International Journal of Forecasting 6 (1990) 503–508
North-Holland

The use of prior information in forecast combination *

Francis X. Diebold
Department of Economics, University of Pennsylvania, Philadelphia, PA 19104-6297, USA

Peter Pauly
Department of Economics, University of Toronto, Ontario, Canada M5S 1A1

Abstract: Simple averages often, but not always, outperform more sophisticated "optimal" forecast composites. We used Bayesian shrinkage techniques to allow the incorporation of prior information into the estimation of combining weights; the estimated combining weights were coaxed or "shrunken" toward equality but were not forced to be exactly equal. The least-squares and prior (i.e., arithmetic average) weights then emerged as polar cases for the posterior mean; the exact location depended on prior precision, which was estimated from the data. In a simple example involving U.S. GNP forecasts, a large amount of shrinkage was found to be optimal.

Keywords: Bayesian, Pooling, Prediction, Shrinkage.

1. Introduction

Numerous forecasts based on partially overlapping information sets are available to decision-makers; the possibility of combining such primary forecasts into an optimal pooled forecast immediately arises. While pooling of forecasts is suboptimal relative to pooling of information sets, it must be recognized that in many forecasting situations, particularly in real time, pooling of information sets is either impossible or prohibitively costly. This pragmatic observation has spurred a large literature, with contributions from diverse areas including economics, management science, mathematical statistics (classical and Bayesian), operations research and psychology. A survey and comprehensive annotated bibliography is given by Clemen (1989).

This paper, like most of the literature, is concerned with the formation of minimum mean-squared-error (MSE) combined forecasts by projecting realizations into the space spanned by the primary forecasts. It is well known that, under certain conditions, linear composite forecasts using weights constructed in this fashion must have asymptotic MSE less than or equal to that of the best primary forecast. In finite samples of the size typically available in econometric forecasting, however, sampling error contaminates the combining weight estimates, and the sampling error problem is exacerbated by the collinearity that typically exists among primary forecasts. Thus, although we hope to reduce out-of-sample MSE by combining, there is no guarantee that we will be able to do so.

It is not surprising, then, that recent attention has been paid to the instability of estimated com-

* We would like to thank the editor, associate editor and the anonymous referees, as well as R.F. Engle, C.W.J. Granger, E.P. Howrey, L.R. Klein, C.A. Sims, A. Zellner and seminar participants at the Federal Reserve Board and the University of Arizona Conference on Econometric Forecasting, for useful comments on earlier drafts. All remaining errors are ours. The data were generously supplied by Robert T. Clemen, Department of Decision Sciences, College of Business Administration, University of Oregon.

0169-2070/91/$03.50 © 1991 – Elsevier Science Publishers B.V. (North-Holland)

bining weights calculated in real time (e.g., Kang, 1986), and that apparently suboptimal combining methods such as simple arithmetic averaging often perform quite well (e.g., Clemen, 1989). A fundamental problem remains, however: a forecaster does not know, ex ante, whether an "optimal" composite or a simple average will perform best in his particular application. In this paper we address this problem by developing a forecast combination methodology that uses Bayesian shrinkage techniques to allow the incorporation of prior information into the estimation of combining weights; least squares and prior weights then emerge as polar cases for the posterior mean. The actual posterior mean combining weights are a matrix weighted average of those for the two polar cases, with the exact location depending on prior precision.

In many respects, this paper is an amplification and extension of the important work of Clemen and Winkler (1986). Unlike Clemen and Winkler, however, we work in the regression-based framework of Granger and Ramanathan (1984). Moreover, we focus explicitly on the usefulness of Bayesian shrinkage procedures in finessing the difficult decision of whether to combine forecasts by the Bates–Granger–Ramanathan approach or by simple averaging. Empirical Bayes procedures, which are used to estimate prior precision from the data, are a key element of the approach. We introduce the methodology in Section 2, followed by an empirical illustration in Section 3. Section 4 concludes.

