Improving Forecast Accuracy by Combination

Technical Report 1395
School of Operations Research and Industrial Engineering
Cornell University, Ithaca, New York 14853

FENG ZHANG ROBIN ROUNDY*

January 15, 2004

Abstract

In this paper we propose an analytical framework that intelligently combines multiple forecasts into a single and more accurate combined forecast. Four different combination methods are derived by using different weight estimation and forecast selection techniques. We examine our methods and four other commonly used combination methods in computational experiments. The results suggest that combination does improve accuracy and our methods have better and more stable performance than the others.

1 Related Research on Forecast Combination

Considerable literature has accumulated over the years regarding the combination of forecasts. The primary conclusion of this line of research is that forecast accuracy can be substantially improved through the combination of multiple individual forecasts. Furthermore, simple combination methods often work reasonably well relative to more complex methods (Clemen, 1989). In this paper we propose several new forecast combination methods and test

*Corresponding author

them and other commonly used methods in computational experiments.

Reid (1968) and Bates and Granger (1969) are considered to be the seminal works in the area of combining forecasts. Since then a large number of combination procedures have appeared in the literature. Many models have been developed to find 'optimal' combinations of forecasts. Both simulation and empirical studies have been conducted to test performance of those models. The results consistently indicate that combining multiple forecasts increases forecast accuracy. Throughout the years, applications of combined forecasts have been found in many fields such as meteorology, economics, insurance and forecasting sales and price. Details can be found in Clemen (1989), an extensive review paper discussing theoretical and empirical work on various methods of combining forecasts. More recently, de Menezes *et al.* (2000) review evidence on the performance of different combining methods with the aim of providing practical guidelines based on three properties of the forecast errors: variance, asymmetry and serial correlation. The evidence indicates that using different criteria leads to distinct preferences, and that the properties of the individual forecast errors can strongly influence the characteristics of the combination's errors.

Well-studied combining methods range from the robust simple average to the far more theoretically complex, such as state-space methods that attempt to model non-stationarity in the combining weights (de Menezes *et al.*, 2000). To get a sense of the methodology five well-studied methods are worth mentioning. All of the methods adopt the linear formulation. The vector $\mathbf{f}$ consists of k individual forecasts, which are combined via a linear weighting vector $\mathbf{w}$. The combined forecast $\mathbf{f}_c$ is equal to $\mathbf{w}'\mathbf{f}$.

- Simple Average: Assign equal weight to each individual forecast. This approach has the virtues of simplicity and robustness. It has consistently been the choice of many researchers. Many studies provide strong support for this method (Clemen, 1989). Gunter (1990) identified analytically the conditions under which the Simple Average outperforms Optimal and OLS, which will be discussed later. A possible answer to the success of simple average may rely on the instability of the combining weights, which results from unsystematic changes over time in the covariance matrix of individual forecast errors.

- Outperformance Probabilities: This method is initially proposed by

Bunn (1975). By this method, each individual weight is an estimate of the probability that its respective individual forecast performs best on the next occasion. Each probability is estimated as the fraction of occurrences in which the respective individual forecast has performed the best in the past. This is a robust, nonparametric method, which performs well when there is relatively little historical data.

- Optimal (Minimum Variance): The combining weights are calculated to minimize the variance of the error of the combined forecast, based on the assumption that each individual forecast is unbiased. Mathematically, the weight vector $\mathbf{w}$ is determined as follows:

$$\mathbf{w} = \frac{\mathbf{S}^{-1}\mathbf{e}}{\mathbf{e}'\mathbf{S}^{-1}\mathbf{e}}$$

  where $\mathbf{e}$ is a unit vector and $\mathbf{S}$ is the covariance matrix of individual forecast errors. As $\mathbf{S}$ is generally unknown in practice, this method requires $\mathbf{S}$ to be properly estimated. Bates and Granger (1969) suggested five procedures to estimate $\mathbf{S}$. In a finite sample of typical size, sampling error and collinearity among individual forecasts contaminate the estimate of combining weights. Thus, while one hopes to reduce out-of-sample forecast mean squared error (MSE) by combination, there is no guarantee that this will happen in practice.

- Ordinary Least Squares (OLS): In this method the individual forecasts are used as regressors in an ordinary least squares (OLS) regression with a constant term. Granger and Ramanathan (1984) showed that if the individual forecasts are biased, this method is better than the Optimal method. Granger and Ramanathan's suggestion has been discussed and contested theoretically (Clemen, 1986; Bordley, 1986) and empirically (Holden and Peel, 1989; Aksu and Gunter, 1992). Recently, MacDonald and Marsh (1994) reported that the presence of substantial biases in individual forecasts led them to use OLS regression to combine exchange rate forecasts.

- Constrained OLS: Here the least squares regression is performed with the inclusion of a group of constraints on the vector of combining weights. Among others see Pindyck and Rubinfeld (1981), Clemen (1989), Gunter (1992), Aksu and Gunter (1992), Dorfman and McIntosh (2001) and, in the context of forecast combination, Chan *et al.*

(1999). All of these authors conclude that introducing inequality constraints into the least squares model improves accuracy.

The empirical work of Gunter (1992) and Aksu and Gunter (1992) compared the accuracy of a wide range of constrained least squares combining procedures, plus Simple Average. They found that constraining the weights to be non-negative was as robust and accurate as Simple Average, and that both of these methods almost always outperform least squares without constraint and least squares with a constraint that weights sum to one but are allowed to be negative. Dorfman and McIntosh (2001) provide evidence that the Bayesian inequality-constrained posterior mean is more efficient than the restricted-MLE for medium sample sizes.

There exist many other combination models in the literature. Interested readers are referred to Clemen (1989), an excellent review on this topic, and the papers referenced there.

