependence between these two nodes. In the second tree T_2, there are only three nodes, each of which corresponds to the edge in the previous tree T_1. The edges in T_2 represent the bivariate conditional copulas with the common predictor between the adjacent nodes being the conditioning variable. The last tree T_3 has two nodes, which are linked by one edge. The conditioning variables in the copula are given by the union of all conditioning variables in the two nodes. The whole vine is characterized by three trees, nine nodes and six edges. Note that the D-vine structure in Figures 1 is uniquely determined once the ordering of the predictors in T_1 is fixed. The question thus boils down to how the ordering in T_1 is determined. To the best of our knowledge, there is no formal strategy in the literature to specify the ordering given a dataset. In practice, this may depend on which bivariate relationships among predictors are the most important to model. Suppose X_1 is more correlated with X_3 than it is with X_2 in the sample. Then it is better to put X_1 closer to X_3 than to X_2. Aas *et al.* (2009) provided an empirical example from finance to implement this strategy in a four-dimensional case. For a high-dimensional illustration, the

readers are referred to Dissmann *et al.* (2013).

Figure 1: A D-vine with 4 variables, 3 trees and 6 edges



The idea of PCC was first explored by Joe (1996) and Bedford and Cooke (2001, 2002). Aas *et al.* (2009) established the main framework and proposed a simulation algorithm, as well as a sequential estimation method of PCC model. However, this methodology has not received much attention in econometrics to date. Both the theoretical and the empirical papers on PCC are scare in economics. Exceptions include Brechmann and Czado (2013), Minand Czado (2010), and Zimmer (2014). Compared with the multivariate elliptical copulas, PCC model provides more flexibility because it enables us to specify a wide range of parametric families of copulas to capture the different (conditional or unconditional) dependence structures in Figures 1. In fact, if each of the pair relationship is modeled by bivariate Gaussian copula or t copula with the same degree of freedom, the PCC reduces to multivariate Gaussian or t copula. However, it is neither necessary nor desirable to require these pair-copulas belong to the same parametric family *a priori*. Suppose, for example, the empirical evidences show that X_1 and X_2 are independent of each other in both tails, while a strong upper tail dependence exists between X_2 and X_3 (but weak lower tail dependence). The Gaussian copula may be suitable for c_{12}, and the Gumbel copula is a good choice for c_{23}. Nevertheless, bivariate copula specification is essential to make PCC model more useful. Many goodness-of-fit tests would help in choosing a series of copulas that best fit the data. As a cost, the PCC model often includes many parameters to estimate, especially when I is large. Fortunately, we do not need to estimate all parameters in one step. Instead, we can estimate the marginal distributions by (1), and then estimate the parameters in each bivariate copula using the algorithm given by Aas *et al.* (2009). Though inefficient by themselves, these sequential estimators are

consistent and asymptotically normally distributed, as shown by Chen and Fan (2006). In order to achieve asymptotic efficiency, we can take these sequential estimates as the starting values, and maximize the likelihood function by a recursive algorithm. The joint maximum likelihood estimator is expected to be efficient.

The main goal of this paper is to construct a new composite index summarizing the idiosyncratic information contained in each indicator in an optimal way. A number of forecast skill measures have been proposed in the literature to quantify the performance of competing forecasts. Here, we use the receiver operating characteristic (ROC) curve and the area under the curve (AUC). ROC visualizes the discriminatory power of a forecasting system in distinguishing between $Z = 1$ and $Z = 0$. If the forecasts are completely insensitive to the value Z would take, they have zero discriminatory power. On the other hand, forecasts which take one value when $Z = 1$ and take another when $Z = 0$ would obviously possess the highest discriminatory power. Most real-life forecasts lie between these two extremes. Without loss of generality, a cutoff w is adopted such that an observation with $X_1 > w$ is predicted as $Z = 1$. We can define two conditional probabilities resulting from this decision rule as,

$$\begin{aligned} H(w) &\equiv P(X_1 > w|Z = 1), \\ F(w) &\equiv P(X_1 > w|Z = 0). \end{aligned} \tag{9}$$

$H(w)$ is referred to as the hit rate and it is the probability of correct forecast when $Z = 1$. $F(w)$ is called the false alarm rate or the probability of false forecast when $Z = 0$. Ideally, we hope $H(w)$ could be as large as possible and $F(w)$ should be as small as possible. Both of them are functions of w. In general, given the forecasting system, it is hard to achieve a high value of $H(w)$ without changing $F(w)$. The trade off between them is depicted by plotting the pair $(F(w), H(w))$ in a unit square for a fine grid of w. The resulting ROC curve is an increasing function from $(0,0)$ to $(1,1)$. The ROC curve for forecasts with zero discriminatory power is represented by the diagonal line in the unit square with its AUC 0.5. Conversely, the ROC curve described by the left and upper boundaries of the square has the highest discriminatory power with its AUC 1. Most real-life forecasts yield an ROC curve lying in the upper triangular area whose AUC is strictly between 0.5 and 1.

When more than one predictors are available, none of them could be maximally utilized unless the information contained in them is processed in an efficient manner so that the hit rate is maximized for any given false alarm rate. The region C_α defined as

$$\{X : \frac{f(X|H_1)}{f(X|H_0)} > w\}$$

plays a critical role in testing $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$. Here, $X = (X_1, X_2, ..., X_I)'$ is the vector of all predictors, $f(X|H_j)$ is the likelihood function under H_j for $j = 0, 1$, and w is a constant such that $P(f(X|H_1)/f(X|H_0) > w|H_0) = \alpha$. Among all tests of H_0 against H_1 with the same size α, Neyman-Pearson lemma states that the power, defined as $P(f(X|H_1)/f(X|H_0) > w|H_1)$, achieves its maximum if we reject H_0 when $f(X|H_1)/f(X|H_0) > w$. Therefore, the likelihood ratio test for the simple hypothesis is the most powerful. The implication of this lemma is that the rule constructed from the likelihood ratio of multiple predictors maximizes the hit rate for any given false alarm rate. This rule can be used to justify the test $H_0 : Z = 0$ against $H_1 : Z = 1$, where both the null and the alternative are simple. When $I = 3$, the combined forecast takes the following form,

$$\frac{\hat{f}_1^1(x_1)\hat{f}_2^1(x_2)\hat{f}_3^1(x_3)\hat{c}_{12}^1(\hat{F}_1^1(x_1), \hat{F}_2^1(x_2))\hat{c}_{23}^1(\hat{F}_2^1(x_2), \hat{F}_3^1(x_3))\hat{c}_{13|2}^1(\hat{F}_{1|2}^1(x_1|x_2), \hat{F}_{3|2}^1(x_3|x_2))}{\hat{f}_1^0(x_1)\hat{f}_2^0(x_2)\hat{f}_3^0(x_3)\hat{c}_{12}^0(\hat{F}_1^0(x_1), \hat{F}_2^0(x_2))\hat{c}_{23}^0(\hat{F}_2^0(x_2), \hat{F}_3^0(x_3))\hat{c}_{13|2}^0(\hat{F}_{1|2}^0(x_1|x_2), \hat{F}_{3|2}^0(x_3|x_2))}, \quad (10)$$

