A Non-linear Forecast Combination Procedure for Binary Outcomes

Kajal Lahiri*and Liu Yang
†Department of Economics
University at Albany, SUNY
NY 12222, USA

Abstract

We develop a non-linear forecast combination rule based on copulas that incorporate the dynamic interaction between individual predictors. This approach is optimal in the sense that the resulting combined forecast produces the highest discriminatory power as measured by the receiver operating characteristic (ROC) curve. Under additional assumptions, this rule is shown to be equivalent to the quintessential linear combination scheme. To illustrate its usefulness, we apply this methodology to optimally aggregate two currently used leading indicators—the ISM new order diffusion index and the yield curve spread—to predict economic recessions in the United States. We also examine the sources of forecasting gains using a counterfactual experimental set up.

JEL Classifications: C11, C15, C38, C53, E37

Key words: Receiver operating characteristic curve, Copula, Bayesian methods, Markov chain Monte Carlo, Yield spread, ISM diffusion index.

*Corresponding author. Tel.: +1 518 442 4758. E-mail address: klahiri@albany.edu.

†The authors are grateful to James Ramsey and other participants at the 22nd SNDE Symposium in New York city for making valuable comments and suggestions

1 Introduction

In economic forecasting, it is not uncommon to have multiple predictors for the single target variable of interest. Each predictor may contain independent information pertinent to the target that others do not have. Instead of focusing on the best predictor, one can achieve diversification gains from combining all of them in an appropriate manner. Bates and Granger (1969) have suggested a linear scheme to combine a set of forecasts with data-driven weights. Granger and Ramanathan (1984) showed that this linear combination method is numerically equivalent to a linear regression of the target variable on the forecasts. Timmermann (2006) provided a comprehensive overview of various combination methods.

This article proposes a non-linear combination procedure to predict a binary outcome variable. Binary events, such as loan defaults, occurrence of recessions, passage of a specific legislation, etc., are frequently involved in numerous economic decisions. The conventional combination methodologies are known to work well for a continuous target variable, such as GDP growth and inflation rates. However, it is possible that they may fail to capture some important features of a binary target variable. Lahiri and Yang (2013b) highlighted the uniqueness of a binary event and summarized a large body of literature to forecast this type of target variables by considering their distinct features.

To quantify the forecast performance so that competing models can be compared, an appropriate criterion is necessary. Due to the very nature of a binary event, the joint distribution between forecasts and actuals are of special forms, which can be utilized to design a wide array of tools for forecast evaluation. In practice, the Brier score (Brier (1950)), or the mean squared error between forecasts and actuals, is by the far the most widely employed. Though useful by itself, this score is influenced by the marginal information regarding the actuals, and as a result is often viewed as a poor metric for measuring performance in many cases. For this reason, this paper uses the receiver operating characteristic (ROC) curve, which is especially designed to assess the efficacy of a forecasting system with a binary target. An attractive property of ROC curves is their insensitivity to changes in the event distribution (Fawcett (2006)). Furthermore, it completely describes the conditional distribution of the

forecast given the actual. This graphic device was originally proposed in the signal detection theory during the 1950s, and has gained increasing popularity in applications, ranging from meteorology to medical sciences, and psychology, among other fields. The readers are referred to Krzanowski and Hand (2009), Pepe (2003), Swets *et al.* (2000), and Zhou *et al.* (2002) for a general introduction to this methodology. Recently, interest in ROC curves has grown among econometricians. Several economic and financial applications of ROC can be found in Berge and Jordà (2011), Drehmann and Juselius (2012), Lahiri and Wang (2013), and Lahiri and Yang (2013a).

Specifically, we separately model the marginal distributions of individual predictors and their dependence structure given each materialized regime of the target. The ratio of the joint densities is taken as the combination rule, which not only augments the predictive accuracy of each single predictor but is optimal within a family of all combination rules in the sense of maximizing the discriminatory power as measured by the ROC curve. The implied ROC curve of the combined forecasts has no known analytic form, a fact that makes the statistical inference inconvenient to be conducted in the standard parametric framework. To address this problem, we will develop a Bayesian variant of this model equipped with a non-informative prior. One of the appealing merits of the proposed procedure is that it allows for any plausible margin and dependence pattern in the predictors. It also nests the linear benchmark as a special case under some additional distributional assumptions. We demonstrate the effectiveness of our approach with an empirical illustration to predict the economic recessions in the United States based on two currently used leading indicators: ISM new order diffusion index and the yield curve spread.

This paper contributes to the literature on forecasting combination in three aspects. First, we use ROC, in lieu of the usual mean squared error, as the measure of predictive performance since the latter tends to reward hedging behavior of a forecaster when it is used to evaluate uncommon event probabilities (Stephenson (2000)). Second, an optimal decision rule to combine multiple predictors to maximize the discrimination capacity is formulated within a Bayesian framework. An algorithm is also given to make the rule practically operational. In addition, a counterfactual exercise is undertaken to identify several important determinants of a better forecast, which in turn suggests possible paths to follow to enhance the reliability

of the combined predictor.

The rest of the paper is organized as follows. In Section 2, the Bayesian parametric model is introduced and the optimal decision rule is derived theoretically. The empirical application and the related computational issues are presented in Section 3, which is followed by a counterfactual experiment in Section 4. Section 5 closes this paper with further remarks.

2 A non-linear combination procedure

We develop a parsimonious, yet flexible, Bayesian parametric model in this section to combine the information inherent in two predictors for a binary event. Throughout this paper, Z_t, X_{1t}, and X_{2t} denote the binary target variable, the first and the second predictor in period t, respectively. We use upper case letters to denote cumulative distribution functions and corresponding lower case letters to denote the density functions.

Our Bayesian combination model consists of two elements: a conditional likelihood and a prior. We denote the joint conditional distribution of (X_{1t}, X_{2t}) given Z_t by $H(x_1, x_2|Z_t)$, which can be further decomposed into three parts: the marginal distribution of X_{1t} given Z_t (denoted by $F(x_1|Z_t)$), the marginal distribution of X_{2t} given Z_t (denoted by $G(x_2|Z_t)$), and the dependence structure between X_{1t} and X_{2t} given Z_t. The latter is characterized by the copula associated with $H(x_1, x_2|Z_t)$. By Sklar's theorem, for each $j \in \{0,1\}$, there exists a unique copula C_j such that $H(x_1, x_2|Z_t = j) = C_j(F(x_1|Z_t = j), G(x_2|Z_t = j))$ for all $(x_1, x_2) \in R^2$ when both X_{1t} and X_{2t} are continuous random variables. Assuming C_j is twice differentiable, the corresponding copula density is $c_j(y_1, y_2) \equiv \partial^2 / \partial y_1 \partial y_2 C_j(y_1, y_2)$. In general, copulas are bivariate distribution functions whose one-dimensional marginals are uniform on the interval $(0,1)$; they enable us to construct large families of joint distributions. The popularity of copulas in empirical macroeconomics and finance owes much to their flexibility by freeing the analyst from considering only existing bivariate distributions. The stream of literature on a general introduction to the modeling strategies based on copulas includes Nelsen (2006), Patton (2012), and Trivedi and Zimmer (2005). Anatolyev(2009), Patton(2006) and Scotti(2011) applied this methodology to predict multiple economic events. Patton (2013) provided a re-

cent survey on copula methods to forecasting multivariate time series.

Let $f(x_1;\alpha_1)$ and $f(x_1;\alpha_0)$ be the parametrized conditional densities of X_{1t} given $Z_t = 1$ and $Z_t = 0$ respectively. Similar notations are used for the conditional densities of X_{2t}, namely, $g(x_2;\beta_1)$ and $g(x_2;\beta_0)$. The likelihood function $L_T(\theta)$ given a sample $X_T = \{(Z_t, X_{1t}, X_{2t}) : t = 1, ..., T\}$ can be written as

$$L_T(\theta) \equiv \prod_{t=1}^{T}\prod_{j=0}^{1}\{c(F(X_{1t};\alpha_j), G(X_{2t};\beta_j);\gamma_j) f(X_{1t};\alpha_j) g(X_{2t};\beta_j)\}^{I(Z_t=j)},$$

where $c(y_1, y_2;\gamma_j)$ is the parametrized density of the copula, and $I(\cdot)$ is the indicator function which equals one if the condition in the parenthesis is true, and zero otherwise. $\theta \equiv (\alpha_1', \alpha_0', \beta_1', \beta_0', \gamma_1', \gamma_0')'$ is the parameter vector. To complete the Bayesian model, a prior for θ is needed. Given a specific functional form for $L_T(\theta)$, it may be cumbersome, if not impossible, to construct the conjugate prior. In this section, we use a non-informative prior $P(\cdot)$ to express our knowledge on θ before X_T is observed. The posterior $\theta|X_T$ is derived by Bayes' theorem, that is,

$$\theta|X_T = \frac{L_T(\theta)P(\theta)}{\int_\theta L_T(\theta)P(\theta)d\theta}. \tag{1}$$

The main goal of this paper is to compare the accuracy of forecasts based on either X_{1t} or X_{2t} with those made using both. 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 $Z_t = 1$ and $Z_t = 0$. If the forecasts are completely insensitive to the value Z_t would take, they have zero discriminatory power. On the other hand, forecasts which take one value when $Z_t = 1$ and take another when $Z_t = 0$ would obviously possess the highest discriminatory power. Most real-life forecasts lie between these two extremes. Suppose we predict $Z_t = 1$ whenever X_{1t} exceeds a threshold w. We can define two conditional probabilities resulting from this

decision rule, namely,

$$H(w) \equiv P(X_{1t} > w | Z_t = 1) = 1 - F(w; \alpha_1),$$
$$F(w) \equiv P(X_{1t} > w | Z_t = 0) = 1 - F(w; \alpha_0).$$

