ELSEVIER

International Journal of Forecasting 14 (1998) 83–93

international journal of forecasting

Combining qualitative forecasts using logit

Mark Kamstra, Peter Kennedy*

Department of Economics, Simon Fraser University, Burnaby, B.C., Canada V5A 1S6

Received 30 June 1996; received in revised form 31 January 1997; accepted 31 July 1997

Abstract

This paper introduces a computationally-convenient means of combining qualitative forecasts, through use of logit regression applied to training set data, applicable in dichotomous, polychotomous and ordered polychotomous contexts. It can be employed in the cases of combining probability forecasts, combining qualitative forecasts which have no associated probability forecasts, and combining both of these types of forecasts, a case for which no combining method currently exists. This methodology offers insights into the suitability of equal-weight averaging of probability forecasts, yields an existing method as a special case, and facilitates associated hypothesis testing. © 1998 Elsevier Science B.V.

Keywords: Probability forecasting; Event forecasting; Logit model; Ordinal data

1. Introduction

Combining forecasts is an accepted means of improving forecasts. A quick glance at the combining literature, however, suggests that it has focused almost exclusively on combining quantitative forecasts, with little attention paid to how combining can be undertaken in the context of forecasting qualitative variables. To the best of our knowledge, Feather and Kaylen (1989) and Fan et al. (1996) are the only work done in this area. In fact, the widely-cited review article on combined forecasts, by Clemen (1989), does not even mention combining qualitative forecasts. Yet, many forecasting problems are qualitative in nature. Examples include diagnosing a patient, granting a loan, and predicting the direction of a price change.

A more careful analysis of the literature reviewed by Clemen (1989) reveals, however, that the literature on combining probability forecasts is under special circumstances relevant to the problem of combining qualitative forecasts. Indeed, if the training set data and the qualitative forecasts to be combined have probability forecasts associated with them, as they often do, then combined qualitative forecasts can be produced by using combined probability forecasts. In general the forecast is chosen to produce the smallest expected cost, implying that, in the typical case of all forecast errors carrying equal costs, the category with the highest estimated probability is forecast.

Unfortunately, not all qualitative forecasts have probability forecasts associated with them. For example, the training set data may provide only the information that a doctor diagnosed a patient as ill, rather than providing an estimated probability that the patient has the disease in question. We refer to

*Corresponding author: Tel.: 604 2914516; fax: 604 2915944; e-mail: kennedy@sfu.ca

PII S0169-2070(97)00056-3

such cases as pure qualitative forecasts. The combining probability literature is not applicable to cases in which all or some of the available forecasts are pure qualitative forecasts. Feather and Kaylen (1989) and Fan et al. (1996) address the case of combining pure qualitative forecasts; to our knowledge there exists no literature addressing the problem of combining a mixture of probability forecasts and pure qualitative forecasts.

In a sense, though, the focus of the literature on probability forecasting is appropriate. Regardless of what type of information is available to a researcher in the training set data, a qualitative forecast can only be undertaken by estimating probabilities attached to alternative categories. Furthermore, the better are these estimated probabilities, in terms of their bias and variance, the better will be the qualitative forecasts that result from them. The methodology we propose in this paper is no exception to this: training set forecasts, some or all of which may not be in the form of probability estimates, are employed via a logit regression to produce probability estimates for new observations which are then used to make qualitative forecasts for these new observations.

The purpose of this paper is to introduce to the forecasting literature a generic method for combining qualitative forecasts, based on applying logit regression to the training set data to produce weights for combining. This methodology is applicable to the case of combining probabilities, to the case of combining pure qualitative forecasts, to the case of combining mixtures of probability and pure qualitative forecasts, to cases in which there are many rather than two qualitative categories, and to cases in which the categories are ordered. The main advantage of this logit regression method compared to existing methods is that, like the linear regression methodology of Granger and Ramanathan (1984) for determining weights to be used for combining quantitative forecasts, it is a computationally-attractive means of producing suitable combinations while alleviating bias. Further advantages are that it facilitates testing of related hypotheses and offers perspective on some existing methods in the literature. Indeed, the method of Feather and Kaylen (1989) is seen to be a special case of our combining method. The method of Fan et al. (1996) is shown later to be questionable because it requires a very restrictive assumption, avoided by using our method.

The paper proceeds by looking first at the case of a binary, or dichotomous, qualitative forecast, investigating how the logit regression can be applied to the combining of probability forecasts, the combining of pure qualitative forecasts and the combining of mixtures of probability and pure qualitative forecasts. Next it looks at the polychotomous case in which there are more than two categories, employing for this case multinomial logit. Last it looks at how the logit methodology can be extended to the ordered polychotomous case by using ordered logit.

2. Dichotomous qualitative forecasts

2.1. Combining probability forecasts

Qualitative forecasts are typically produced by using a classification procedure such as probit, logit, linear discriminant analysis, k-nearest neighbor or goal programming. They also may be produced by less formal means involving subjective expertise. Most of the formal procedures produce probability forecasts associated with their qualitative forecasts, whereas many of the less formal techniques do not. We begin our exposition by looking at the case in which probability forecasts are available so the problem is simply how best to combine these probability forecasts. To avoid later confusion, we note that one of the forecasts to be combined may have resulted from using logit regression in some suitable fashion. This is distinct from our use of logit regression as a combining mechanism.

