

## Prediction with Misspecified Models

John Geweke\* and Gianni Amisano\*\*

Many decision-makers in the public and private sectors routinely consult the implications of formal economic and statistical models in their work. Especially in large organizations and for important decisions, there are often competing models. Of course no model under consideration is a literal representation of reality for the purposes at hand – more succinctly, no model is “true” – and different models focus on different aspects of the relevant environment. This fact can often be supported by formal econometric tests concluding that the models at hand are, indeed, misspecified in various dimensions.

There are well-developed practical Bayesian procedures, with solid theoretical foundations, for combining competing models in decision-making (Geweke, 2005, Section 1.5). But these procedures all condition on one of the models under consideration being true. Non-Bayesian model selection procedures may be less formal but proceed under the same maintained hypothesis. This leads to the common situation in which decision-makers are left to balance informally the various predictions and decisions that are formal consequences of competing economic and statistical models. In this circumstance serious attention is typically granted to the implications of models to which formal econometric procedures assign very little credence.

The approach taken here draws on a literature long established in statistics and recently emerging in econometrics that dispenses with the condition that one of the

models under consideration is true. In its place it substitutes a log scoring rule for predictive distributions, described in Section I. A linear prediction pool, described in Section II, enlarges the set of model predictive densities to include weighted means of these densities. The weights can be fixed, or they can be chosen using the history of model predictions together with the log scoring rule. Both theory and implementation show that these weights are quite different from the weights that emerge from Bayesian model averaging. Section III applies these procedures to models representing three different classes of macroeconomic models used by many central banks. Linear prediction pools of these models substantially outperform each of the three constituent models as well as Bayesian model averaging.

## I. Models and prediction

Attention focuses on predicting a vector of economic time series over a sequence of periods  $t = 1, 2, \dots$ . Prior to time  $t$  this is a random vector  $Y_t$ , and after time  $t$  is a vector of observed values  $y_t$ . Let  $Y_{t_1:t_2}$  denote the set  $\{Y_{t_1}, \dots, Y_{t_2}\}$  and  $y_{t_1:t_2}$  the realization of these random vectors. Each of a set of models  $\{A_1, \dots, A_n\}$  provides predictive densities  $p(Y_t; y_{1:t-1}, A_i)$ . These models could use more or less formal procedures for inference; they could involve undocumented and unreproducible personal judgment; they could use covariates not in  $Y_t$ ; or observations  $y_t$  ( $t \leq 0$ ). All that matters is that each model provide one-step-ahead predictive densities in real time.

The log score for model  $A_i$  at time  $t$  is

$$LS(y_{1:t}, A_i) = \sum_{s=1}^t \log p_s(y_s; y_{s-1}, A_i). \quad (1)$$

This is a measure of forecast accuracy. There are many measures of forecast accuracy, of which mean square error is perhaps the best known, but there are some compelling reasons to use log score. (1) Because the scoring rule at time  $t$  depends on  $p(Y_t; y_{1:t-1}, A_i)$  only through  $p(y_t; y_{1:t-1}, A_i)$  it is said to be local (Bernardo, 1979). A scoring rule is said to be proper if a forecaster maximizing expected score is led to report her true subjective distribution (Winkler and Murphy, 1968). The log scoring rule is the only proper local scoring rule (Shuford et al., 1966; Bernardo, 1979). (2) Log score is monotonically related to the probability (density) assigned to events that actually occurred:  $\exp[t^{-1}LS(y_{1:t}, A_i)]$  is the geometric mean of these probabilities for model  $A_i$ . (3) If  $A_i$  is formally subjective Bayesian then  $LS(y_{1:t}, A_i)$  is its log marginal likelihood in the sample  $y_{1:t}$ .

If models  $A_i$  and  $A_j$  are both formally subjective Bayesian then  $LS(y_{1:t}, A_i) - LS(y_{1:t}, A_j)$  is the logarithm of the Bayes factor in favor of model  $A_i$  over model  $A_j$ . Under standard asymptotic conditions (e.g., Geweke, 2005, Theorem 3.4.2)  $[LS(y_{1:t}, A_i) - LS(y_{1:t}, A_j)]/t$  converges to a finite constant as  $t \rightarrow \infty$  and so the posterior odds ratio either converges to zero or diverges to  $+\infty$  as  $t \rightarrow \infty$ . This leads to a phenomenon well known to Bayesian econometricians: as sample size increases Bayesian model averaging places weight entirely on a single model and consequently

only one model remains relevant for prediction and decision-making. However in these same circumstances decision-makers do not confine their attention to this one model – Section III provides an example. As will be seen in the next section, this disjuncture may be attributed to the fact that Bayesian procedures condition on one of models  $A_1, \dots, A_n$  being true while actual decision-makers do not do this.