$H(w)$ is referred to as the hit rate and it is the probability of correct forecast when $Z_t = 1$. $F(w)$ is called the false alarm rate or the probability of false forecast when $Z_t = 0$. 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 tradeoff between them is depicted by plotting the pair $(F(w), H(w))$ in a unit square for every 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 two predictors are available, none of them could be maximally utilized unless the information contained in them is processed in an efficient manner. Suppose we only use X_{1t}. We may consider predicting $Z_t = 1$ when $X_{1t} > w$ or when $\Lambda(X_{1t}) > w$, where $\Lambda(\cdot)$ is a known function. Krzanowski and Hand (2009) showed that the ROC curves generated by the rules $\Lambda(X_{1t}) > w$ and $X_{1t} > w$ will be the same if $\Lambda(\cdot)$ is a strictly increasing function. In addition, if multiple predictors are available, it is natural to combine them for the optimal performance in the sense that the hit rate is maximized for any given false alarm rate. The region C_α defined as

$$\{X_t : \frac{h(X_t|H_1)}{h(X_t|H_0)} > w\}$$

plays a critical role in testing $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$. Here, $X_t = (X_{1t}, X_{2t})$ is the observed data for predictors, $h(X_t|H_j)$ is the likelihood function under H_j for $j = 0, 1$, and w is a constant such that $P(h(X_t|H_1)/h(X_t|H_0) > w|H_0) = \alpha$. Among all tests of H_0 against H_1 with the same size α, Neyman-Pearson lemma states that the power,

defined as $P(h(X_t|H_1)/h(X_t|H_0) > w|H_1)$, achieves its maximum if we reject H_0 when $h(X_t|H_1)/h(X_t|H_0) > w$. Therefore, the likelihood ratio test for 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 $H_0 : Z_t = 0$ against $H_1 : Z_t = 1$. Given the parametric specification, the combined forecast takes the following form,

$$\frac{c(F(X_{1t};\alpha_1), G(X_{2t};\beta_1);\gamma_1) f(X_{1t};\alpha_1) g(X_{2t};\beta_1)}{c(F(X_{1t};\alpha_0), G(X_{2t};\beta_0);\gamma_0) f(X_{1t};\alpha_0) g(X_{2t};\beta_0)}, \tag{2}$$

and we predict $Z_t = 1$ if and only if (2) exceeds w. In this case, the size is $P(h(X_{1t}, X_{2t}|Z_t = 1)/h(X_{1t}, X_{2t}|Z_t = 0) > w|Z_t = 0)$, while the power is $P(h(X_{1t}, X_{2t}|Z_t = 1)/h(X_{1t}, X_{2t}|Z_t = 0) > w|Z_t = 1)$. Therefore, the size corresponds to the false alarm rate $F(w)$, while the power corresponds to the hit rate $H(w)$. Given $F(w)$, $H(w)$ is maximized among all possible combination rules based on (X_{1t}, X_{2t}). The optimality of (2) was also established in McIntosh and Pepe (2002).

The likelihood rule in (2) is in the same spirit as in Graham (1996). His binary combination scheme is also built upon the likelihood of receiving a particular forecast prior to a specific outcome, and the reason of doing so, as argued by him, is that $f(X_t|Z_t)$ contains information about the discrimination capacity of each predictor, as well as the correlation between predictors. However, Graham (1996) only considered combining binary forecasts, that is, each variable in X_t is binary. Though Graham (1996) applied this rule to combine probability forecasts as well, he rounded the probability into a binary format by choosing a threshold, which would lead to information loss. The likelihood ratio rule in (2) is more general in that both continuous and discrete predictors are allowed. In addition, we established the optimality of (2) in terms of maximizing the discriminatory power, while it is not clear if the binary combination scheme in Graham (1996) is optimal in some sense.

It is probably worth mentioning at this point that some popular forecast combination rules

can be shown to be special cases of (2). According to Bayes' theorem,

$$\begin{aligned} P(Z_t=1|X_t) &= \frac{h(X_t|Z_t=1)P(Z_t=1)}{h(X_t|Z_t=0)P(Z_t=0)+h(X_t|Z_t=1)P(Z_t=1)} \\ &= \frac{\frac{h(X_t|Z_t=1)}{h(X_t|Z_t=0)}P(Z_t=1)}{P(Z_t=0)+\frac{h(X_t|Z_t=1)}{h(X_t|Z_t=0)}P(Z_t=1)}. \end{aligned} \tag{3}$$

If $h(X_t|Z_t=1)$ and $h(X_t|Z_t=0)$ are bivariate normal, it follows that the conditional log odds ratio $ln(\frac{P(Z_t=1|X_t)}{P(Z_t=0|X_t)})$ is equal to

$$\begin{aligned} & ln(\frac{P(Z_t=1)}{P(Z_t=0)})+\frac{1}{2}(ln|\Sigma_0|-ln|\Sigma_1|)+\frac{1}{2}((X_t-\mu_0)'\Sigma_0^{-1}(X_t-\mu_0)-(X_t-\mu_1)'\Sigma_1^{-1}(X_t-\mu_1)) \\ = \; & ln(\frac{P(Z_t=1)}{P(Z_t=0)})+\frac{1}{2}(ln|\Sigma_0|-ln|\Sigma_1|)+\frac{1}{2}(X_t'\Sigma_0^{-1}X_t-X_t'\Sigma_1^{-1}X_t-2X_t'\Sigma_0^{-1}\mu_0 \\ & +2X_t'\Sigma_1^{-1}\mu_1+\mu_0'\Sigma_0^{-1}\mu_0-\mu_1'\Sigma_1^{-1}\mu_1) \\ = \; & A_0+A_1'X_t+X_t'A_2X_t, \end{aligned} \tag{4}$$

where μ_j is the mean of $h(X_t|Z_t=j)$, and Σ_j is the variance of $h(X_t|Z_t=j)$. In (4), $A_0 = ln(\frac{P(Z_t=1)}{P(Z_t=0)})+\frac{1}{2}(ln|\Sigma_0|-ln|\Sigma_1|)+\frac{1}{2}(\mu_0'\Sigma_0^{-1}\mu_0-\mu_1'\Sigma_1^{-1}\mu_1)$, $A_1=\Sigma_1^{-1}\mu_1-\Sigma_0^{-1}\mu_0$, and $A_2=\Sigma_0^{-1}-\Sigma_1^{-1}$. As a result,

$$p_t \equiv P(Z_t=1|X_t)=\frac{exp(A_0+A_1'X_t+X_t'A_2X_t)}{1+exp(A_0+A_1'X_t+X_t'A_2X_t)}, \tag{5}$$

which implies that $P(Z_t=1|X_t)$, as a regression function, is consistent with a logit specification with quadratic index $A_0+A_1'X_t+X_t'A_2X_t$. Since (2) is a strictly increasing transformation of (5), both would yield identical ROC curves. Note that $\Sigma_1=\Sigma_0$ implies $A_2=0$, and $P(Z_t=1|X_t)$ reduces to the regular logit specification with linear index $A_0+A_1'X_t$, which is the dichotomous combination rule introduced by Kamstra and Kennedy (1998). This relationship offers insights into effectiveness of the logit regression as a type of non-linear combination scheme. Although Kamstra and Kennedy (1998) used this approach as an alternative means of combining qualitative forecasts, we have shown that it is indeed optimal under certain simplifying assumptions including $\Sigma_1=\Sigma_0$. The implication of (5) is that one can regress Z_t on a constant, X_{1t}, X_{2t}, X_{1t}^2, X_{2t}^2 and $X_{1t}X_{2t}$ using the logit model, get the maximum likelihood esti-

mates for A_0, A_1 and A_2, and take the fitted probability $P(Z_t = 1|X_t)$ as the combined forecast. Following the same approach as in Ramsey (1969), the appropriateness of non-linearity can be assessed by examining the size and statistical significance of estimates of A_2. If the linear index is adequate, the joint test should be insignificant. If the test is significant, the non-linear combination scheme in (5) should be used. However, normality assumption is essential for the equivalence between (2) and (5) to hold.

Another popular option is the linear combination method proposed by Bates and Granger (1969), which often serves as a useful benchmark in practice. Given $P(Z_t = 1|X_{1t})$ and $P(Z_t = 1|X_{2t})$, both of which are perfectly calibrated in the sense of Dawid (1984), the linearly combined forecast can be constructed by taking the weighted average, that is,

$$q_t \equiv \omega P(Z_t = 1|X_{1t}) + (1-\omega)P(Z_t = 1|X_{2t}), \tag{6}$$

where $\omega \in (0,1)$ is a properly selected weight. De Luca and Carfora (2014) provided an example where q_t is used to combine two binary regression models for predicting recessions. It is interesting to know the difference between (5) and (6). Ranjan and Gneiting (2010) proved that (6) lacks calibration even though both $P(Z_t = 1|X_{1t})$ and $P(Z_t = 1|X_{2t})$ are calibrated. By construction, (5) is calibrated. Moreover, for every strictly proper scoring rule S defined by Gneiting and Raftery (2007), such as the usual mean squared error, $E(S(p_t, Z_t)) < E(S(q_t, Z_t))$, which can be shown by observing that

$$\begin{aligned} E(S(p_t,Z_t)) &= E(E(S(p_t,Z_t)|X_t)) \\ &= E(p_t S(p_t,1) + (1-p_t)S(p_t,0)) \\ &< E(p_t S(q_t,1) + (1-p_t)S(q_t,0)) \\ &= E(E(S(q_t,Z_t)|X_t)) \\ &= E(S(q_t,Z_t)), \end{aligned}$$

where the inequality uses the property of the strictly proper scoring rule. Therefore, (5) outperforms (6) in a general sense. As (2) and (5) share the same ROC curve, (5) is also preferred to (6) in terms of the discriminatory capacity as measured by ROC curve.