Consider the problem of variable selection in forecast combination. Compared with classical statistical approaches, in practical business settings there is often a somewhat different set of potential benefits and risks to consider. The classical statistical criterion for including an individual forecast in a combination approach is the probability that using the forecast will reduce the error. From the business point of view, when forecasting demand for a large number of products on a weekly basis, the expected size of the decrease in forecast error is at least as important as important as the probability that the decrease is real. The economic benefit of including a forecast in a combination technique (measured by the forecast error) is balanced against the risks involved - the risk that random effects in the data will result in false signals, and the risk inherent in using historical data to forecast in a dynamic business climate, where the past may misrepresent the future.

We propose several approaches for dealing with these issues. To address the magnitude of the benefit of including a forecast, as well as the probability that the benefit is real, we propose a new Economic Significance Test for selecting which forecasts should be used. To ameliorate the impact of random effects we constrain the combination coefficients, using constraints that

differ from what has been done in the past. Finally, we perform simulation tests of forecast combination methods. To test the robustness of the methods proposed we use both dynamic and stationary environments. Whereas this paper does simulation experiments with simulated data, a companion paper tests these methods on an extensive data set from a large semiconductor manufacturing company (Zhang et al 2004).

2 Our Combination Methods

The major problems to be solved in a forecast combination process are how to choose which individual forecasts will be combined (i.e., choose a sub-model), and how to estimate combining weights. This section will answer these two questions. Note that these two questions are not isolated. In order to choose the best sub-model [1], an error measurement for each sub-model is required, which in turn is based on the estimation of combining weights for the given sub-model. The combination process can be summarized as follows:

- Step 1: Estimation of combining weights and computation of error measurement for all sub-models [2].
- Step 2: Sub-model selection (or variable selection).
- Step 3: Generation of combined forecasts based on the previous two steps.

In section 2.1 we discuss the constraints that we place on the combining weights. Section 2.2 addresses estimation of the weights. It is followed section 2.3, in which we discuss two methods for sub-model selection: The Economic Significance Test, and Bayesian Model Averaging.

[1] A sub-model defines a subset of k candidate individual forecasts to use. Thus, in total, 2^k sub-models are under consideration. A forecast selection problem is essentially to determine which sub-model is the best, or is true in some sense.

[2] This exhaustive enumeration approach does not work well when the number of individual forecasts is big, because it is time-consuming. However this is not a problem for the applications that we have in mind.

2.1 Constraints on Combining Weights

We assume that $y_t = \mathbf{x}_t\beta + \epsilon_t$ or, in matrix format,

$$\mathbf{y} = \mathbf{X}\beta + \epsilon \tag{1}$$

where $\mathbf{y}$ is an $n \times 1$ vector of actual demand in periods $1 \ldots n$, $\mathbf{X}$ is an $n \times k$ forecast matrix, β is a $k \times 1$ vector of unknown weights and ϵ is an $n \times 1$ vector of errors with distribution $N(\mathbf{0}, \sigma^2\mathbf{I})$.

As we mentioned in section 1, there is a substantial literature on using linear inequality constraints in linear regression model. Specific constraints that have been studied include:

- No intercept
- Non-negative weights
- Sum of weights is 1
- Lower and upper bounds on weights

As we mentioned earlier, Gunter (1992) and Aksu and Gunter (1992) examined various combinations of linear constraints. They concluded that constraining the weights to be non-negative was as robust and accurate as anything, and comparable to Simple Average. In a forecasting context, Chan *et al.* (1999) conclude that it is important to use the constraint that the weights are nonnegative and sum to unity.

Considering past research, and as a result of a series of preliminary experiments, we impose the linear inequality constraints on the combining weights β described below. Intuitive reasoning to support each constraint is provided.

- Nonnegativity constraint: $\beta_i \geq 0 \; \forall i$. It is believed that each individual forecast should be positively correlated with the underlying demand model. Although positive correlation between the demand and each individual forecast does not guarantee that the optimal weights are non-negative, we require nonnegative combining weights.

- Unbiasedness constraint: $\sum_i \beta_i = 1$. This constraint is based on the assumption that each individual forecast is unbiased. However, this is a rather restrictive, sometimes unrealistic, assumption. To make the proposed combination model robust, the Unbiasedness constraint is often relaxed to a band constraint.

- Band constraint: $1-\gamma \leq \sum_i \beta_i \leq 1+\gamma$. This constraint is based on the assumption that there is some bias in the individual forecasts, but the bias is limited. The user will use either the Unbiasedness constraint or the Band constraint, not both. The parameter γ, $0 < \gamma \leq 1$ is user-selected.

- Diversification constraint: $\beta_i \leq 1 \ \forall i$. This constraint is used in conjunction with the Nonnegativity constraint and the Band constraint. Without this constraint, it is possible for one β_i to become very large while the others are very small. This is risky because the individual forecast with the large combining weight is likely to have experienced a string of random under-predictions in the recent past. It is likely to bounce back, or even to over-predict, in subsequent time periods.

Hereafter these linear constraints on β are represented by an indicator function $q(\beta)$, which takes value 1 when all constraints are satisfied, and value 0 otherwise.

2.2 Estimation of Combining Weights

The use of Bayes's theorem in statistical inferences has been thoroughly studied. One advantage of the Bayesian approach is that prior knowledge about parameters of interest can be combined in a well-defined mathematical way with information obtained from observed data. Inference in the normal linear regression model subject to inequality constraints is one of the most common tasks in applied econometrics. Geweke (1986) solved this problem using a Bayesian approach. However, Geweke only considered an uninformative prior distribution where there is very limited prior information available. Our approach is a generalization to Geweke's.

First we present the prior and posterior distributions. Then we discuss hyperparameter selection for the prior distribution, and finally, our estimation procedures.