where the numerator is the fitted joint density of (X_1, X_2, X_3) given $Z = 1$, while the denominator is the fitted joint density of (X_1, X_2, X_3) given $Z = 0$. In (10), $\hat{f}_i^j(x_i)$ is given by (1), and $\hat{F}_i^j(x_i) = \int_{-\infty}^{x_i} \hat{f}_i^j(u)du$ for $i = 1, 2, 3$ and $j = 0, 1$. All $\hat{c}$'s denote the copulas with the parameters being replaced by their joint maximum likelihood estimates, as mentioned before.[1] According to Joe (1996), all conditional distribution functions can be derived by the partial derivative of the selected copulas. For example,

$$F_{1|2}^j(x_1|x_2) = \frac{\partial C_{12}^j(F_1^j(x_1), F_2^j(x_2))}{\partial F_2^j(x_2)}, \quad (11)$$

[1] In general, $c_{13|2}$ is a conditional copula whose parameters depend on x_2 in an unknown way. To simplify the analysis, we will avoid this complexity and assume all parameters in $c_{13|2}$ are constants. See Acar *et al.* (2011, 2012) for details on a general PCC model.

and $\hat{F}^{j}_{1|2}(x_1|x_2)$ denotes the estimate of (11) when all unknown objects in (11) are substituted with their estimates.

We predict $Z = 1$ if and only if (10) exceeds w. In this case, the size is the false alarm rate $F(w)$, while the power is the hit rate $H(w)$. Given $F(w)$, $H(w)$ is maximized among all possible combination rules based on (X_1, X_2, X_3).[2] Suppose Z is independent of (X_1, X_2, X_3), that is, the information contained in (X_1, X_2, X_3) is completely irrelevant for the purpose of predicting Z. In this special case, the three predictors (X_1, X_2, X_3) cannot be utilized to distinguish between $Z = 1$ and $Z = 0$. In this case, (10) is always equal to 1 and the resulting ROC curve is the diagonal. As long as (X_1, X_2, X_3) is of some relevance to predicting Z, the two joint densities in (10) cannot be identical. In general, the larger the gap between the numerator and the denominator in (10), the higher will be the ROC curve. In the simple case of two predictors, Lahiri and Yang (2015b) used a counterfactual experiment to show that a higher ROC curve can be obtained if either the means of the two predictors or the correlation coefficient between the two predictors vary considerably as Z changes. The same conclusion is expected to hold in the current multivariate circumstance also.[3]

The likelihood ratio in (10) is of some interest in itself. However, a more interpretable statistic is the predictive probability of $Z = 1$ given all predictors, namely, $P(Z = 1|X)$. We predict $Z = 1$ if and only if $P(Z = 1|X)$ exceeds a threshold value. An ROC curve can be generated by using $P(Z = 1|X)$ instead of (10). Lahiri and Yang (2015b) showed that both ROC curves are identical since $P(Z = 1|X)$ is a strictly increasing transformation of (10) due to the Bayes' theorem. Note that

$$\begin{aligned} P(Z=1|X) &= \frac{f(X|Z=1)P(Z=1)}{f(X|Z=0)P(Z=0)+f(X|Z=1)P(Z=1)} \\ &= \frac{\frac{f(X|Z=1)}{f(X|Z=0)}P(Z=1)}{P(Z=0)+\frac{f(X|Z=1)}{f(X|Z=0)}P(Z=1)}, \end{aligned} \tag{12}$$

which is obviously monotonic with the likelihood ratio $f(X|Z = 1)/f(X|Z = 0)$. Two strands of literature exist with regard to combining multiple predictors for the purpose of

[2]The optimality of (10) was also established in McIntosh and Pepe (2002).

[3]Due to the high dimension, we do not pursue a full-fledged counterfactual analysis to isolate the marginal contribution of each predictor in our empirical application.

detecting a given disease in medical statistics. The first approach focuses on modeling $P(Z=1|X)$, resembling the binary response regression model in econometrics. For example, Pepe *et al.* (2006) recommended the generalized linear model with unknown link function for $P(Z=1|X)$. The estimator is obtained as the maximizer of the empirical AUC, and can be taken as a special case of maximum rank correlation estimator proposed by Han (1987). The only drawback of this estimator is the discontinuity of its objective function, which was replaced by a kernel-based smoothed version by Brown and Wang (2005) and Lin *et al.* (2011). Our paper follows the second approach and we try to model the two conditional distributions $f(X|Z=1)$ and $f(X|Z=0)$ using the vine copula technique. In a similar vein, Jafarzadeh *et al.* (2016) developed a random effects model to capture the multivariate dependence structure among several predictors, each of which is assumed to be normally distributed. These two approaches are mathematically equivalent in that the restrictions imposed on $f(X|Z=1)$ and $f(X|Z=0)$ can be translated into those imposed on $P(Z=1|X)$ through (12). A well-known example can be found in Hastie et al. (2001), which proved that if the two conditional distributions, i.e. $f(X|Z=1)$ and $f(X|Z=0)$, are $I-$dimensional normal[4], the corresponding $P(Z=1|X)$ can be described by a logit regression model with a quadratic index. In addition, if these two normal distributions share the same covariance matrices, $P(Z=1|X)$ reduces to a logit regression with a linear index.[5] In general, the implied $P(Z=1|X)$ from arbitrary distributions $f(X|Z=1)$ and $f(X|Z=0)$ is not likely to take a well-known form, especially if all marginal distributions are non-parametrically estimated and some pairwise copulas are not Gaussian in our vine copula construction. Nevertheless, (12) provides a tool to translate the estimated likelihood ratio (10) into a predictive probability of a binary event like an incoming recession.

Given a time series data $\{(Z_t, X_t) : t = 1, ..., T\}$, we first estimate all unknown quantities in (10) using the previous procedure, and evaluate the estimated (10) at each observation. Given a threshold w, we can calculate the proportion of cases when $Z_t = 1$ is correctly predicted,

[4]This holds if the marginal distribution of each predictor given Z is normal and all pairwise copulas in (10) are Gaussian.