Existing methods for combining probability forecasts range from average-the-probabilities to formal Bayesian techniques. A good discussion of these methods can be found in Winkler et al. (1977). French (1985) and Genest and Zidek (1986) are good surveys of combining probability forecasts which reflect individual beliefs. This paper does not distinguish between forecasts made by experts and forecasts made by classification methods such as logit and discriminant analysis.

For expositional reasons we look at the problem of combining probability forecasts from two techniques, A and B, and begin with the dichotomous case in which for convenience we call the two categories success and failure. We specify that the true probability of success for the ith observation is equal to a cumulative distribution F evaluated at an index value θ_i. Technique A's and B's probability-of-success estimates p_{Ai} and p_{Bi} for this ith observation, however produced, are viewed as having associated with them implicit estimated index values θ_{Ai} and θ_{Bi} so that

$$p_{Ai} = F(\theta_{Ai}) \text{ and } p_{Bi} = F(\theta_{Bi}).$$

Existing probability forecast combining methods combine in probability space so that p_{Ai} and p_{Bi} are combined to produce $(p_{Ai}+p_{Bi})/2$ for the case of equal weights, for example. In contrast, our combining method operates in θ space, so that θ_{Ai} and θ_{Bi} are combined to produce $F[(\theta_{Ai}+\theta_{Bi})/2]$ for the case of equal weights, for example. Our method takes the implicit θ estimates associated with the competing techniques and combines them in some suitable way to produce a combined θ estimate which can be translated through F into an estimated probability and hence a qualitative choice.

Choosing a specific F creates an operational framework for this methodology. F could be linear, truncated from below at zero and from above at one, corresponding to the linear probability model. This specification was abandoned in empirical work long ago as being unreasonable. It can create potentially embarrassing forecasts of certainty, and does not allow for what many believe are necessary nonlinearities associated with ceilings and floors. Modern empirical work has adopted an S-shaped curve for F, the most popular forms being the logistic function and the cumulative normal distribution. We adopt the logistic function because of its computational simplicity, its expositional clarity and the fact that, as shown below, for a special case it produces results identical to those of Feather and Kaylen (1989).

When F is the logit function we have for the example described above

$$\text{prob(success)} = \frac{e^{\theta}}{1+e^{\theta}}, \qquad (1)$$

$$\text{prob(failure)} = \frac{1}{1+e^{\theta}}.$$

For the ith observation the implicit θ_{Ai} and θ_{Bi} can be calculated as the log odds ratio so that $\theta_{Ai} = \ln[p_{Ai}/(1-p_{Ai})]$ and $\theta_{Bi} = \ln[p_{Bi}/(1-p_{Bi})]$. Our proposal amounts to combining probability forecasts by using a logit regression on the training set data to estimate weights for a weighted average of these two θ values.

Our method is similar in spirit to that of Granger and Ramanathan (1984) who show that the estimated optimal weighting of Bates and Granger (1969) for quantitative forecasts can be found by using the training set data to run a regression of actual value on forecasted values, constraining the intercept to be zero and the slope coefficients to sum to one. They note that removing these constraints allows an automatic adjustment for possible bias in the individual forecasts. Unfortunately, the Granger and Ramanathan technique is not applicable to most probability forecasting problems. This is because in probability forecasting the true probability being forecast is usually not known in the training set, so that it is not possible to run the regression which is the heart of their technique. Even if it were possible to run this regression, their method would not be suitable because unconstrained regression could produce forecast probabilities outside the zero-one interval. A nonlinear combining estimating form must be adopted to exploit knowledge that the estimated probabilities should lie between zero and one.

In place of the linear regression of Granger and Ramanathan, we employ a logit regression. In particular, we specify

$$\text{prob(success)} = \frac{e^{\mu+\alpha\theta_A+\beta\theta_B}}{1+e^{\mu+\alpha\theta_A+\beta\theta_B}},$$

$$\text{prob(failure)} = \frac{1}{1+e^{\mu+\alpha\theta_A+\beta\theta_B}}.$$

Estimation is by maximum likelihood and makes use of the success/failure information in the training set data, and θ values calculated for each technique as the log odds ratio using that technique's estimated probability. Once the θ values have been calculated estimation is easily performed using any software with a logit regression routine.

This method is an obvious extension of the Granger and Ramanathan technique, but we make no claims that the resulting probability forecasts are the best possible. Such optimality would only occur asymptotically (because of the nonlinearities) and in the absence of an appeal to a quasi-likelihood argument, would require that the actual probability be determined through a logit function by an underlying θ index of which the log odds ratios of competing probability forecasts are suitable estimates. Although this may be a more reasonable approximation to the actual underlying process than the linear F associated with the current practice of averaging probabilities, we only claim that this methodology is a means of combining individual probability forecasts in a computationally-attractive manner, while alleviating bias. Allowing the intercept in the exponent to be nonzero and the slopes to sum to other than unity alleviates bias for the same reason that it does so in the regression context of Granger and Ramanathan, and simultaneously produces weights which can adjust to reflect different forecast error variances.