2.2.1 Prior and Posterior Distributions

Both for reasons of computational simplicity and for the interpretability of results, the most obvious choice for the prior distribution is a natural conjugate distribution. In this paper, a Normal-Gamma conjugate prior distribution is

adopted, which is the most commonly used one in Bayesian research.

Let $p(\beta, \sigma)$ denote the prior distribution of β and σ [3]. By Bayes Theorem

$$p(\beta, \sigma) = p(\beta|\sigma)p(\sigma)$$

where

$$p(\beta|\sigma) = f_N^{(k)}(\beta|\mu, \sigma^2\mathbf{V})$$

Here, $f_N^{(k)}(\beta|\mu, \sigma^2\mathbf{V})$ denotes the p.d.f. of a k-variate normal distribution with mean μ and covariance matrix $\sigma^2\mathbf{V}$. Also,

$$p(\sigma^{-2}) = f_G(\sigma^{-2}|c, d)$$

which corresponds to a Gamma distribution with mean c/d and variance c/d^2. Thus σ has an inverse Gamma distribution, i.e.,

$$p(\sigma) = f_{IG}(\sigma|v', s') \propto e^{-\frac{v's'^2}{2\sigma^2}}(\frac{v's'^2}{2\sigma^2})^{\frac{v'}{2}+\frac{1}{2}}$$

where $v' = 2c$ and $s'^2 = \frac{d}{2c}$. Therefore the joint prior distribution is

$$\begin{aligned} p(\beta, \sigma) &= f_N^{(k)}(\beta|\mu, \sigma^2\mathbf{V}) f_{IG}(\sigma|v', s') \\ &\propto (\sigma^{-2})^{\frac{1}{2}(k+v'+1)} e^{-\frac{1}{2\sigma^2}[(\beta-\mu)'\mathbf{V}^{-1}(\beta-\mu)+v's'^2]} \end{aligned} \tag{2}$$

It is assumed that the error term ϵ in model (1) has a multi-variate normal distribution with mean $\mathbf{0}$ and covariance matrix $\sigma^2\mathbf{I}$. Consequently, the likelihood function is

$$\begin{aligned} l(\mathbf{y}|\beta, \sigma) &= f_N^{(n)}(\mathbf{y}|\mathbf{X}\beta, \sigma^2\mathbf{I}) \\ &\propto \sigma^{-n} e^{-\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\beta)'(\mathbf{y}-\mathbf{X}\beta)} \\ &= (\sigma^{-2})^{\frac{1}{2}(k+v)} e^{-\frac{1}{2\sigma^2}[(\beta-\hat{\beta})'\mathbf{X}'\mathbf{X}(\beta-\hat{\beta})+vs^2]} \end{aligned}$$

where $v = n-k$, $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$, and s is given by $vs^2 = (\mathbf{y}-\mathbf{X}\hat{\beta})'(\mathbf{y}-\mathbf{X}\hat{\beta})$.

By Bayes Theorem again, the joint posterior distribution is given by

$$\begin{aligned} p(\beta, \sigma|\mathbf{y}) &\propto p(\beta, \sigma) l(\mathbf{y}|\beta, \sigma) \\ &\propto (\sigma^{-2})^{\frac{1}{2}(k+v'+1+k+v)} e^{-\frac{1}{2\sigma^2}[vs^2+v's'^2+(\beta-\hat{\beta})'\mathbf{X}'\mathbf{X}(\beta-\hat{\beta})+(\beta-\mu)'\mathbf{V}^{-1}(\beta-\mu)]} \\ &= (\sigma^{-2})^{\frac{1}{2}(k+v''+1)} e^{-\frac{1}{2\sigma^2}[v''s''^2+(\beta-\tilde{\mu})'\mathbf{W}(\beta-\tilde{\mu})]} \end{aligned} \tag{3}$$

[3]For the time being, β is unconstrained. Appropriate constraints will be applied later

where

$$\begin{aligned}
\mathbf{W} &= \mathbf{V}^{-1} + \mathbf{X}'\mathbf{X} \\
\tilde{\mu} &= \mathbf{W}^{-1}(\mathbf{X}'\mathbf{X}\hat{\beta} + \mathbf{V}^{-1}\mu) \\
v'' &= v' + v + k \\
s''^2 &= \frac{(v's'^2 + \mu'\mathbf{V}^{-1}\mu) + (vs^2 + \hat{\beta}'\mathbf{X}'\mathbf{X}\hat{\beta}) - \tilde{\mu}'\mathbf{W}\tilde{\mu}}{v''}
\end{aligned}$$

It is obvious that posterior distribution is of the same form as the prior distribution. Taking the integral of $p(\beta, \sigma|\mathbf{y})$ over σ, one can get the marginal posterior distribution of β (see Appendix A for the derivation).

$$p(\beta|\mathbf{y}) \propto \frac{1}{2}(\frac{2}{v''s''^2 + (\beta - \tilde{\mu})'\mathbf{W}(\beta - \tilde{\mu})})^{\frac{v''+k}{2}}\Gamma((v'' + k)/2)$$

This is a multi-variate t distribution.

2.2.2 Prior Hyperparameters

Generally, the choice of the prior hyperparameters μ, $\mathbf{V}$, v' and s' in (2) is not trivial in the absence of prior information. In the proposed combination framework, it is assumed that every individual forecast is assigned equal weight in the prior, i.e.,

$$\mu = (\frac{1}{k}, \frac{1}{k}, \ldots, \frac{1}{k})'$$

where $k = \dim(\mu)$. The reason for choosing equal weights is that many empirical studies provide strong support for the Simple Average method (Clemen, 1989). If there exists relatively strong prior evidence against equal weights, other choices of μ should be considered.