2. Methodology

Consider m competing unbiased forecasts $f_t^1, \ldots, f_t^m$ of a variable y_t made at time $t-1$, from which we form a composite as

$$C_t = \omega_1 f_t^1 + \omega_2 f_t^2 + \cdots + \left(1 - \sum_{i=1}^{m-1} \omega_i\right) f_t^m.$$

It is easily verified that the forecast errors satisfy the same equality, from which the variance of the combined forecast error is obtained and found to be minimized at

$$\omega^* = (\Sigma^{-1} i)/(i' \Sigma^{-1} i),$$

where Σ is the variance–covariance matrix of one-step-ahead forecast errors and i is a conformable column vector of ones. The minimum variance combined forecast is also the minimum MSE combined forecast under the maintained assumption of unbiased primary forecasts. This approach, which for obvious reasons we refer to as the variance–covariance method, was developed by Bates and Granger (1969) and has been extended and succesfully applied by numerous authors. Granger and Ramanathan (1984), for example, show that the optimal combining weight vector has a regression interpretation as the coefficient vector of a linear projection of y onto the primary forecasts, subject to two constraints: the weights sum to unity, and an intercept is not included.

A flexible unconstrained regression-based forecast combination framework, which readily allows us to incorporate prior information, makes use of the g-prior model of Zellner (1986). Consider a standard linear forecast combination,

$$\underset{(T\times 1)}{Y} = \underset{(T\times K)}{F}\ \underset{(K\times 1)}{\beta} + \underset{(T\times 1),}{\varepsilon} \tag{1}$$

where ε is distributed as $N(0, \sigma^2 I)$, with the natural conjugate normal-gamma prior,

$$P_0(\beta, \sigma) \propto \sigma^{-K-v_0-1} \exp\{(-1/2\sigma^2) \times [v_0 s_0^2 + (\beta - \beta_0)' M (\beta - \beta_0)]\}.$$

Typically, $K = m + 1$, because an intercept is included in the regression. The likelihood is

$$L(\beta, \sigma/y, F) \propto \sigma^{-T} \exp\{(-1/2\sigma^2)(y - F\beta)'(y - F\beta)\}.$$

Combining the prior and likelihood in the usual fashion yields the joint posterior for (β, σ). Integrating over σ then yields the marginal posterior of β, which is multivariate t:

$$P_1(\beta/y, F) \propto [1 + (1/v_1)(\beta - \beta_1)' s_1^{-2} \times (M + F'F)(\beta - \beta_1)]^{-(K+v_1)/2},$$

with mean vector β_1 and covariance matrix

$$s_1^2 (M + F'F)^{-1} (v_1/(v_1 - 2)),$$

where

$$\beta_1 = (M + F'F)^{-1}(M\beta_0 + F'F\hat{\beta}),$$
$$\hat{\beta} = (F'F)^{-1}F'y,$$
$$v_1 = T + v_0,$$
$$s_1^2 = (1/v_1)[v_0 s_0^2 + y'y + \beta_0' M \beta_0 - \beta_1'(M + F'F)\beta_1].$$

For g-prior analysis we take M to be $M = gF'F$, $0 < g < \infty$, which yields the Bayes rule

$$\beta_1 = (gF'F + F'F)^{-1}(gF'F\beta_0 + F'F\hat{\beta}), \qquad (2)$$

the shrinkage characteristics of which are easily seen by rewriting it as

$$\beta_1 = \beta_0 + (1/(1+g))(\hat{\beta} - \beta_0). \qquad (3)$$

The g-prior approach amounts to a particular and convenient choice of natural conjugate prior, and may be derived formally as an application of rational expectations to data generated from a conceptual sample. (For details, see Zellner (1986).) The power of the approach stems from the fact that specification of the prior covariance structure, which is typically very difficult, is reduced to the choice of a single parameter. A large value of g implies, ceteris paribus, high prior precision and hence substantial shrinkage toward β_0. The smaller is g, the less shrinkage occurs.

