Combining Probability Forecasts

Michael P. Clements  David I. Harvey*
Department of Economics  School of Economics
University of Warwick  University of Nottingham

May 3, 2007

Abstract

We consider different methods for combining probability forecasts. In empirical exercises, the data generating process giving rise to the forecasts and the event being forecast is not known, and therefore the optimal form of combination will also be unknown. We consider the properties of various combination schemes for a number of plausible data generating processes, and indicate which types of combination are likely to be useful. We also show that whether forecast encompassing is found to hold between two rival sets of forecasts may depend on the type of combination adopted. The relative performance of the different combination methods is illustrated with an application to predicting recession probabilities using leading indicators.

JEL classification: C12, C15, C53.

Keywords: Probability forecasts; Forecast combinations; Recession probabilities.

*Corresponding author: School of Economics, University of Nottingham, University Park, Nottingham NG7 2RD, UK. Tel.: +44 (0)115 9515481. Fax.: +44 (0)115 9514159. Email: dave.harvey@nottingham.ac.uk
1 Introduction

In this paper we consider different ways of combining probability forecasts of the same event. As noted by Diebold and Lopez (1996), forecasts of economic and financial variables often take the form of probabilities, and there are good reasons to believe that probability forecasts will become increasingly prominent. For example, in a macroeconomic policy setting, a forecast of the probability that the target rate of inflation will be exceeded next year, or of the probability that the economy will contract, may be markedly more informative than simple point forecasts of the expected rates of inflation and output growth, especially in the absence of any indication of the degree of uncertainty to be attached to the point forecasts.

An extensive literature in economic and management science attests to the usefulness of forecast combination for point forecasts, where by a point forecast we mean a forecast defined on $\mathbb{R}$ of an outcome that is also defined on $\mathbb{R}$, in contrast to probability forecasts defined on the interval $[0, 1]$ of a binary outcome variable. The recent literature on combining point forecasts covers a number of areas, including the specification of the combination weights; testing for forecast encompassing; the impact of parameter estimation uncertainty; and the limiting distributions of the tests when the forecasts are from models which are nested. With few exceptions the standard assumption in the literature has been that the forecaster has a squared error loss function, and most of the work on the combination of point forecasts has been based on linear combinations of the individual forecasts. The focus on linear combinations is readily justifiable when in addition to squared-error loss we assume that the variable being forecast ($y_{t+1}$) and the forecasts (given by the vector $y_{t+1|t}$) follow a joint gaussian distribution. Then standard results indicate that the conditional expectation $E(y_{t+1} | y_{t+1|t})$ is the optimal predictor (the conditional expectation minimizes expected squared error loss), and the joint normality of $[y_{t+1} \; y_{t+1|t}]'$ entails that the conditional expectation of $y_{t+1}$ is a linear combination of the elements of $y_{t+1|t}$ (see, for example, Timmermann (2006, p.144-5)).

But for probability forecasts, the limited support of $(y_{t+1}, y_{t+1|t})$ suggests that the justifica-

---

1 In other spheres, the combination of probability assessments is commonplace, although the emphasis tends to be different from ours. The literature on the combination of experts’ subjective probability distributions (see, e.g., Genest and Zidek (1986) and Clemen and Winkler (1999)) looks at ways of aggregating individual assessments such that the aggregate possesses desirable properties, rather than focusing on accuracy. See also Dawid (1986) and Winkler (1996) on probability forecasting and evaluation from a meteorological perspective, as well as a discussion of earlier contributions.


3 Elliott and Timmermann (2004) is a notable exception.
tion for considering linear combinations of forecasts is more problematic, and alternatives such as the logarithmic opinion pool (LoOP) figure prominently in the literature. We investigate different forms of combination for probability forecasts, in terms of the accuracy of the combined forecasts, and in terms of the implications for forecast encompassing. Because in practice the data generating process of the forecasts and the event being forecast will typically be unknown, we consider the properties of various combination schemes for a number of plausible data generating processes, to see whether some tend to dominate the others. We establish the large-sample properties of the combination schemes, but at least as importantly, we investigate their relative performances when the combination weights are estimated, as is often done in practice.

The plan of the rest of the paper is as follows. In section 2 we describe the forecast combinations we consider, as well as the loss functions that are typically used for evaluating probability forecasts. Section 3 presents a discussion of optimal combinations and the estimation of combination parameters under different loss functions, as well as establishing the relationship between forecast encompassing and the optimal form of combination. Section 4 derives the properties of the different forms of combination for two representative data generating processes. Section 5 illustrates the different types of combinations of probability forecasts with an application based on combining forecast recession probabilities generated using two leading indicators from the Conference Board’s Composite Leading Indicator. Section 6 offers some concluding remarks.

2 Combinations and scoring rules for probability forecasts

2.1 Forecast combinations

The first combination method is the linear combinations of forecasts. In the literature on combining experts’ subjective probability distributions this is commonly referred to as the ‘linear opinion pool’: LiOP. In the econometrics literature, the most general form of linear combination of two forecasts, \( f_{1t} \) and \( f_{2t} \), is given by:

\[
C_t(f_{1t}, f_{2t}; \beta) = \alpha + \beta_1 f_{1t} + \beta_2 f_{2t} \tag{1}
\]

where \( \beta \) is used in a generic sense to denote the combination parameters, and where no restrictions on \( \alpha \) or the weights \( \{\beta_1, \beta_2\} \) are imposed. The weights are typically estimated by applying OLS to:

\[
y_t = \alpha + \beta_1 f_{1t} + \beta_2 f_{2t} + \varepsilon_t
\]
as this corresponds to minimizing the sum of squared forecast errors of the combined forecast. This is the form of the regression to test for forecast encompassing advocated by Fair and Shiller (1990), while restricted versions have been considered by Chong and Hendry (1986) and West (2001) (setting $\beta_1 = 1$ when testing that $\beta_2 = 0$) and West and McCracken (1998) and Harvey et al. (1998) (restricting $\beta_1 + \beta_2 = 1$). Forecast encompassing holds when a combination of two (or more) forecasts does not result in a statistically significant reduction in the forecast loss relative to just one of the forecasts: here $\beta_2 = 0$ implies that $f_1$ forecast encompasses $f_2$.

Although tests of forecast encompassing assess whether ex post a linear combination of forecasts results in a statistically significant reduction in forecast loss (such as mean squared error) relative to using a particular forecast, they may also convey information on when combinations might improve accuracy ex ante. Note that neither the general form nor either of the restricted forms ensures that $f_t \in [0, 1]$.

The second form of combination is the multiplicative combination:

$$C_t (f_{1t}, f_{2t}; \beta) = \alpha f_1^{\beta_1} f_2^{\beta_2}$$

(2)

known as the ‘logarithmic opinion pool’ (LoOP) when applied to experts’ subjective probability distributions. This is a simplified version of the following formula for combining discrete probability distributions:

$$f^j = \frac{\prod_{i=1}^{N} (f^j_i)^{\beta_i}}{\sum_{j=1}^{M} \prod_{i=1}^{N} (f^j_i)^{\beta_i}} = \frac{\exp \left( \sum_{i=1}^{N} \beta_i \log f^j_i \right)}{\sum_{j=1}^{M} \exp \left( \sum_{i=1}^{N} \beta_i \log f^j_i \right)}$$

where $f^j_i$ is individual $i$’s probability of ‘class $c_j$’, where there are $M$ classes. The denominator is a scaling factor. Typically $\sum_{i=1}^{N} \beta_i = 1$, and if the weights are equal, i.e. $\beta_i = N^{-1}$, the LoOP delivers the geometric mean. Assuming a binary variable ($M = 2$), and with $N = 2$ we obtain:

$$f = \frac{f_1^{\beta_1} f_2^{\beta_2}}{f_1^{\beta_1} f_2^{\beta_2} + (1 - f_1)^{\beta_1} (1 - f_2)^{\beta_2}}$$

where we drop the $j$ superscript on $f$ and $f_i$, which are now the combined and individual probabilities of the event occurring. The multiplicative combination (2) follows directly by setting the denominator equal to $\alpha^{-1}$. As was the case with (1), without restrictions on $\alpha$, $\beta_1$ and $\beta_2$, there is no guarantee that $f_t \in [0, 1]$.