The ROC curve corresponding to the optimal rule (2) can be obtained by calculating $H(w)$ and $F(w)$ for each w. Suppose X_{1t} is the only predictor available. The optimal rule says that $Z_t = 1$ is predicted if $f(X_{1t};\alpha_1) > wf(X_{1t};\alpha_0)$. Given w,

$$\begin{aligned} H(w) &= P(f(X_{1t};\alpha_1) > wf(X_{1t};\alpha_0)|Z_t = 1) \\ F(w) &= P(f(X_{1t};\alpha_1) > wf(X_{1t};\alpha_0)|Z_t = 0). \end{aligned} \tag{7}$$

Sometimes, evaluation of (7) is intractable due to the parametric assumptions imposed on $F(x_1|Z_t)$. Consequently, we use simulation to approximate both $H(w)$ and $F(w)$. For example, $H(w)$ should be close to $H^s(w) \equiv \sum_{s=1}^{S} I(f(h_s;\alpha_1) > wf(h_s;\alpha_0))/S$ if $\{h_s : s = 1,...,S\}$ is a large sequence of draws from $f(x_1;\alpha_1)$. Given a posterior point estimate of θ, such as the posterior mean, we can get the ROC curve by plotting simulated pairs $(F^s(w), H^s(w))$ over a fine grid of w. The AUC value is calculated by numerical integration of the simulated ROC curve over $[0,1]$. The same procedure can be applied when X_{2t} is considered. Evaluating (7) for combined forecasts involves simulating random samples from bivariate copulas, and the details are contained in the appendix of Trivedi and Zimmer (2005). The inference is based on the posterior distribution of AUC derived by repeating this process for each draw of θ.

In standard Bayesian analysis, $\theta|X_T$ summarizes all we know about θ after observing X_T if and only if $L_T(\theta)$ is correct. When X_T is a time series, as it often is in economic forecasting, this requirement seems restrictive and will generally not be fulfilled. For example, although $H(x_1,x_2|Z_t)$ appears to be correctly specified, dynamic misspecification might be present. This implies that $L_T(\theta)$, obtained by assuming the absence of serial correlation in X_T, is not valid and $\theta|X_T$ does not reflect all information about θ when X_T is available. We do not seek to model the dynamic structure in X_T, which is thought of as a nuisance parameter since our analysis only requires the systematic component $H(x_1,x_2|Z_t)$ be correct. There is a great deal of literature which validates this conjecture. For instance, Lahiri and Wang (1994) showed that explicitly modeling the serial correlation in the context of Markov Switching model of leading indicators jeopardizes turning point predictions. Fortunately, $\theta|X_T$ under this type of model misspecification can be asymptotically approximated by a multivariate normal distribution with the maximum likelihood estimator $\hat{\theta}$ as its mean and the inverse of

the estimated negative Hessian matrix as its covariance (Berk (1966), Bunke and Milhaud (1998), and Geweke (2005)). The well-known quasi-maximum likelihood theory by White (1982) implies that $\theta|X_T$ approaches towards a distribution degenarated at the pseudo-true value of θ. Provided the pseudo-truth is the same as the population parameter of interest, the inferential procedure based on (1) is still consistent. This is the case in our context since $H(x_1, x_2|Z_t)$ is assumed to be correctly specified and the pseudo-true value and the true value are identical. However, Müller (2013) demonstrated that the asymptotic frequentist risk associated with the pseudo-truth is systematically lower if the asymptotic variance of the normal posterior is replaced by a sandwich covariance matrix. The "bread" is the Hessian matrix of the logarithm of $L_T(\theta)$ and the "meat" is the long run variance of the score functions—the latter is usually estimated by a kernel-based approach, as in Andrews (1991) and Newey and West (1987). Given that the inference based on the sandwich posterior is of better quality only in large samples, we will implement both procedures. To save space, only the results obtained from the normal posterior with sandwich covariance are reported in Section 3. Fortunately, very similar conclusions were drawn by sampling from $\theta|X_T$.

3 Application to recession prediction

3.1 Data description

In this part, we will present an empirical illustration to showcase the usefulness of our methodology. The task is to predict future U.S. economic recessions. The monthly data we use consists of 629 observations on two leading indicators for U.S. recessions—ISM diffusion index of new orders and the yield curve spread—from August 1959 to December 2011 (cf. Levanon *et al.* (2011)). The binary target event Z_t is the recession indicator that is one, if the recession occurred, and zero otherwise. The sample proportion of months that were in recession over this period is about 14.8%, indicating that it is a relatively uncommon event. The first predictor X_{1t} is the manufacturing new orders diffusion index compiled by the Institute for Supply Management (ISM), which reflects the number of manufacturers reporting

decreased orders during the previous month compared to the number reporting increased orders. A higher value of this index tends to signify a future economic recession. The yield spread, as the second predictor X_{2t}, is defined as the difference between the constant maturity yields on a 3-month T-bill and the 10-year Treasury note. Previous literature has found that the yield spread is the single indicator which has the highest predictive power in terms of forecasting economic recessions at the fourth quarter horizon. There are many possible reasons why this is the case. See Estrella and Mishkin (1996,1998) for comprehensive explanations. For the purpose of exposition, we use these two indicators to predict recessions six and nine months ahead.

3.2 Model specification and computation

For simplicity, we consider a binormal specification, in which all marginal distributions, including $F(x_1|Z_t)$ and $G(x_2|Z_t)$, are assumed to be normal with different means and variances. Let $\mu_{j,ISM}$ ($\mu_{j,YS}$) and $\sigma^2_{j,ISM}$ ($\sigma^2_{j,YS}$) be the conditional mean and variance of ISM diffusion index (yield spread) given $Z_t = j$. Thus, $\alpha_j = (\mu_{j,ISM}, \sigma^2_{j,ISM})'$ and $\beta_j = (\mu_{j,YS}, \sigma^2_{j,YS})'$ for $j = 0, 1$.

We employ the copula corresponding to the bivariate t-distribution with two parameters: correlation coefficient ρ and degrees of freedom df. Therefore, $\gamma_j = (\rho_j, df_j)'$. The t-copula is often used to model the dependence structure among returns of multiple financial assets (see for example Mashal *et al.* (2003) and Breymann *et al.* (2003)). Gaussian copula and t-copula belong to the so-called elliptical class. The correlation coefficient in both copulas captures the dependence between two random variables, say X_1 and X_2, in the center of their distributions, while the degrees of freedom in the t-copula controls for the tail dependence. The upper tail dependence coefficient λ_u 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 λ_l 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. A strong dependence in the center does not necessarily translate into the strong dependence in both tails. For instance, $\lambda_u = \lambda_l = 0$ for any Gaussian copula without perfect correlation, but they are positive for the t-copula (Demarta and McNeil (2005)). In general, the degrees of freedom controls how quickly the tail dependence shrinks

towards zero. When the degrees of freedom approach infinity, the tail dependence vanishes. In this sense, the t-copula is more general. However, it requires symmetric dependence, namely, $\lambda_u = \lambda_l$, which may be too restrictive in some cases. To overcome this drawback, Demarta and McNeil (2005) has constructed t-copulas with asymmetric tail behaviors by introducing more parameters. Besides the elliptical class, there exist other flexible copulas, some of which may be more than or as flexible as t-copula, including many members within the Archimedean family. The specific choice of copulas crucially rests on the empirical setting. In our circumstance, we favor t-copula because it captures the dependence structure both in the center and tails of the distributions in a relatively parsimonious fashion. This can be seen from Figure 1, which presents the scatter plots of $F(X_{1t}|Z_t)$ and $G(X_{2t}|Z_t)$, the CDF-scaled predictors for each combination of regimes and horizons. We have plotted five contours of bivariate densities of the fitted t-copulas, with all points on a contour representing the same density value. The inner contours represent higher densities. Our framework can accommodate virtually all reasonable copulas, provided some mild regularity conditions, like smoothness, are met.

The prior $P(\cdot)$ is specified as normal. To this end, all parameters with limited range are rescaled by suitable transformations. For example, we impose a prior on $log(\sigma^2_{j,YS})$ rather than $\sigma^2_{j,YS}$. To be more specific, the mean of $P(\cdot)$ is 0 and the variance is a constant multiple (s) of the sandwich covariance matrix associated with $\hat{\theta}$. In order for the prior to be relatively flat, s must be a large positive number. Here, we choose $s = 1,000$ since the empirical results are nearly the same for any s higher than $1,000$. In this case, the assumed value of the mean vector is of little relevance.

Simulating from the sandwich posterior is straightforward as long as the quasi-maximum likelihood estimator $\hat{\theta}$ is available. The attractive feature of copula facilitates the numerical computation substantially in that we can use a sequential procedure to get a preliminary estimator. In the first step, the parameters in marginal distributions $(\alpha_1', \beta_1', \alpha_0', \beta_0')'$ are estimated. The dependence parameters $(\gamma_1', \gamma_0')'$ are estimated in the second step after the estimated marginal distributions have been substituted into $L_T(\theta)$. Finally, we set the initial values of θ to be those obtained in the first two steps and then maximize $L_T(\theta)$ again to get $\hat{\theta}$. The long run covariance matrix of the score functions is estimated via quadratic spectral ker-

nel after the data is filtered by the AR(1) prewhitening procedure as advocated by Andrews and Monahan (1992).