Even the choice of g can present difficulties, however, unless one truly has subjective prior information regarding the covariance structure. It has been our experience that (in the context of forecast combination), although we often have rather strong prior information regarding location, we are agnostic regarding scale. We therefore have investigated various empirical Bayes procedures that enable *estimation* of g from observable data. An obvious approach is MSE or MAE optimization of g by cross validation. Alternatively, we might make use of a closed-form estimator of g. Such an estimator is now discussed.

We work entirely in terms of conditional densities, eventually replacing unknown parameters with sample estimates. We again consider the linear combining regression (1). If we take the usual normal prior for β conditional upon σ, [1]

$$P_0(\beta/\sigma) = N(\beta_0, \tau^2 A^{-1}),$$

we obtain the posterior

$$P_1(\beta/\sigma, y) = N(\beta_1, (\tau^{-2}A + \sigma^{-2}F'F)^{-1}),$$

where

$$\beta_1 = (\tau^{-2}A + \sigma^{-2}F'F)^{-1}(\tau^{-2}A\beta_0 + \sigma^{-2}F'F\hat{\beta}).$$

Under quadratic loss, the posterior mean β_1 is of course the Bayes rule, which we make operational by taking $A = F'F$ and replacing σ^2 and τ^2 with the estimators suggested by Judge and Bock (1978):

$$\hat{\sigma}^2 = [(y - F\hat{\beta})'(y - F\hat{\beta})]/T,$$
$$\hat{\tau}^2 = ([(\hat{\beta} - \beta_0)'(\hat{\beta} - \beta_0)]/\mathrm{tr}(F'F)^{-1}) - \hat{\sigma}^2.$$

After substituting $\hat{\sigma}^2$ and $\hat{\tau}^2$, it is easily seen that the empirical Bayes rule is Stein-like, shrinking $\hat{\beta}$ toward β_0, because

$$\beta_1 = \beta_0 + [1 - \hat{\sigma}^2/(\hat{\sigma}^2 + \hat{\tau}^2)](\hat{\beta} - \beta_0). \qquad (4)$$

Furthermore, this expression makes clear the fact that the g-prior Bayes rule (3) is equivalent to the empirical Bayes rule (4) when $g = \hat{\sigma}^2/\hat{\tau}^2$. (Of course, *both* estimators (3) and (4) are empirical Bayes; in order to distinguish them, we adopt the convention of referring to (3) as the "g-prior estimator" and (4) as the "empirical Bayes estimator".)

Finally, we note that the g-prior posterior mean estimator, and therefore the empirical Bayes estimator, is equivalent to the least squares estimator on transformed data, i.e., $\beta_1 = (W'W)^{-1}W'\lambda$, where

$$\lambda = \begin{bmatrix} m^{1/2}\beta_0 \\ y \end{bmatrix} \quad \text{and} \quad W = \begin{bmatrix} M^{1/2} \\ F \end{bmatrix},$$

and $M^{1/2}$ is a symmetric matrix such that $M^{1/2}M^{1/2} = M$. Thus, the usual recursions for real-time parameter (i.e., combining weight) estimation are valid.

[1] We do not restrict the extracted constant τ^2 to equal the model disturbance variance σ^2, which is why, in general, $A \neq M$.

To summarize the results, we have proposed a simple normal-gamma approach for the incorporation of prior information in regression-based forecast combination. Two data-based methods for assessing the prior covariance matrix were discussed and contrasted, and it was shown that each of the resulting posterior-mean estimators fits naturally in a real-time recursive estimation framework.

3. An empirical illustration

In order to explore the usefulness of the incorporation of prior information in forecast combination, we explore the possibilities for combining four econometric model-based (and judgmentally augmented) forecasts of both nominal and real GNP; the forecasts are those of Wharton Econometrics, Chase Econometrics, Data Resources, and the Bureau of Economic Analysis. These ex ante real-time quarterly one-step-ahead forecasts cover the period 1971–1982, yielding a total of 46 observations. [2] All variables are expressed as percentage change, at annual rates. Analysis of such a data set is interesting for a number of reasons. First, its two variables reflect the differing financial and real sides of the economy, both of which are of importance in planning and decision making. It is interesting to see whether forecast enhancement due to combining differs across the real and financial sectors. Second, this data set is representative of the type of forecast combination likely to be attempted in practice. One can easily imagine a corporate decision-maker receiving the forecasts of more than one service. Third, the stochastic properties of real and nominal GNP (in terms of trend and cyclical components, for example) are likely to be representative of economic variables in general, and the stochastic properties of the four forecasts (e.g. high contemporaneous correlation) are likely to be indicative of the properties of sets of other economic forecasts. Finally, the fact that the data have been used previously in published work enables replication of existing results, hence establishing a useful baseline. [3]