The final form of combination is that of Kamstra and Kennedy (1998) (henceforth, KK).
KK suggest combining log odds ratios by logit regressions, without claiming any optimality properties for the resulting combination: “we only claim that this methodology is a means of combining individual probability forecasts in a computationally attractive manner, while alleviating bias” (p.86). Specifically, the KK combination of $f_1t$ and $f_2t$ is:

$$C_t(f_{1t}, f_{2t}; \beta) = \frac{\exp \left[ \alpha + \beta_1 \ln \left( \frac{f_{1t}}{1 - f_{1t}} \right) + \beta_2 \ln \left( \frac{f_{2t}}{1 - f_{2t}} \right) \right]}{1 + \exp \left[ \alpha + \beta_1 \ln \left( \frac{f_{1t}}{1 - f_{1t}} \right) + \beta_2 \ln \left( \frac{f_{2t}}{1 - f_{2t}} \right) \right]}$$

$$= \frac{\exp (\alpha) \left( \frac{f_{1t}}{1 - f_{1t}} \right)^{\beta_1} \left( \frac{f_{2t}}{1 - f_{2t}} \right)^{\beta_2}}{1 + \exp (\alpha) \left( \frac{f_{1t}}{1 - f_{1t}} \right)^{\beta_1} \left( \frac{f_{2t}}{1 - f_{2t}} \right)^{\beta_2}}$$

(3)

where $\beta_1$ and $\beta_2$ are the maximum likelihood estimates of the slope coefficients from a logit regression of $y_t$ on a constant, $\ln \left( \frac{f_{1t}}{1 - f_{1t}} \right)$ and $\ln \left( \frac{f_{2t}}{1 - f_{2t}} \right)$. The KK combination does ensure that $f_t \in [0, 1]$.

### 2.2 Scoring rules

In contrast to the evaluation of point forecasts, the actual probabilities are not observed, so that $f$, the forecast probability that $Y = 1$, is compared to the realized value of $Y$. The two main ways of scoring probability forecasts are the quadratic and logarithmic scores. The Brier or quadratic probability score (QPS: Brier (1950)) is simply $(f - Y)^2$, corresponding to the usual notion of squared-error loss. The logarithmic probability score (LPS: see Brier (1950) and Good (1952)) is defined as: $-Y \log(f) - (1 - Y) \log(1 - f)$.

For a sequence of probability forecasts and outcomes, $\{f_t, y_t\}, t = 1, \ldots, n$, these scores are calculated as:

$$\text{QPS} = \frac{1}{n} \sum_{t=1}^{n} 2 (f_t - y_t)^2$$

(4)

and:

$$\text{LPS} = -\frac{1}{n} \sum_{t=1}^{n} [y_t \ln f_t + (1 - y_t) \ln (1 - f_t)].$$

(5)

Clearly, QPS is proportional to the standard mean squared error measure commonly calculated for point forecasts. These two well-known measures of scoring probability forecasts have been used in economic applications by Diebold and Rudebusch (1989) and Anderson and Vahid (2001), inter alia.

---

4 As written, both scores give possible values on $[0, +\infty)$. 

4
3 Optimal combinations and loss functions

Both of the DGPs we consider in sections 4.1 and 4.2 can be written in the following form:

$$y_t = 1 \left( g(f_{1t}, f_{2t}) > v_t \right)$$

where $v_t \sim U(0,1)$ and $g(.)$ is some function of the individual probability forecasts $f_{1t}, f_{2t}$. That is, $\Pr (y_t = 1) = \Phi_u (g(f_{1t}, f_{2t}))$, where $\Phi_u$ is the cdf of a $U(0,1)$, and so in general terms the form of the optimal combination is:

$$C^*_t (f_{1t}, f_{2t}; \beta) = \Pr (y_t = 1 \mid f_{1t}, f_{2t}) = \max [g(f_{1t}, f_{2t}), 0]$$

where the max function is employed as the form of $g(f_{1t}, f_{2t})$ for one of our specific DGPs in the next section (CJ-DGP) does not ensure $g(f_{1t}, f_{2t}) > 0$, although for both DGPs, $g(f_{1t}, f_{2t}) < 1$. Thus when the DGP for $y_t$ is specified in terms of the forecasts $f_{1t}, f_{2t}$ as here, the optimal or correct form of combination is given by the DGP. It can also be noted that when the DGP relates $y$ to the individual information sets (as in the case of the L-DGP of the next section), one can alternatively establish that a form of combination is optimal if it is equivalent to the DGP ‘super model’ of $y$ on the variables contained in the combination of the individual information sets.

Consider selecting combination weights $\beta$ to minimize the LPS of the combined forecast $C^*_t (f_{1t}, f_{2t}; \beta)$. This is equivalent to maximum likelihood estimation of $\beta$ as the likelihood function (for iid data on a Bernoulli random variable with probability $C^*_t(\beta)$ that $Y_t = 1$) is given by:

$$L = \prod_{y_t=1} C^*_t (f_{1t}, f_{2t}; \beta)^{y_t} \prod_{y_t=0} [1 - C^*_t (f_{1t}, f_{2t}; \beta)]^{1-y_t}$$

and LPS is proportional to:

$$\ln L = - \sum [y_t \ln C^*_t (f_{1t}, f_{2t}; \beta) + (1 - y_t) \ln (1 - C^*_t (f_{1t}, f_{2t}; \beta))].$$

Alternatively, $\Pr (y_t = 1 \mid f_{1t}, f_{2t}) = C^*_t (f_{1t}, f_{2t}; \beta)$ can instead be thought of as the conditional expectation:

$$E (y_t \mid f_{1t}, f_{2t}) = C^*_t (f_{1t}, f_{2t}; \beta)$$

giving rise to the (generally nonlinear) regression model:

$$y_t = C^*_t (f_{1t}, f_{2t}; \beta) + \epsilon_t$$

(7)
to be estimated by (nonlinear) least squares. Nonlinear least squares estimation of (7) is clearly
equivalent to estimating $\beta$ on the basis of minimizing QPS.

Given that both (6) and (7) are correctly specified in the sense that they incorporate the
optimal form of combination, both maximum likelihood estimation of (6) and nonlinear least
squares estimation of (7) will provide consistent estimates of $\beta$ under standard assumptions. By
implication, the parameters of the optimal combination will be consistently estimated regardless
of whether we use QPS or LPS. Tests of forecast encompassing based on both QPS and LPS will
therefore be valid, in the sense that when $f_{2t}$ does not enter the optimal combination, both QPS
and LPS will indicate that $f_{1t}$ encompasses $f_{2t}$ in population. Conversely, when (6) and (7) are
not correctly specified, as when $C_t(.)$ is replaced by some other type of combination $C_t(\cdot)$, we
cannot establish in general that the inference we make concerning forecast encompassing will
match that made when the correct form of combination is used—even asymptotically.