## II. Prediction pooling

A linear pool of the predictive densities  $p(Y_t; y_{1:t-1}, A_j)$  ( $j = 1, \dots, n$ ) is

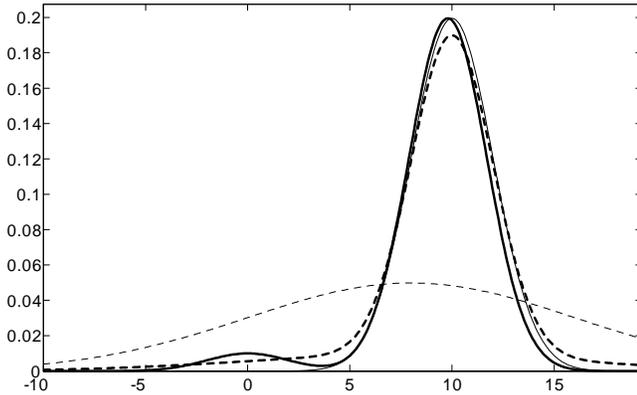
$$p(Y_t; y_{1:t-1}, w_{t-1}) = \sum_{i=1}^n w_{t-1,i} p(Y_t; y_{1:t-1}, A_i) \quad (2)$$

where the weight vector  $w_{t-1}$  has all elements nonnegative and summing to one, so that a linear pool of predictive densities is itself a predictive density. The idea dates at least to Stone (1961). The subscript  $t - 1$  for  $w$  emphasizes the requirement that the weight vector must be computed in real time if a linear pool is actually to be used. Geweke and Amisano (2011) show that finding the vector  $w_{t-1}$  that maximizes the historical log score  $\sum_{s=1}^{t-1} \log [\sum_{i=1}^n w_{t-1,i} p(y_s; y_{1:s-1}, A_i)]$ , henceforth the real-time optimal weight vector, is a trivial computational task. The weights could also be fixed at some reasonable value (e.g.,  $w_{t-1,i} = 1/n$  for all  $t$  and  $i$ ), and Section III utilizes both schemes.

Geweke and Amisano (2011) shows that under standard conditions  $w_t \rightarrow w^*$ . In the empirically irrelevant but theoretically illuminating case that one of the models

( $A_j$ , say) is true  $w_j^* = 1$ . In the more general and relevant case that all models are false, several (and potentially all) elements  $w_i^* > 0$ . In these same circumstances Bayesian model averaging will assign a limiting weight of one to one model and zero to the others.

To provide some intuition for these facts, consider the very simple case of predicting a single scalar time series  $y_t$  that is in fact independent and identically distributed. Each of two models  $A_1$  and  $A_2$  makes this i.i.d. specification, and each has a fixed predictive density – neither learns from the past, making it possible to abstract from predictive densities that change over time. The history of prediction and realization leads to real-time optimal weights updated each period.



CAPTION: Figure 1. The heavy solid line is the predictive density of the data generating process. The light solid and dashed lines are the predictive densities of models  $A_1$  and  $A_2$ , respectively. The heavy dashed line is the predictive density of the optimal pool.

Figure 1 provides an example, in which the thick solid line is the true popu-

lation density and the thin lines provide the predictive densities of the two models. The model  $A_1$  (solid thin line) predicts well most of the time, but it assigns  $P(y \leq 3) < 0.001$  whereas in fact  $P(y \leq 3) \cong 0.05$ . Model  $A_2$  (dashed thin line) would be regarded as inferior to  $A_1$ , assigning probability 0.25 to events whose actual probability is less than 0.001. The optimal pool incorporates  $A_2$  because that model assigns substantial positive probability to the event  $(-4, 4)$  that is assigned almost no probability by  $A_1$  and because actual events assigned very low probabilities are heavily penalized in the log score function.

Formal analysis supports these observations. The expected log score of  $A_1$  in a sample of size 100 is -269.1 whereas that of  $A_2$  is -307.5. The corresponding weight for  $A_2$  in Bayesian model averaging with sample size  $T = 100$  is  $1.92 \times 10^{-17}$ , so that essentially all of the weight is assigned to  $A_1$ . The limiting value  $w^*$  of the real-time optimal weight vector  $w_{t-1}$  is  $w^* = (0.908, 0.092)'$ . In a sample of size  $T = 100$  the expected log score in the optimal linear pool is -230.3. This log score is much closer to the expected log score from the data generating process in a sample of size 100, which is -225.3, than it is to the expected log score of the better model  $A_1$ .

