"On Bayesian Forecast Combination Procedures," in *Economic Structural Change: Analysis and Forecasting* (A. Westlund and P. Hackl, eds.), 225-232, 1991. New York: Springer-Verlag.

(17)

Peter Hackl · Anders H. Westlund (Eds.)

Economic Structural Change

Analysis and Forecasting

(1991) pp. 225-232

With 101 Figures and 56 Tables

Springer-Verlag
Berlin Heidelberg NewYork
London Paris Tokyo
Hong Kong Barcelona Budapest

CHAPTER 15

A Note on Bayesian Forecast Combination Procedures

Francis X. Diebold

Summary

The properties of Bayesian composite forecasts are studied. It is argued, and illustrated with an example, that the asymptotic performance of such composite forecasts depends on the validity of a maintained assumption, namely, that one of the models among those whose forecasts are combined is the true data-generating process. The implications of this phenomenon are explored.

15.1 Introduction

Scientific knowledge obtained from research in one area often proves useful in others. Such has been the case with the Bayesian theory of econometric model selection, as developed by Geisel (1970, 1974) and Zellner (1971, 1972, 1979, 1984), which has generated insights useful not only for model selection but also for prediction. In particular, it is now known that under certain conditions the posterior probabilities associated with various forecasting models may be used as weights in forming a linear composite forecast, and that the resulting composite forecast is optimal, in the sense of minimizing posterior expected loss.

In this chapter, I focus on one of those "certain conditions"—in particular, the assumption that one of the models among those whose forecasts are combined is the true data-generating process (DGP)—and I explore the effects of its failure on the performance of Bayesian composite forecasts. I argue that, if the assumption is satisfied, such Bayesian composite forecasts will have certain desirable asymptotic properties relative to their classical counterparts, but that the result is reversed if the assumption is not satisfied.

In Section 15.2, I give an explicit derivation of the Bayesian composite forecast, and I show that (under certain conditions) it minimizes posterior expected loss. I explore the linkage between the maintained assumption of truth of one of the models and the resulting good performance of the Bayesian composite forecast. In Section 15.3, I illustrate the argument with a simple example involving the combination of forecasts from two linear regression models. In Section 15.4, I conclude with a summary and directions for future research.

15.2 Bayesian Model Selection and Forecast Combination

The Bayesian solution to the model selection problem (under symmetric loss) is well known: It is optimal to choose the model with highest posterior probability. Zellner (1972, 1984), Zellner *et al.* (1989), and others have suggested that the posterior probabilities may be used fruitfully not only for model selection, but also for forecast combination. Forming a composite forecast with weights equal to the posterior probabilities seems reasonable, and the case for doing so is easily formalized.

Consider two models, M_1 and M_2, with associated posterior probabilities p_1 and p_2, respectively, where $p_1+p_2 = 1$. [The generalization to the case of more than two competing models is immediate.] Then posterior expected loss is given by

$$E(y-\hat{y})^2 = p_1[E(y-\hat{y})^2|M_1] + p_2[E(y-\hat{y})^2|M_2], \tag{15.1}$$

where $\hat{y}$ is any point forecast. Let $\bar{y}_1$ ($\bar{y}_2$) be the mean of the predictive probability density function for M_1 (M_2). Then we can write

$$\begin{aligned} E[(y-\hat{y})^2|M_i] &= E\{[(y-\bar{y}_i)-(\hat{y}-\bar{y}_i)]^2|M_i\} \\ &= E\{[(y-\bar{y}_i)^2+(\hat{y}-\bar{y}_i)^2-2(y-\bar{y}_i)(\hat{y}-\bar{y}_i)]|M_i\} \\ &= E[(y-\bar{y}_i)^2|M_i]+(\hat{y}-\bar{y}_i)^2 \\ &= C_i+(\hat{y}-\bar{y}_i)^2, \quad i=1,2, \end{aligned} \tag{15.2}$$

where $C_i = E[(y-\bar{y}_i)^2|M_i]$. But then

$$\begin{aligned} E(y-\hat{y})^2 &= p_1[C_1+(\hat{y}-\bar{y}_1)^2]+p_2[C_2+(\hat{y}-\bar{y}_2)^2] \\ &= C+p_1(\hat{y}-\bar{y}_1)^2+p_2(\hat{y}-\bar{y}_2)^2, \end{aligned} \tag{15.3}$$

where $C \equiv p_1C_1+p_2C_2$. The first-order condition for minimization of (15.3) with respect to $\hat{y}$ is