[2] The exact sample is 1970-IV through 1982-II. There is a missing observation corresponding to 1981-I; following Clemen and Winkler (1986), we simply omit it. This is admissible because our combining regressions are not dynamic. For a treatment of dynamic aspects see Diebold (1988).

[3] We successfully replicated the results of Clemen and Winkler (1986) for each of the four primary forecasts, the variance–covariance combination and the regression-based combination, but the truncated-sample weighting scheme that Clemen and Winkler used led to little enhancement of combined forecast performance. Because the weighting made little difference, we focused on unweighted combining in this paper.

Table 1
Performance of GNP forecasts. [a]

	Nominal GNP		Real GNP	
	RMSE	MAE	RMSE	MAE
Uncombined				
WHARTON	4.35	3.68	3.41	2.86
CHASE	4.64	3.77	3.54	2.95
DRI	4.43	3.83	3.78	3.12
BEA	4.52	3.75	3.52	2.91
Combined				
Variance–covariance	4.67	3.91	3.82	3.24
Empirical Bayes	4.26(36.4)	3.55(36.4)	3.72(0.7)	3.06(0.7)
g-prior:				
$g = 0$ (OLS regression)	4.46	3.77	4.19	3.43
$g = 2$	4.29	3.54	3.50	2.88
$g = 8$	4.27	3.54	3.39	2.88
$g = 25$	4.26	3.55	3.36	2.88
$g = \infty$ (arithmetic avg.)	4.26	3.55	3.35	2.88

[a] Weights are calculated using observations 1–20, while forecast comparisons are based on observations 21–46. The empirical Bayes RMSE and MAE scores are followed (in parentheses) by the estimated g-value.

We calculate combining weights using the first 20 observations, and observations 21–46 are then used for forecast comparison; results appear in Table 1. Note that there are substantial differences in primary model forecasting performance for nominal versus real GNP; the greater RMSE for nominal GNP forecasts reflects errors in the prediction of inflation. The generally poor performance of the variance–covariance method is also apparent. Regression-based combining (equivalent to g-prior combining for $g = 0$) yields some improvement in the case of nominal GNP, but no improvement for real GNP. This is due to the upward bias in the primary model nominal GNP forecasts, caused in turn by consistent underprediction of inflation; the non-zero intercept of the regression-based approach removes this bias. No such bias is found in the primary real GNP forecasts, and OLS regression-based combining (i.e., g-prior combining with $g = 0$) fails to improve on variance–covariance combining. The important point, however, is that *both* variance–covariance and OLS regression-based composites yield disappointing results.

Our Bayesian regression-based composites fare much better, however. All reported results make use of a prior mean corresponding to forecast exchangeability (i.e., $\beta_0 = (0,\ 0.25,\ 0.25,\ 0.25,\ 0.25)$). (Alternative shrinkage directions produced no improvement.) The empirical Bayes procedure, which corresponds to data-based estimation of g, always improves upon both the variance–covariance and the OLS regression-based composites, sometimes drastically so. Some insight into this phenomenon may be gained by examining the performance of the g-prior combinations as g ranges from 0 (OLS regression-based) to ∞ (arithmetic average), also shown in the table. In each case, shrinkage produces substantial improvement, resulting in a combined forecast with lower RMSE and MAE than each primary forecasts in all cases but one. Two facts stand out. First, the forecast enhancement is highly nonlinear in g, with the vast majority of forecast improvement occurring by the time g is in the 1–5 range. (This is to be expected, of course, since the support of g_0 is the half-interval $[0,\infty]$. Thus, for example, $g_0 = 1$ implies a weight on $\hat{\beta}$ of $\frac{1}{2}$, $g_0 = 2$ implies a weight of $\frac{1}{3}$, and so forth.) Second, the global optimum tends to correspond to large amounts of shrinkage, for which most weight is placed on the simple average.