[5]This implies that we can estimate the logit model (with or without quadratic terms) by maximum likelihood to get the optimal rule. Since ROC curve is invariant to the logit transformation, an even simpler combination rule is given by the linear index of the estimated logit model. However, the coefficients in the linear index must be estimated by maximum likelihood, rather than by OLS as we often do in a typical linear probability model.

which is the empirical hit rate $\hat{H}(w)$. Analogously, the empirical false alarm rate $\hat{F}(w)$ can be calculated as well. By plotting the pair $(\hat{F}(w), \hat{H}(w))$ for a range of w, the empirical ROC curve is obtained. The empirical AUC can be derived numerically as in Fawcett (2006). All of these empirical quantities are subject to sampling variability, which should be accounted for properly in order to draw a statistically meaningful conclusion. Lahiri and Yang (2015a) derived the asymptotic confidence bands for ROC curves using time series data. It is cumbersome, if not impossible, to get the asymptotic distribution of the empirical ROC curve given the above two-step estimation procedure. We favor the bootstrap due to its convenience. To mimic the serial dependence in the original sample, a circular block bootstrap is used. The block length is either fixed or can be chosen randomly from a geometric distribution. We implement the whole estimation process on each bootstrap replicate of the original sample. The confidence intervals are constructed using these bootstrap replicates of ROC curve. The details can be also found in Lahiri (2003). This method is expected to work well when T is sufficiently large.

3 An Empirical illustration

In this section, we will demonstrate the usefulness of the methodology developed in Section 2 by constructing a new Leading Economic Index for predicting business cycles in the United States. The business cycle is not about any single variable, like GDP, or any other observable variable. Rather, the business cycle is about the dynamics and interactions of a set of relevant variables. The comovement of many individual economic series over the cycles leads naturally to the creation of composite index, which summarizes the idiosyncratic information encompassed in each individual indicator in an easy digestible way. It agrees with Burns and Mitchell's view and is thought of as a useful tool for monitoring and forecasting the unobserved business cycles.

Currently, the widely used Leading Economic Index (LEI) constructed by The Conference Board (TCB) is based on the following ten predictors: ISM Index of New Orders (X_1), Interest rate spread (10-year Treasury bonds less federal funds) (X_2), Average weekly hours

(manufacturing) (X_3), Average weekly initial claims for unemployment insurance (X_4), Manufacturers' new orders (X_5), Consumer goods and materials, Manufacturers' new orders (non-defense capital goods excluding aircraft orders) (X_6), Building permits, New private housing units (X_7), S&P500 (X_8), Leading Credit Index (X_9), and Average consumer expectations for business conditions (X_{10}). The monthly series of these predictors run from 1959:08 to 2011:12. TCB aggregates these 10 predictors by following four steps: first, compute the month-to-month symmetric percent change for each predictor; second, adjust these changes to equalize the volatility of each predictor; third, add the adjusted changes to obtain the growth rate of the composite index; finally, compute the level of the index for the current month by using the growth rate and the last month index.[6] The reference variable Z is the recession indicator defined by the National Bureau of Economic Research (NBER) business cycle dating committee. It is one if the recession occurred, and zero otherwise. The sample proportion of months that were in recession is about 14.8%, indicating that it is a relatively uncommon event. Levanon *et al.* (2014) assessed the performance of the LEI in terms of predicting Z. They found the LEI, by utilizing diverse information implicit in the 10 predictors, performed remarkably well in signalling the incoming recessions both in- and out-of-sample. Though popular by itself, this linear combination scheme is far from optimal. The primary reason is that the four-step aggregation procedure is completely isolated from the target variable. Consequently, the resulting LEI is unlikely to provide an efficient summary. Indeed, one of our aims is to show how large the improvement over the current LEI can be gained through our non-linear combination scheme. For the sake of brevity, we merely consider six month ahead forecast, that is, X_t is used to predict Z_{t+6}.

In order to implement our method, all individual series must be stationary. Some predictors can be regarded as stationary in their level form, such as the interest spread. For others, we convert them into the month-to-month symmetric percent change to achieve stationarity. The p-values of a formal Augmented Dickey-Fuller test are uniformly lower than 0.01, confirming the stationarity of all series we use. Figure 2 presents the two scatter plot matrices of the 10 predictors (after stationarity transformation) during recessions and expansions. We observe the following: (i) some of the leading indicators do not seem to be normally dis-

[6] In the base year 2004, the average value of LEI is fixed at 100.

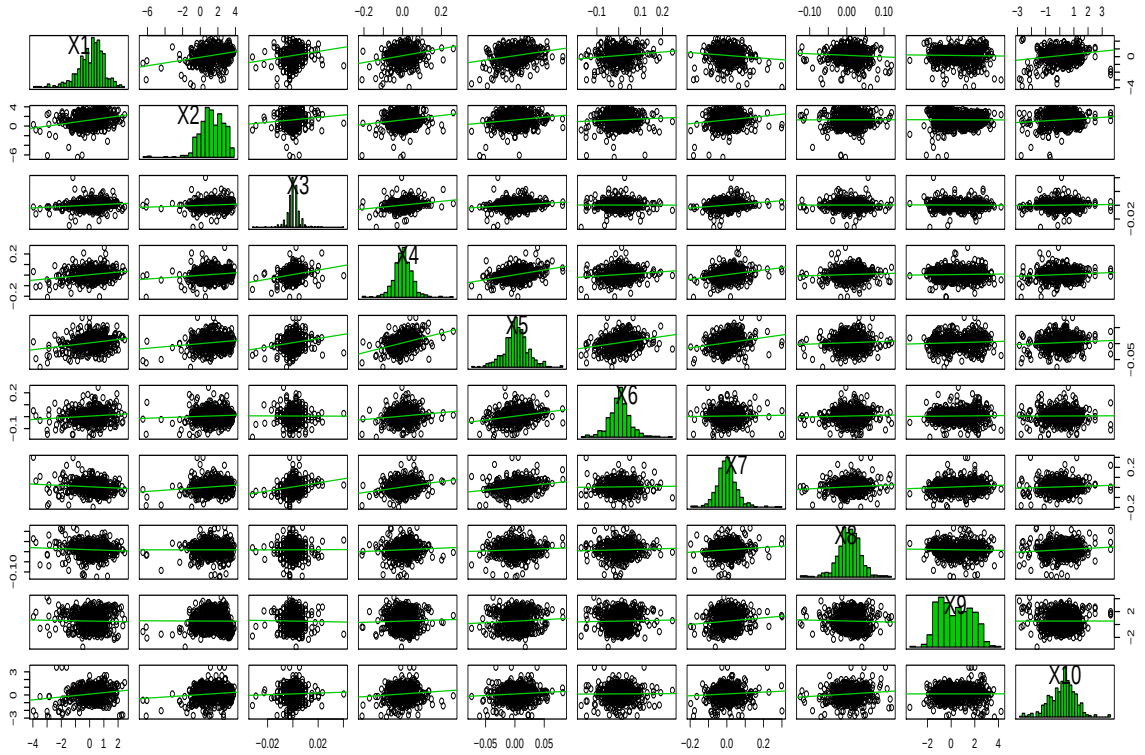
tributed as depicted by the histograms along the diagonal, especially during recessions, (ii) the dependence pattern between pairs of indicators for each regime seems to be quite different, and (iii) the correlation between a given pair seems to be different across regimes. Fortunately, our combination scheme allows for all of these features observed in the data. For each marginal distribution, we do not impose normality assumption. Instead, it is estimated by non-parametric method as in (1). In a given regime, the PCC construction facilitates modeling diverse dependence structures across pairs, as argued in Section 2. Moreover, we estimate all unknown quantities for recession and expansion separately to accommodate (iii).