A convenient way of simulating from $\theta|X_T$ is by using Markov chain Monte Carlo (MCMC) method. We consider the random walk chain constructed by a Metropolis-Hastings algorithm. Specifically, suppose a draw θ^b in step b is given. A candidate draw θ^* is sampled from $\theta^b + cN$, where N is a multivariate normal vector with mean zero and variance V, and c is a positive scale number. We use the sandwich covariance matrix as V. θ^* is accepted as the next draw θ^{b+1} if it lies within the area of higher posterior density relative to θ^b. The details can be found in Koop (2003). The value of c is determined in such a way that the resulting acceptance rate is about 25%, as suggested by Albert (2009). The first $10,000$ draws are discarded to remove the impact of the initial value. A variety of diagnostic plots and formal statistical tests based on the remaining $90,000$ draws are not reported here due to brevity. All of them tend to indicate successful convergence of the chain towards $\theta|X_T$.

3.3 Empirical results

Table 1 presents the posterior mean of θ together with two bounds of the highest posterior density (HPD) interval. This is the parameter interval of a given size in which any point delivers a higher posterior density than an arbitrary point outside the interval. It has lower and upper bounds if the posterior is unimodal. Here, the nominal size is fixed at 95%. Analogous to the role a confidence interval plays in a frequentist framework, an HPD interval accounts for the uncertainty associated with a point estimate of θ.

From this table, we see that $|\mu_{1,ISM} - \mu_{0,ISM}|$ shrinks towards zero as the forecast horizon goes from 6 to 9 months. This is just opposite with the yield spread. On the other hand, the yield spread tends to be more volatile than the diffusion index although the volatility of the former decreases with horizon. The two conditional distributions for the diffusion index ($f(x_1|Z_t = 1)$ and $f(x_1|Z_t = 0)$) and the yield spread ($g(x_2|Z_t = 1)$ and $g(x_2|Z_t = 0)$) are shown in Figure 2. For ISM index, $f(x_1|Z_t = 1)$ and $f(x_1|Z_t = 0)$ get closer to each other when we predict recession at the longer horizon, indicating a weakening capacity of distinguishing two regimes defined by Z_t. For yield spread, these two densities remain

roughly unchanged at both horizons. The dependence structure between the two predictors is captured by ρ_j and df_j. The correlation coefficient ρ_1 is negative, while ρ_0 is positive (both are significant).[1] This pattern is also revealed in Figure 1, where both Figures 1(a) and 1(c) for $Z_t = 1$ show the presence of negative relationship between the two predictors while their correlation reverses sign in Figures 1(b) and 1(d) for $Z_t = 0$. Table 1 also reports two additional dependence measures: Kendall's tau and Spearman's rho, which are denoted by τ_j and rho_j respectively. Unlike the standard correlation coefficient ρ_j, these are computed based on ranks of two random variables, and hence are unaffected by the marginal distribution of each variable. For t-copula, we have $\rho_j = \sin(\frac{\pi}{2}\tau_j)$ and $\rho_j = 2\sin(\frac{\pi}{6}rho_j)$. Therefore, all three measures have one-to-one relationship with each other, and they provide essentially the same information. As shown in this table, ρ_j, τ_j and rho_j share the identical sign and significance despite the differing magnitudes. The degrees of freedom when $Z_t = 1$ is large, implying that the Gaussian copula serves as a good approximation and the tail dependence between two predictors should be quite weak. This is consistent with Figures 1(a) and 1(c), where the number of points that lie in the upper-right and lower-left corners is very small. However, this is not the case during the recession months $(Z_t = 0)$, and there are clearly more points in Figures 1(b) and 1(d) located in the corresponding regions. This can be thought of as a consequence of the positive tail dependence in the t-copula with moderate degrees of freedom, which is ruled out by any Gaussian copula.

ROC curves evaluated at the posterior means of θ are displayed in Figures 3(a) and 3(b). The dashed lines are based on (2) with a single predictor in X_t. The explicit analytic forms for these curves are available in Lahiri and Yang (2013a) for a binormal model. All other curves are approximated by simulation. To generate the blue solid lines (Lin ISM+YS), we refit the model by assuming a bivariate normal distribution and forcing two covariance matrices to be equal. As argued in Section 2, this amounts to using a linear combination scheme. We also tried the linear opinion pool in (6) with roughly the same results.[2] ROC curves of the optimal non-linear combination scheme (2), Opt ISM+YS, are represented by green solid lines. Figures 3(a) and 3(b) show that the predictive power of ISM index deteriorates

[1] The 95% HPD intervals of $\rho_0 - \rho_1$ are $[0.382, 0.896]$ and $[0.352, 0.745]$ for 6 and 9-month-ahead forecasts respectively, indicating the differences between correlations across regimes are highly significant.

[2] Results based on (6) are available upon request.

as horizon gets longer. The performance of yield spread, in contrast, is slightly better in predicting recessions 9 months ahead compared with 6 months ahead. These findings are consistent with the evidence in Figure 2. We will examine several determinants of forecasts' accuracy through a counterfactual exercise in Section 4.

The solid ROC curves based on combined forecasts uniformly dominate the dashed counterparts, implying that forecast combination, as an effective way to integrate the useful information contained in (X_{1t}, X_{2t}), leads to substantial improvement in predictive ability over each predictor, particularly at the 6-month horizon. However YS individually performs better than ISM at the 9-month horizon, and is close to Lin ISM+YS. Although the linearly combined forecasts are never better than the optimally combined ones, they are overall close to each other. The difference between these two schemes depends on whether the bivariate normal distribution with homogeneous covariance matrix is a good approximation. For this particular example, it is reasonable to assume (X_{1t}, X_{2t}) to be normally distributed when $Z_t = 1$ but it does not seem to be valid when $Z_t = 0$. Furthermore, virtually all of the second conditional moments in one regime are far away from their counterparts in the other regime, as shown in Table 1. For instance, $|\rho_1 - \rho_0| = 0.6$ and $|\sigma^2_{1,YS} - \sigma^2_{0,YS}| = 1.6$ for 6-month-ahead forecast. This provides a good explanation for the discernible gap between linear and non-linear schemes. As expected, the optimally combined forecasts perform the best for both horizons. For 6-month-ahead forecasts, each predictor contains useful information the other one does not and neither of them dominates the other over the entire range of the ROC curve. By exploiting independent information contained in both predictors, forecast combination is able to achieve a dramatic improvement over each single predictor. A different scenario occurs for the 9-month-ahead forecasts, in which the yield spread is significantly better than the diffusion index. Once the yield spread is known, the additional information provided by the diffusion index is marginally inconsequential. As a result, little gain is achieved through forecast combination over YS. In an extreme case, if the ISM index were completely random and as a result its ROC curve were the diagonal line, the combined forecast would have offered zero improvement over the yield spread since the information contained in the ISM index is redundant.

Table 2 highlights the magnitude of improvement resulting from forecast combinations.

The AUC value for each curve in Figure 3 is shown on the top of the panel of this table, and the percentage of improvement is on the bottom. On average, the AUC values of the combined forecasts are strikingly higher than those of the ISM diffusion index, and this considerable gain in discriminatory power is significant, as suggested by its 95% HPD interval. However, the combined forecasts are not much better than the yield spread. For the 9-month-ahead forecasts, ImpvYS, which is defined as the proportionate improvement of the optimally combined forecast over the yield spread [i.e., ImpvYS$\equiv$ (AUC(Opt ISM+YS)-AUC(YS))/AUC(YS)], is only 3.9% because the marginal contribution from the diffusion index is found to be small. However, for the 6-month-ahead forecasts, this gain is 11.5% and significant.

As a single index to summarize the predictive accuracy of a forecasting system, AUC can be used to compare different forecasting systems. However, one is likely to miss important information by merely relying on AUC exclusively. For example, two ROC curves may cross at an interior point in the unit square with the same AUC value. On the left hand side of this point, one curve is higher than the other, which is reversed on the right hand side. By the AUC criterion, both curves are equally good. However, a decision maker having appetite for a higher hit rate compared to a lower false alarm rate will opt for the ROC curve, which is higher on the right hand side of the crossing point. Different decision makers may have different preferences. To fix the idea, we consider the linear score indexed by $m \in [0,1]$, i.e. $S(m) = mH(w) + (1-m) * (1-F(w))$. A higher m means that the decision maker places more weight on $H(w)$ than he puts on $1-F(w)$. Given m, the problem faced by the decision maker is to choose a threshold value w to maximize $S(m)$. We denote the maximizer and maximum value of this problem by $w^*(m)$ and $S^*(m)$, respectively. Both are functions of m. For a particular value of m, the decision maker may look at Figure 4 to appreciate how large the improvement in $S^*(m)$ can be by using the optimal non-linear scheme. In these graphs, the three lines trace out the relative improvements due to Opt ISM+YS over ISM, YS and Lin ISM+YS as m varies over $[0,1]$. If m is close to 0 or 1, all models are seen to be roughly equivalent. When m lies in the middle, say 0.5, the non-linear combined forecasts offer a 17.5%(27.2%) improvement over the diffusion index while predicting recessions 6 (9) months ahead. Note that the score $S(0.5)$ is proportional to the Peirce skill score often used in the literature; See Manzato (2007) and Granger and Pesaran (2000). Over these m

values, at the 6-month horizon, the non-linear scheme offers $13\% - 18\%$ improvement over ISM and YS. At the 9-month horizon, the non-linear scheme offers nearly 4% improvement over the linear scheme when m is around 0.7. Given that the latter is already a highly efficient classifier, this improvement can be considered economically significant. One implication is that the non-linear scheme seems to be a sensible choice for the target events that are relatively rare and unwarranted like impending financial meltdown, malignant tumour, tsunami, and the like. In these circumstances, the losses associated with miss signals are certainly much larger than those due to false alarms.