4. Summary and conclusions

We have developed and illustrated a number of simple techniques for the incorporation of prior information into unrestricted regression-based composite forecasts. The techniques are likely to be particularly useful when combining forecasts based on overlapping (but nevertheless different) information sets, as is so often the case.

The results are important because they enable forecasters to address and exploit a common, significant and surprising phenomenon: the simple average often provides accuracy comparable to forecasts from sophisticated "optimally" pooled forecasts. The techniques advocated here, in an important sense, deliver the best of both worlds: estimated combining weights are coaxed, or shrunken toward equality but are not forced to be exactly equal. Instead, the combining weights emerge as a weighted average of those for two polar cases, OLS regression and the simple average. The exact location depends on prior precision, which can be estimated from the data. In this way the data are allowed to speak, when (and if) they have something to say.

The results are illustrated by combining four major real-time ex ante forecasts of both nominal and real GNP. A number of conclusions emerge for this data set.

(1) Shrinkage enhances the ex ante predictive performance of regression-based composite forecasts and leads to pooled forecasts that dominate all primary forecasts.

(2) The arithmetic average prior is the shrinkage direction that produces the greatest forecast enhancement.

(3) The optimal amount of shrinkage is typically large. This result provides support for the frequently-reported good performance of simple arithmetic averages.

(4) MSE and MAE drop at a decreasing rate with the amount of shrinkage. Achieving nearly opti-

mal forecast enhancement requires only about 50% shrinkage. [4]

These conclusions are conditional on the small data set that we examined. Further empirical analysis on a wide range of economic forecasts is needed. It should also prove useful to explore shrinkage toward robust measures of location, such as the median forecast.

References

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

Clemen, R.T., 1989, "Combining forecasts: A review and annotated bibliography" (with discussion), *International Journal of Forecasting*, 5, 559–583.

Clemen, R.T. and R.L. Winkler, 1986, "Combining economic forecasts", *Journal of Business and Economic Statistics*, 4, 39–46.

Diebold, F.X., 1988, "Serial correlation and the combination of forecasts", *Journal of Business and Economic Statistics*, 6, 105–111.

Granger, C.W.J. and R. Ramanathan, 1984, "Improved methods of combining forecasts", *Journal of Forecasting*, 3, 197–204.

Judge, G.G. and M.E. Bock, 1978, *Statistical Implications of Pre-Test and Stein Rule Estimators in Econometrics* (North-Holland, Amsterdam).

Kang, H., 1986, "Unstable weights in the combination of forecasts", *Management Science*, 32, 683–695.

Zellner, A., 1986, "On assessing prior distributions and bayesian regression analysis with g-prior distributions", in: P. Goel and A. Zellner, eds., *Bayesian Inference and Decision Techniques* (North-Holland, Amsterdam).

[4] By "50% shrinkage", for example, we refer to a posterior mean such as eq. (3) or (4) with weights of $\frac{1}{2}$ on each of the least squares and prior parameter vectors.

Biographies: Francis X. DIEBOLD received his Ph.D. in Economics from the University of Pennsylvania in 1986, after which he joined the research staff of the Federal Reserve Board in Washington, D.C. In 1989 he returned to Penn's economics department as Joseph M. Cohen Assistant Professor of Economics. He has published widely in econometrics, macroeconomics and international economics.

Peter PAULY received his Ph.D. in Economics from the University of Hamburg, and he was Associate Professor of Economics at the University of Pennsylvania from 1981–1989. He is currently Professor of Economics at the University of Toronto and Associate Director of Project LINK. He has published widely in econometrics, macroeconomics and international economics.