An important lesson from the combining literature is that estimation of suitable weights for combining probability forecasts (in probability space) often gives rise to inferior forecasts because these weights are estimated so poorly. See, for example, Einhorn and Hogarth (1975); Winkler et al. (1977); Kang (1986); Blattberg and Hoch (1990) and Schmittlein et al. (1990). DeWispelare et al. (1995) summarize this literature by noting that 'the simple average of point forecasts has tended to do as well, (and often better) than more complex methods'. If F had been linear instead of logistic, it is easily seen that averaging the θ estimates produces the raw average of the p estimates: the popular method of combining by averaging probability estimates is implicitly assuming a linear F! This offers useful perspective on the popular method of combining by averaging probabilities, information which Winkler (1989) claims would be of particular value in matching combining rules to forecasting situations.

Because in general a linear F is unrealistic relative to a logistic F, a method which exploits the logistic should be an attractive alternative to the method of combining by averaging probabilities. Empirical results reported later in this paper indicate that using fixed, equal weights in our logistic combining method produces results virtually identical to those produced by a simple average of probability estimates. Furthermore, when sample sizes are small our logit combining method suffers from the same problem as methods using estimated weights for averaging in probability space: the weights are often estimated poorly and this causes these methods to be outperformed by methods using fixed, equal weights. The main advantage of our logistic combining method is that it is theoretically more satisfactory and in large samples provides a simple, effective means of finding appropriate weights which automatically alleviate bias.

An added advantage of the logit combining method is that hypothesis testing is easily undertaken, conditional on the training set data and assuming that the logit specification is appropriate. The rationale for this follows that used by Fair and Schiller (1990) for testing in the context of a linear combination of quantitative forecasts. For example, examining whether technique A contributes beyond technique B in Eq. (1) above can be tested by testing α against zero using a likelihood ratio test or an asymptotic t test. This feature is very important. Combining is attractive so long as all the forecasts to be combined are 'good', but existing combining methods for qualitative forecasting do not have an easy way of identifying 'bad' forecasts. We turn now to another attractive feature of this combining method: it nests an existing method as special case.

2.2. Combining pure qualitative forecasts

Let us move now to the case in which the competing techniques provide only qualitative forecasts, without any indication of what probability estimate has given rise to that forecast. Feather and Kaylen (1989) provide several examples of such situations, and analyze a case of forecasting the direction of hog price changes. To our knowledge Feather and Kaylen have produced the only combining methodology for this case in the literature. Their method begins by subdividing the training set into groups corresponding to unique combinations of competing forecasts. For our example there are four groups:

1. both A and B forecast success;
2. both A and B forecast failure;

3. A forecasts success and B forecasts failure; and
4. A forecasts failure and B forecasts success.

Feather and Kaylen then assume that given the group to which an observation belongs the probabilities of success and failure (and other outcomes, if relevant) are random variables having a joint Dirichlet distribution. Estimates of these probabilities are obtained by estimating their expected values through estimation of the parameters of this Dirichlet distribution. For each group this produces a combined estimate of the probability of success equal to the fraction of successes in that group in the training set. For small samples in which a group may have no successes or no failures, an uninformative prior is added to avoid probability forecasts of zero or one. This prior is equivalent to adding to each group two additional training set observations, one a success and one a failure.

A natural extension of the logit methodology developed earlier is to use as explanatory variables in the logit specification four dummy variables representing the four groups defined above, dropping the intercept to avoid perfect multicollinearity. (We have dropped the intercept here for expositional convenience; in actual applications estimation may be facilitated by retaining the intercept and dropping one of the dummies.) This creates the specification

$$\text{prob(success)} = \frac{e^{\delta_1 D_1 + \delta_2 D_2 + \delta_3 D_3 + \delta_4 D_4}}{1 + e^{\delta_1 D_1 + \delta_2 D_2 + \delta_3 D_3 + \delta_4 D_4}},$$

$$\text{prob(failure)} = \frac{1}{1 + e^{\delta_1 D_1 + \delta_2 D_2 + \delta_3 D_3 + \delta_4 D_4}},$$

where D_k is a dummy variable taking the value one for observations in group k above and zero otherwise. Probability forecasts can be made by finding the maximum likelihood estimates of the δ_k and substituting them into this logit specification. Straightforward algebra shows that doing so produces for observations in group k a probability forecast equal to the fraction of successes in group k in the training set. This is exactly the result of Feather and Kaylen! The modification of the Feather and Kaylen result produced by adding the uninformative prior can be obtained by adding to the training set data an artificial success and an artificial failure observation for each group and including these artificial observations in the logit regression.

