Department of Economics

Forecasting Using Predictive Likelihood Model Averaging

George Kapetanios, Vincent Labhard and Simon Price

Working Paper No. 567 September 2006 ISSN 1473-0278

Queen Mary
University of London

Forecasting Using Predictive Likelihood Model Averaging

George Kapetanios*
Queen Mary, University of London and
Bank of England

Vincent Labhard†
Bank of England

Simon Price‡
Bank of England and
City University

August 3, 2006

Abstract

Recently, there has been increasing interest in forecasting methods that utilise large datasets. We explore the possibility of forecasting with model averaging using the out-of-sample forecasting performance of various models in a frequentist setting, using the predictive likelihood. We apply our method to forecasting UK inflation and find that the new method performs well; in some respects it outperforms other averaging methods.

Keywords: forecasting, inflation, Bayesian model averaging, Akaike criterion, forecast combining

JEL: C110, C150, C530

1 Introduction

Recently, there has been increasing interest in forecasting methods that utilise large datasets. There are two main methodologies that can be applied: factor modelling where

*Department of Economics, Queen Mary, University of London, Mile End Road, London E1 4NS. Tel. No.: 00-44-20-78825097. Fax No.: 00-44-20-89833580. Email: G.Kapetanios@qmul.ac.uk.

†Monetary Analysis, Bank of England, Threadneedle Str. London EC2R 8AH. Tel. No. 00-44-20-76013267. Email: vincent.labhard@bankofengland.co.uk. This paper represents the views and analysis of the authors and should not be thought to represent those of the Bank of England or Monetary Policy Committee members.

‡Monetary Analysis, Bank of England, Threadneedle Str. London EC2R 8AH. Tel. No. 00-44-20-76014259. Email: simon.price@bankofengland.co.uk.

factor summaries of the dataset are used for forecasting; and forecast combination or averaging, where information in the form of forecasts from many forecasting models, typically simple and incomplete, are combined in some manner.

We focus on forecast combining. This idea grew out of the observation that for whatever reasons, combining forecasts produced a forecast superior to any element in the combined set. Of course, if it were possible to identify the correctly specified model and the data generating process (DGP) is unchanging, then the forecast from the correct model should be used. But the weight of evidence dating back to Bates and Granger (1969) and Newbold and Granger (1974) reveals that combinations of forecasts often outperform individual forecasts. Models may be incomplete, in different ways; they employ different information sets. Forecasts might be biased, and biases can offset each other. Even if forecasts are unbiased, there will be covariances between forecasts which should be taken into account. Thus combining misspecified models may, and often will, improve the forecast.

In this context forecast combining is viewed as a stop-gap measure that works in practice but would be surpassed by an appropriate model that addressed the underlying misspecification. A further practical problem is that with standard combining methods, based on regressions, the forecast weights can only be reliably constructed for a relatively small number of models. Nevertheless, given that the true DGP may involve a vast number of variables, it is clear that forecast combination is a route into the combining of information: and this is how it is interpreted in the literature relating to large data sets.

There is an alternative way of looking at this problem, most clearly seen from a Bayesian perspective. Here, it is assumed that there is a distribution of models, thus delineating the concept of model uncertainty more rigorously. There is also a frequentist information theoretic approach in an analogous vein. Model weights within this framework have been suggested by Akaike in a series of papers (see, e.g., Akaike (1978, 1979)) and expounded further by Burnham and Anderson (1998). Kapetanios, Labhard, and Price (2005) have shown that this approach might be useful for forecasting.

This paper extends the above approach in a significant way. Standard information criteria are usually constructed by combining a measure of fit with a penalty term for model complexity. The measure of fit is usually an in-sample measure. However, given the use of such criteria for constructing forecasting weights, it is reasonable to substitute the measure of fit by some measure of predictive ability such as the predictive likelihood[1]. We pursue this idea in the paper.

[1] A similar approach but adopting a Bayesian perspective is presented in Eklund and Karlsson (2005).