$$2\hat{y}^* - 2(p_1\bar{y}_1+p_2\bar{y}_2) = 0 \tag{15.4}$$

or

$$\hat{y}^* = p_1\bar{y}_1+p_2\bar{y}_2\,. \tag{15.5}$$

On what does this standard Bayesian result depend? Most important is the assumption that the posterior probabilities associated with the models being combined sum to 1. This is equivalent to the assumption (often made explicitly) that one of the models is true. To see this, note that $p_1 + p_2 = 1$ is equivalent to $p_1 + p_2 - P(M_1 \cap M_2) = 1$, because $P(M_1 \cap M_2) = 0$ so long as $M_1 \neq M_2$. But $p_1 + p_2 - P(M_1 \cap M_2) = 1$ is equivalent to $P(M_1 \cup M_2) = 1$. The motivation for the assumption seems to be the idea that it should be possible to write down an exhaustive listing of candidate models, one of which must (by construction) be the true DGP. In practice, of course, the forecasts of only a small number of models are combined; enunciation of an exhaustive set of candidate models for the true DGP is always infeasible and probably impossible. In short, it seems difficult to take seriously the assumption that the true DGP is among the candidate models whose forecasts are being combined.

To better understand the effects of the assumption, let us first suppose that it *is* true. Without loss of generality, assume that M_1 is true. Then, if the Bayesian model selection procedure is consistent, p_1 will approach 1 as sample size (T) gets large. The implication of consistency of the Bayesian model selection procedure for Bayesian forecast combination, of course, is that progressively more weight is placed on M_1 as T gets large; in the limit, M_1 receives unit weight and M_2 receives zero weight. In other words, the Bayesian model selection and Bayesian forecast combination procedures coincide asymptotically. This result is natural and desirable, *if* the true DGP is among the models whose forecasts are combined.

But what happens when the true DGP is *not* among those whose forecasts are combined, as is likely to be the case in practice? Is there any harm in maintaining the assumption of truth of one of the models, in order to make the Bayesian analysis operational? Recent work on estimation and testing in misspecified models [e.g., White (1982), Gourieroux *et al.* (1984), Vuong (1989)] furnishes a useful perspective on this question. We now know that, under general conditions, an estimator of the parameter of a misspecified model will converge to a pseudo-true value, that is, to a parameter configuration closest (within the misspecified class) to the true DGP. Furthermore, the metric defining "closeness" is induced by the estimation procedure.

Now, if the true DGP is not among the models entertained, it is of course impossible for any model selection procedure—Bayesian or otherwise—to be consistent for the true model. But, as the discussion above indicates, we might expect the model selected by the Bayesian procedure to be consistent for *something*, namely, the model *closest* to the true DGP. Without loss of generality, assume that M_1 is closer. Then it is reasonable to expect that p_1 will converge to 1, as was the case when M_1 was true. For model selection, such a property is very useful—it is often desired to determine which among a set of models provides the best approximation to the true DGP. The implications for forecast combination, however, appear less desirable. Consistency of the Bayesian model selection procedure for the closest model implies that Bayesian composite forecasts will asymptotically place unit weight on M_1 and zero weight on M_2. But M_1 and M_2 are *both* false models; the fact that M_1 is closer to the true DGP does not mean that the information contained in M_2 cannot be usefully combined with that in M_1 to produce

a superior composite forecast. (Contrast this with the case where M_1 is in fact the true DGP.) This insight, of course, is the entire motivation for forecast combination [see Clemen (1990)].

In summary, then, it would appear that Bayesian composite forecasts will perform well in large samples, placing all weight on the forecast of the true model, *if* the true model is among those whose forecasts are combined. Otherwise, it would appear that Bayesian composite forecasts will perform poorly in large samples, placing all weight on the forecast of one false model, and thereby discarding the potentially valuable information contained in other false models. In the next section, the truth of these conjectures in the context of simple linear regression is illustrated.

15.3 Combination of Forecasts from Regression Models

Consider a simple comparison of two regression models, as in Zellner (1971, pp. 306-312),

$$M_1: \quad y_t = X_{1t}\beta_1 + \mu_{1t}, \qquad t = 1, \ldots, T \tag{15.6}$$

$$M_2: \quad y_t = X_{2t}\beta_2 + \mu_{2t}, \qquad t = 1, \ldots, T \tag{15.7}$$

one of which is the true model, where X_1 and X_2 are nonstochastic matrices of maximum and equal column rank, β_1 and β_2 contain no common elements, and the disturbances of the true model are i.i.d. Gaussian with zero mean and constant variance. Then, in a Bayesian analysis with diffuse priors over parameters and models, the posterior odds for M_1 versus M_2 are given by

$$\frac{p_1}{p_2} = \left[\frac{s_2}{s_1}\right]^T, \tag{15.8}$$

where s_i is the square root of the usual unbiased estimator of $\sigma_i^2, i = 1, 2$. [The result also holds for informative-prior Bayesian analyses if T is large and certain other regularity conditions are satisfied.]