Eliciting a prior covariance $\mathbf{V}$ is even more difficult. We adopt the convenient g-prior (Zellner, 1986), which corresponds to taking

$$\mathbf{V}^{-1} = g\mathbf{X}'\mathbf{X}$$

with $g > 0$. The power of this g-prior 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. Consequently, $\tilde{\mu}$ in (3) becomes

$$\frac{g}{1+g}\mu + \frac{1}{1+g}\hat{\beta}$$

This is a convex combination of the prior mean μ and OLS estimator $\hat{\beta}$. A large value of g implies high prior precision and hence substantial shrinkage toward μ. A smaller g leads to less shrinkage.

Fernandez *et al.* (2001) make g a function on the sample size n and the number of regressors k and examine all candidates in an extensive simulation. We adopt one of the functions studied in Fernandez *et al.* (2001), namely $g = \frac{1}{k^2}$ in the proposed combination framework.

As for σ, a widely accepted non-informative improper prior distribution is actually used, which is the limiting distribution of the inverse Gamma prior in (2) when $v' \to 0$. The density is given by

$$p(\sigma) \propto \sigma^{-1}$$

This distribution is invariant under scale transformations. Thus the joint prior (2) simplifies to

$$p(\beta, \sigma) \propto (\sigma^{-2})^{\frac{1}{2}(k+1)} e^{-\frac{1}{2\sigma^2}[(\beta-\mu)'\mathbf{V}^{-1}(\beta-\mu)]} \tag{4}$$

If there is a group of constraints, say $q(\beta)$ on β, then the prior distribution becomes proportional to $p(\beta, \sigma)q(\beta)$, where $p(\beta, \sigma)$ is defined in (2). The derivations remain the same. Therefore, the joint posterior distribution becomes proportional to $p(\beta, \sigma|\mathbf{y})q(\beta)$ and the marginal posterior distribution of β is proportional to $p(\beta|\mathbf{y})q(\beta)$. Further details on algebraic simplifications, and on how they relate to our code, are found in Appendix B.

2.2.3 Estimating The Weights

Based on the posterior distribution, there are many possible ways to estimate the combining weights β. Two techniques are discussed here and tested in numerical experiments later. The first one is generalized maximum likelihood estimator (GMLE) or maximum a posteriori. Let $\mathbf{b}$ denote the estimate of β. The GMLE method sets

$$\mathbf{b} = argmax\ p(\beta|\mathbf{y})q(\beta) \tag{5}$$

Obviously $\mathbf{b}$ has the interpretation of being the "most likely" value of β, given the prior and the observations. Note that $p(\beta|\mathbf{y})q(\beta)$, the posterior

density function of β, is a function of $[(\beta - \hat{\mu})'\mathbf{W}(\beta - \hat{\mu})]$. Therefore this method reduces to solving a convex quadratic minimization problem on the feasible region defined by $q(\beta)$. Fortunately there exist many well-studied algorithms which can solve quadratic programming problems efficiently.

The other way to estimate β is to take the expectation of the posterior distribution, which is proportional to $p(\beta|y)q(\beta)$, i.e.,

$$\mathbf{b} = E(\beta|\mathbf{y}) = \frac{\int \beta p(\beta|\mathbf{y})q(\beta)d\beta}{\int p(\beta|\mathbf{y})q(\beta)d\beta} \tag{6}$$

This method is referred as EPost (*E*xpectation of *Post*erior distribution).

We now argue that EPost is equivalent to minimizing the mean squared forecasting error in the next period. To see this, suppose that b is the estimate of β based on the first n data points. Then the combined forecast for period $n+1$ is $\hat{y}_{n+1} = \mathbf{x}_{n+1}\mathbf{b}$. Recall that the unknown actual demand in period $n+1$ is $y_{n+1} = \mathbf{x}_{n+1}\beta + \epsilon_{n+1}$. The forecasting error in period $n+1$ is $e_{n+1} = y_{n+1} - \hat{y}_{n+1}$, and the mean squared forecasting error in is

$$\begin{aligned} E[e_{n+1}^2] &= E[(\mathbf{x}_{n+1}(\beta - \mathbf{b}) + \epsilon_{n+1})^2] \\ &= E[(\mathbf{x}_{n+1}(\beta - E(\beta|\mathbf{y}) + E(\beta|\mathbf{y}) - \mathbf{b}) + \epsilon_{n+1})^2] \\ &= \sigma^2 + [\mathbf{x}_{n+1}(E(\beta|\mathbf{y}) - \mathbf{b})]^2 + \mathbf{x}_{n+1}Cov(\beta)\mathbf{x}_{n+1}^{'} \end{aligned}$$

Only the second term can be influenced by the choice of $\mathbf{b}$. Obviously when $\mathbf{b} = E(\beta|\mathbf{y})$, $E[e_{n+1}^2]$ is minimized.

Because $q(\beta)$ captures linear inequalities, analytical integration of $p(\beta|\mathbf{y})q(\beta)$ in (6) is usually impossible. Numerical integration is challenging when the dimension of β is greater than 3 or 4. Therefore importance sampling, a Monte Carlo integration technique, is adopted to compute the posterior mean of β. This method can be described as follows. Suppose that β_i $i = 1 \ldots N$ is a random sample from a distribution with p.d.f. $I(\beta)$, called the importance function. The support of $I(\beta)$ must include the support of the posterior density of β. Then almost surely

$$E(\beta|\mathbf{y}) = \lim_{N\to\infty} \frac{\sum_{i=1}^{N} \beta_i p(\beta_i|\mathbf{y})q(\beta_i)/I(\beta_i)}{\sum_{i=1}^{N} p(\beta_i|\mathbf{y})q(\beta_i)/I(\beta_i)}$$

The rate of almost sure convergence depends critically on the choice of the importance sampling density $I(\beta)$. Loosely speaking, the tails of $I(\beta)$ should not decay more quickly than the tails of $p(\beta|\mathbf{y})q(\beta)$. We generate the random samples β_i uniformly over the feasible region defined by $q(\beta)$.