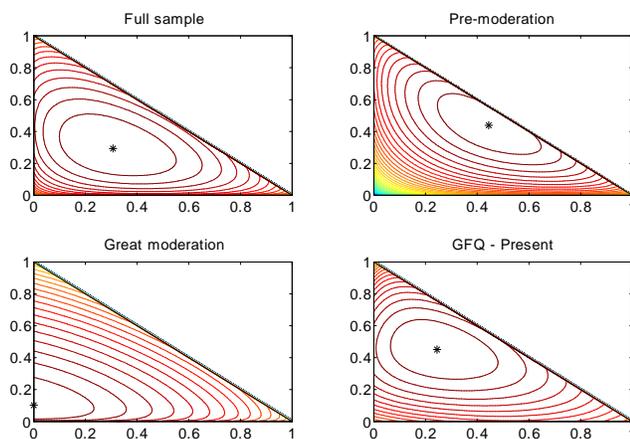
The formal Bayesian calculus conditions on one of  $A_1$  and  $A_2$  being true. The Bayes factor is the ratio of their predictive scores. Since log predictive scores (1) are proportional to sample size, model probabilities move in a regular and often rapid fashion toward their limiting probabilities of 0 for one model and 1 for all others. This is a consequence of the maintained condition one of the models must be true.

Model pooling weakens this assumption by introducing the continuum of prediction models defined by the set of weights  $w_{t-1}$  in the unit simplex in (2). When none of the constituent models are true, the consequence is that positive weights are usually assigned to several models. As illustrated in Figure 1, models that are clearly inferior to others in isolation typically have a role to play in these pools.

### **III. An application: Macroeconomic forecasting**

This section illustrates these points using a  $7 \times 1$  vector of US economic time series and three macroeconometric forecasting models. The time series is the same as that in Smets and Wouters (2007): growth rates in per capita real consumption, investment, and output; growth rate in the real wage; log per capita weekly hours worked; the inflation rate; and the Federal funds rate. The three models are the dynamic stochastic general equilibrium (DSGE) model of Smets and Wouters (2007); a conventional vector autoregression (VAR) with a Minnesota prior using four lags of each variable; and a dynamic factor model (DFM) following the specification of Stock and Watson (2005) with three common factors. The DFM includes five additional macroeconomic time series: growth rate in the S&P 500 index, the unemployment rate, the spread between 3-month and 10-year Treasury rates, the BAA - AAA corporate bond rate spread, and growth rate in the M2 money supply. The predictive densities for the DFM are the marginal densities in the same seven time series common to all three models. Inference and prediction are fully Bayesian, using conventional proper prior distributions and Markov chain Monte Carlo algorithms.

The analysis covers the quarters 1966:1 ( $t = 1$ ) through 2010:3 ( $t = 179$ ) and all prediction is fully out-of-sample. The first predictive distribution, for 1966:1, follows from the posterior distribution for 1951:1 - 1965:4 and the last, for 2010:3, follows from the posterior distribution for 1951:1 - 2010:2,. Log scores and optimal prediction pool weights are updated each quarter. Within the analysis period we focus on three subperiods: pre-moderation, 1966:1 - 1984:4; the great moderation, 1985:1 - 2007:4; and post global financial crisis, 2008:1 - 2010:3.



CAPTION: Figure 2. Log scores of model pools as a function of model weights: horizontal axis is  $w_{DSGE}$ , vertical axis is  $w_{DFM}$ , and the balance is  $w_{VAR} = 1 - w_{DSGE} - w_{DFM}$ .

Figure 2 conveys the mean log score of linear pools as functions of their weight vectors  $w$ ,  $LS(w) = (t_2 - t_1 + 1)^{-1} \sum_{s=t_1}^{t_2} \log [\sum_{i=1}^3 w_i p(y_s; y_{1:s-1}, A_i)]$ , where  $t_1$  and  $t_2$  are the start and end of the analysis period or subperiod. In each panel the separation between contours is 0.025. The lower left corner is the VAR model alone, the upper left is the DFM model alone, and the lower right is the DSGE model alone.

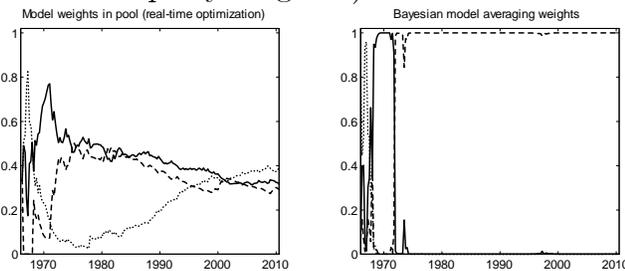
The first three rows of Table 1 provide the values at these points. The differences in these log scores are consistent with dominance of one model in Bayesian model averaging. Treating the pre-1966 period as part of the prior distribution and assigning equal prior probabilities to the three models, the log-odds ratio in favor of DFM over DSGE for the entire period 1966:1 - 2010:3 is  $179 \cdot (-6.10650 - (-6.1866)) = 14.3379$ , which assigns all but about  $6 \times 10^{-7}$  of the weight to DFM. Bayesian model averaging yields similarly one-sided results for most of the subperiods as well.