As for the ordering of predictors in T_1 of the D-Vine tree, we take the strategy of Aas *et al.* (2009). The principle is that the more dependent the two predictors are, the closer they should be placed in T_1. That is, more emphasis is put on modeling stronger pair dependence, and all of other dependence structures are implied from the selected PCC. We did try other possibilities, and the results are quite robust. Having determined the ordering of predictors in T_1, the entire tree is constructed. The next step is to choose an appropriate bivariate copula family associated with each edge. This can be handled in the R system with the aid of **CDVine** package. Brechmann and Schepsmeier (2013) provided a comprehensive introduction to various computational facilities in this package. There are 41 built-in bivariate copula families available to use, encompassing virtually all of the usual copulas that capture a wide variety of dependence patterns. Among them, the one that fits data best in terms of *Bayesian Information Criterion* is chosen for each pair in the vine structure so that the built 10-dimensional copula function stands as a close approximation to the true dependence pattern observed in Figure 2. The complete D-Vine trees when $Z = 0$ (expansion) and when $Z = 1$ (recession) are listed in Table 1. Note that there are 90 pair-copulas to be specified, but only 11 of them are Gaussian, indicating the prevalence of asymmetric tail dependence in the data. Recently, Schepsmeier (2016) developed a goodness-of-fit test for regular vine copula models based on the well-known information matrix equality. With 200 bootstraps, the calculated p-value for the chosen vine copula is 0.26 (0.4) for expanionary (recessionary) periods, indicating that our model fits data pretty well.[7]

[7]These p-values should be interpreted with caution. As Schepsmeier (2016) pointed out, the goodness-of-fit test merely works when the data is i.i.d., which is apparantly not the case in the current circumstance. To the