2 Theory

Bayesian model averaging is widely used in the literature and so we refer to work by, among others, Koop and Potter (2003), Draper (1995) and Wright (2003) for details. However, model averaging is not confined to the Bayesian approach. In the context of forecasting the idea of model averaging (*i.e.*, forecast combination) has a long tradition starting with Bates and Granger (1969). The main suggestion of this line of work is to use forecasts obtained during some forecast evaluation period to determine optimal weights, via, usually, a regression approach, from which a forecast can be constructed. A problem with this class of methods arises if the number of models, N, is large.

An alternative can be based on the analogue of Bayesian model probabilities for frequentist statistics. Such a weight scheme has been implied in a series of papers by Akaike and others (see, *e.g.*, Akaike (1978, 1979) and Bozdogan (1987)) and expounded further by Burnham and Anderson (1998). Akaike's suggestion derives from the Akaike information criterion (AIC). AIC is an asymptotically unbiased measure of minus twice the log likelihood of a given model. It contains a term in the number of parameters in the model, which may be viewed as a penalty for over-parameterization. From an information theoretic point of view, AIC is an unbiased estimator of the Kullback and Leibler (1951) (KL) distance of a given model where the KL distance is given by

$$I(f,g) = \int f(x) \log \left(\frac{f(x)}{g(x|\hat{\theta})} \right) dx.$$

$f(x)$ is the unknown true model generating the data, $g(x|.)$ is the entertained model and $\hat{\theta}$ is the estimate of the parameter vector for $g(x|.)$. The KL distance is an influential concept in the model selection literature and forms the basis of the development of AIC. Within a given set of models, the difference of the AIC for two different models can be given a precise meaning. It is an estimate of the difference between the KL distance for the two models. Further, $exp\,(-1/2\Psi_i)$ is the relative likelihood of model i where $\Psi_i = AIC_i - \min_j AIC_j$ and AIC_i denotes the AIC of the i-th model in $\mathcal{M}$. Thus $exp\,(-1/2\Psi_i)$ can be thought of as the odds for the i model to be the best KL distance model in $\mathcal{M}$. In other words, this quantity can be viewed as the weight of evidence for model i to be the KL best model given that there is some model in $\mathcal{M}$ that is KL best as a representation of the available data. It is natural to normalise $exp\,(-1/2\Psi_i)$ so that

$$w_i = \frac{exp\,(-1/2\Psi_i)}{\sum_{i=1}^{N} exp\,(-1/2\Psi_i)} \tag{1}$$

where $\sum_i w_i = 1$. We refer to these as AIC weights.

More generally any information criterion can be used as a basis for constructing weights. Criteria, in general, have the following structure

$$IC_i = L_i - C_{T,i}$$

where L_i is the estimated loglikelihood and $C_{T,i}$ is a penalty term for model complexity. Usually, the estimated loglikelihood is calculated in-sample from observed data. However, given the aim of constructing weights for forecasting, it is reasonable to consider likelihood measures that are directly related to forecasting such as predictive likelihood (see, e.g, Bjornstad (1990), Davison (1986) and Butler (1986) for the motivation of this likelihood concept in a forecasting framework). Given the observed data y, the set of random variables to be forecast z and a vector of parameters θ, predictive likelihood can be generally defined as

$$l_y(z, \theta) = f_\theta(y, z)$$

where $f_\theta(y, z)$ is the joint pdf of y and z. This general concept has been operationalised in a number of ways such as, e.g., the profile predictive likelihood given by $L_p(z|y) = sup_\theta f_\theta(y, z)$. Theoretically, our suggestion abstracts from the principles underlying the derivations of criteria such as AIC[2] and basically considers alternative likelihood concepts for L_i. This can be justified by noting that many asymptotic properties of the criteria such as model selection consistency (see, e.g., Sin and White (1996)) are retained when predictive likelihood measures are used. This will be clear from the practical implementation we suggest below.