One advantage of the logit regression approach to computing the Feather and Kaylen estimates is that hypotheses of interest can easily be tested. One may be interested in whether method B makes any difference when combined with method A. This could be tested by using a likelihood ratio test to test the joint hypothesis that $\delta_1 = \delta_3$ and $\delta_2 = \delta_4$. Such hypothesis tests are of value because they can help avoid unsuitable forecasts, as noted earlier. They also can serve as a means of simplifying the combining method by reducing the number of regressors in the logit regression. In our example every additional forecasting method doubles the number of groups in the Feather and Kaylen method, so an easy means of moving towards a more parsimonious specification is attractive.

2.3. Combining using probability forecasts and pure qualitative forecasts

A third generic combining problem in the qualitative forecasting context arises when some of the individual techniques provide associated probability estimates but others provide only the qualitative forecasts. Dawes et al. (1989) provide an example of such a problem in their discussion of combining clinical and actuarial judgements regarding medical diagnoses. To our knowledge there does not exist in the literature a means of combining probability forecasts and pure qualitative forecasts. By now it should be obvious how the logit regression technique introduced in this paper can be employed in this context—the regressors in the logit regression must include both θ estimates and group dummies.

The simplest example occurs when technique A yields a probability estimate and technique B produces only a qualitative forecast. In this case there is one continuous variable θ_A, and one dummy D taking the value one for observations for which B forecasts success and zero otherwise. The combining specification is

$$\text{prob(success)} = \frac{e^{\mu + \alpha\theta_A + \delta D}}{1 + e^{\mu + \alpha\theta_A + \delta D}},$$

$$\text{prob(failure)} = \frac{1}{1 + e^{\mu + \alpha\theta_A + \delta D}}.$$

3. Polychotomous qualitative forecasts

3.1. Combining probability forecasts

We turn now to the case in which there are more than two categories so that there are more than two probabilities to be forecast. To simplify exposition, we consider the case of three probabilities to be forecast, those associated with success, failure and no change, and continue to combine the forecasts of only two techniques, A and B. The fundamental logic of our approach remains the same, but now we employ the multinomial logit so that we specify

$$\text{prob(success)} = p_{S,i} = \frac{e^{\theta_i}}{1 + e^{\theta_i} + e^{\phi_i}} \quad i = A, B,$$

$$\text{prob(failure)} = p_{F,i} = \frac{e^{\phi_i}}{1 + e^{\theta_i} + e^{\phi_i}} \quad i = A, B,$$

$$\text{prob(no change)} = p_{N,i} = \frac{1}{1 + e^{\theta_i} + e^{\phi_i}} \quad i = A, B,$$

Estimated probabilities $p_{S,A}$, $p_{F,A}$ and $p_{N,A}$ for technique A correspond to implicit θ and ϕ values given by the log odds ratios $\theta_A = \log[p_{S,A}/p_{N,A}]$ and $\phi_A = \log[p_{F,A}/p_{N,A}]$. The multinomial logit can be used to combine these probabilities using the specification

$$\text{prob(success)} = \frac{e^{\mu + \alpha\theta_A + \beta\theta_B}}{1 + e^{\mu + \alpha\theta_A + \beta\theta_B} + e^{\mu + \alpha\phi_A + \beta\phi_B}},$$

$$\text{prob(failure)} = \frac{e^{\mu + \alpha\phi_A + \beta\phi_B}}{1 + e^{\mu + \alpha\theta_A + \beta\theta_B} + e^{\mu + \alpha\phi_A + \beta\phi_B}},$$

$$\text{prob(no change)} = \frac{1}{1 + e^{\mu + \alpha\theta_A + \beta\theta_B} + e^{\mu + \alpha\phi_A + \beta\phi_B}}.$$

Estimation can be undertaken by maximum likelihood using a multinomial logit estimation routine. Different variants of multinomial logit routines are available for different types of multinominal specifications. A distinguishing feature of this specification is that the explanatory variables are different for each category (i.e. θ versus ϕ values) and the parameters the same for each category. Software packages may not refer to this case as multinomial logit. In LIMDEP and TSP, for example, two popular software packages for estimating qualitative choice models, this specification is referred to as 'discrete choice' and 'conditional logit', respectively, to distinguish it from the case which they call 'multinomial logit' in which the set of explanatory variables is the same for all categories, with parameters differing across categories.

Forcing the parameters to be identical across categories is specifying that the averaging weights are the same across categories. In theory we could allow the parameters to differ across categories, but in light of so much evidence in the literature suggesting that simple specifications (such as equal weights) are better than more sophisticated specifications, it seems reasonable to impose this constraint.

This combining methodology is bound to be plagued by the independence of irrelevant alternatives problem (see Kennedy, 1992). In the multinomial logit model the relative probability of choosing two alternatives is unaffected by the presence of additional alternatives, a constraint which is inappropriate if any of these additional alternatives are close substitutes. If the forecasts to be combined are highly correlated, which will often be the case, this combining technique may not perform well. Combining should only be done with alternatives which are not highly correlated. The only way around this independence of irrelevant alternatives problem is to use multinomial probit in which the logit is replaced by a cumulative normal distribution. In this specification correlations between alternatives can be accounted for, but at a very high computational cost requiring specialized software.