Figure 2: Scatter plot matrices of the 10 predictors in two regimes

(a) Expansion



(b) Recession

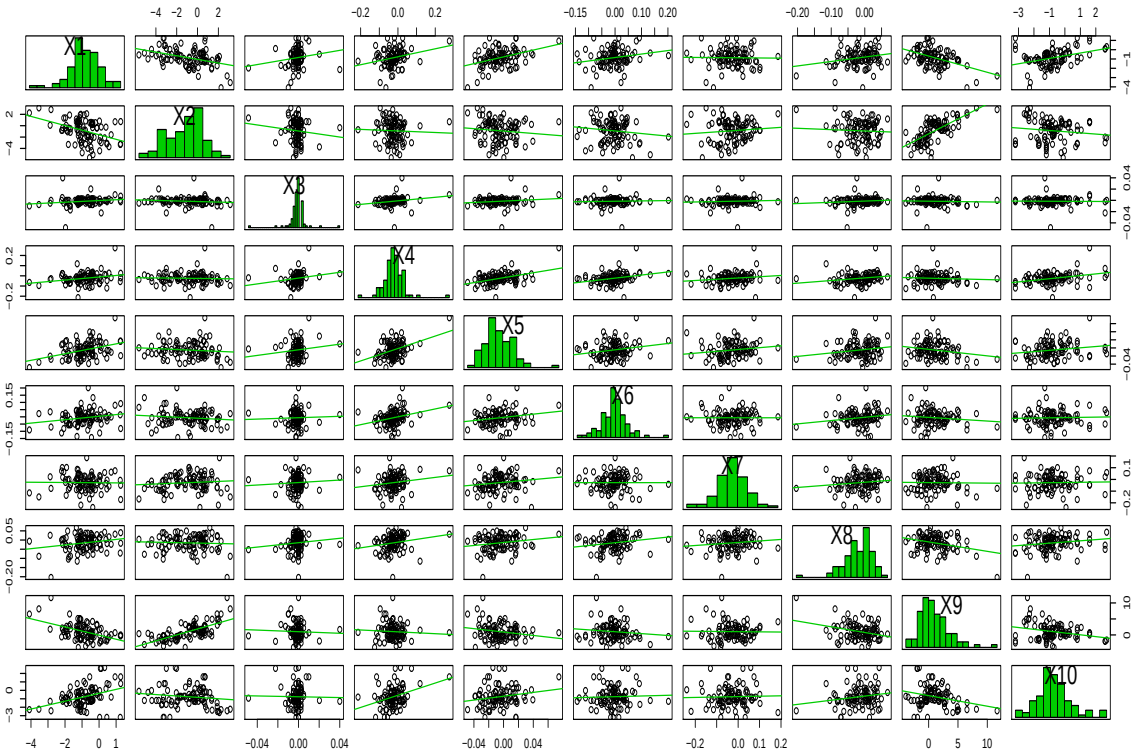


Table 1: 10−dimensional D-Vine trees and parameter estimates

	expansion		recession	
pair	family	parameter	family	parameter
$c_{1,2}$	Gaussian	0.284	270-rotated Clayton	-0.685
$c_{2,3}$	Gumbel	1.087	90-rotated Clayton	-0.247
$c_{3,4}$	Survival Gumbel	1.172	Frank	2.616
$c_{4,5}$	Gaussian	0.386	Gumbel	1.269
$c_{5,6}$	Survival Gumbel	1.198	Gaussian	0.221
$c_{6,7}$	t	0.038(5.828)	Frank	0.624
$c_{7,8}$	Survival Clayton	0.111	Frank	1.036
$c_{8,9}$	90-rotated Clayton	-0.059	90-rotated Joe	-1.291
$c_{9,10}$	Survival Clayton	0.020	270-rotated Joe	-1.355
$c_{1,3\|2}$	Frank	1.134	Gumbel	1.167
$c_{2,4\|3}$	Survival Gumbel	1.083	270-rotated Clayton	-0.083
$c_{3,5\|4}$	Gaussian	0.121	Frank	0.417
$c_{4,6\|5}$	Frank	0.419	Frank	1.876
$c_{5,7\|6}$	Frank	1.581	Frank	1.516
$c_{6,8\|7}$	Survival Clayton	0.037	Survival Clayton	0.321
$c_{7,9\|8}$	Gaussian	0.109	Clayton	0.130
$c_{8,10\|9}$	Gumbel	1.077	Clayton	0.136
$c_{1,4\|2,3}$	Survival Gumbel	1.112	Survival Clayton	0.199
$c_{2,5\|3,4}$	Clayton	0.138	90-rotated Clayton	-0.319
$c_{3,6\|4,5}$	270-rotated Gumbel	-1.062	Frank	-0.607
$c_{4,7\|5,6}$	Frank	1.124	Gumbel	1.140
$c_{5,8\|6,7}$	Survival Joe	1.036	Survival Clayton	0.146
$c_{6,9\|7,8}$	Survival Clayton	0.079	270-rotated Joe	-1.105
$c_{7,10\|8,9}$	Survival Joe	1.104	90-rotated Clayton	-0.085
$c_{1,5\|2-4}$	Survival Gumbel	1.079	Survival Gumbel	1.204
$c_{2,6\|3-5}$	Survival Joe	1.051	90-rotated Joe	-1.197
$c_{3,7\|4-6}$	Gaussian	0.134	Survival Joe	1.057
$c_{4,8\|5-7}$	Frank	0.332	Frank	1.292
$c_{5,9\|6-8}$	Survival Clayton	0.105	90-rotated Joe	-1.301
$c_{6,10\|7-9}$	Survival Joe	1.040	90-rotated Clayton	-0.018
$c_{1,6\|2-5}$	t	0.065(5.511)	Survival Joe	1.137
$c_{2,7\|3-6}$	Survival Joe	1.116	Gumbel	1.081
$c_{3,8\|4-7}$	90-rotated Gumbel	-1.024	Gaussian	0.032
$c_{4,9\|5-8}$	Gaussian	-0.016	Frank	-0.114
$c_{5,10\|6-9}$	Clayton	0.088	Gumbel	1.043
$c_{1,7\|2-6}$	90-rotated Gumbel	-1.145	270-rotated Clayton	-0.205
$c_{2,8\|3-7}$	Frank	-0.387	Joe	1.117
$c_{3,9\|4-8}$	90-rotated Clayton	-0.095	270-rotated Clayton	-0.091
$c_{4,10\|5-9}$	Clayton	0.098	Gumbel	1.204
$c_{1,8\|2-7}$	90-rotated Joe	-1.094	Frank	0.879
$c_{2,9\|3-8}$	270-rotated Joe	-1.199	Frank	5.685
$c_{3,10\|4-9}$	270-rotated Joe	-1.015	Gaussian	-0.026
$c_{1,9\|2-8}$	t	0.039(6.740)	Gaussian	-0.115
$c_{2,10\|3-9}$	Gaussian	0.026	Survival Clayton	0.108
$c_{1,10\|2-9}$	t	0.122(4.415)	Survival Clayton	0.294

Notes: $c_{i,j|k-q} \equiv c_{i,j|k,k+1,k+2,\ldots,q}$ for $q > k$. For t copula, the first parameter is the correlation coefficient, while the second one in the parenthesis is the degree of freedom. The details of all other copulas are presented in Brechmann and Schepsmeier (2013).

Figure 3 depicts the empirical ROC curve produced by our optimal procedure based on the D-Vine trees in Table 1. For the sake of comparison, we also plot the empirical ROC curve of the current LEI, as well as those generated by multivariate Gaussian copula, t copula, dynamic factor model[8] and the ten individual predictors. As expected, the linear composite index LEI, as a combination scheme, outperforms the best single predictor (the interest spread) by exploiting the useful information contained in other predictors. However, it fails to use this additional information in an efficient manner, and the amount of efficiency loss can be represented by the non-trivial gap between the ROC curve of the current LEI with that of the optimal scheme. Even a simple combination rule based on multivariate Gaussian or t copula is able to make a substantial improvement over the current LEI. In this particular example, the ROC curves generated by multivariate Gaussian and t copulas approximate the optimal one fairly well and the differences amongst them are visually indistinguishable.[9] However, this might not hold in general, particularly when the baseline ROC is not very discriminatory such that here is scope for further improvement. With this idea in mind, we conducted another experiment where the worst three predictors in Figure 3 (Invest, Credit and Hours) are combined to predict the economic recession in the next 9 months. In this example, the optimal ROC curve based on vine copula is considerably higher than those based on Gaussian and t copulas,[10] indicating that accounting for non-symmetric dependence structure does matter. Surprisingly, the dynamic factor model ranks lowest among all of the combination schemes in terms of the ROC curve. Probably, this poor performance might be caused by the inherent ignorance of the target when the model is fitted.

To summarize our findings in a single statistic, Table 2 reports the AUC for the curves in Figure 3, together with their bootstrap 95% confidence intervals. The number of replicates in the circular bootstrap is 200. In the last row of Table 2, we report Impv, (the improvement of the optimal procedure over the current TCB LEI), which is 9.2%, and significant at 5% level as its confidence interval excludes zero. Given that the AUC for LEI is already high (0.884),

best of our knowledge, no goodness-of-fit test robust to serial correlation exists in the literature of vine copula.

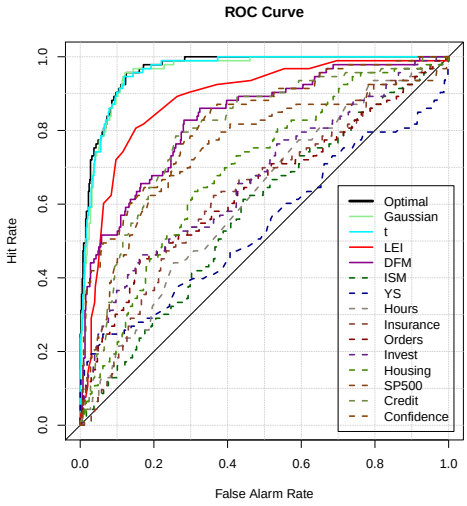
[8]Following Stock and Watson (2010), we assume the comovement of all predictors is driven by an underlying factor, which follows AR(2) process. This model is fitted by maximum likelihood method.

[9]For this reason, we focus on our optimal scheme in the subsequent analysis. The results with Gaussian or t copula are quite similar and thus are omitted to save space.

[10]The areas under ROC curve are 0.731(Optimal), 0.632(Gaussian) and 0.651(t) respectively.

this improvement is remarkable.