To assess the parametric model outlined above in terms of its goodness-of-fit, we compute the posterior predictive p-values for a variety of statistics. Suppose $\tilde{X}_T$ is the another dataset of size T generated from the model under study. We are able to draw a sample from $g(\tilde{X}_T)|X_T$ by sequentially simulating from $\theta|X_T$ and $g(\tilde{X}_T)|\theta$, where $g(\cdot)$ contains some statistics of interest, $g(\tilde{X}_T)|X_T$ is the predictive distribution of $g(\tilde{X}_T)$ after observing X_T, and $g(\tilde{X}_T)|\theta$ is the likelihood function of $g(\tilde{X}_T)$. We can calculate the statistics $g(X_T)$ using the current sample X_T. If the model fits data well, $g(X_T)$ is fairly unlikely to lie too far away from the center of $g(\tilde{X}_T)|X_T$. The predictive p-value is the relative frequency of those more extreme $g(\tilde{X}_T)$ (larger than $g(X_T)$ or small than $g(X_T)$). A small predictive p-value is taken as evidence against a model, and the rule-of-thumb is to reject a model when the p-value is below 0.05. For our purpose, we select eight statistics in $g(\cdot)$: the sample mean and variance for each predictor in each regime, with the results presented in Table 3. Confronted with the data, the t-copula with binormal margins is not rejected for all statistics. Indeed, the minimum p-value is 0.922, meaning that the model is broadly consistent with the actual data.

Using (3), we can generate the probability of an impending recession, a probability that is of interest on its own right. Once the posterior distribution of θ is obtained and $P(Z_t = 1)$ is estimated by $\sum_{t=1}^{T} Z_t/T$, the posterior behavior of $P(Z_t = 1|X_{1t}, X_{2t})$ is also completely determined. In particular, we can get a path of the combined probability forecasts $\{P(Z_t = 1|X_{1t}, X_{2t}) : t = 1, ..., T\}$ for each horizon when the posterior mean of θ is plugged in. Since ISM and YS variables are never revised, except for the fact that the parameters are estimated using the whole sample, these generated probabilities of recessions can be considered to be real-time forecasts. The recession probabilities are presented in Figure 5, together with

the forecasts based on each predictor. With a few exceptions, almost every recession since 1969 is accompanied by a higher-than-usual probability generated by the combined forecasts. The ISM diffusion index tends to generate high (low) probabilities 6 months ahead of recessions (expansions). However, it seems to be too conservative to give high probabilities of recessions 9 months ahead. Irrespective of which regime materializes, the ISM diffusion index always fluctuates around its average value, suggesting it lacks the ability to identify forthcoming economic recessions at the longer horizon. The yield spread is superior at the 9-month horizon compared to 6-month horizon. The value of the combined forecasts are borne out remarkably well across the whole sample period with the linear and non-linear schemes, even though at the 9-month horizon YS performs very close to the combined forecasts. As we have mentioned before, the non-linear forecasts have a slight edge over those generated by the linear scheme.

4 Counterfactual analysis

To identify the determinants of accuracy gain through forecast combination, it is worthwhile to break down the overall improvement of the combined forecasts into several components. For the sake of brevity, we merely consider 6-month-ahead prediction in this section. Ideally, the forecasts should behave quite distinctly across regimes in order to have a high discriminatory power. Put it differently, the conditional distribution of the predictor given $Z_t = 1$ must be strongly separated from that given $Z_t = 0$. Not only does this require the two conditional means be different from each other but the conditional variances cannot be very high. Intuitively speaking, higher variance makes the two distributions overlap to a larger extent, and thereby dilutes a given mean difference. For instance, the AUC(YS) in Table 2 can be written as

$$AUC(YS) = \Phi\left(\frac{\mu_{1,YS} - \mu_{0,YS}}{\sqrt{\sigma^2_{1,YS} + \sigma^2_{0,YS}}}\right) \tag{8}$$

for the binormal specification. The implication of (8) is that the ability of the yield spread in discriminating between two regimes, as measured by the AUC, depends positively on the difference between two conditional means $\mu_{1,YS} - \mu_{0,YS}$ but negatively on the magnitude of two conditional variances: $\sigma^2_{1,YS}$ and $\sigma^2_{0,YS}$. They correspond to two terms in the decomposition of the mean squared error between forecasts and actuals as suggested by Yates (1982). Given $\mu_{1,YS} - \mu_{0,YS}$, the minimum forecast variance is equal to $(\mu_{1,YS} - \mu_{0,YS})^2 P(Z_t = 1)P(Z_t = 0)$, which will only be achieved when the predictor takes $\mu_{1,YS}$ on all occasions of $Z_t = 1$, and it takes $\mu_{0,YS}$ on other occasions. Under this circumstance, the variability of forecasts is completely due to the event's occurrence. Thus, the minimum variance is the smallest variance necessary to support a given wedge between $\mu_{1,YS}$ and $\mu_{0,YS}$. The actual variance that is beyond this minimum value is called excess variability, which is equal to $P(Z_t = 1)\sigma^2_{1,YS} + P(Z_t = 0)\sigma^2_{0,YS}$ and reflects how responsive the predictor is to information that is not related to the event's occurrence. When $\sigma^2_{1,YS} = \sigma^2_{0,YS} = 0$, the excess variability becomes zero. To maximize the discriminatory power, a higher value of the minimum variance (hence a higher $|\mu_{1,YS} - \mu_{0,YS}|$) and a lower value of the excess variability (hence a lower $\sigma^2_{1,YS}$ and $\sigma^2_{0,YS}$) are desirable. In sum, a forecast with high discriminatory ability is expected to be highly sensitive to relevant information, but insensitive to irrelevant information related to the event's occurrence. In terms of the mean difference in Table 1, the yield spread is better than the ISM diffusion index. However, the diffusion index contributes because it has less excess variability compared to YS at the shorter horizon. Hence, these two predictors complement one another to produce more accurate forecasts, which absorbs the strength and abandons the weakness in each of them. In other words, the improvement of the combined forecast relative to any one predictor may stem from the marginal information in the other predictor.

As is obvious in Table 1, not only do the two conditional distributions for each predictor differ, but the dependence structure between the two predictors varies across regimes. The diffusion index is negatively correlated with the yield spread when $Z_t = 1$ (recessions) but remarkably the correlation becomes positive when $Z_t = 0$ (expansions). The combined forecast also benefits from the inclusion of this additional information regarding the dependence structure, which is not available for use by the single predictor or the linear scheme. The

more distinct the values of ρ_1 and ρ_0 are, the higher is the improvement that can be made through non-linear forecast combination.

Figure 6 illustrates the sensitivity of ImpvISM and ImpvYS to the various determinants of the discriminatory ability. Specifically, it shows the quantitative importance of each of the factors that contributes towards the improved forecasting ability of the combined forecasts relative to ISM and YS individually. That is, it shows what happens to ImpvISM and ImpvYS when one of the parameters deviates from its posterior mean. Panel (a) of Figure 6 depicts how the relative superiority of the combined forecast in terms of AUC changes compared to ISM (ImpvISM) and to YS (ImpvYS) as the difference in the means of the conditional distributions of ISM diffusion index decreases from its observed value of 1.02 in the sample (see Table 1). This value is the right most point in Figure 6(a). As $\mu_{1,ISM} - \mu_{0,ISM}$ decreases (i.e., ISM is deteriorating as a classifier), the combined forecast also gets worse, but relatively less than the decline in that of ISM because YS continues to be the same. Thus, relatively speaking, the combined forecast improves upon ISM more because of YS. Simultaneously, the relative superiority of the combined forecast over YS decreases, albeit slowly, because YS by itself continues to be as good, and relatively speaking the combined forecast has less scope to improve upon YS. Figure 6(b) is a mirror image of Figure 6(a) as $\mu_{1,YS} - \mu_{0,YS}$ decreases to 0 from its sample value of 2.18. Note that the simulated changes in the mean differences of ISM and YS between the two regimes do not affect the distributions of ISM and YS nor their dependence structure. Figures 6(c) and 6(d) trace out the effects of changing the variances of ISM and YS respectively from their observed values. As the variance of ISM during recessions increases, the relative improvement of the combined forecast over ISM increases as a result of the deteriorating quality of ISM as an individual predictor with the quality of YS remaining the same. But ImpvYS seems to be largely insensitive to changes in $\sigma^2_{1,ISM}$. Figure 6(d) that traces out the effect of changes in $\sigma^2_{1,YS}$ is again a mirror image of Figure 6(c). The slopes of these curves are consistent with expectations, but their quantitative magnitudes are functions of the data structure and the forecast combination procedure.

Of particular interest is the role of changes in correlations between the two predictors as the regime changes. Figure 6(e) depicts ImpvISM and ImpvYS as $|\rho_1 - \rho_0|$ increases. Note that $|\rho_1 - \rho_0| = 0.67$ in the sample. As this value goes from 0 to 1.2, we find that the gain in

the non-linear combination scheme over ISM and YS monotonically increases, as expected, in a parallel fashion. This gain is coming entirely from the increased value of the correlation during the two regimes with the individual predictors being the same. This affirms that the combined forecasts are able to explore the differential in dependence structure as well. When two conditional distributions of each predictor are fixed, more distinct correlations give rise to higher AUC(Opt ISM+YS), which further raises ImpvISM and ImpvYS. This is an noteworthy implication of our non-linear forecast combination framework.

5 Conclusion

This paper has proposed an optimal non-linear procedure to combine multiple predictors for a binary target event within a Bayesian framework. The resulting likelihood ratio rule, which has a solid rationale from Neyman-Pearson lemma, maximizes the hit rate for any given false alarm rate among all competing schemes. We show that under additional simplifying assumptions, the rule reduces to the linear combination scheme extensively used in practice. To illustrate the usefulness of this method, we use a binormal specification with t-copulas to characterize the marginal distributions and the contemporaneous dependence structure of the ISM new order diffusion index and the yield curve spread in predicting U.S. recessions defined by NBER. The merit of our approach is that the discriminatory ability of the individual predictors and the optimal combination rule can be uniquely determined by a few fundamental parameters, each of which controls a specific aspect of forecast skill. Evaluated at the posterior means of the estimated parameters, the aggregated forecasts have considerably higher AUC values than the individual predictors. Given that recessions are hard to predict, this new approach offers a noteworthy improvement over the existing approaches. To better understand the underlying sources, we decomposed the overall improvement of the combined forecasts relative to each predictor into gains from using the differentials in the dependence structure of the two predictors in the two regimes, and gains from using the differentials in the means and variances of the conditional distributions of the other predictor. The two predictors seem to complement one another to produce a more accurate forecast, absorbing the

strengths but abandoning the weaknesses in each of them.