3.2. Combining pure qualitative forecasts

Suppose now that the training set data for our success versus failure versus no change example provides only the qualitative forecast for each of techniques A and B, and not the associated probability estimate. For this case there are nine possible combinations of competing forecasts, three different A forecasts for each of the three different B forecasts. Using dummy variables D_k for $k = 1,...,9$ to represent each of these combinations we use the specification

$$\text{prob(success)} = \frac{e^{\delta_1 D_1 + \cdots + \delta_9 D_9}}{1 + e^{\delta_1 D_1 + \cdots + \delta_9 D_9} + e^{\lambda_1 D_1 + \cdots + \lambda_9 D_9}},$$

$$\text{prob(failure)} = \frac{e^{\lambda_1 D_1 + \cdots + \lambda_9 D_9}}{1 + e^{\delta_1 D_1 + \cdots + \delta_9 D_9} + e^{\lambda_1 D_1 + \cdots + \lambda_9 D_9}},$$

prob(no change)

$$= \frac{1}{1 + e^{\delta_1 D_1 + \cdots + \delta_9 D_9} + e^{\lambda_1 D_1 + \cdots - \lambda_9 D_9}}.$$

Estimation can be undertaken by using maximum likelihood via a multinomial logit software package. In this case the explanatory variables have the same values for all categories and the parameters differ across categories. This is the case called 'multinomial logit' by LIMDEP and TSP, noted earlier. Probability forecasts can be made by substituting estimates of the δ_k and λ_k into this logit specification. Straightforward algebra shows that doing so produces for observations in group k a probability forecast equal to the fraction of successes in group k in the training set, once again exactly the result of Feather and Kaylen.

3.3. Combining using probability forecasts and pure qualitative forecasts

Suppose now that some of the forecasting techniques provide associated probability estimates but others provide only the qualitative forecasts. In this case the regressors in the multinomial logit regression must include θ and ϕ estimates as well as group dummies. Consider the simplest case in which technique A yields probability estimates and technique B produces only qualitative forecasts. Define D_S as a dummy taking value one when B forecasts success, otherwise zero, and D_F as a dummy taking value one when B forecasts failure, otherwise zero. Then the combining specification is

prob(success)

$$= \frac{e^{\mu + \alpha\theta_A + \delta_1 D_S + \gamma_1 D_F}}{1 + e^{\mu + \alpha\theta_A + \delta_1 D_S + \gamma_1 D_F} + e^{\mu + \alpha\phi_A + \delta_2 D_S + \gamma_2 D_F}},$$

prob(failure)

$$= \frac{e^{\mu + \alpha\phi_A + \delta_2 D_S + \gamma_2 D_F}}{1 + e^{\mu + \alpha\theta_A + \delta_1 D_S - \gamma_1 D_F} + e^{\mu + \alpha\phi_A + \delta_2 D_S + \gamma_2 D_F}},$$

prob(no change)

$$= \frac{1}{1 + e^{\mu + \alpha\theta_A + \delta_1 D_S + \gamma_1 D_F} + e^{\mu + \alpha\phi_A + \delta_2 D_S + \gamma_2 D_F}},$$

4. Ordered polychotomous forecasts

4.1. Combining probability forecasts

Sometimes it is known that the polychotomous outcomes are ordered. Bond ratings, for example, are such that a triple A rating is superior to a double A rating which in turn is superior to a single A rating, and so on. For this case we specify that there is an index θ called, say, creditworthiness, which determines classification. Suppose there are J categories, ordered from 1, the lowest, to J, the highest. As θ increases and exceeds progressively larger unknown threshold values α_j, $j = 1,\ldots,J-1$, classification changes from category j to category $j+1$. The true probability that the ith observation belongs to category j is given by the integral of a standard logit from $\alpha_{j-1} - \theta_i$ to $\alpha_j - \theta_i$. (For $j=1$ the lower limit is minus infinity and for $j=J$ the upper limit is infinity.) This reflects the thinking behind the ordered logit model (see Greene, 1993 for details).

We view each forecasting technique as producing for each observation $J-1$ measures $\omega_{ji} = a_j - \theta_i$, $j = 1,\ldots,J-1$. Straightforward algebra shows that these measures can be estimated as

$$\omega_{ji} = \ln\left[\frac{p_{li} + \cdots + p_{ji}}{1 - p_{li} - \cdots - p_{ji}}\right],$$

where p_{ji} is an estimate of the probability that the ith observation falls in the jth category. Thus technique A's estimated probability that the ith observation belongs to category j is given by the integral of a standard logit from $w_{j-1,i,A}$ to $w_{j,i,A}$ where the undefined $w_{0,i,A}$ and $w_{J,i,A}$ are minus infinity and plus infinity, respectively. The ordered logit combining method we propose involves suitably weighting different techniques' ω values. In this case combining occurs in the space of integral limits rather than probability or index (θ) space.