Formally, for a loss function $L = L(e_{t+1})$, let $e_{j,t+1} = y_{t+1} - C_{jt} (y_{t+1}; \beta_j)$ denote the
combined forecast error associated with a form of combination given by $C_{jt}(\cdot)$, for $j = 1, 2, \ldots, J$
combination methods (in our case $J = 3$: LiOP, LoOP and KK). Conditional on the form of
combination, the optimal combination weights are given by:

$$
\beta^*_j = \arg \min_{\beta_j} E[L (e_{j,t+1} (\beta_j))]
$$

(8)

where the expectation is over the conditional distribution of $e_{t+1}$ given past forecasts and
outcomes. Whereas the optimal weight on $f_{2t}$ may be zero in the optimal form of combination,
indicating forecast encompassing of $f_{2t}$ by $f_{1t}$, there is no guarantee that the vector $\beta^*_j$, for
combination method $j$, will also indicate forecast encompassing.

Further, although:

$$
E[L (e_{1,t+1} (\beta^*))] \leq E[L (e_{j,t+1} (\beta^*_j))] \forall j
$$

(9)

where $e_{t+1} (\beta^*)$ denotes the forecast error using the optimal form of combination, we could have
any accuracy ranking among the different non-optimal combination methods, i.e.:

$$
E[L (e_{1,t+1} (\beta^*_1))] \leq E[L (e_{2,t+1} (\beta^*_2))] \leq \ldots \leq E[L (e_{J,t+1} (\beta^*_J))].
$$

(10)

Moreover, in practice the population parameters $\{\beta^*, \beta^*_1, \ldots, \beta^*_J\}$ will be replaced by estimates
of the combination weights. In that case, the inequality in (9) may no longer hold, and the
relationships between the different forms of combination in (10) based on population values may
be overturned. The rankings may also depend on the loss function, $L(\cdot)$. In what follows, we
report results for a single loss function, QPS, to focus on the form of the combination, \( C_t(\cdot) \), and the impact of uncertainty about \( \beta \). In the next section, we investigate the accuracy of optimal and non-optimal methods of forecast combination for both model-based and non-model-based types of DGP.

### 4 Accuracy of combination methods for two DGPs

#### 4.1 L-DGP

The ‘logit-based’ DGP (henceforth, L-DGP) is the DGP considered by Clements and Harvey (2006), which relates \( y_t \) to the forecast models’ explanatory variables via a logit-type transformation, and where the forecasts are also logit functions:

\[
y_t = \begin{cases} 1 & \left( \frac{\exp(\delta_1 X_{1t} + \delta_2 X_{2t})}{1 + \exp(\delta_1 X_{1t} + \delta_2 X_{2t})} > v_t \right) \\ \text{expression} & \end{cases}
\]

\[
f_{1t} = \frac{\exp(\theta_{11} X_{1t})}{1 + \exp(\theta_{11} X_{1t})}
\]

\[
f_{2t} = \frac{\exp(\theta_{12} X_{2t})}{1 + \exp(\theta_{12} X_{2t})}.
\]

This DGP reflects the common practice of using logit models to obtain probability forecasts.\(^5\) It allows for an investigation of forecast encompassing (when, say, \( \delta_2 = 0 \)), and the effects of parameter estimation uncertainty on predictive accuracy (i.e., the effects of replacing the population \( \theta_{ij} \) by the estimator \( \hat{\theta}_{ij} \)), although our focus will be on comparing the accuracy of different combination methods and also the effect of uncertainty when estimating the combination weights. If we assume that \( \{X_{1t}, X_{2t}\} \) has a normal distribution:

\[
\begin{bmatrix} X_{1t} \\ X_{2t} \end{bmatrix} \sim N \left( \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho_{X_1,X_2} \\ \rho_{X_1,X_2} & 1 \end{bmatrix} \right)
\]

then \( f_{1t} \) and \( f_{2t} \) are correlated provided \( \rho_{X_1,X_2} \neq 0 \).

Assuming \( X_{1t} \) and \( X_{2t} \) are scalars, the DGP can be re-written for \( y_t \) in terms of \( f_{1t} \) and \( f_{2t} \) as:

\[
y_t = 1 \left( \frac{\left( f_{1t} \right)^{\delta_1} \left( f_{2t} \right)^{\delta_2}}{1 + \left( f_{1t} \right)^{\delta_1} \left( f_{2t} \right)^{\delta_2}} > v_t \right)
\]

\(^5\)For example, logit models have been used to obtain forecasts of the probabilities of recessions from information sets that include leading indicators such as the yield curve; see, e.g., Estrella and Mishkin (1998).
so that the true Pr \( (y_t = 1) \) in terms of the individual forecasts is given by:

\[
Pr (y_t = 1) = \frac{\left( \frac{f_{1t}}{1-f_{1t}} \right)^{\delta_1} \left( \frac{f_{2t}}{1-f_{2t}} \right)^{\delta_2}}{1 + \left( \frac{f_{1t}}{1-f_{1t}} \right)^{\delta_1} \left( \frac{f_{2t}}{1-f_{2t}} \right)^{\delta_2}}
\]

(12)

which defines the optimal form of combination:

\[
C_t^* (f_{1t}, f_{2t}; \beta) = \frac{\left( \frac{f_{1t}}{1-f_{1t}} \right)^{\beta_1} \left( \frac{f_{2t}}{1-f_{2t}} \right)^{\beta_2}}{1 + \left( \frac{f_{1t}}{1-f_{1t}} \right)^{\beta_1} \left( \frac{f_{2t}}{1-f_{2t}} \right)^{\beta_2}}.
\]

(13)

This combination is equivalent to the ‘super model’ of \( y_t \) on \( X_{1t} \) and \( X_{2t} \) given by (11). It is a multiplicative combination of the odds ratios of the two models’ forecasts. In terms of the general DGP introduced in section 3, \( g(f_{1t}, f_{2t}) = C_t^* (f_{1t}, f_{2t}; \beta) \).

It is also readily apparent that (13) is the form of combination suggested by Kamstra and Kennedy (1998), which we termed the KK combination in section 2 (since (13) is equivalent to (3) with \( \alpha = 0 \)). Because (13) is the optimal form of combination, it follows from the arguments of section 3 that estimation of the combination weights \( \beta \) by QPS in the optimal combination:

\[
C_t^* (f_{1t}, f_{2t}; \beta) = \frac{\left( \frac{f_{1t}}{1-f_{1t}} \right)^{\beta_1} \left( \frac{f_{2t}}{1-f_{2t}} \right)^{\beta_2}}{1 + \left( \frac{f_{1t}}{1-f_{1t}} \right)^{\beta_1} \left( \frac{f_{2t}}{1-f_{2t}} \right)^{\beta_2}}
\]

(14)

will result in:

\[
\hat{\beta}_1 \overset{p}{\rightarrow} \frac{\delta_1}{\theta_{11}}, \quad \hat{\beta}_2 \overset{p}{\rightarrow} \frac{\delta_2}{\theta_{12}}.
\]

Based on this optimal type of combination, it is apparent that when \( \delta_2 = 0 \), so that \( y_t \) only depends on \( X_{1t} \) in the DGP, testing for forecast encompassing, in terms of whether \( \beta_2 = 0 \), will give the correct inference as \( \hat{\beta}_2 \overset{p}{\rightarrow} 0 \). We investigate whether the same is true for tests based on other forms of combination (such as LiOP and LoOP).