Period	1966:1- 2010:3	1966:1- 1984:4	1985:1- 2007:4	2008:1- 2010:3
DSGE	-6.1866	-7.3596	-4.7604	-8.9153
VAR	-6.2945	-8.0163	-4.4767	-8.9459
DFM	-6.1065	-7.1169	-4.8089	-9.0647
Optimal	-5.8008	-6.8178	-4.3879	-8.6498
Equal	-5.8028	-6.8530	-4.4933	-8.6604

The asterisk in each panel of Figure 2 indicates the weight vector  $w$  that maximizes log score, and Table 1, row 4, presents the average log predictive score for this combination. In all four panels the maximum is very shallow, consistent with the intuition of Figure 1: the marginal contribution of a model to a pool is greatest at its introduction (weight zero) where it can assign some probability to historical events otherwise very improbable in the pool. As a corollary substantial departure in pool

weights from the optimal value has only modest impact on log score of prediction so long as they remain well within the unit simplex. The last row of Table 1 indicates the average log predictive score of a pool with equal model weights. It is very close to that with the optimal weight vector in each case.

Pooling improves substantially on the predictive performance of the best model. For the entire period the average per period improvement in log score over the DFM is 0.3057. By implication, the geometric mean of the ratio of the probability density assigned by the DFM to actual events one-quarter-ahead to that assigned by the pool (either optimal or equally weighted) is less than 0.74.



CAPTION: Figure 3. Real-time pool and Bayesian model averaging weights for  $t = 1$  corresponding to 1966:1 through  $t = 179$  corresponding to 2010:3: DSGE solid line, VAR dotted line, DFM dashed line.

Figure 3 tracks model weights through the entire analysis period from 1966:1 through 2010:3, with optimal pool weights in the left panel and Bayesian model averaging weights in the right panel. In both cases the weights for the first period predictive density, 1966:1, are equal, and for roughly the first dozen quarters they fluctuate rapidly. But the pattern soon settles down. In the case of the optimal pool

weights (left panel) the VAR steadily gains through the great moderation, where its log score dominates that of the other two models (Table 1). The weights at the right of this graph are the same as those indicated by the asterisk in the upper left panel of Figure 2. In contrast Bayesian model averaging (right panel) assigns essentially all weight to DFM from the early 1970's onward. This is consistent with the logic of Bayesian model averaging: given the maintained hypothesis that one of the three models is true, the evidence that this one model must be the DFM is overwhelming ten years into the sample and beyond.

#### IV. Conclusion

The assumption that one of a set of prediction models is a literal description of reality formally underlies many formal econometric methods, including Bayesian model averaging and most approaches to model selection. Prediction pooling does not invoke this assumption and leads to predictions that improve on those based on Bayesian model averaging. In many cases, including the prediction of macroeconomic aggregates using leading macroeconometric models examined here, the improvement is substantial.

#### REFERENCES

- Bernardo JM (1979). Expected information as expected utility. *The Annals of Statistics* 7: 686-690.
- Geweke J (2005). *Contemporary Bayesian Econometrics and Statistics*. Englewood Cliffs NJ: Wiley.

Geweke J, Amisano G (2011). Optimal prediction pools. *Journal of Econometrics* 164: 130-141.

Shuford EH, Albert A, Massengill HE (1966). Admissible probability measurement procedures. *Psychometrika* 31: 125-145.

Smets F, Wouters R (2007). Shocks and frictions in US business cycles: A Bayesian DSGE approach. *American Economic Review* 97: 586-606.

Stock JH, Watson MW (2005). Implications of Dynamic Factor Models for VAR Analysis. NBER Working Paper No. 11467.

Stone M (1961). The opinion pool. *Annals of Mathematical Statistics* 32: 1339-1342.

Winkler RL, Murphy AM (1968). "Good" probability assessors. *Journal of Applied Meteorology* 7: 751-758.

#### FOOTNOTE

\*University of Technology Sydney (Australia), Erasmus University (The Netherlands) and University of Colorado (US). Support from Australian Research Council grant 110104732 is gratefully acknowledged. Corresponding author: John.Geweke@uts.edu.au

\*\*European Central Bank, Frankfurt. The views expressed do not represent those of the ECB.