Figure 3: In-sample ROC curves



To appreciate the economic benefit from our scheme, we consider the choice of a forecaster facing two forecasting systems, namely, the optimal procedure vs. the current LEI. Suppose the utility function of the forecaster is described by a weighted average of the hit rate and false alarm rate as $U(m,w) = mH(w) + (1-m)(1-F(w))$, where $m \in [0,1]$ is the weight attached to the hit rate. Given m, the forecaster would choose w to maximize $U(m,w)$. Let $w^*(m) \equiv argmax_w U(m,w)$ and $U^*(m) \equiv U(m,w^*(m))$ be the optimal solution and the maximized utility function. Of particular interest is $U^*(0.5)$, which represents the maximized utility when H and $1-F$ are treated as equally important. In the biostatistics literature, $2U^*(0.5)-1$ is also called "Youden's index" (cf. Youden (1950)).[11] Table 3 presents the Youden's index for each curves in Figure 3, along with the corresponding threshold value $w^*(0.5)$. Both Table 2 and Table 3 are qualitatively similar, where our optimal procedure always performs best. In terms of Youden's index, the improvement over the TCB's LEI is as high as 30.4%, considerably higher than that in terms of AUC.

In general, the forecasting system yielding the higher $U^*(m)$ is preferred for given m. For another forecaster with different m, $U^*(m)$ corresponding to the two systems would be different as well. Figure 4 plots $U^*(m)$ for the two systems as functions of m, as well as the improvement of the optimal scheme over the current LEI in terms of the utility. Both

[11] $2U^*(0.5)-1$ is represented graphically as the height of the ROC curve above the diagonal. See Schisterman *et al.* (2009).

Table 2: The area under the ROC curve

Object	estimate	L	U
AUC(Optimal)	0.965	0.964	0.996
AUC(LEI)	0.884	0.811	0.940
AUC(ISM)	0.795	0.678	0.901
AUC(YS)	0.823	0.693	0.933
AUC(Hours)	0.617	0.575	0.647
AUC(Insurance)	0.690	0.642	0.746
AUC(Orders)	0.631	0.570	0.681
AUC(Invest)	0.577	0.531	0.619
AUC(Housing)	0.641	0.568	0.722
AUC(SP500)	0.668	0.588	0.719
AUC(Credit)	0.542	0.301	0.744
AUC(Confidence)	0.750	0.653	0.846
Impv	0.092	0.051	0.204

Notes: The top of the table contains the AUC value computed for each curve in Figure 3. Impv=(AUC(Optimal)-AUC(LEI))/AUC(LEI). Column "L" is the lower bound of the interval, while column "U" is the upper bound.

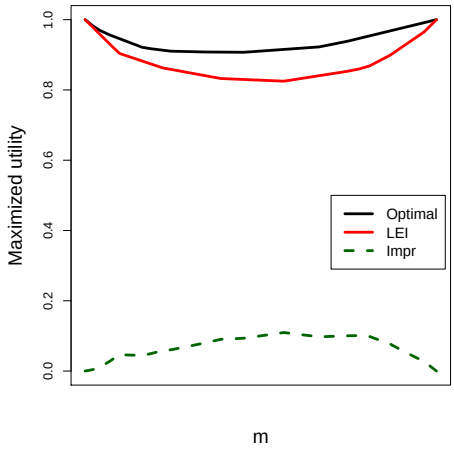
Table 3: Youden's index

Object	w	$H(w)$	$F(w)$	Youden's index
Optimal	0.108	0.925	0.084	0.841
LEI	0.130	0.817	0.172	0.645
ISM	0.151	0.785	0.265	0.520
YS	0.107	0.860	0.358	0.502
Hours	0.159	0.441	0.263	0.178
Insurance	0.157	0.634	0.306	0.328
Orders	0.151	0.624	0.364	0.260
Invest	0.147	0.613	0.444	0.169
Housing	0.188	0.452	0.164	0.288
SP500	0.186	0.462	0.162	0.300
Credit	0.194	0.247	0.082	0.165
Confidence	0.159	0.677	0.259	0.418
Impv				0.304

Notes: For each individual predictor, a logit model is fitted to generate the predictive probabilities of recession, which are used to plot its ROC curve. Impv=(Youden's index(Optimal)-Youden's index(LEI))/Youden's index(LEI).

forecasting systems are akin to each other when m is around zero or one. That is, when either $H(w)$ or $F(w)$ matters (but not both), the forecaster is indifferent between two systems. Our method performs best relative to the LEI when $m = 0.565$. The gain in utility by using our method is about 11%. Since a great amount of loss is associated with recession, this improvement can be economically significant.

Figure 4: The maximized utility as a function of the weight m



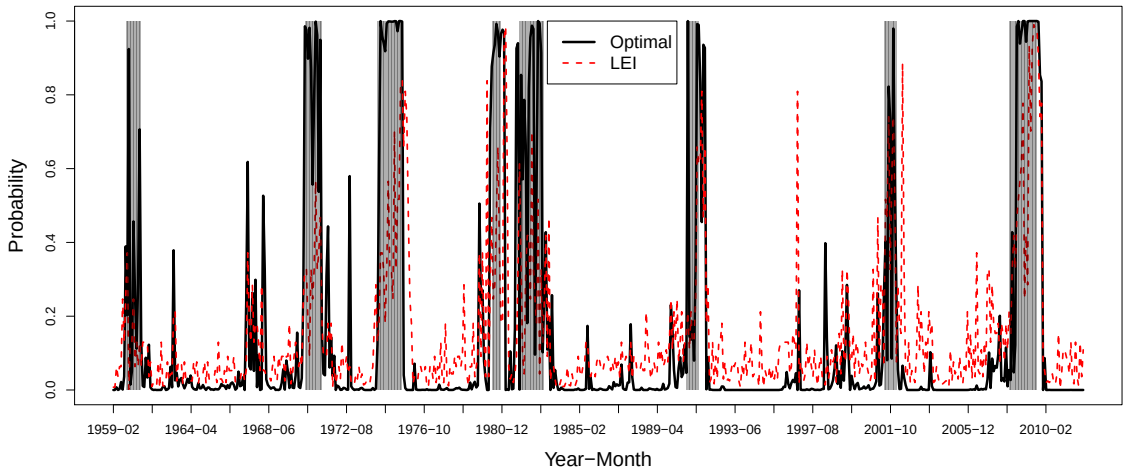
It follows from (12) that the likelihood ratio $f(X|Z=1)/f(X|Z=0)$ can be translated into a predictive probability of economic recession given the 10 predictors. Similarly, we can use a univariate logit model to regress Z on the LEI, and compute the fitted probabilities. Figure 5 displays the behavior of the probabilistic series over time, along with the recessionary phases identified by NBER. It is evident that every recession since 1959 is accompanied by a higher-than-usual forecast probability for both series.[12] The current TCB LEI seems to be slightly more conservative to give high probabilities during recessions. During non-recessionary periods, we observe that, prior to 1985, the optimal approach tends to generate a number of relatively high probabilities not accompanied by recessions. After 1985, its predictive probabilities always fluctuated between 0 and 0.2, with 1997 as an exception when the probabilities were as high as 0.4.[13] Therefore, compared with the currently used leading index, our procedure is more capable of identifying the incoming economic recessions throughout the sample period. Since 1985, it has also become more reliable to discriminate

[12]The average lead time for our optimal index is 3.25 and 3 months for peak and trough of business cycle respectively.