2.3 Forecast Selection

As mentioned before, the other problem that the proposed combination framework must address is selecting which individual forecasts to combine. Obviously no one wants to omit any potentially beneficial individual forecast. However a simple model is preferred since it is easier to understand and maintain. Also, unnecessary individual forecasts introduce additional uncertainty into the combined forecast. When many forecasts are used, collinearity among forecasts may affect the estimate of combining weights. Apparently very limited attention has been paid to this problem in the forecast combination literature, although it has been addressed in the broader statistics literature.

In classical statistics, the variable selection problem in regression has been thoroughly studied. The usual approach is to include an individual forecast if the probability that using it makes the combined forecast more accurate, exceeds a given threshold. We note that many classical statistical significance tests do not fit naturally into the proposed combination framework. The constraints on the combining weights β effect the posterior distribution, invalidating critical assumptions. This is one of the reasons that we propose a simple, new forecast selection procedure.

2.3.1 The Test of Economic Significance

In practical business settings there is often a different measure of benefit and a different element of risk to consider. Business environments are continually changing. So a method for combining individual forecasts may see a sharp degradation in performance when the relationship between the forecasts being used and the true demand shifts. When the business environment does not change, the classical statistical criterion for including an individual forecast is the probability that its inclusion will reduce the error. From the

business point of view, the size of the decrease in error is at least as important as the probability that the decrease is real. The trade-off between benefit and risk associated with each individual forecast should reflect both the expected decrease in forecast error that the forecast is capable of delivering, and the risk associated with using the forecast in both stationary and dynamic business climates.

One of the research goals is to find a systematic way to determine which individual forecasts should be combined and which should not, reflecting the considerations described above. To this end we propose a new significance test, named the "economic significance test", in contrast to the regular statistical significance test.

Each one of those k individual forecasts may or may not be used in the final combination. There are a total of $2^k - 1$ different subsets of the individual forecasts to consider. For the applications we have in mind k is unlikely to be greater than 5 or 6, so total enumeration of the 2^k subsets is manageable. The proposed economic significance test considers all subsets one by one. For each one, the estimate of β is obtained using one of estimation methods discussed before. Note that some elements of the estimate of β are constrained to be zero because the corresponding individual forecasts are not used in the combination. For each subset, after β has been estimated, the in-sample squared forecast error is calculated.

Let S be one of the $2^k - 1$ non-empty subsets of individual forecasts, and let $i \in S$. Individual forecast i is *economically significant in* S if

$$\frac{ERR_{S_{-i}} - ERR_S}{ERR_S} > \alpha \tag{7}$$

where

- ERR_S is the in-sample squared error associated with subset S,
- S_{-i} is the subset where the ith forecast is dropped from S, and
- α is a pre-determined economical significance threshold.

Intuitively, given a subset S, an individual forecast is economically significant if its presence in the combination can reduce the in-sample forecast error by

more than a certain percentage. Otherwise we conclude that it is not wise to include this forecast because the benefit of using it is too low to offset the risk associated with using it.

A subset S of individual forecasts is *admissible* if all forecasts in S are economically significant in S. Singleton sets are automatically admissible. The economic significance test algorithm enumerates the non-empty subsets S of the individual forecasts, identifies the admissible ones, and selects the admissible subset with the smallest in-sample squared forecast error.

The economic significance test differs from one classical model selection method, the Schwarz Information Criteria (SIC), in following senses:

- The threshould α is flexible and has actual business meanings. In contrast, it is implicitly determined by sample size in SIC.
- Economic significance does not apply a dimension penalty when comparing admissible subsets.
- Other measurements, besides the in-sample squared error, can be used in the economic significance test.[4]

2.3.2 Bayesian Model Averaging

There is another way to tackle the model selection problem. Both the classic statistical approach and the the economic significant test described before only choose one single sub-model from $2^k - 1$ candidates. To what extent is one confident in that sub-model? Model risk is present, and basing inferences on a single sub-model might be risky. The Bayesian approach to this problem is conceptually straightforward: model selection is treated as a further parameter which lies in the model space. In addition to selecting priors for the weights β and the error variance σ, we select a prior for the sub-models themselves. Bayesian inference can be conducted in the usual way, with one level (the prior on the sub-model space) added to the hierarchy.

Bayesian model averaging (BMA) takes a weighted average of the weights for each sub-model, using the posterior probabilities as weights. Min and

[4]We tried using the out-of-sample squared error rather than the in-sample error, but the in-sample error worked better.

Zeller [18] show that such mixing over models minimizes the expected predictive squared error loss, provided that the set of models under consideration is exhaustive. Also see Leamer [16].