When the \( X_{it} \) are vectors of variables, we cannot in general establish the optimality of KK combination. For vector information sets, the log odds ratios are given by:

\[
\ln \left( \frac{f_{1t}}{1-f_{1t}} \right) = \theta'_{11} X_{1t}, \quad \ln \left( \frac{f_{2t}}{1-f_{2t}} \right) = \theta'_{12} X_{2t}
\]

where \( \theta_{11} \) and \( \theta_{12} \) are the population values for the two individual models. Then the logit
regression of \( y_t \) on the two log odds ratios results in:

\[
\Pr(y_t = 1) = \frac{\exp(\beta_1 \theta_{11} X_{1t} + \beta_2 \theta_{12} X_{2t})}{1 + \exp(\beta_1 \theta_{11} X_{1t} + \beta_2 \theta_{12} X_{2t})}.
\]  

(15)

This will only match the ‘super model’ of the logit regression of \( y_t \) on \( X_{1t} \) and \( X_{2t} \) (given by (11)) if the following are satisfied:

\[
\beta_1 \theta_{11} = \delta_1, \quad \beta_2 \theta_{12} = \delta_2.
\]  

(16)

When these restrictions hold, KK is the optimal form of combination.

### 4.1.1 Simulations for L-DGP

Because deriving analytical expressions appears to be intractable, we investigate the accuracy of forecast combinations using LiOP, LoOP and KK for the L-DGP by simulation. We ignore parameter estimation uncertainty in the forecast model parameters \( \theta \) to focus on uncertainty in the estimation of the combination weights. Simulated population parameters \( \theta \) were used, based on the average values across 10,000 replications of the logit model estimates on samples of size 10,000. We generated samples of size \( n = \{25, 50, 100, 200, \infty\} \) to estimate the combination weights for a given replication (the case \( n = \infty \) indicates the optimal population weights were used, which were simulated using a sample size of \( n = 10,000 \)). We then used these estimated weights to compute out-of-sample combined forecasts and the corresponding QPS for that replication, using 10,000 out-of-sample observations so as to approximate the population QPS associated with the given set of combination weights. The tables then report the mean and standard deviation of the simulated QPS values across replications. Here and throughout the paper, simulations were performed in Gauss 7.0 using 10,000 Monte Carlo replications.

Table 1 reports the results for scalar \( X_{1t} \) and \( X_{2t} \), \( \delta_1 = 1 \) and \( \delta_2 = 0 \) (so that encompassing holds in the KK combination), and for correlation parameter values \( \rho_{X_{1t}X_{2t}} = \{0.2, 0.5, 0.8\} \). Focusing firstly on the results for the population combination weights (\( n = \infty \)), unreported results confirm that the KK combination attaches zero weight to \( f_2 \), and the same is found to be true for LiOP and LoOP. All methods result in the same limiting value of QPS, regardless of the value of \( \rho_{X_{1t}X_{2t}} \). Even though KK is the optimal form of combination, when the combination weights are estimated using small samples of forecasts, the estimation uncertainty of this method is greater than that for LiOP, and the latter method is shown to be the most accurate in terms of QPS, with a lower mean and standard deviation obtained for all correlation values. The accuracy of the LoOP method lies between that of KK and LiOP in small samples, although
for some replications, difficulties were encountered in obtaining convergence of the combination weights to values that gave sensible results, highlighting a potential drawback of this approach in practical applications (these anomalous replications were resampled to obtain the simulation results reported). Unreported results for a non-encompassing case ($\delta_1 = 0.5$ and $\delta_2 = 0.5$) again show that all three combinations give essentially the same limiting value of QPS, but that LiOP is most accurate for small $n$.

We also report results when the two models’ explanatory variables are vectors. In this case, combining forecasts will in general be less accurate than building a ‘super model’, and KK is not necessarily the optimal type of combination. To explore the relative performances of the three combination methods in these cases, we consider the following generalization of (11). Let $X_{1t}$ and $X_{2t}$ be $(2 \times 1)$ vectors:

$$
X_{1t} = \begin{bmatrix} X_{11t} \\ X_{12t} \end{bmatrix}, \quad X_{2t} = \begin{bmatrix} X_{21t} \\ X_{22t} \end{bmatrix}
$$

so that:

$$
y_t = 1 \left( \frac{\exp(\delta_{11} X_{11t} + \delta_{12} X_{12t} + \delta_{21} X_{21t} + \delta_{22} X_{22t})}{1 + \exp(\delta_{11} X_{11t} + \delta_{12} X_{12t} + \delta_{21} X_{21t} + \delta_{22} X_{22t})} > v_t \right)
$$

$$
f_{1t} = \frac{\exp(\theta_{11} X_{11t} + \theta_{12} X_{12t})}{1 + \exp(\theta_{11} X_{11t} + \theta_{12} X_{12t})},
$$

$$
f_{2t} = \frac{\exp(\theta_{21} X_{21t} + \theta_{22} X_{22t})}{1 + \exp(\theta_{21} X_{21t} + \theta_{22} X_{22t})}.
$$

We assume that $X_t = (X_{1t}^T, X_{2t}^T)^T$ is jointly normally distributed, with zero means and unit variances, and all correlations are equal to zero except for $\rho_{X_{11},X_{21}}$. A non-zero value of $\rho_{X_{11},X_{21}}$ generates correlated forecasts, and in addition each forecast has an idiosyncratic component ($X_{12t}$ and $X_{22t}$ respectively), independent of the other variables.

Table 2 reports results for the design parameters $\{\delta_{11} = 2, \delta_{12} = 5, \delta_{21} = 4, \delta_{22} = 1\}$, which are chosen such that the restrictions given by (16) for KK to be the optimal form of combination are not satisfied. Despite this, in the limiting case KK is still the most accurate forecast combination, followed by LiOP and lastly LoOP, with the differences in accuracy being large—of the order of 40% for $\rho_{X_{11},X_{21}} = 0.2$. For small $n$ the performance of LiOP is more competitive, with the estimation uncertainty having a considerably greater impact on the accuracy of KK than for LiOP. While the LiOP mean QPS remains higher than that for KK, the standard deviation is substantially smaller, and there is little to choose between the methods.

Finally, we consider a specification where bias is introduced into the forecasts. The DGP is
now:

\[ y_t = \begin{cases} 1 & \left( \frac{\exp(\delta_{11}X_{11t} + \delta_{12}X_{12t} + \delta_{21}X_{21t} + \delta_{22}X_{22t} + \delta_3Z_t)}{1 + \exp(\delta_{11}X_{11t} + \delta_{12}X_{12t} + \delta_{21}X_{21t} + \delta_{22}X_{22t} + \delta_3Z_t)} > v_t \right) \\ \end{cases} \]

\[ f_{1t} = \frac{\exp(\theta_{11}X_{11t} + \theta_{12}X_{12t})}{1 + \exp(\theta_{11}X_{11t} + \theta_{12}X_{12t})} \]

\[ f_{2t} = \frac{\exp(\theta_{21}X_{21t} + \theta_{22}X_{22t})}{1 + \exp(\theta_{21}X_{21t} + \theta_{22}X_{22t})} \]

with \( X_t = (X'_{1t} \quad X'_{2t} \quad Z_t)' \) jointly normally distributed with the \( X \)'s as defined above, and with \( Z_t \) having a unit variance and non-zero mean \( \mu_Z \). The inclusion of the non-zero mean variable \( Z_t \) in the DGP but not the forecasts ensures that the forecasts will be biased. We set \( \delta_3 = 1, \mu_Z = 1, \rho_{X_{12},X_{21}} = 0.5 \), and consider the parameter combinations \( (\rho_{X_{11},Z},\rho_{X_{21},Z}) = \{(0,0),(0,0.5),(0.5,0),(0.5,0.5)\} \), for \( \{\delta_{11} = 2, \delta_{12} = 5, \delta_{21} = 4, \delta_{22} = 1\} \). Table 3 again indicates that KK is markedly more accurate when the combination weights are known (and optimal for the combination method) but that LiOP combinations can be more accurate when the weights are estimated using small samples of forecasts.