Consider now the implications of the earlier-derived Bayesian forecast combination procedure. Rearranging (15.8) yields

$$p_2 = \left[\frac{s_1}{s_2}\right]^T p_1 \tag{15.9}$$

or, because $p_1 + p_2 = 1$ by assumption,

$$p_1 + \left[\frac{s_1}{s_2}\right]^T p_1 = 1. \tag{15.10}$$

Thus,

$$p_1 = \frac{s_2^T}{s_1^T + s_2^T}. \tag{15.11}$$

Therefore, p_1 depends only on the ratio s_1/s_2, which is emphasized by writing

$$p_1 = \frac{1}{1 + \left[\frac{s_1}{s_2}\right]^T} . \tag{15.12}$$

Note that

$$\lim_{T \to \infty} p_1 = \begin{cases} 0 & \text{if} \quad \sigma_1 > \sigma_2 \\ 1/2 & \text{if} \quad \sigma_1 = \sigma_2 \\ 1 & \text{if} \quad \sigma_1 < \sigma_2 . \end{cases} \tag{15.13}$$

Now compare the weight arising from the Bayesian forecast combination procedure (15.5) with the weight arising from the classical variance-covariance forecast combination procedure. [By classical variance-covariance combining weight I mean the weight that minimizes combined prediction error variance, as developed by Bates and Granger (1969) and discussed in Granger and Newbold (1987).] As is well known, the classical forecast combination is

$$\hat{y} = \phi^* f_1 + (1 - \phi^*) f_2 , \tag{15.14}$$

where f_1 and f_2 are forecasts (possibly but not necessarily the means of predictive probability density functions),

$$\phi^* = \frac{1 - s_{12}/s_2^2}{1 + s_1^2/s_2^2 - 2s_{12}/s_2^2} \tag{15.15}$$

and s_{12} is the usual estimator of the covariance of the forecast errors associated with M_1 and M_2. If $s_{12} = 0$, the classical weight is

$$\phi^* = \frac{1}{1 + \left[\frac{s_1}{s_2}\right]^2} . \tag{15.16}$$

While the Bayesian and classical combining weights are very similar, several interesting differences are apparent. For example, the Bayesian weights are required to be convex, while the classical weights need not be. The convexity restriction is not necessarily beneficial. A forecast with a negative weight can play the same useful role in producing a combined forecast as an asset sold short plays in producing the return on a portfolio. Convexity of the Bayesian weights follows immediately from the definition of probability and the assumption that one of the two models is true. In addition, the Bayesian weights do not exploit covariance information, while the classical weights (in general) do. Presumably this again reflects the assumption that one, and only one, of the models is true.

These differences are of minor importance, however, compared with those associated with the nature of dependence on sample size. Both the classical and Bayesian weights change implicitly with sample size, as the underlying estimators of the relevant variances

(and, in the classical case, covariances) converge to their pseudo-true values. The Bayesian weight changes explicitly with sample size, however, as is made clear by the appearance of T in (15.12).

To understand the significance of the role played by sample size in the construction of the Bayesian combining weight, it will again prove useful to segment the discussion into two parts, according to the truth of the assumption that one of the two models is the true DGP. Suppose first that the assumption is true, and with no loss in generality assume that M_1 is true; then by (15.12) and (15.13) the Bayesian weight placed on M_1 converges to unity, while that placed on M_2 converges to zero. (The truth of M_1 implies that it has a smaller disturbance variance.) As argued earlier, it is desirable that this should happen, and it is made possible by virtue of the validity of the assumption that one of the two models is true. The desirability follows from the fact that the true model encompasses all rivals. [For a discussion of encompassing in its relation to forecast combination see Diebold (1990).]

Suppose now that neither M_1 nor M_2 is the true DGP. Suppose also, without loss of generality, that M_1 is closer than M_2 to the true DGP, in the sense that $\text{plim}(s_1) < \text{plim}(s_2)$. As before, the Bayesian weight placed on M_1 converges to unity, while that placed on M_2 converges to zero. Such convergence is no longer desirable, however, because asymptotically all weight is placed on a false model, M_1. The essential insight of forecast combination, of course, is that loss may be reduced by pooling the information contained in false models, all of which represent different approximations to reality.