Although the method presented in this paper is conceptually and computationally straightforward, it is subject to possible misspecification error. Its validity requires both the marginal distributions and the copulas be correctly specified. If a forecaster has a large training sample to estimate the model, we can circumvent this problem by using non-parametric or semi-parametric methods. For example, if we want to maintain a parametric form for copulas, we may estimate the marginal distributions non-parametrically by kernel smoothing. The parameters in copulas are estimated by plugging in the non-parametric estimates. Asymptotic properties of this semi-parametric two-step estimator have been established in Chen *et al.* (2010). If and how the existing results can be modified to make inference on the resulting ROC curve is still an open question. Generalization of our combination scheme along this line would be a promising topic for future research.

Tables and Figures

Table 1: Posterior summary of θ

	6 months ahead			9 months ahead		
θ	Mean	LHPD	UHPD	Mean	LHPD	UHPD
$\mu_{1,ISM}$	0.865	0.415	1.357	0.510	0.077	0.941
$\mu_{0,ISM}$	-0.153	-0.371	0.082	-0.096	-0.350	0.194
$\mu_{1,YS}$	0.890	-0.684	2.309	1.076	-0.256	2.374
$\mu_{0,YS}$	-1.294	-1.555	-0.980	-1.310	-1.794	-0.848
$\sigma^2_{1,ISM}$	0.886	0.481	1.557	0.734	0.408	1.154
$\sigma^2_{0,ISM}$	0.902	0.626	1.219	1.026	0.657	1.446
$\sigma^2_{1,YS}$	3.651	2.384	5.492	3.308	0.396	13.120
$\sigma^2_{0,YS}$	2.051	1.202	3.105	1.967	1.346	2.675
ρ_1	-0.398	-0.624	-0.181	-0.247	-0.258	-0.234
ρ_0	0.272	0.130	0.441	0.306	0.104	0.495
τ_1	-0.260	-0.408	-0.097	-0.159	-0.166	-0.150
τ_0	0.175	0.083	0.291	0.198	0.066	0.330
rho_1	-0.382	-0.606	-0.174	-0.237	-0.248	-0.224
rho_0	0.260	0.124	0.424	0.293	0.099	0.478
df_1	8008.878	7942.348	8085.989	27392.215	26900.526	27899.045
df_0	70.974	65.942	75.709	28.565	11.829	53.228

Table 2: Comparison of posterior AUC

	6 months ahead			9 months ahead		
Object	Mean	LHPD	UHPD	Mean	LHPD	UHPD
AUC(ISM)	0.777	0.671	0.887	0.676	0.535	0.803
AUC(YS)	0.820	0.613	0.956	0.851	0.705	0.958
AUC(Lin ISM+YS)	0.891	0.738	0.978	0.862	0.706	0.942
AUC(Opt ISM+YS)	0.914	0.835	0.983	0.884	0.808	0.966
ImpvISM	0.176	0.008	0.381	0.307	0.087	0.664
ImpvYS	0.115	0.011	0.386	0.039	0.002	0.209

Notes: The top of the table contains the AUC value computed for each curve in Figure 3. ImpvISM=(AUC(Opt ISM+YS)-AUC(ISM))/AUC(ISM). ImpvYS=(AUC(Opt ISM+YS)-AUC(YS))/AUC(YS).

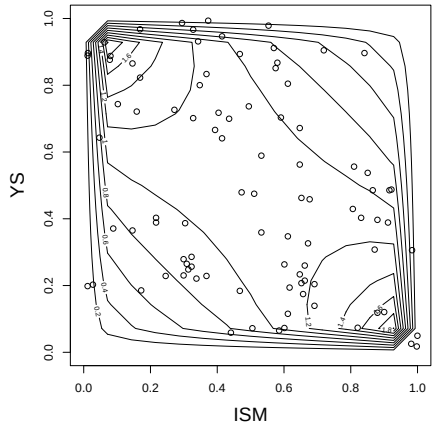
Table 3: Model comparison in terms of the posterior predictive p-values

$g(\cdot)$	6 months ahead	9 months ahead
$p(\mu_{1,ISM})$	0.984	0.956
$p(\sigma^2_{1,ISM})$	0.980	0.970
$p(\mu_{0,ISM})$	0.984	0.988
$p(\sigma^2_{0,ISM})$	0.994	0.922
$p(\mu_{1,YS})$	0.974	0.986
$p(\sigma^2_{1,YS})$	0.996	0.976
$p(\mu_{0,YS})$	0.984	0.970
$p(\sigma^2_{0,YS})$	0.956	0.994

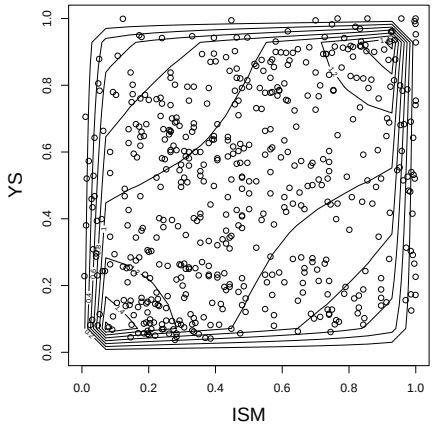
Notes: $p(\mu_{1,ISM})$ is the posterior predictive p-value when $\mu_{1,ISM}$ is the statistic of interest. Any other p-value in this table is self-explained.

Figure 1: Scatter plots of $F(X_{1t}|Z_t)$ and $G(X_{2t}|Z_t)$

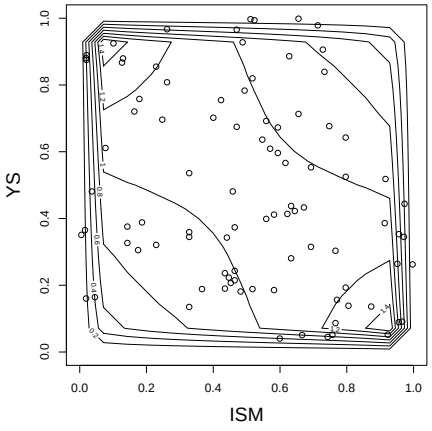
(a) 6 months ahead ($Z_t = 1$)



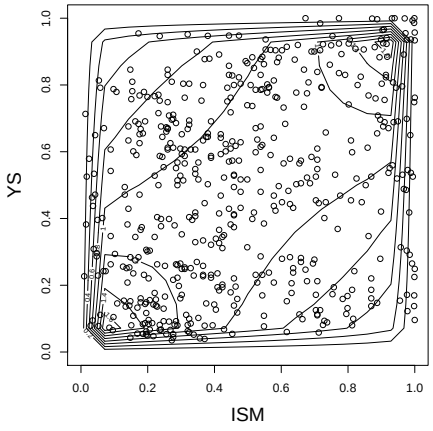
(b) 6 months ahead ($Z_t = 0$)



(c) 9 months ahead ($Z_t = 1$)

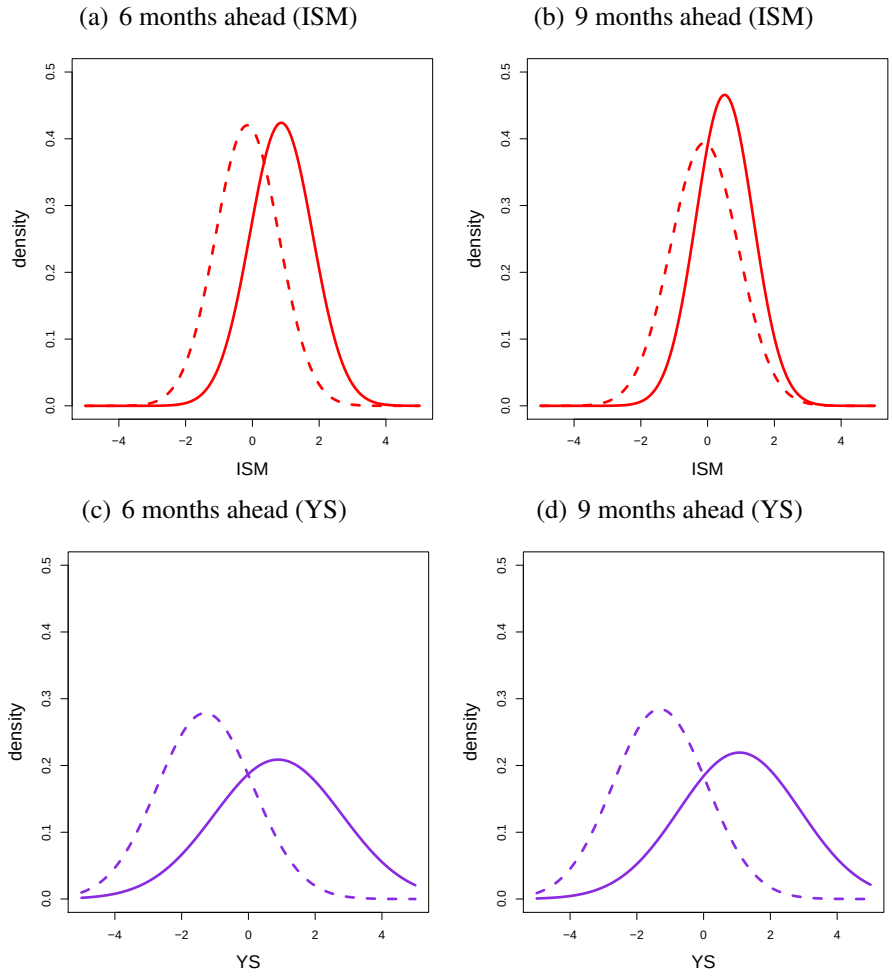


(d) 9 months ahead ($Z_t = 0$)



Notes: The solid curves represent the contours of the fitted t-copulas.