[13]The predictive probabilities of the LEI are even higher during 1997.

between the two regimes. Admittedly, for a true comparison we need to generate the probabilities recursively using real time data. The main purpose of this section, however, was to show that our algorithm works.

Figure 5: In-sample probability series



Notes: The shaded bars mark the months of economic recessions defined by NBER.

4 Concluding remarks

This paper makes a multivariate extension of the non-linear forecast combination scheme developed by Lahiri and Yang (2015b). We provide a convenient, computationally affordable method to combine multiple predictors. Unlike many previously proposed alternatives, the resulting combined forecasts are optimal to maximize the discriminatory power as measured by the ROC curve. In order to implement this scheme, one has to estimate the conditional distributions of predictors given the binary target. When the number of predictors are larger than two, vine-copula facilitates the modeling of the marginal distribution of each predictor and their dependence structure separately. A two-step procedure is advocated in practice. In the first step, we estimate the density of each predictor by the standard non-parametric method. The copula dependence parameters are estimated sequentially in the second step by plugging in the first-step estimates. To account for the sampling variability, a circular block

bootstrap is used to construct the confidence interval of the area under the ROC curve. The relevance of our approach is demonstrated in an empirical illustration, where we combine the 10 predictors of the Leading Economic Index of The Conference Board to construct a new leading index for U.S. business cycles. Compared with the currently used LEI, the predictive probabilities implied by our model tend to distinguish the underlying economic states significantly better.

Although a D-Vine structure with an appropriate ordering is used in our application, we do not claim this particular specification to be universally applicable. Other structures, such as a C-Vine, may also provide satisfactory fits. With many predictors at hand, numerous vine structures are available to choose from. Given a sample, specifying a different structure may lead to different estimates, which in turn may imply different relationships among predictors. To avoid this arbitrariness in specification, a model selection criterion or test is necessary. Despite a large number of papers dealing with copula specification, to the best of our knowledge, there is little prior research providing guidance to choose the best structure in the context of vine copulas. It is unclear if the misspecification of the vine structure and pairwise copulas would result in a sizeable bias for the estimated ROC curve. Furthermore, this paper is primarily concerned with the in-sample comparison between competing combination schemes to serve as an illustration of our methodology on non-linear forecast combination. In particular, in-sample predictive superiority may not translate into a similar out-of-sample performance. We intend to leave these interesting questions for future investigation as a real-time vintage dataset becomes available.

References

Aas, K., Czado, C., Frigessi, A. and Bakken, H. (2009), 'Pair-Copula Constructions of Multiple Dependence', *Insurance: Mathematics and Economics* **44**, 182–198.

Acar, E. F., Craiu, R. V. and Yao, F. (2011), 'Dependence Calibration in Conditional Copulas: A Nonparametric Approach', *Biometrics* **67**, 445–453.

Acar, E. F., Genest, C. and Nešlehová, J. (2012), 'Beyond Simplified Pair-Copula Constructions', *Journal of Multivariate Analysis* **110**, 74–90.

Anatolyev, S. (2009), 'Multi-Market Direction-of-Change Modeling Using Dependence Ratios', *Studies in Nonlinear Dynamics & Econometrics* **13**, Article 5.

Bedford, T. and Cooke, R. M. (2001), 'Probability Density Decomposition for Conditionally Dependent Random Variables Modeled by Vines', *Annals of Mathematics and Artificial intelligence* **32**, 245–268.

Bedford, T. and Cooke, R. M. (2002), 'Vines—A New Graphical Model for Dependent Random Variables', *Annals of Statistics* **30**, 1031–1068.

Berg, D. and Aas, K. (2009), 'Models for Construction of Higher-Dimensional Dependence: A Comparison Study', *European Journal of Finance* **15**, 639–659.

Berge, T. J. and Jordà, Ò. (2011), 'Evaluating the Classification of Economic Activity into Recessions and Expansions', *American Economic Journal: Macroeconomics* **3**, 246–277.

Brechmann, E. C. and Czado, C. (2013), 'Risk Management with High-Dimensional Vine Copulas: An Analysis of the Euro Stoxx 50', *Statistics & Risk Modeling* **30**, 307–342.

Brechmann, E. C. and Schepsmeier, U. (2013), 'Modeling Dependence with C- and D-Vine Copulas: The R Package CDVine', *Journal of Statistical Software* **52**, 1–27.

Brown, B. M. and Wang, Y. G. (2005), 'Standard Errors and Covariance Matrices for Smoothed Rank Estimators', *Biometrika* **92**, 149–158.

Burns, A. F. and Mitchell, W. C. (1946), 'Measuring Business Cycles', New York: National Bureau of Economic Research, New York.

Chen, X. and Fan, Y. (2006), 'Estimation and Model Selection of Semiparametric Copula-Based Multivariate Dynamic Models under Copula Misspecification', *Journal of Econometrics* **135**, 125–154.

Chen, X. and Fan, Y. (2007), 'A Model Selection Test for Bivariate Failure-Time Data', *Econometric Theory* **23**, 414–439.

Czado, C. (2010), Pair-Copula Constructions of Multivariate Copulas, *in* P. Jaworski, F. Durante, W. K. Härdle and T. Rychlik, eds, 'Copula Theory and its Applications', Springer.

Demarta, S. and McNeil, A. J. (2005), 'The t Copula and Related Copulas', *International Statistical Review* **73**, 111–129.

Dissmann, J., Brechmann, E. C., Czado, C. and Kurowicka, D. (2013), 'Selecting and Estimating Regular Vine Copulae and Application to Financial Returns', *Computational Statistics and Data Analysis* **59**, 52–69.

Drehmann, M. and Juselius, M. (2014), 'Evaluating Early Warning Indicators of Banking Crises: Satisfying Policy Requirements', *International Journal of Forecasting* **30**, 759–780.

Fawcett, T. (2006), 'An Introduction to ROC Analysis', *Pattern Recognition Letters* **27**, 861–874.

Fermanian, J. (2005), 'Goodness-Of-Fit Tests for Copulas', *Journal of Multivariate Analysis* **95**, 119–152.

Genest, C., Rémillard, B. and Beaudoin, D. (2009), 'Goodness-Of-Fit Tests for Copulas: A Review and a Power Study', *Insurance: Mathematics and Economics* **44**, 199–213.