15.4 Concluding Remarks

I have conjectured that the asymptotic performance of Bayesian composite forecasts is intimately linked to the truth of a maintained assumption, namely, that the true DGP is among the models whose forecasts are combined. The conjecture was verified in the context of a particular linear regression example. The argument points to the desirability of exploring Bayesian approaches to forecast combination that do not assume the truth of one of the underlying models. Is such a problem well-posed? If so, how would such an analysis proceed? [The difficulty is related to the fact that the calculations for posterior expected loss, (15.1)–(15.3), are apparently not meaningful unless one of the models is assumed true.] What relationship would the resulting combining weights have to the Bayesian and classical weights discussed in this chapter?

It is worth noting that, regardless of the answers to the questions posed above, Bayesian insights will likely contribute in other ways to the advancement of forecasting methodology and forecast combination methodology. Shrinkage techniques, for example, have been used advantageously by Zellner and Hong (1987) and Zellner *et al.* (1988, 1989) to forecast international growth rates and turning points, and by Diebold and Pauly (1990) to shrink classical composite forecasts toward a prior location, such as the sample mean.

Finally, I am happy to report on concurrent and independent work by Zellner (1989) who has recently initiated development of Bayesian methods for combining forecasts from

sets of models whose posterior probabilities do not sum to unity. I hope that my paper will stimulate additional work along similar lines.

Acknowledgments

I gratefully acknowledge financial support from the National Science Foundation and the University of Pennsylvania Research Foundation. John Geweke, Bruce Mizrach, and especially Arnold Zellner provided useful comments that improved the content and exposition of this chapter. All remaining errors, omissions, and inaccuracies are mine.

References

Bates, J.M. and Granger, C.W.J. (1969), The combination of forecasts. *Operations Research Quarterly*, **20**, 451-468.

Clemen, R.T. (1990), Combining forecasts: A review and annotated bibliography (with discussion). Forthcoming in *International Journal of Forecasting.*

Diebold, F.X. (1990), Forecast combination and encompassing: Reconciliation of two divergent literatures. Forthcoming in *International Journal of Forecasting.*

Diebold, F.X. and Pauly P. (1990), The use of prior information in forecast combination. Forthcoming in *International Journal of Forecasting.*

Geisel, M.S. (1970), Comparing and choosing among parametric statistical models: A Bayesian analysis with macroeconomic applications. Ph.D. dissertation, University of Chicago.

Geisel, M.S. (1974), Bayesian comparisons of simple macroeconomic models, pp. 227-256 in S. Feinberg and A. Zellner (eds.), *Studies in Bayesian Econometrics and Statistics.* Amsterdam: North-Holland.

Gourieroux, C., Monfort A., and Trognon, A. (1984), Pseudo maximum likelihood method theory. *Econometrica*, **52**, 681-700.

Granger, C.W.J. and Newbold, P. (1987), *Forecasting Economic Time Series*, second edition. New York: Academic Press.

Vuong, Q.H. (1989), Likelihood ratio tests for model selection and nonnested hypotheses. *Econometrica*, **57**, 307-334.

White, H. (1982), Maximum likelihood estimation of misspecified models. *Econometrica*, **50**, 1-25.

Zellner, A. (1971), *An Introduction to Bayesian Inference in Econometrics.* New York: John Wiley.

Zellner, A. (1972), A note on optimal prediction and control with two alternative regression models. Unpublished manuscript, Graduate School of Business, University of Chicago.

Zellner, A. (1979), Posterior odds ratios for regression hypotheses: General considerations and some specific results. Invited paper presented to the Econometric Society; reprinted in Zellner (1984).

Zellner, A. (1984), *Basic Issues in Econometrics.* Chicago: University of Chicago Press.

Zellner, A. (1989), Bayesian and non-Bayesian methods for combining models and forecasts. Unpublished manuscript, Graduate School of Business, University of Chicago.

Zellner, A. and Hong, C. (1987), Forecasting international growth rates using Bayesian shrinkage and other procedures. *Journal of Econometrics*, **40**, 183-202.

Zellner, A. and Hong, C. (1988), Bayesian methods for forecasting turning points in economic time series: Sensitivity of forecasts to asymmetry of loss structures, forthcoming in K. Lahiri and G. Moore (eds.), *Leading Economic Indicators: New Approaches and Forecasting Records*. Cambridge: Cambridge University Press.

Zellner, A., Hong, C., and Gulati, G.M. (1988), Turning points in economic time series, loss structures and Bayesian forecasting, forthcoming in J. Hodges, S. Press, and A. Zellner (eds.), *Bayesian Likelihood Methods in Statistics and Econometrics: Essays in Honor of George A. Barnard*. Amsterdam: North-Holland.

Zellner, A., Hong, C., and Min, C. (1989), Forecasting turning points in international output growth rates using Bayesian exponentially weighted autoregression, time-varying parameter, and pooling techniques. Manuscript, Graduate School of Business, University of Chicago.