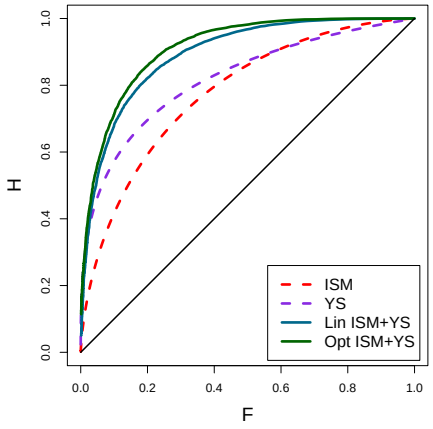
Figure 2: Conditional densities of two predictors



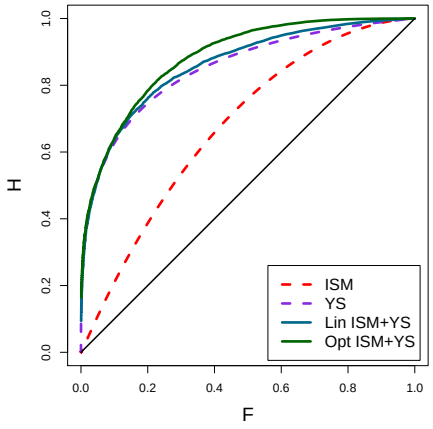
Notes: The solid curve represents the conditional density given $Z_t = 1$ and the dotted curve is the conditional density given $Z_t = 0$.

Figure 3: ROC curves evaluated at posterior mean

(a) 6 months ahead



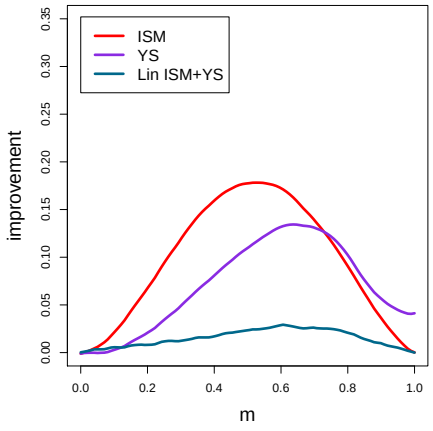
(b) 9 months ahead



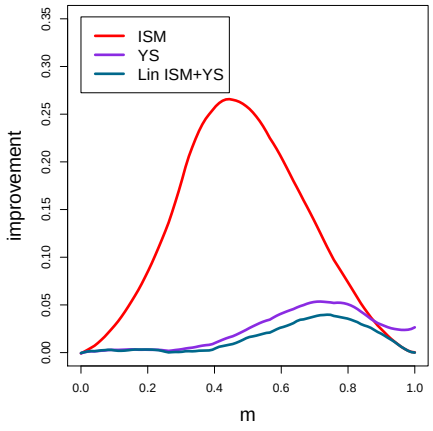
Notes: "Lin ISM+YS" ("Opt ISM+YS") is the ROC curve of the linearly (optimally) combined forecasts.

Figure 4: Improvements of the optimally combined forecast in terms of the linear score

(a) 6 months ahead



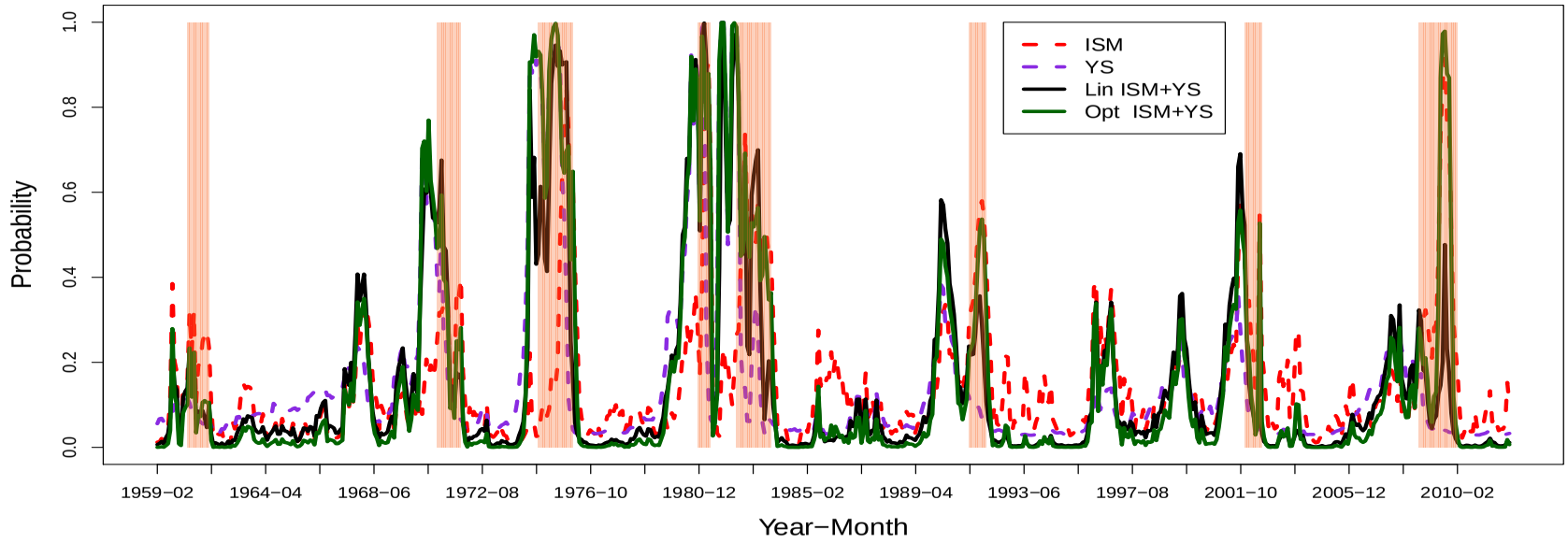
(b) 9 months ahead



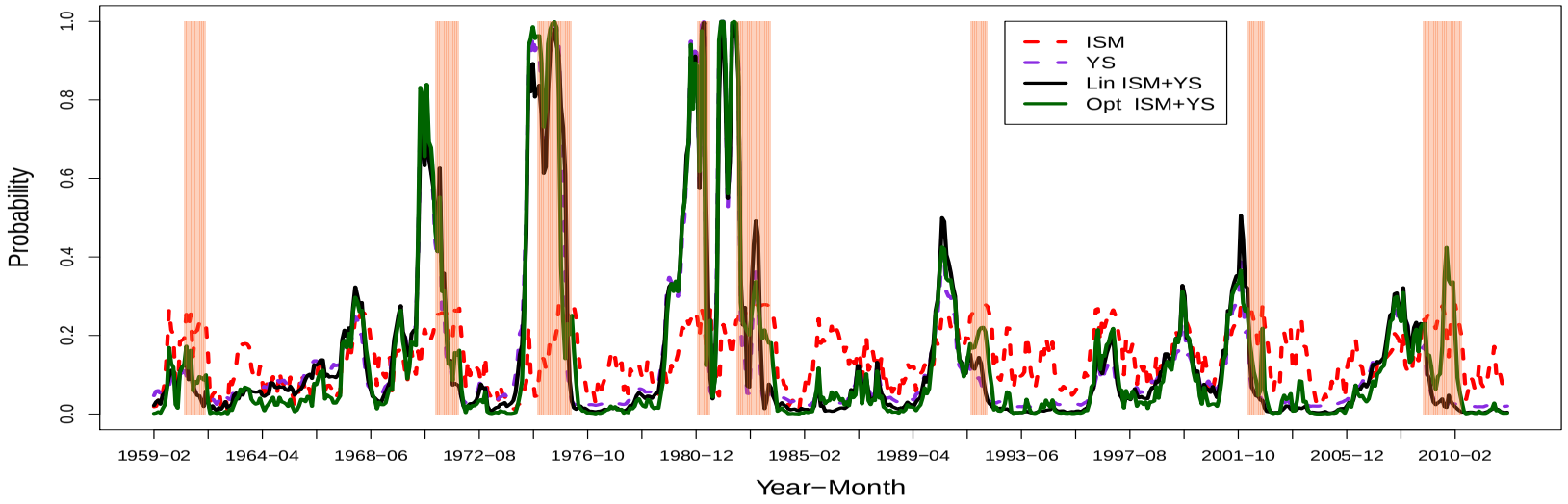
Notes: "ISM" represents the improvement of the optimally combined forecast over ISM diffusion index, that is, $(S^*(m)$(Opt ISM+YS)-$S^*(m)$(ISM))/$S^*(m)$(ISM). All of the other curves are calculated in the same way.