In practical terms, we suggest that forecast errors from regression models are used in the construction of L_i rather than in-sample residuals. To fix ideas consider the regression model

$$y_t = \alpha' x_t + \epsilon_t$$

The concentrated log-likelihood of this model is given by $-T/2ln(\hat{\sigma}^2)$ where $\hat{\sigma}^2 = 1/T \sum_{t=1}^{T} \hat{\epsilon}_t^2$, $\epsilon_t = y_t - \hat{\alpha}^{(1,T)} x_t$ and $\hat{\alpha}^{(1,T)}$ denotes the estimate of α using data from $t = 1$ to $t = T$. The predictive likelihood measure we suggest replaces $\hat{\epsilon}_t$ with $\tilde{\epsilon}_t$ for $t = t_0, \ldots, T$, where $\tilde{\epsilon}_t = y_t - \hat{\alpha}^{t-t_0,t-1} x_t$. In other words we use out-of-sample forecast errors rather than residuals. Interestingly, this implies that the predictive likelihood measure will change depending on the forecast horizon. Clearly, due to the recursive nature of the scheme there are fewer out-of-sample errors than residuals since one has to have an original sample for the first estimate of α, $\hat{\alpha}_{1,t_0}$, where t_0 has to be chosen a priori. Note that if we

[2]It is worth noting that the derivation of AIC involves a predictive likelihood concept (see, e.g. Burnham and Anderson (1998, pp. 242)). In practice, however, AIC is calculated in sample.

set $t_0 = bT$, where $0 < b < 1$, the model selection consistency properties of the various information criteria are retained.

3 Empirical Application

We focus on inflation forecasting using the new model averaging scheme. The regressions we consider are k lags autoregressive processes augmented with a single predictor variable ($ARX(k)$) (see also Stock and Watson (2004)). The number of lags is either set to 1 or 4. Different models are specified for each forecasting horizon. Model i for forecasting horizon h is given by

$$\pi_{t+h} = \alpha + \sum_{j=1}^{k} \beta_j \pi_{t-j+1} + \gamma x_{it} + \epsilon_t \tag{2}$$

where π_t is either UK year-on-year CPI or RPIX inflation, x_{it} is the i-th predictor variable at time t and ϵ_t is the error term, with variance σ^2. We consider 58 predictor variables, where the data span 1980Q2-2004Q1.[3]

Where we average models, we consider predictive likelihood (PLMA), Bayesian (BMA), information theoretic (AITMA) and equal-weight (AV) model averaging. The information theoretic weights are given by (1). We include the AR forecast, making a total of 59 forecasts to combine. The information criterion considered is AIC. The Bayesian weights are set following Wright (2003). In particular, we set the model prior probabilities $P(M_i)$ to the uninformative priors $1/N$. The prior for the regression coefficients is chosen to be given by $N\left(0, \phi\sigma^2(X'X)^{-1}\right)$, conditional on σ^2, where X is the $T \times p$ regressor matrix for a given model and p is the numbers of regressors. The improper prior for σ^2 is proportional to $1/\sigma^2$. Following Wright (2003) we consider the conventional choice of $\phi = 2$. Then, the model weights are proportional to $(1+\phi)^{-p/2} S^{-(T+1)}$ where $S^2 = Y'Y - Y'X(X'X)^{-1}X'Y\frac{\phi}{1+\phi}$ and Y is the $T \times 1$ regressand vector. We also consider two factor model forecasts. As discussed in the introduction, these are widely used alternatives to forecast combination in large data sets. In this case we specify models of the form given by (2) where the exogenous variables are replaced by either the first or the first five principal components of the dataset as estimated in the full sample. For the recursive calculation of the out-of-sample forecast errors we set $t_0 = 10$ which we think is a reasonable compromise between data availability and the length of the forecast evaluation period.

[3]The UK Office of National Statistics (ONS) codes for these variables and a brief description including the source are given in the appendix of Kapetanios, Labhard, and Price (2005).

Table 1: Relative RMSE of Out-of-Sample CPI Forecasts using ARX(1) Models (Period: 1997Q2-2004Q1)

Horizon	$BMA(\phi = 2)$	$AITMA$	AV	1 Factor	5 Factors	PLMA
1	1.016^{*}	1.122	1.015	1.047^{*}	1.140^{*}	0.909
2	0.990	1.264	0.992	1.151	1.114^{*o}	0.804^{*o}
3	0.951^{*}	1.125	0.984	1.159^{o}	1.100^{*o}	0.726
4	0.881^{*}	0.992	0.974	1.164^{*o}	1.047^{*o}	0.776^{o}
8	0.804^{*}	0.725	0.952	1.210^{*o}	0.959^{*o}	0.779
12	0.824^{*}	0.662^{*}	0.946^{*}	1.125^{*o}	0.843^{*o}	0.798