### 4.2 CJ-DGP

A simple DGP for probability forecasts is:

\[ y_t = \begin{cases} 1 & \left( u_{1t}^{-\lambda}u_{2t}^{(1-\lambda)} > v_t + c \right) \\ \end{cases} \]

\[ f_{1t} = u_{1t} \]

\[ f_{2t} = u_{2t} \]

The forecasts \( u_{1t} \) and \( u_{2t} \) are drawn from the Cook and Johnson (1981) bivariate distribution:

\[ f(u_{1t},u_{2t}) = \frac{\delta + 1}{\delta} \left( u_{1t}u_{2t} \right)^{-\frac{\delta + 1}{2}} \left( u_{1t}^{-\frac{1}{\delta}} + u_{2t}^{-\frac{1}{\delta}} - 1 \right)^{-\frac{\delta + 2}{2}}. \]

This has the property that the random variables are correlated, and have marginal distributions which are \( U(0,1) \). The degree of correlation between \( u_{1t} \) and \( u_{2t} \), \( 0 \leq \rho \leq 1 \), is determined by the parameter \( \delta \), with \( \rho \to 0 \) as \( \delta \to \infty \) and \( \rho \to 1 \) as \( \delta \to 0 \). The equation for \( y_t \) indicates that \( y_t \) depends on \( f_{1t} \) and \( f_{2t} \), such that when the constant \( c (|c| < 1) \) is zero, \( \Pr(y_t = 1) = u_{1t}^{-\lambda}u_{2t}^{(1-\lambda)} \), as \( v_t \) is \( U(0,1) \), drawn independently of \( u_{1t} \) and \( u_{2t} \) for all \( t \).\(^6\) The form of the DGP ensures that the forecasts have the characteristics of probability forecasts (i.e., they are defined on the

\(^6\)Note that \( y_t \) can also be written as \( y_t = 1(u_{1t} > w_t) \), where \( w_t \sim U(c,1+c) \).
unit interval), while also allowing for contemporaneous correlation among the predictions (a property frequently observed in practice). By setting $\lambda = 1$, $\Pr(y_t = 1) = u_{1t}$ and so depends only on $f_{1t}$, so that $f_{2t}$ conveys no useful information given that we have $f_{1t}$. When in addition $c \neq 0$, $f_{1t}$ is biased (as is $f_{2t}$). It is straightforward to simulate $\{y_t, f_{1t}, f_{2t}\}$ from (17), and the form of the DGP also allows analytical results to be obtained.

Since from (17):

$$
\Pr(y_t = 1 \mid u_{1t}, u_{2t}) = \Pr\left(u_{1t}^{\lambda}u_{2t}^{(1-\lambda)} > v_t + c\right)
= \Pr\left(v_t < u_{1t}^{\lambda}u_{2t}^{(1-\lambda)} - c\right)
= \max(u_{1t}^{\lambda}u_{2t}^{(1-\lambda)} - c, 0)
$$

it follows immediately that the optimal type of combination is given by, for $c \neq 0$:

$$
C^*_t(f_{1t}, f_{2t}; \lambda, c) = \max(f_{1t}^{\lambda}f_{2t}^{(1-\lambda)} - c, 0)
$$

with $f_{1t} = u_{1t}$ and $f_{2t} = u_{2t}$, or simply $C^*_t(f_{1t}, f_{2t}; \lambda) = f_{1t}^{\lambda}f_{2t}^{(1-\lambda)}$ when $c = 0$ as $u_{1t}^{\lambda}u_{2t}^{(1-\lambda)} > 0$.

In terms of the general DGP of section 3, $g(f_{1t}, f_{2t}) = f_{1t}^{\lambda}f_{2t}^{(1-\lambda)} - c$.

When $\lambda = 1$, we find that the values of the parameters $\{\lambda, c\}$ that minimize QPS, i.e.:

$$
\arg \min_{\{\lambda, c\}} \left[y_t - \max(u_{1t}^{\lambda}u_{2t}^{(1-\lambda)} - c, 0)\right]^2
$$

has $\hat{\lambda} \overset{p}{\to} 1$, so that $u_{2t}$ receives zero weight, thus the DGP given by (17) exhibits forecast encompassing when $\lambda = 1$, as expected, when the optimal type of combination is used. As we show below, this result does not hold in general, underlining the dependence of the finding of forecast encompassing on the form of combination.

### 4.2.1 Simulations for CJ-DGP

We simulated data from (17) and (18) and computed the out-of-sample QPS mean and standard deviation for each combination method, where the combination parameters were estimated using $n = \{25, 50, 100, 200, \infty\}$ observations (the case $n = \infty$ again indicates use of the optimal population weights, simulated using $n = 10,000$). The simulations were conducted using the same methodology as for the L-DGP, and the settings employed were $\lambda = \{0.5, 1\}$, where $\lambda = 1$ corresponds to $f_1$ encompassing $f_2$ using the optimal type of combination, as well as $\rho \equiv \text{Corr}(u_{1t}, u_{2t}) = \{0, 0.8\}$ and $c = \{0, 0.5\}$ to allow for correlated and biased forecasts (when $\rho \neq 0$ and $c \neq 0$ respectively). The results are recorded in Tables 4 and 5.
Consider first the case where the population weights are used. When $\lambda = 1$, so that $f_1$ encompasses $f_2$ using the optimal combination, unreported results show that all three combinations attach zero weight to $f_2$ when $c = 0$ and/or $\rho = 0$, but not in the case where the forecasts are both correlated and biased. Instead, when both $\rho \neq 0$ and $c \neq 0$, we find $\beta_2$ values for LiOP, LoOP and KK of $-0.106$, $0.015$ and $0.082$, respectively, although the LoOP value of $0.015$ is found to be insignificantly different from zero across replications. This highlights the result that when non-optimal combination methods are employed, an inference of forecast encompassing can depend on the form of the combination adopted. In terms of the most accurate combination, when we abstract from estimation uncertainty ($n = \infty$), LoOP is always at least as good as the other two methods. This is true whether $\lambda = 1$ or $\lambda = 0.5$, although the differences are small in magnitude. The form of the DGP is such that the entries in the tables for LiOP and LoOP can be calculated analytically for some combinations of values of $\{c, \rho\}$. For example, we can show that $QPS_{LiOP} > QPS_{LoOP}$ for the encompassing specification $\lambda = 1$, $c = 0.5$ and $\rho = 0.7$.

When we allow for finite sample estimation uncertainty, the accuracy of LiOP relative to LoOP and KK improves considerably, especially for the smaller sample sizes. For small $n$, the LoOP combination method becomes very susceptible to the effects of estimation uncertainty, with a large QPS mean and standard deviation. Moreover, this method again exhibited convergence problems in the simulations, to a greater extent even than was observed for the L-DGP.