Figure 5: In-sample probability forecasts

(a) 6 months ahead

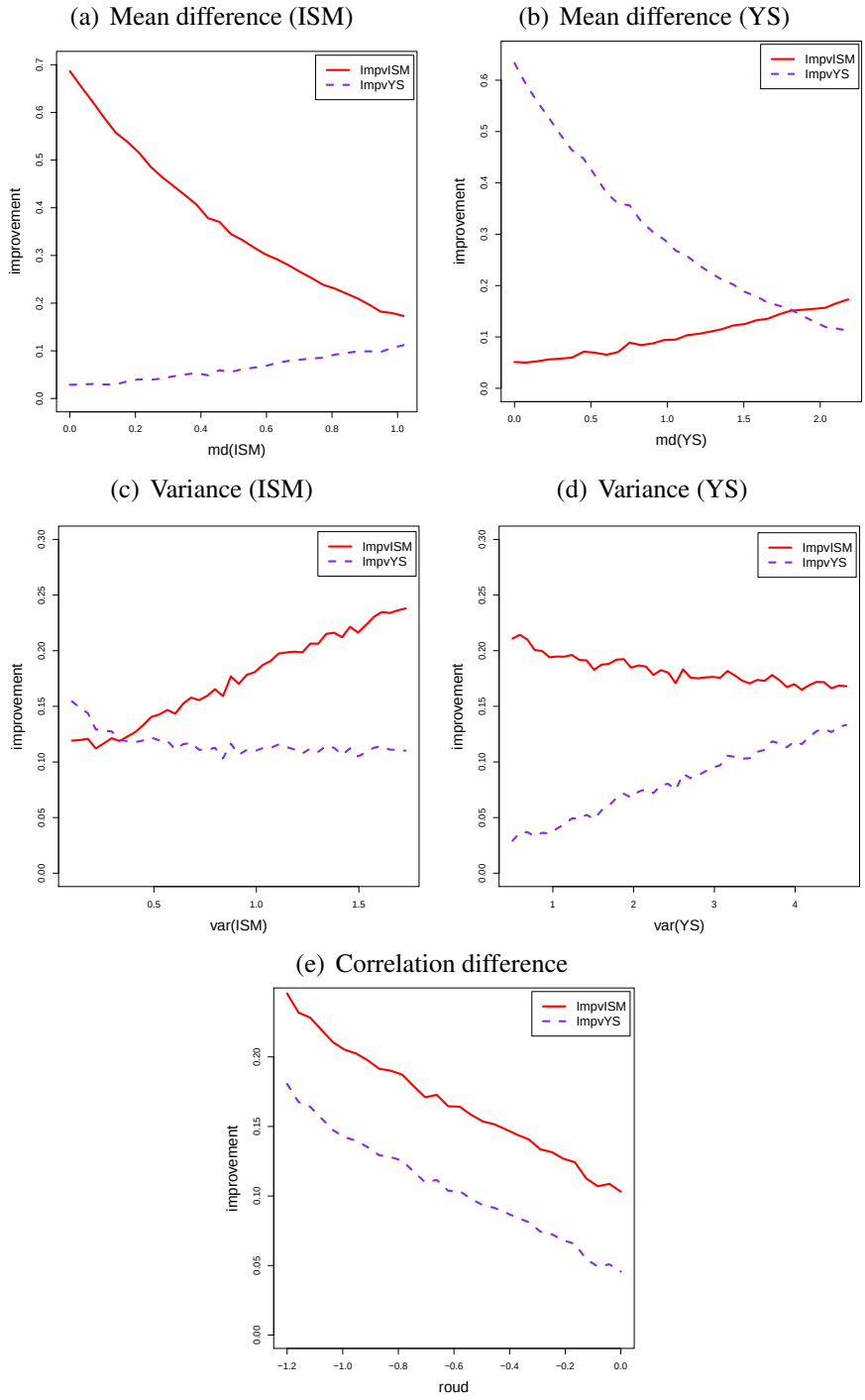


(b) 9 months ahead



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

Figure 6: Determinants of improvements in the combined forecasts over ISM and YS



Notes: For the ISM, md(ISM)= $\mu_{1,ISM} - \mu_{0,ISM}$ and var(ISM)= $\sigma^2_{1,ISM}$. The similar notations apply to the yield spread. roud= $\rho_1 - \rho_0$.

References

Albert, J. (2009), *Bayesian Computation with R*, Springer.

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

Andrews, D. W. K. (1991), 'Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation', *Econometrica* **59**, 817–858.

Andrews, D. W. K. and Monahan, J. C. (1992), 'An Improved Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimator', *Econometrica* **60**, 953–966.

Bates, J. M. and Granger, C. W. J. (1969), 'The Combination of Forecasts', *Operational Research Quarterly* **20**, 451–468.

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

Berk, R. H. (1966), 'Limiting Behavior of Posterior Distributions when the Model is Incorrect', *The Annals of Mathematical Statistics* **37**, 51–58.

Breymann, W., Dias, A. and Embrechts, P. (2003), 'Dependence Structures for Multivariate High-Frequency Data in Finance', *Quantitative Finance* **3**, 1–14.

Brier, G. W. (1950), 'Verification of Forecasts Expressed in Terms of Probability', *Monthly Weather Review* **78**, 1–3.

Bunke, O. and Milhaud, X. (1998), 'Asymptotic Behavior of Bayes Estimates under Possibly Incorrect Models', *The Annals of Statistics* **26**, 617–644.

Chen, X., Fan, Y., Pouzo, D. and Ying, Z. (2010), 'Estimation and Model Selection of Semiparametric Multivariate Survival Functions under General Censorship', *Journal of Econometrics* **157**, 129–142.

Dawid, A. P. (1984), 'Present Position and Potential Developments: Some Personal Views: Statistical Theory: The Prequential Approach', *Journal of the Royal Statistical Society, Series A* **147**, 278–292.

De Luca, G. and Carfora, A. (2014), 'Predicting U.S. Recessions through a Combination of Probability Forecasts', *Empirical Economics* **46**, 127–144.

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

Drehmann, M. and Juselius, M. (2012), Improving Early Warning Indicators for Banking Crises Satisfying Policy Requirements. Paper presentation at: Understanding Macroprudential Regulation, Norges Bank, Oslo, 2930 November 2012.

Estrella, A. and Mishkin, F. S. (1996), 'The Yield Curve as a Predictor of U.S. Recessions', *Current Issues in Economics and Finance* **2**, 41–51.

Estrella, A. and Mishkin, F. S. (1998), 'Predicting U.S. Recessions: Financial Variables as Leading Indicators', *The Review of Economics and Statistics* **80**, 45–61.

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

Geweke, J. (2005), *Contemporary Bayesian Econometrics and Statistics*, John Wiley & Sons.

Gneiting, T. and Raftery, A. E. (2007), 'Strictly Proper Scoring Rules, Prediction, and Estimation', *Journal of the American Statistical Association* **102**, 359–378.

Graham, J. R. (1996), 'Is a Group of Economists Better Than One? Than None?', *The Journal of Business* **69**, 193–232.

Granger, C. W. J. and Pesaran, M. H. (2000), 'Economic and Statistical Measures of Forecast Accuracy', *Journal of Forecasting* **19**, 537–560.

Granger, C. W. J. and Ramanathan, R. (1984), 'Improved Methods of Combining Forecasts', *Journal of Forecasting* **3**, 197–204.

Kamstra, M. and Kennedy, P. (1998), 'Combining Qualitative Forecasts Using Logit', *International Journal of Forecasting* **14**, 83–93.

Koop, G. (2003), *Bayesian Econometrics*, John Wiley & Sons.

Krzanowski, W. J. and Hand, D. J. (2009), *ROC Curves for Continuous Data*, Chapman & Hall.

Lahiri, K. and Wang, J. G. (1996), Interest Rate Spreads as Predictors of Business Cycles, *in* G. S. Maddala and C. R. Rao, eds, 'Handbook of Statistics 14 (Statistical Methods in Finance)', North-Holland Amsterdam, pp. 297–315.

Lahiri, K. and Wang, J. G. (1994), 'Predicting Cyclical Turning Points with Leading Index in a Markov Switching Model', *Journal of Forecasting* **13**, 245–263.

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. (2013a), Confidence Bands for ROC Curves with Serially Dependent Data. Working paper, Department of Economics, State University of New York at Albany.

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

Levanon, G., Ozyildirim, A., Schaitkin, B. and Zabinska, J. (2011), Comprehensive Benchmark Revisions for The Conference Board Leading Economic Index for the United States. EPWP 11-06, The Conference Board, New York, December 2011.

Manzato, A. (2007), 'A Note On the Maximum Peirce Skill Score', *Weather and Forecasting* **22**, 1148–1154.

Mashal, R., Naldi, M. and Zeevi, A. (2003), 'On the Dependence of Equity and Asset Returns', *Risk* **16**, 83–87.

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

Müller, U. K. (2013), 'Risk of Bayesian Inference in Misspecified Models, and the Sandwich Covariance Matrix', *Econometrica* **81**, 1805–1849.

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

Newey, W. K. and West, K. D. (1987), 'A Simple, Positive Semi-definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix', *Econometrica* **55**, 703–708.

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.

Pepe, M. S. (2003), *The Statistical Evaluation of Medical Tests for Classification and Prediction*, Oxford University Press.

Ramsey, J. B. (1969), 'Tests for Specification Errors in Classical Linear Least-Squares Regression Analysis', *Journal of the Royal Statistical Society, Series B* **31**, 350–371.

Ranjan, R. and Gneiting, T. (2010), 'Combining Probability Forecasts', *Journal of the Royal Statistical Society, Series B* **72**, 71–91.

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

Stephenson, D. B. (2000), 'Use of the 'Odds Ratio' for Diagnosing Forecast Skill', *Weather Forecasting* **15**, 221–232.

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

Timmermann, A. (2006), Forecast Combinations, *in* G. Elliott, C. W. J. Granger and A. Timmermann, eds, 'Handbook of Economic Forecasting', North-Holland Amsterdam, pp. 135–196.

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

White, H. (1982), 'Maximum Likelihood Estimation of Misspecified Models', *Econometrica* **50**, 1–25.

Yates, J. F. (1982), 'External Correspondence: Decompositions of the Mean Probability Score', *Organizational Behavior and Human Performance* **30**, 132–156.

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