BMA solves the variable selection problem as follows. Suppose that $M_1, \ldots, M_K$ are the models under consideration. For each model M_j, $\theta_j = (\beta^j, \sigma^j)$ consists of weights β^j and the error variance σ^j. One selects a prior probability $\pi(M_j)$ that model M_j is selected, and a prior distribution $p(\theta_j|M_j)$ on the parameter vector θ_j for model M_j. Then the posterior probability of model M_j, given data $\mathbf{y}$, is

$$p(M_j|\mathbf{y}) = \frac{f(\mathbf{y}|M_j)\pi(M_j)}{\sum_j f(\mathbf{y}|M_j)\pi(M_j)} \tag{8}$$

where

$$f(\mathbf{y}|M_j) = \int f(\mathbf{y}|\theta_j, M_j)p(\theta_j|M_j)d\theta_j \tag{9}$$

is the integrated likelihood function of model M_j. If Δ is the quantity of interest, and if $\Delta_j = E(\Delta|M_j, \mathbf{y})$ is the estimate of Δ obtained from model M_j, then the BMA estimate of Δ is

$$\Delta = \sum_j \Delta_j p(M_j|\mathbf{y}) \tag{10}$$

We assume that there is no useful prior information, and assign equal prior probabilities to each model ($\pi(M_j) = \frac{1}{2^k-1}$). If prior information is available, it can be captured by adjusting $\pi(M_j)$. Recall that $\theta_j = (\beta^j, \sigma^j)$. Setting σ to σ^j and β^j to the appropriate subset of $[\beta_1, \beta_2, \ldots, \beta_k]$ for model j, (4) becomes

$$p(\theta_j|M_j) = C_j\{\sigma^{-(k_j+1)}e^{-\frac{1}{2\sigma^2}(\beta^j-\mu)'\mathbf{V}^{-1}(\beta^j-\mu)}q(\beta^j)$$

where $k_j = dim(\beta^j)$ and

$$C_j = \frac{1}{\int \sigma^{-(k_j+1)}e^{-\frac{1}{2\sigma^2}(\beta^j-\mu)'\mathbf{V}^{-1}(\beta^j-\mu)}q(\beta^j)d\sigma d\beta^j} \tag{11}$$

The likelihood function, given θ_j and M_j, is

$$f(\mathbf{y}|\theta_j, M_j) = f_N^{(n)}(\mathbf{y}|\mathbf{X}^j\beta^j, \sigma^2\mathbf{I})$$

Therefore,

$$f(\mathbf{y}|M_j) = C_j(2\pi)^{-\frac{n}{2}} \int \sigma^{-(n+k_j+1)} e^{-\frac{1}{2\sigma^2}[(\beta^j-\mu)'\mathbf{V}^{-1}(\beta^j-\mu)+(\mathbf{y}-\mathbf{X}\beta)'(\mathbf{y}-\mathbf{X}\beta))]} q(\beta^j) d\beta^j d\sigma$$

It is impossible to get $f(\mathbf{y}|M_j)$ analytically because of the presence of $q(\beta)$. We turn to the importance sampling method again to compute the integral numerically. Similarly, C_j can be calculated numerically as well by applying the importance sampling technique to the integral in (11). In this manner one is able to obtain the posterior model probability $p(M_j|\mathbf{y})$ as given in (8), and the BMA combined forecast as in (10).

3 Computational Experiments

The primary goal of the following numerical experiments is to test the performance of our combination methods, along with other methods from the literature. Four methods are derived from the previously proposed combination framework. Epost(BMA) estimates combining weights by taking the expectation of the posterior, and selects forecasts using the BMA approach. GMLE(ES) estimates weights using maximum likelihood and forecasts using economic significance. EPost(ES) and GMLE(BMA) are similarly defined. All of them exhibit very similar performance. We show results from Epost(BMA) because it is usually the best of the four, and from GMLE(ES) because it is the fastest[5]. Four other popular combination methods are tested: Simple Average (SA), Ordinary Least Squares (OLS), Outperformance (Outper) and Optimal (Minimum Variance).

3.1 Data Generation Process

In the experiments we use randomly generated data. Computer-generated data is easier to understand and to manipulate, and enables sensitivity analysis. Recall that in time period n, we assume $y_n = \mathbf{x}_n\beta + \epsilon_n$. Forecast vectors $\mathbf{x}_n$ are drawn from a multi-variate normal distribution $N(\mu, \Sigma)$, and the error ϵ_n is drawn from $N(0, \sigma^2)$. The parameters to be specified are the

[5]Average running time to generate a combined forecast is about 2.5 seconds for the EPost-based methods, and 1.5 seconds for the MLE-based methods, on a good PC.

true weights β, the mean of individual forecasts μ, the covariance matrix of individual forecasts Σ, and the variance of the error term σ^2.

There are two rounds of computational experiments. In the first round we test performance in a stationary environment. The second round is designed to examine the effect of an undetected shift in forecast characteristics. Shifts might occur for a variety of reasons. Consider the following scenarios:

1. An unanticipated competitor aggressively enters the market.
2. A sales forecaster modifies his forecasting technique or adopts a new one.

In both scenarios historical data on may not reflect current conditions, potentially impacting the statistical properties of the forecasts. In our experiments, there is a pre-specified change point. Different sets of parameters are used before and after the change point.

3.2 Experiments on Stationary Data

In the experiments on stationary data a Base case is studied first. The performance of each combination method is measured by its Mean Squared Forecast Error (MSFE). Suppose that at the beginning of period $n+1$, the estimate of β is b. The out-of-sample forecast error is $x_{n+1}(\beta - b) + \epsilon_{n+1}$. The MSFE is given as

$$E[x_{n+1}(\beta - b) + \epsilon_{n+1}]^2 = (\beta - b)'\Sigma(\beta - b) + \sigma^2 + ((\beta - b)'\mu)^2$$

Since the historical data is computer-generated, β, Σ and σ^2 are known. The smallest possible out-of-sample MSFE is σ^2 (attained by estimating β perfectly). To obtain a unitless measure with intuitive meaning, we define the MSFE by the MSFE of a perfect forecast (with $b = \beta$). Thus, the Mean Squared Forecast Error Ratio(MSFER) is

$$\frac{(\beta - b)'\Sigma(\beta - b) + \sigma^2 + ((\beta - b)'\mu)^2}{\sigma^2}$$

There are three individual forecasts in the Base Case. They are unbiased, have similar variance, and are slightly correlated. The parameters are:

- true weights $\beta = [\frac{7}{14},\ \frac{5}{14},\ \frac{2}{14}]'$
- mean forecast $\mu = [700,\ 700,\ 700]$
- variance of individual forecasts $300^2 * [\phi_1,\ \phi_2,\ \phi_3]$ where ϕ_i is randomly generated using $\phi_i \sim U(0.85,\ 1.15)$.
- correlation among individual forecasts $[\rho_{12},\ \rho_{13},\ \rho_{23}]$ where ρ_{ij} is randomly generated according to $\rho_{ij} \sim U(0,\ \frac{1}{3})$ $1 \leq i \neq j \leq 3$.
- residual uncertainty [6] $\frac{\sigma^2}{\sigma^2+\beta'\Sigma\beta} = 0.3$

Demand and individual forecast data for up to 36 historical time periods is generated and fed into the combination methods.