In short, our results indicate that for large samples of forecasts, corresponding to the known combination weight case in the limit, LiOP is dominated by KK for the L-DGP, and by LoOP for the CJ-DGP. However, estimation uncertainty has a dramatic effect on the accuracy of KK and LoOP combination forecasts when the sample of forecasts is small, and in such small $n$ cases, the LiOP combination method may be expected to generate forecasts on a par with LoOP and KK.

## 5 Empirical illustration

Our empirical illustration of the use of different types of combinations of probability forecasts is based on recession probability forecasting, where recessionary periods are those defined by the NBER’s Business Cycle Dating Committee. Two logit forecasting models are considered—one

---

7 An Appendix to this paper is available from the corresponding author’s webpage. This outlines the calculation of QPS for LiOP and LoOP for the CJ-DGP.

uses the ‘spread’ or yield curve, and the other uses average weekly hours.\textsuperscript{9} The usefulness of
the spread for predicting US recessions has been established by Estrella and Mishkin (1998),
Anderson and Vahid (2001) and Hamilton and Kim (2000), and is a component of the Confer-
ence Board’s Composite Leading Indicator (CLI). Average weekly hours is also a component of
the CLI.\textsuperscript{10} Stock and Watson (2003, Table 3, p.77) found that hours was on a par with spread
in terms of predicting output growth around the time of the 2001 recession.

After constructing the annual percentage change in hours, we have monthly data from
1965:01 to 2007:02. We estimate logit models from around 1965 (the precise start point depends
on the lag selected for the leading indicator) to a fixed point, and then use the in-sample
predicted recession probabilities over this period to estimate optimal weights for LiOP, LoOP
and KK forecast combinations, where by optimal is meant weights that minimize QPS. We
then generate out-of-sample 1-step ahead forecast recession probabilities, without updating
the logit model parameter estimates as the forecast origin is moved forward in time (usually
known as a ‘fixed forecasting scheme’ in the literature). Using the weights estimated for the in-
sample predictions, the three forecast combinations are constructed out-of-sample. The forecast
combinations are evaluated out-of-sample by QPS, and we report pairwise comparisons of equal
forecast accuracy using the modified Diebold and Mariano (1995) statistic of Harvey, Leybourne
and Newbold (1997). We consider two splits into the in-sample and out-of-sample forecast
periods. For the first, the first period of the out-of-sample period begins in 1986:01, and for
the second, 1996:01. For each split, the appropriate lag of the indicator variable (the spread or
the hours variable) is selected to maximize the model likelihood value, in-sample.

A possible criticism is that the out-of-sample combination weights are calculated from in-
sample predictions rather than out-of-sample forecasts. This could be remedied by introducing
an additional data split, so that between the current in- and out-of-sample periods there is
a period over which out-of-sample forecasts are generated, which are then used to estimate
weights. It was felt this would complicate the illustration with little gain in terms of insight.
The in-sample fits are not the result of elaborate model search/specification procedures—in
each case we simply choose the best lag from a maximum lag of twelve—so there is little reason
to suppose the in-sample fits are very different from their out-of-sample counterparts.

For both in-sample periods, a nine month lag was selected for the spread, and a one month

\textsuperscript{9}The spread data is the 10-year treasury constant maturity rate (series identifier GS10) less the 3-
month treasury bill at secondary market rate (series identifier TB3MS). The hours variable is average weekly
hours, total private industries (series identifier AWHNONAG). The data were taken from the FRED website
http://research.stlouisfed.org/fred2.

\textsuperscript{10}See http://www.conference-board.org/economics/bci/serieslist01.cfm. The Conference Board uses
average weekly hours in manufacturing.
lag for hours. Both leading indicators have significant in-sample predictive power for recessions for both in-sample periods. We obtained pseudo-$R^2$ statistics of 0.288 and 0.290 for the spread, and of 0.131 and 0.138 for hours, using the measure of Estrella and Mishkin (1998). Their pseudo-$R^2$ statistic is defined as:

$$R^2 = 1 - \left( \frac{ll_u}{ll_r} \right)^{-\left( \frac{2}{ll_r} \right)}$$

where $ll_u$ and $ll_r$ are the unrestricted maximised value of the log likelihood, and the value imposing the restriction that the slope coefficients are zero. In all cases we rejected the null that the slope coefficient (on the spread, or on hours) could be omitted.

Table 6 records the values obtained for QPS out-of-sample using the individual indicator logit models, and the LiOP, LoOP and KK combinations of these forecasts. Of the two individual model forecasts, the spread model is more accurate than that using hours, matching the higher $R^2$ in-sample for the spread, but combination improves upon the best individual set of forecasts, with KK performing markedly better than LiOP and LoOP in both forecast periods. Table 7 indicates that the spread model forecasts are statistically significantly more accurate than those of the hours model at the 1% level for the forecast period 1986:01 to 2007:02, and at the 10% level for the shorter forecast period. The KK forecast combination is significantly more accurate than both LiOP and LoOP at any level of significance for the second forecast period, and at the 10% level for the first forecast period. This result is consistent with the simulation findings of section 4.1 for logit-based DGPs, given that the combination weights are estimated using samples of size in excess of $n = 200$.

### 6 Conclusions

Forecasts of economic and financial variables that take the form of probabilities are becoming increasingly common. As there is an extensive literature in economic and management science that suggests that forecast combination can improve on the best individual forecast, we consider ways of combining probability forecasts. For probability forecasts, the justification for considering linear combinations is weaker than in the case of standard point forecasts, and we consider three combination schemes, one of which is linear combination. Our simulation results indicate that linear combination may work reasonably well when the optimal combination weights are estimated for a small sample of forecasts, but for moderate sample sizes, combinations such as those proposed by Kamstra and Kennedy (1998) are likely to prove superior. This is borne out in our empirical illustration based on combining US recession probability forecasts, where we
have reasonable numbers of forecast observations, and the Kamstra-Kennedy combination is found to be more accurate than both the individual forecasts and the other types of combination we consider, including linear combination.

For standard point forecasts, where the literature focuses almost exclusively on linear combinations of forecasts and squared error loss, the notion of forecast encompassing is well-defined, and is a useful test of predictive accuracy to be reported alongside related tests, such as tests of whether two sets of forecasts are equally accurate (e.g., Diebold and Mariano, 1995). We have shown that for probability forecasts there are a number of types of combination that might be considered, and that forecast encompassing may hold for one type of combination, but not another. Thus while probability forecast encompassing tests can still be conducted in a linear combination setting (see Clements and Harvey, 2006), when a broader range of combination methods is admitted, the notion of forecast encompassing appears to be less useful due to the dependence on the form of the combination. For this reason we have mainly focused on the relative accuracy of the different types of combination, paying particular attention to the need to estimate combination weights using what may on occasion be relatively small samples of forecasts.
References


Table 1. Means and standard deviations of QPS across Monte Carlo replications:
L-DGP, $\delta_1 = 1$, $\delta_2 = 0$