Han, A. K. (1987), 'Non-parametric Analysis of a Generalized Regression Model', *Journal of Econometrics* **35**, 303–316.

Harding, D. and Pagan, A. (2011), 'An Econometric Analysis of Some Models for Constructed Binary Time Series', *Journal of Business & Economic Statistics* **29**, 86–95.

Hastie, T., Tibshirani, R. and Friedman, J. (2001), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer.

Jafarzadeh, S. R., Johnson, W. O. and Gardner, I. A. (2016), 'Bayesian Modeling and Inference for Diagnostic Accuracy and Probability of Disease Based on Multiple Diagnostic Biomarkers with and without a Perfect Reference Standard', *Statistics in Medicine* **35**, 859–876.

Joe, H. (1996), Families of m-Variate Distributions with Given Margins and m(m-1)/2 Bivariate Dependence Parameters, *in* L. Rüschendorf, B. Schweizer and M. D. Taylor, eds, 'Distributions with Fixed Marginals and Related Topics'.

Joe, H. (1997), *Multivariate Models and Dependence Concepts*, Chapman & Hall.

Jordà, Ò. and Taylor, A. M. (2012), 'The Carry Trade and Fundamentals: Nothing to Fear but FEER Itself', *Journal of International Economics* **88**, 74–90.

Klugman, S. A. and Parsa, R. (1999), 'Fitting Bivariate Loss Distributions with Copulas', *Insurance: Mathematics and Economics* **24**, 139–148.

Kurowicka, D. and Cooke, R. M. (2004), Distribution-Free Continuous Bayesian Belief Nets, *in* Fourth International Conference on Mathematical Methodsin Reliability Methodology and Practice. Santa Fe, New Mexico.

Kurowicka, D. and Joe, H. (2011), *Dependence Modeling—Handbook on Vine Copulae*, Singapore: World Scientic Publishing Co..

Lahiri, S. N. (2003), *Resampling Methods for Dependent Data*, Springer.

Lahiri, K. and Wang, J. G. (2013), 'Evaluating Probability Forecasts for GDP Declines using Alternative Methodologies', *International Journal of Forecasting* **29**, 175–190.

Lahiri, K. and Yang, L. (2013), Forecasting Binary Outcomes, *in* A. Timmermann and G. Elliott, eds, 'Handbook of Economic Forecasting Volume 2B', North-Holland Amsterdam, pp. 1025–1106.

Lahiri, K. and Yang, L. (2015a), 'Confidence Bands for ROC Curves with Serially Dependent Data', *Journal of Business & Economic Statistics* forthcoming.

Lahiri, K. and Yang, L. (2015b), 'A Nonlinear Forecast Combination Procedure for Binary Outcomes', *Studies in Nonlinear Dynamics and Econometrics* **29**, 175–190.

Levanon, G., Manini, J., Ozyildirim, A., Schaitkin, B. and Tanchua, J. (2014), 'Using Financial Indicators to Predict Turning Points in the Business Cycle: The Case of the Leading Economic Index for the United States', *International Journal of Forecasting* forthcoming.

Li, Q. and Racine, J. S. (2008), *Nonparametric Econometrics: Theory and Practice*, Princeton University Press.

Lin, H., Zhou, L., Peng, H. and Zhou, X. (2011), 'Selection and Combination of Biomarkers Using ROC Method for Disease Classification and Prediction', *The Canadian Journal of Statistics* **39**, 324–343.

Liu, W. and Moench, E. (2014), What Predicts U.S. Recessions?. Federal Reserve Bank of New York Staff Reports No. 691.

McIntosh, M. W. and Pepe, M. S. (2002), 'Combining Several Screening Tests: Optimality of the Risk Score', *Biometrics* **58**, 657–664.

Min, A. and Czado, C. (2010), 'Bayesian Inference for Multivariate Copulas Using Pair-Copula Constructions', *Journal of Financial Econometrics* **8**, 511-546.

Nelsen, R. B. (2006), *An Introduction to Copulas*, Springer.

Patton, A. J. (2006), 'Modelling Asymmetric Exchange Rate Dependence', *International Economic Review* **47**, 527–556.

Patton, A. J. (2012), 'A Review of Copula Models for Economic Time Series', *Journal of Multivariate Analysis* **110**, 4–18.

Patton, A. J. (2013), Copula Methods for Forecasting Multivariate Time Series, *in* A. Timmermann and G. Elliott, eds, 'Handbook of Economic Forecasting Volume 2B', North-Holland Amsterdam, pp. 899–960.

Patton, A. J. and Fan, Y. (2014), 'Copulas in Econometrics', *Annual Review of Economics* **6**, 179–200.

Pepe, M. S. (2000), 'Receiver Operating Characteristic Methodology', *Journal of the American Statistical Association* **95**, 308–311.

Pepe, M. S., Cai, T. and Longton, G. (2006), 'Combining Predictors for Classification Using the Area under the Receiver Operating Characteristic Curve', *Biometrics* **62**, 221–229.

Schepsmeier, U. (2016), 'A Goodness-Of-Fit Test for Regular Vine Copula Models', *Econometric Reviews*, forthcoming.

Schisterman, E. F., Perkins, N. J., Liu, A. and Bondell, H. (2005), 'Optimal Cut-Point and Its Corresponding Youden Index to Discriminate Individuals Using Pooled Blood Samples', *Epidemiology* **16**, 73–81.

Scotti, C. (2011), 'A Bivariate Model of Federal Reserve and ECB Main Policy Rates', *International Journal of Central Banking* **7**, 37–78.

Sklar, A. (1973), 'Random Variables, Joint Distributions, and Copulas', *Kybernetica* **9**, 449–460.

Stock, J. H. and Watson, M. W. (2010), Dynamic Factor Models, *in* M. P. Clements, and D. F. Hendry, eds, 'Oxford Handbook of Economic Forecasting', Oxford: Oxford University Press, pp. 35–60.

Swets, J. A., Dawes, R. M. and Monahan, J. (2000), 'Better Decisions through Science', *Scientific American* **283**, 82–87.

Trivedi, P. K. and Zimmer, D. M. (2005), 'Copula Modeling: An Introduction for Practitioners', *Foundations and Trends in Econometrics* **1**, 1–111.

Youden, W. J. (1950), 'Index for Rating Diagnostic Tests', *Cancer* **3**, 32–35.

Zhou, X. H., Obuchowski, N. A. and McClish, D. K. (2002), *Statistical Methods in Diagnostic Medicine*, John Wiley & Sons.

Zimmer, D. M. (2014), 'Analyzing Comovements in Housing Prices Using Vine Copulas', *Economic Inquiry*, forthcoming.