A good forecast combination method is expected to perform well across all scenarios of demand and forecasts. Therefore more cases are generated by changing the parameters of the Base case (β, Σ, σ, etc), to test all combination methods. To limit the dimensionality of the search space and test sensitivity, only one parameter is perturbed at a time. For example, in the HeterCor case, only the correlations among individual forecasts are different from the Base case. Other parameters are unchanged. Table 1 summarizes all 8 cases tested in the experiments and their differences from the Base case.

Table 2 presents the MSFER of all combination methods and of the best individual forecast, averaged over all cases listed in Table 1. For each case 2500 different historical data sets are generated and tested. After every 25 data sets, a new group of parameters ϕ_i, ρ_{ij} is randomly drawn. The number 25 is chosen so that the standard deviation of the sample mean of these 25 instances is within 5% of the mean. T is the number of historical time periods used to estimate the combining weights β. It starts from 6 and goes up to 36.

Table 2 clearly shows that combination does improve forecast accuracy. Each combined forecast is much more accurate than the best individual forecast and the improvement grows as more periods of historical data are used. However the gain is marginal more than 24 time periods are used. Among

[6] The residual uncertainty is defined as the ratio between the variance of ϵ and the variance of demand. It is the portion of the demand variance that can not be eliminated, no matter how well one forecasts.

Table 1: Summary of Cases in Computational Experiments on Stationary Data

Case	Difference from the Base case
Base	NA
HeterCor	$\rho_{12} \sim U(0,\ \frac{1}{3}),\ \rho_{13} \sim U(\frac{1}{3},\ \frac{2}{3}),\ \rho_{23} \sim U(\frac{2}{3},\ 1)$
HighCor	$\rho_{12} \sim U(\frac{2}{3},\ 1),\ \rho_{13} \sim U(\frac{2}{3},\ 1),\ \rho_{23} \sim U(\frac{2}{3},\ 1)$
HeterBeta	$\beta = [\frac{2}{3},\ \frac{1}{4},\ \frac{1}{12}]$
ZeroBeta	$\beta = [\frac{2}{3},\ \frac{1}{3},\ 0]$
Ratio=0.2	$\frac{\sigma^2}{\sigma^2+\beta'\Sigma\beta} = 0.2$
Ratio=0.4	$\frac{\sigma^2}{\sigma^2+\beta'\Sigma\beta} = 0.4$
HeterVar	$\phi_i \sim U(0.4,\ 1.6),\ 1 \leq i \leq 3$

Table 2: Mean (Standard Deviation) of MSFER in All Cases

Methods	T = 6	T = 12	T = 18	T = 24	T = 30	T = 36
EPost(ES)	1.52 (0.55)	1.26 (0.27)	1.17 (0.16)	1.13 (0.12)	1.10 (0.09)	1.09 (0.07)
Epost(BMA)	1.48 (0.50)	1.27 (0.28)	1.18 (0.17)	1.13 (0.12)	1.10 (0.09)	1.09 (0.08)
GMLE(ES)	1.59 (0.63)	1.27 (0.28)	1.17 (0.16)	1.13 (0.12)	1.10 (0.09)	1.09 (0.07)
GMLE(BMA)	1.57 (0.58)	1.29 (0.31)	1.19 (0.18)	1.13 (0.13)	1.11 (0.10)	1.09 (0.08)
SA	1.32 (0.23)	1.32 (0.23)	1.32 (0.23)	1.32 (0.23)	1.32 (0.23)	1.32 (0.23)
OLS	2.37 (3.39)	1.36 (0.40)	1.21 (0.20)	1.14 (0.13)	1.11 (0.10)	1.09 (0.08)
Outper.	1.46 (0.56)	1.29 (0.34)	1.23 (0.26)	1.20 (0.22)	1.18 (0.19)	1.17 (0.17)
Optimal	2.02 (2.77)	1.25 (0.33)	1.14 (0.17)	1.10 (0.11)	1.08 (0.08)	1.06 (0.07)
Best Single	2.12 (0.78)	2.12 (0.78)	2.12 (0.78)	2.12 (0.78)	2.12 (0.78)	2.12 (0.78)

the combination methods, Simple Average (SA) is the most accurate when we only use very recent historical data (T=6). As more historical data is available, Optimal becomes the best. All of our methods perform similarly. They do not dominate, but they are clearly robust. For most T's they are second best. For small amounts of historical data Epost is slightly better than GMLE, and both Optimal and OLS perform poorly.