<table>
<thead>
<tr>
<th>n</th>
<th>$\rho_{X_1,X_2} = 0.2$</th>
<th>$\rho_{X_1,X_2} = 0.5$</th>
<th>$\rho_{X_1,X_2} = 0.8$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>S.d.</td>
<td>Mean</td>
</tr>
<tr>
<td>25</td>
<td>LiOP</td>
<td>0.461</td>
<td>0.042</td>
</tr>
<tr>
<td></td>
<td>LoOP</td>
<td>0.469</td>
<td>0.067</td>
</tr>
<tr>
<td></td>
<td>KK</td>
<td>0.486</td>
<td>0.076</td>
</tr>
<tr>
<td>50</td>
<td>LiOP</td>
<td>0.437</td>
<td>0.021</td>
</tr>
<tr>
<td></td>
<td>LoOP</td>
<td>0.438</td>
<td>0.022</td>
</tr>
<tr>
<td></td>
<td>KK</td>
<td>0.444</td>
<td>0.032</td>
</tr>
<tr>
<td>100</td>
<td>LiOP</td>
<td>0.425</td>
<td>0.010</td>
</tr>
<tr>
<td></td>
<td>LoOP</td>
<td>0.425</td>
<td>0.011</td>
</tr>
<tr>
<td></td>
<td>KK</td>
<td>0.427</td>
<td>0.013</td>
</tr>
<tr>
<td>200</td>
<td>LiOP</td>
<td>0.419</td>
<td>0.006</td>
</tr>
<tr>
<td></td>
<td>LoOP</td>
<td>0.419</td>
<td>0.006</td>
</tr>
<tr>
<td></td>
<td>KK</td>
<td>0.420</td>
<td>0.007</td>
</tr>
<tr>
<td>$\infty$</td>
<td>LiOP</td>
<td>0.413</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>LoOP</td>
<td>0.413</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>KK</td>
<td>0.413</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 2. Means and standard deviations of QPS across Monte Carlo replications:
L-DGP, $\delta_{11} = 2$, $\delta_{12} = 5$, $\delta_{21} = 4$, $\delta_{22} = 1$

<table>
<thead>
<tr>
<th>n</th>
<th>$\rho_{X_{11},X_{21}} = 0.2$</th>
<th>$\rho_{X_{11},X_{21}} = 0.5$</th>
<th>$\rho_{X_{11},X_{21}} = 0.8$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>S.d.</td>
<td>Mean</td>
</tr>
<tr>
<td>25</td>
<td>LiOP</td>
<td>0.209</td>
<td>0.022</td>
</tr>
<tr>
<td></td>
<td>LoOP</td>
<td>0.243</td>
<td>0.076</td>
</tr>
<tr>
<td></td>
<td>KK</td>
<td>0.191</td>
<td>0.061</td>
</tr>
<tr>
<td>50</td>
<td>LiOP</td>
<td>0.196</td>
<td>0.010</td>
</tr>
<tr>
<td></td>
<td>LoOP</td>
<td>0.219</td>
<td>0.016</td>
</tr>
<tr>
<td></td>
<td>KK</td>
<td>0.152</td>
<td>0.036</td>
</tr>
<tr>
<td>100</td>
<td>LiOP</td>
<td>0.190</td>
<td>0.005</td>
</tr>
<tr>
<td></td>
<td>LoOP</td>
<td>0.211</td>
<td>0.008</td>
</tr>
<tr>
<td></td>
<td>KK</td>
<td>0.134</td>
<td>0.016</td>
</tr>
<tr>
<td>200</td>
<td>LiOP</td>
<td>0.188</td>
<td>0.003</td>
</tr>
<tr>
<td></td>
<td>LoOP</td>
<td>0.207</td>
<td>0.004</td>
</tr>
<tr>
<td></td>
<td>KK</td>
<td>0.126</td>
<td>0.008</td>
</tr>
<tr>
<td>$\infty$</td>
<td>LiOP</td>
<td>0.185</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>LoOP</td>
<td>0.204</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>KK</td>
<td>0.119</td>
<td>-</td>
</tr>
</tbody>
</table>
Table 3. Means and standard deviations of QPS across Monte Carlo replications:
L-DGP, biased case, $\delta_{11} = 2$, $\delta_{12} = 5$, $\delta_{21} = 4$, $\delta_{22} = 1$, $\rho_{X_{11},X_{21}} = 0.5$

<table>
<thead>
<tr>
<th>$n$</th>
<th>LiOP Mean (S.d.)</th>
<th>LoOP Mean (S.d.)</th>
<th>KK Mean (S.d.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>25</td>
<td>$\rho_{X_{11},z} = 0$</td>
<td>$\rho_{X_{21},z} = 0$</td>
<td>$\rho_{X_{11},z} = 0.5$</td>
</tr>
<tr>
<td></td>
<td>Mean (S.d.)</td>
<td>Mean (S.d.)</td>
<td>Mean (S.d.)</td>
</tr>
<tr>
<td>25</td>
<td>0.214 0.024</td>
<td>0.217 0.024</td>
<td>0.202 0.022</td>
</tr>
<tr>
<td></td>
<td>0.232 0.045</td>
<td>0.238 0.053</td>
<td>0.219 0.035</td>
</tr>
<tr>
<td></td>
<td>0.223 0.064</td>
<td>0.223 0.064</td>
<td>0.209 0.060</td>
</tr>
<tr>
<td>50</td>
<td>0.200 0.011</td>
<td>0.203 0.011</td>
<td>0.189 0.010</td>
</tr>
<tr>
<td></td>
<td>0.214 0.014</td>
<td>0.218 0.015</td>
<td>0.201 0.014</td>
</tr>
<tr>
<td></td>
<td>0.182 0.035</td>
<td>0.181 0.035</td>
<td>0.170 0.035</td>
</tr>
<tr>
<td>100</td>
<td>0.195 0.006</td>
<td>0.197 0.006</td>
<td>0.183 0.006</td>
</tr>
<tr>
<td></td>
<td>0.207 0.007</td>
<td>0.210 0.007</td>
<td>0.194 0.007</td>
</tr>
<tr>
<td></td>
<td>0.164 0.017</td>
<td>0.162 0.016</td>
<td>0.152 0.017</td>
</tr>
<tr>
<td>200</td>
<td>0.192 0.004</td>
<td>0.194 0.004</td>
<td>0.181 0.004</td>
</tr>
<tr>
<td></td>
<td>0.203 0.004</td>
<td>0.207 0.004</td>
<td>0.191 0.004</td>
</tr>
<tr>
<td></td>
<td>0.155 0.008</td>
<td>0.154 0.008</td>
<td>0.143 0.008</td>
</tr>
<tr>
<td>$\infty$</td>
<td>0.189 -</td>
<td>0.191 -</td>
<td>0.178 -</td>
</tr>
<tr>
<td></td>
<td>0.200 -</td>
<td>0.203 -</td>
<td>0.187 -</td>
</tr>
<tr>
<td></td>
<td>0.148 -</td>
<td>0.148 -</td>
<td>0.137 -</td>
</tr>
</tbody>
</table>

Table 4. Means and standard deviations of QPS across Monte Carlo replications:
CJ-DGP, $\lambda = 1$