Let us exposit all this in terms of a simple example. Suppose there are only three categories, ordered from lowest to highest as 1 = failure, 2 = no change and 3 = success. Then for the ith observation we have

$$\text{prob(failure)} = \frac{e^{\alpha_1-\theta_i}}{1+e^{\alpha_1-\theta_i}} = \frac{e^{\omega_{1i}}}{1+e^{\omega_{1i}}},$$

$$\text{prob(no change)} = \frac{e^{\alpha_2-\theta_i}}{1+e^{\alpha_2-\theta_i}} - \frac{e^{\alpha_1-\theta_i}}{1+e^{\alpha_1-\theta_i}} = \frac{e^{\omega_{2i}}}{1+e^{\omega_{2i}}} - \frac{e^{\omega_{1i}}}{1+e^{\omega_{1i}}},$$

$$\text{prob(success)} = \frac{1}{1+e^{\alpha_2-\theta_i}} = \frac{1}{1+e^{\omega_{2i}}}.$$

Now suppose there are two forecasting techniques A and B. We specify that for the ith observation

$$\text{prob(failure)} = \frac{e^{\pi_1+\pi_A\omega_{1iA}+\pi_B\omega_{1iB}}}{1+e^{\pi_1+\pi_A\omega_{1iA}+\pi_B\omega_{1iB}}},$$

$$\text{prob(no change)} = \frac{e^{\pi_2+\pi_A\omega_{2iA}+\pi_B\omega_{2iB}}}{1+e^{\pi_2+\pi_A\omega_{2iA}+\pi_B\omega_{2iB}}} - \frac{e^{\pi_1+\pi_A\omega_{1iA}+\pi_B\omega_{1iB}}}{1+e^{\pi_1+\pi_A\omega_{1iA}+\pi_B\omega_{1iB}}},$$

$$\text{prob(success)} = \frac{1}{1+e^{\pi_2+\pi_A\omega_{2iA}+\pi_B\omega_{2iB}}}.$$

This can be estimated using an ordered logit software package with the ω_A and ω_B values as explanatory variables and the π_1 and π_2 parameters playing the role of the unknown threshold values.

4.2. Combining pure qualitative forecasts

Suppose now that the training set data for our success versus failure versus no change example provides only the qualitative forecast for each of techniques A and B, and not the associated probability estimates. For this case there are nine possible combinations of competing forecasts, three different A forecasts for each of the three different B forecasts. Using dummy variables D_k for $k = 1,...,9$ to represent each of these combinations we use the specification

$$\text{prob(failure)} = \frac{e^{\varphi_1 D_1+\cdots+\varphi_9 D_9}}{1+e^{\varphi_1 D_1+\cdots+\varphi_9 D_9}},$$

$$\text{prob(no change)} = \frac{e^{\psi_1 D_1+\cdots+\psi_9 D_9}}{1+e^{\psi_1 D_1+\cdots+\psi_9 D_9}} - \frac{e^{\varphi_1 D_1+\cdots+\varphi_9 D_9}}{1+e^{\varphi_1 D_1+\cdots+\varphi_9 D_9}},$$

$$\text{prob(success)} = \frac{1}{1+e^{\psi_1 D_1+\cdots+\psi_9 D_9}}.$$

We saw earlier that ordered logit combines probability forecasts by combining integral limits which have come from probability estimates. In the case of pure qualitative forecasts these probability estimates are unavailable and so it would be surprising if this ordered logit combining technique could improve upon multinomial logit. Indeed, it does not. Using the equations above, we can form the likelihood function, maximize with respect to the unknown φ and ψ parameters, and make probability forecasts by substituting estimates of the φ and ψ parameters into this ordered logit specification. Straightforward algebra shows that doing so produces for observations in group k a probability forecast for category j equal to the fraction of j observations in group k in the training set, once again exactly the result of Feather and Kaylen.

This case of combining ordered polychotomous forecasts when only pure qualitative forecasts are available is addressed by Fan et al. (1996). They propose three combining methods, one using multinomial logit, one using linear programming and one using mixed integer programming. Unfortunately, all three are based on arbitrarily assigned constants whose choice affects the combining outcome. Let us illustrate this using our example of failure versus no change versus success. Fan et al. assume that each of the three categories has a latent index associated with it, and the actual outcome for an observation is determined by the highest of that observation's three index values. Forecasters are viewed as contributing information on the values of these latent indices. In particular, those forecasting failure for an observation are assumed to have a conditional expectation (conditional on the data) of these indices of H for the failure index, M for the no-change index and L for the success index. Those forecasting success are assumed to have a conditional expectation of L for the failure index, M for the no-change index and H

for the success index. And those forecasting no change are assumed to have a conditional expectation of $(M+L)/2$ for the success and the failure indices and H for the no-change index. Unfortunately, Fan et al. require knowledge of H, M and L to be able to use their three combining methods (multinomial logit, linear programming and mixed integer programming). In their example, they arbitrarily choose H, M and L to be 3, 2 and 1. This introduces additional information into the combining procedure, namely that, for example, someone forecasting failure for an observation believes that the conditional expectation of the latent failure index value for that observation is three times its success index value and one-and-a-half times its no-change index value. There is no justification for this. Our use of ordered logit circumvents this problem.