*: 10% rejection of Diebold-Mariano test that the forecast differs from the benchmark
^{o}: 10% rejection of Diebold-Mariano test that the forecast differs from BMA forecast

Table 2: Relative RMSE of Out-of-Sample CPI Forecasts using ARX(4) Models (Period: 1997Q2-2004Q1)

Horizon	$BMA(\phi = 2)$	$AITMA$	AV	1 Factor	5 Factors	PLMA
1	0.986	0.926	0.988	0.975^{o}	0.960	0.892
2	0.945^{*}	0.771^{*o}	0.961^{*o}	0.996^{*o}	0.797^{o}	0.812^{*o}
3	0.898^{*}	0.669^{*o}	0.955^{*o}	1.003^{*o}	0.818^{o}	0.839^{*}
4	0.839^{*}	0.734	0.945	1.031^{*o}	0.845^{*o}	0.853^{*}
8	0.790^{*}	0.716	0.927	1.097^{o}	0.873^{*o}	0.839^{*}
12	0.824^{*}	0.665^{*}	0.934^{*}	1.062^{*o}	0.797	0.820^{*}

*: 10% rejection of Diebold-Mariano test that the forecast differs from the benchmark
^{o}: 10% rejection of Diebold-Mariano test that the forecast differs from BMA forecast

We evaluate the forecasts over 1997Q2-2004Q1, which marks the period of the Bank of England independence. We consider $h = 1, 2, 3, 4, 8, 12$. We report the relative RMSE, compared to the benchmark AR model for $k = 1, 4$ in Tables 1-4. Entries with asterisks (respectively ^{o}) in the Tables indicate cases where the relative mean square error with respect to the benchmark AR model (respectively BMA) forecast is significantly different from 1 according to the Diebold-Mariano test at the 10% significance level.[4]

The results make interesting reading. Most averaging techniques can beat the AR model most of the time. However, only PLMA can beat it for all cases considered. Further, PLMA seems to have the most consistent performance across horizons and inflation

[4]The Diebold-Mariano test is not valid for comparisons of nested models. However, all comparisons we carry out are not nested apart from those involving both the benchmark AR model and the factor models. To see this note that with positive probability, asymptotically, all model averaging methods will give a non-zero weight to a model other than the AR model even under the null hypothesis of equal predictive ability and therefore validity of the AR model since the information criterion used is the inconsistent AIC. In the case of BMA a cursory examination of the BMA weights indicates that this is the case for BMA too. We include the Diebold-Mariano tests for the factor-AR comparisons for completeness.

Table 3: Relative RMSE of Out-of-Sample RPIX Forecasts using ARX(1) Models (Period: 1997Q2-2004Q1)

Horizon	$BMA(\phi = 2)$	$AITMA$	AV	1 Factor	5 Factors	PLMA
1	1.021	1.146	1.020	1.277	1.364^{*o}	0.902^{*o}
2	0.992	1.334	0.994	1.490	1.436^{*o}	0.727^{*o}
3	0.967	1.450	0.991	1.530	1.404^{*o}	0.693^{*o}
4	0.927	1.368	0.989	1.427	1.330^{*o}	0.738^{*o}
8	0.818^{*}	0.840	0.953	1.405^{*o}	1.075^{*o}	0.743^{o}
12	0.838	0.698	0.936	1.197^{*o}	0.855^{o}	0.748

*: 10% rejection of Diebold-Mariano test that the forecast differs from the benchmark
^{o}: 10% rejection of Diebold-Mariano test that the forecast differs from BMA forecast

Table 4: Relative RMSE of Out-of-Sample RPIX Forecasts using ARX(4) Models (Period: 1997Q2-2004Q1)