<table>
<thead>
<tr>
<th>$n$</th>
<th>LiOP Mean (S.d.)</th>
<th>LoOP Mean (S.d.)</th>
<th>KK Mean (S.d.)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$c = 0$, $\rho = 0$</td>
<td>$c = 0$, $\rho = 0.8$</td>
<td>$c = 0.5$, $\rho = 0$</td>
</tr>
<tr>
<td></td>
<td>Mean (S.d.)</td>
<td>Mean (S.d.)</td>
<td>Mean (S.d.)</td>
</tr>
<tr>
<td>25</td>
<td>0.372 0.037</td>
<td>0.375 0.039</td>
<td>0.206 0.027</td>
</tr>
<tr>
<td></td>
<td>0.417 0.308</td>
<td>0.397 0.203</td>
<td>0.377 0.792</td>
</tr>
<tr>
<td></td>
<td>0.406 0.073</td>
<td>0.405 0.070</td>
<td>0.236 0.061</td>
</tr>
<tr>
<td>50</td>
<td>0.352 0.017</td>
<td>0.352 0.017</td>
<td>0.189 0.013</td>
</tr>
<tr>
<td></td>
<td>0.360 0.107</td>
<td>0.356 0.023</td>
<td>0.276 0.524</td>
</tr>
<tr>
<td></td>
<td>0.365 0.034</td>
<td>0.365 0.034</td>
<td>0.207 0.035</td>
</tr>
<tr>
<td>100</td>
<td>0.342 0.009</td>
<td>0.343 0.009</td>
<td>0.183 0.007</td>
</tr>
<tr>
<td></td>
<td>0.344 0.011</td>
<td>0.344 0.010</td>
<td>0.203 0.232</td>
</tr>
<tr>
<td></td>
<td>0.347 0.013</td>
<td>0.347 0.014</td>
<td>0.190 0.016</td>
</tr>
<tr>
<td>200</td>
<td>0.338 0.005</td>
<td>0.338 0.006</td>
<td>0.180 0.004</td>
</tr>
<tr>
<td></td>
<td>0.338 0.006</td>
<td>0.338 0.006</td>
<td>0.176 0.025</td>
</tr>
<tr>
<td></td>
<td>0.340 0.007</td>
<td>0.340 0.007</td>
<td>0.183 0.007</td>
</tr>
<tr>
<td>$\infty$</td>
<td>0.333 -</td>
<td>0.333 -</td>
<td>0.177 -</td>
</tr>
<tr>
<td></td>
<td>0.333 -</td>
<td>0.333 -</td>
<td>0.168 -</td>
</tr>
<tr>
<td></td>
<td>0.333 -</td>
<td>0.333 -</td>
<td>0.176 -</td>
</tr>
</tbody>
</table>
Table 5. Means and standard deviations of QPS across Monte Carlo replications: CJ-DGP, $\lambda = 0.5$

<table>
<thead>
<tr>
<th>n</th>
<th>LiOP</th>
<th>LoOP</th>
<th>KK</th>
<th>LiOP</th>
<th>LoOP</th>
<th>KK</th>
<th>LiOP</th>
<th>LoOP</th>
<th>KK</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>S.d.</td>
<td>Mean</td>
<td>S.d.</td>
<td>Mean</td>
<td>S.d.</td>
<td>Mean</td>
<td>S.d.</td>
<td>Mean</td>
</tr>
<tr>
<td>25</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.448</td>
<td>0.043</td>
<td>0.397</td>
<td>0.042</td>
<td>0.135</td>
<td>0.021</td>
<td>0.198</td>
<td>0.029</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.460</td>
<td>0.199</td>
<td>0.415</td>
<td>0.177</td>
<td>0.166</td>
<td>0.238</td>
<td>0.287</td>
<td>0.546</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.477</td>
<td>0.077</td>
<td>0.426</td>
<td>0.071</td>
<td>0.155</td>
<td>0.051</td>
<td>0.223</td>
<td>0.060</td>
<td></td>
</tr>
<tr>
<td>50</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.423</td>
<td>0.021</td>
<td>0.372</td>
<td>0.019</td>
<td>0.124</td>
<td>0.010</td>
<td>0.181</td>
<td>0.013</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.415</td>
<td>0.026</td>
<td>0.375</td>
<td>0.023</td>
<td>0.135</td>
<td>0.072</td>
<td>0.207</td>
<td>0.266</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.433</td>
<td>0.034</td>
<td>0.385</td>
<td>0.034</td>
<td>0.138</td>
<td>0.030</td>
<td>0.196</td>
<td>0.034</td>
<td></td>
</tr>
<tr>
<td>100</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.410</td>
<td>0.011</td>
<td>0.361</td>
<td>0.010</td>
<td>0.119</td>
<td>0.006</td>
<td>0.175</td>
<td>0.007</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.401</td>
<td>0.011</td>
<td>0.362</td>
<td>0.011</td>
<td>0.119</td>
<td>0.025</td>
<td>0.174</td>
<td>0.033</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.414</td>
<td>0.013</td>
<td>0.366</td>
<td>0.015</td>
<td>0.126</td>
<td>0.015</td>
<td>0.180</td>
<td>0.016</td>
<td></td>
</tr>
<tr>
<td>200</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.405</td>
<td>0.006</td>
<td>0.356</td>
<td>0.006</td>
<td>0.117</td>
<td>0.004</td>
<td>0.172</td>
<td>0.005</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.395</td>
<td>0.006</td>
<td>0.356</td>
<td>0.006</td>
<td>0.112</td>
<td>0.007</td>
<td>0.166</td>
<td>0.007</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.407</td>
<td>0.007</td>
<td>0.359</td>
<td>0.007</td>
<td>0.120</td>
<td>0.007</td>
<td>0.173</td>
<td>0.007</td>
<td></td>
</tr>
<tr>
<td>∞</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.399</td>
<td>-</td>
<td>0.351</td>
<td>-</td>
<td>0.115</td>
<td>-</td>
<td>0.169</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.389</td>
<td>-</td>
<td>0.350</td>
<td>-</td>
<td>0.107</td>
<td>-</td>
<td>0.160</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.400</td>
<td>-</td>
<td>0.352</td>
<td>-</td>
<td>0.115</td>
<td>-</td>
<td>0.167</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>
Table 6. QPS for out-of-sample recession probability forecasts

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Spread</td>
<td>0.0871</td>
<td>0.0733</td>
</tr>
<tr>
<td>Hours</td>
<td>0.108</td>
<td>0.0931</td>
</tr>
<tr>
<td>LiOP</td>
<td>0.0821</td>
<td>0.0673</td>
</tr>
<tr>
<td>LoOP</td>
<td>0.0844</td>
<td>0.0711</td>
</tr>
<tr>
<td>KK</td>
<td>0.0764</td>
<td>0.0521</td>
</tr>
</tbody>
</table>

Table 7. Tests for equal QPS-forecast accuracy

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Test stat.</td>
<td>p-value</td>
</tr>
<tr>
<td>Spread &amp; Hours</td>
<td>-2.89</td>
<td>0.00213</td>
</tr>
<tr>
<td>Spread &amp; LiOP</td>
<td>3.80</td>
<td>1.00</td>
</tr>
<tr>
<td>Spread &amp; LoOP</td>
<td>4.58</td>
<td>1.00</td>
</tr>
<tr>
<td>Spread &amp; KK</td>
<td>2.86</td>
<td>0.998</td>
</tr>
<tr>
<td>LiOP &amp; KK</td>
<td>1.49</td>
<td>0.932</td>
</tr>
<tr>
<td>LoOP &amp; KK</td>
<td>2.18</td>
<td>0.985</td>
</tr>
<tr>
<td>LiOP &amp; LoOP</td>
<td>-2.67</td>
<td>0.00398</td>
</tr>
</tbody>
</table>

Notes: The test statistic is the modified Diebold-Mariano statistic of the null of equal accuracy, assessed by QPS. For x & y, a p-value of less than 0.05 indicates rejection of the null of equal accuracy in favour of x being more accurate at the 5% level, and a p-value greater than 0.95 indicates rejection of the null of equal accuracy in favour of y being more accurate at the 5% level.