4.3. Combining using probability forecasts and pure qualitative forecasts

Suppose now that some of the forecasting techniques provide associated probability estimates but others provide only the qualitative forecasts. In this case the regressors in the ordered logit regression must include the w values plus group dummies. Consider the simplest case in which technique A yields probability estimates and technique B produces only qualitative forecasts. As earlier, define D_F as a dummy taking value one when B forecasts failure otherwise zero, and D_S as a dummy taking value one when B forecasts success, otherwise zero. Then the combining specification is

$$\text{prob(failure)} = \frac{e^{\pi_1 + \alpha\omega_{1A} + \delta_1 D_S + \gamma_1 D_F}}{1 + e^{\pi_1 + \alpha\omega_{1A} + \delta_1 D_S + \gamma_1 D_F}},$$

$$\text{prob(no change)} = \frac{e^{\pi_2 + \alpha\omega_{2A} + \delta_2 D_S + \gamma_2 D_F}}{1 + e^{\pi_2 + \alpha\omega_{2A} + \delta_2 D_S + \gamma_2 D_F}} - \frac{e^{\pi_1 + \alpha\omega_{1A} + \delta_1 D_S + \gamma_1 D_F}}{1 + e^{\pi_1 + \alpha\omega_{1A} + \delta D_S + \gamma_1 D_F}},$$

$$\text{prob(success)} = \frac{1}{1 + e^{\pi_2 + \alpha\omega_{2A} + \delta_2 D_S + \gamma_2 D_F}}.$$

5. Empirical illustration

To illustrate the logit combining technique we use data on Canadian female labor force participation from Atkinson et al. (1977). For 263 observations we have data on a three-category dependent variable—no participation, part-time participation and full-time participation. We analyse these data first as a polychotomous qualitative forecasting problem, and then as a binary qualitative forecasting problem by collapsing part-time and full-time participation into a single category. As explanatory variables we use data on husband's income, a dummy for presence or absence of children, and five dummies for region.

We compare four forecasting methods using these data to classify observations. Method KNN uses the k-nearest-neighbor technique, method LDA uses linear discriminant analysis, method AVG averages the KNN and LDA probability forecasts, and method LOG combines the KNN and LDA probability forecasts using the logit combining technique introduced in this paper. In the polychotomous case we force the parameters to be identical across categories, as recommended earlier. Forecasting success is measured by two popular measures, the error rate (ERR) and the mean probability score (MPS) of Brier (1950). Both are estimated using the leaving-one-out method, i.e. each observation is forecast using all the other observations as data for the forecasting method. ERR is calculated as the percentage of the observations incorrectly forecast. MPS is calculated as the average over all observations of the square of one minus the forecasted probability of the event that actually occurred, plus the sum of the squares of the forecasted probabilities for all the other events. For both ERR and MPS small values reflect better forecasting. The results, reported in Table 1, indicate that for this example the logit combining method is superior on all scores.

Likelihood ratio tests were performed to test if either of the forecasting methods should be ignored.

Table 1
Comparing forecasting methods

Method	Polychotomous		Binary	
	ERR	MPS	ERR	MPS
KNN	0.411	0.562	0.217	0.359
LDA	0.548	0.594	0.202	0.326
AVG	0.384	0.568	0.202	0.337
LOG	0.323	0.486	0.198	0.278

The LR statistics for the null that KNN should be excluded were 4.95 and 14.63 for the binary and polychotomous cases, respectively, implying that at the 1% significance level KNN should be retained for both. The LR statistics for the null that LDA should be excluded were 4.58 and 1.44 for the binary and polychotomous cases, respectively, implying that at the 1% significance level LDA should be retained for the binary case but discarded for the polychotomous case.

To lend some perspective to this example, we summarize results of some ongoing Monte Carlo work investigating contexts in which LOG is superior to LAVG, the LOG method with fixed, equal weights which in practice performs very much like AVG. In this Monte Carlo work the base case is two forecasters with unbiased log-odds estimates, both with errors drawn randomly from a normal distribution with standard error 0.5. (For true probability one-half, a log odds error of 0.5 means a probability estimation error of 12 percentage points, a magnitude that drops steadily as the true probability moves closer to its extremes.) For sample size 263 in this base case the mean squared error of the probability estimate for LAVG, which for this case is in theory the method of choice, is 68% of that of LOG. This advantage disappears when the standard error of one method becomes twice that of the other, if a collective log odds estimation bias of 0.5 is introduced, or some combination of these two events occurs. This simply reflects the fact that LOG can benefit by using different combining weights and by absorbing bias into its constant term.

In general our preliminary Monte Carlo results show that LAVG, and its counterpart calculated by averaging the log odds estimates, are remarkably robust in small sample sizes. Only if the sample size is reasonably large (50 is definitely too small; 263 is large enough to make possible gains in the circumstances described above) can sufficient accuracy be obtained for LOG to have an advantage for reasonable error variance differences and bias magnitudes. The same is true of using the logit method to combine a probability forecast with a pure qualitative forecast. If the probability forecast is reasonably good and the sample size is modest, no advantage is gained by using the logit combining method: a researcher is better off ignoring the extra information inherent in the pure qualitative forecast. However if the sample size is reasonably large, or if the probability forecast is poor (high error variance and/or substantive bias), the logit method can be superior.