Horizon	$BMA(\phi = 2)$	$AITMA$	AV	1 Factor	5 Factors	PLMA
1	0.989	0.955	0.990	1.035	1.157^{*o}	0.976^{o}
2	0.950^{*}	0.950	0.962	1.131^{*o}	1.115^{*o}	0.834^{o}
3	0.910^{*}	0.894	0.952	1.221^{*o}	1.150^{*o}	0.754^{o}
4	0.883^{*}	1.121	0.948	1.213^{*o}	1.103^{*o}	0.820^{o}
8	0.807^{*}	0.819	0.921	1.280^{*o}	0.969^{*o}	0.754^{o}
12	0.831^{*}	0.746	0.922	1.162^{*o}	0.799	0.763^{*}

*: 10% rejection of Diebold-Mariano test that the forecast differs from the benchmark
^{o}: 10% rejection of Diebold-Mariano test that the forecast differs from BMA forecast

measures. While AITMA does extremely well for long horizons but not for short horizons, PLMA does very well overall. Further, it is almost always better than BMA and some times significantly so.

4 Conclusion

Recently, there has been rapid growth of interest in forecasting methods that utilise large datasets, driven partly by the recognition that policymaking institutions process large quantities of information, which might be helpful in the construction of forecasts.

This paper focuses on model averaging. It suggests a new model averaging scheme that utilises the out-of-sample forecasting performance of the competing models to determine the weights used in model averaging. An empirical application suggests that the new averaging method is of significant potential interest.

References

Akaike, H. (1978): "A Bayesian Analysis of the minimum AIC Procedure," *Annals of the Institute of Statistical Mathematics*, 30.

——— (1979): "A Bayesian Extension of the Minimum AIC Procedure of Autoregressive Model Fitting," *Biometrika*, 66.

Bates, J. M., and C. W. J. Granger (1969): "The Combination of Forecasts," *Operations Research Quarterly*, 20, 451–68.

Bjornstad, J. F. (1990): "Predictive Likelihood: A Review," *Statistical Science*, 5, 242–265.

Bozdogan, H. (1987): "Model Selection and Akaike's Information Criterion (AIC): the General Theory and its Analytical Extensions," *Psychometrika*, 52(3), 345–70.

Burnham, K. P., and D. R. Anderson (1998): *Model selection and inference*. Berlin: Springer Verlag.

Butler, R. W. (1986): "Predictive Likelihood Inference with Applications," *Journal of the Royal Statistical Society Series B*, 48, 1–38.

Davison, A. C. (1986): "Approximate Predictive Likelihood," *Biometrika*, 73, 323–332.

Draper, D. (1995): "Assessment and propagation of model uncertainty," *Journal of the Royal Statistical Society Series B*, 57, 45–97.

Eklund, J., and S. Karlsson (2005): "Forecast Combination and Model Averaging using Predictive Measures," *Sveriges Riksbank Working Paper No. 191.*

Kapetanios, G., V. Labhard, and S. Price (2005): "Forecasting using Bayesian and Information Theoretic Model Averaging: an application to UK inflation," *Bank of England Working Paper No. 268.*

Koop, G., and S. Potter (2003): "Forecasting in Large Macroeconomic Panels Using Bayesian Model Averaging," *Federal Reserve Bank of New York Report 163.*

Kullback, S., and R. A. Leibler (1951): "On Information and Sufficiency," *Annals of Mathematical Statistics*, 22, 79–86.

Newbold, P., and C. W. J. Granger (1974): "Experience with Forecasting Univariate Time Series and the Combination of Forecasts," *Journal of the Royal Statistical Society, Series A*, 137, 131–65.

Sin, C. Y., and H. White (1996): "Information Criteria for Selecting Possibly Misspecified Parametric Models," *Journal of Econometrics*, 71(1–2), 207–225.

Stock, J. H., and M. W. Watson (2004): "Combination Forecasts of Output Growth in a Seven Country Dataset," *Journal of Forecasting*, 23, 405–30.

Wright, J. H. (2003): "Bayesian Model Averaging and Exchange Rate Forecasts," *Board of Governors of the Federal Reserve System, International Finance Discussion Papers No 779.*

This working paper has been produced by
the Department of Economics at
Queen Mary, University of London

Department of Economics
Queen Mary, University of London
Mile End Road
London E1 4NS
Tel: +44 (0)20 7882 5096
Fax: +44 (0)20 8983 3580
Web: www.econ.qmul.ac.uk/papers/wp.htm