Large sample sizes offset the loss of information in the logit regression procedure caused by collinearity between competing methods' log odds ratios, a problem which is worsened by adding more competing methods. In general, we have found that with adequate sample size, restricting the logit combining process to combining the two best individual methods can produce a small improvement in forecasting over individual methods, but adding additional individual methods usually wipes out this improvement by augmenting the collinearity. This reflects the general result noted earlier that situations which cause weights to be poorly estimated should be avoided. As a last example of this, in the polychotomous case reported in Table 1, had we not forced the parameters to be identical across categories, our logit combining method would not have outperformed its fixed-weight competition.

6. Conclusion

This paper has shown how logit regression can be used to combine qualitative forecasts and facilitate related hypothesis testing. It is similar in spirit to the regression method of Granger and Ramanathan (1984) which provides a computationally-convenient method for finding appropriate weights for combining while at the same time alleviating bias. Several results of note have emerged.

1. A new means of combining probability forecasts was introduced.
2. New light has been shed on the circumstances in which averaging probability forecasts with equal weights is an appropriate means of combining.
3. For combining pure qualitative forecasts, in the dichotomous, polychotomous and ordered polychotomous cases the logit, multinomial logit and ordered logit combining methods introduced here produced results identical to those of Feather and Kaylen (1989), the only existing means of combining pure qualitative forecasts without making arbitrary assumptions.

4. The logit regression technique can combine a mixture of probability forecasts and pure qualitative forecasts, something no existing combining method can do.
5. The logit combining method facilitates related hypothesis testing.

References

Atkinson, T., Blishen, B., Ornstein B., Stevenson, H., 1977. Social Change in Canada Project. Institute for Behavioral Research of York University, Toronto.

Bates, J.M., Granger, C.W.J., 1969. The combination of forecasts. Operational Research Quarterly 20, 451–468.

Blattberg, R., Hoch, S., 1990. Database models and managerial intuition: 50% model +50% manager. Management Science 36, 887–899.

Brier, G.W., 1950. Verification of forecasts expressed in terms of probability. Monthly Weather Review 78, 1–3.

Clemen, R.R., 1989. Combining forecasts: a review and annotated bibliography. International Journal of Forecasting 5, 559–583.

Dawes, R.M., Faust, D., Meehl, P.A., 1989. Clinical versus actuarial judgement. Science 243, 1668–1673.

DeWispelare, A.R., Herren, L.T., Clemen, R.T., 1995. The use of probability elicitation in the high-level nuclear waste regulation program. International Journal of Forecasting 11, 5–24.

Einhorn, H.J., Hogarth, R.M., 1975. Unit weighting schemes for decision making. Organizational Behavior and Human Performance 13, 171–192.

Fair, R.C., Schiller, R.J., 1990. Comparing information in forecasts from econometric models. American Economic Review 80, 375–389.

Fan, D.K., Lau, K.-N., Leung, P.-L., 1996. Combining ordinal forecasts with an application in a financial market. Journal of Forecasting 15, 37–48.

Feather, P.M., Kaylen, M.S., 1989. Conditional qualitative forecasting. American Journal of Agricultural Economics 71, 195–201.

French, S., 1985. Group consensus probability distributions: a critical survey. In: Bernardo, J.M., DeGroot, M.H., Lindley, D.V., Smith, A.F.M. (Eds.), Bayesian Statistics, vol. 2. North Holland, Amsterdam, pp. 183–197.

Genest, C., Zidek, J.V., 1986. Combining probability distributions: a critique and annotated bibliography. Statistical Science 1, 114–135.

Granger, C.W.J., Ramanathan, R., 1984. Improved methods of forecasting. Journal of Forecasting 3, 197–204.

Greene, W.H., 1993. Econometric Analysis. Macmillan, New York.

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

Kennedy, P.E., 1992. A Guide to Econometrics. MIT Press, Cambridge, MA.

Schmittlein, D.C., Kim, J., Morrison, D.G., 1990. Combining forecasts: operational adjustments to theoretically optimal rules. Management Science 36, 1044–1056.

Winkler, R.L., 1989. Combining forecasts: a philosophical basis and some current issues. International Journal of Forecasting 5, 605–609.

Winkler, R.L., Murphy, A.H., Katz, R.W., 1977. The consensus of subjective probability forecasts: are two, three,...., heads better than one? Preprint volume Fifth Conference on Probability and Statistics, Nov. 15–18, 1977, Las Vegas, Nevada, American Meteorological Society, Boston, MA, pp. 57–62.

Biographies: Mark KAMSTRA received his PhD in economics from UCSD, and is currently Assistant Professor of Economics at Simon Fraser University. Recent research interests include empirical asset pricing models, conditional variance forecasting, and the combination of forecasts, both quantitative and qualitative. A common thread through much of his work (though not this article) is the application of artificial neural networks.

Peter KENNEDY is Professor of Economics at Simon Fraser University. He is an associate editor of the International Journal of Forecasting and the Journal of Economic Education. His publications are mainly in the areas of econometrics and economics education; he is best known for his book A Guide to Econometrics.