We now turn to sensitivity analysis with respect to the parameters (β, Σ, σ, etc). Figure 1 illustrates the impact of correlation among individual forecasts on the performance of combination methods. It shows 3 different cases, with an increasing degree of correlation, and with 12 and 24 time periods of historical data. Only the most interesting combination methods are included. Correlation among individual forecasts dramatically improves the

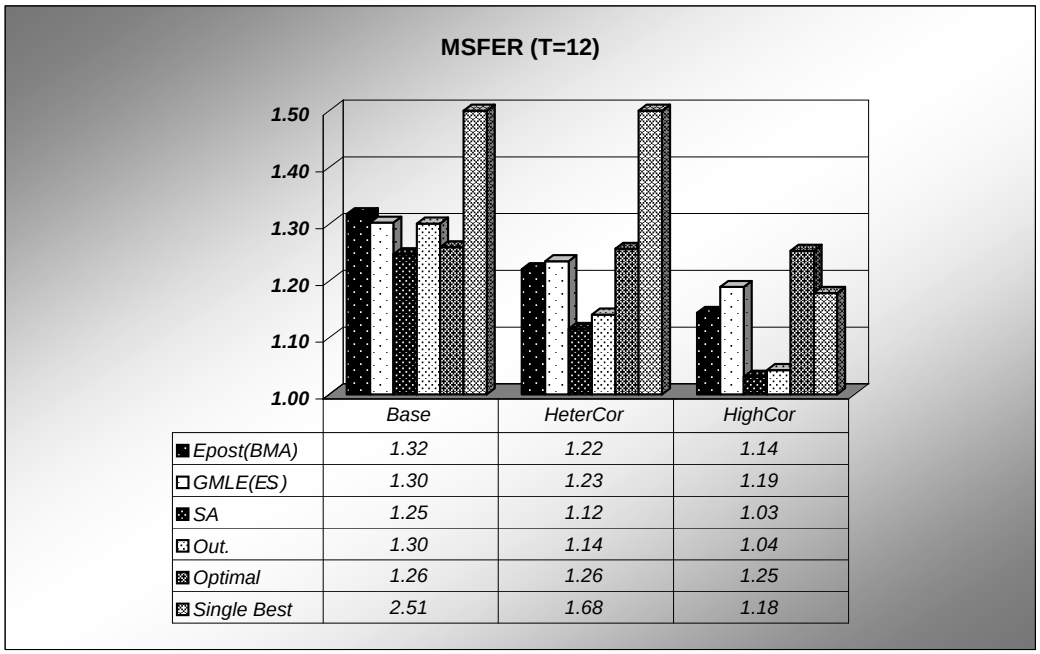
performance of Simple Average and Outperformance, substantially improves the performance of Epost(BMA) and GMLE(ES), and has virtually no impact on Optimal. Simple Average and Outperformance perform very well when there is strong positive correlation among individual forecasts, especially when historical data is limited.

As Figure 2 shows, the structure of the true weights β has a great impact on Simple Average and Outperformance. As the true weights become more unbalanced, the performance of SA and Outperformance suffers dramaticaly. Simple Average and Outperformance are very risky when forecasts of dubious value are used. Epost(BMA), GMLE(ES) and Optimal are robust to the true weights' structure. Figure 3 illustrates how the residual uncertainty affects performance. SA and Outperformance get much better when the residual uncertainty increases, and they get much worse when it goes down. So do our methods, but the changes are small. Again,Optimal is insensitive to the ratio. The variance structure of individual forecasts has no significant impact on the performance of the different methods.

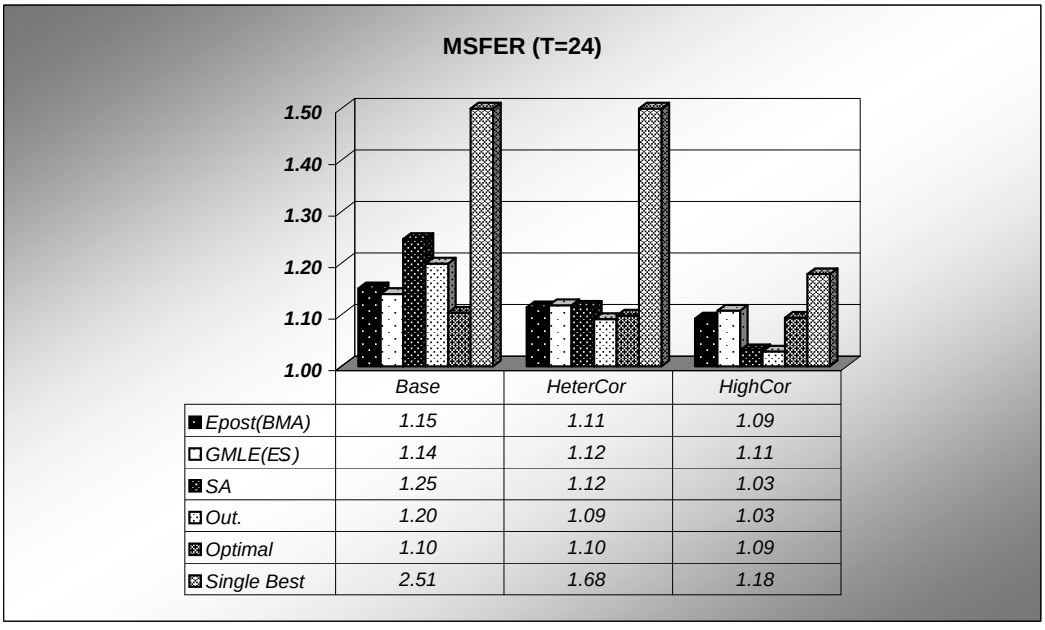
In summary, combination does improve accuracy. The methods proposed by the authors are the most robust ones. Simple Average and Outperformance do very well with small amounts of data and when forecast correlation is high, but they perform very poorly when the true weights are far from uniform and when the residual uncertainty is low. OLS and Optimal beat our methods by a small margin when sufficient data is available, but they require much more historical data to calibrate. This causes problems with highly the skewed demands and short product life spans that characterize the semiconductor industry.

3.3 Experiments on Non-stationary Data

The motivation behind the experiments on non-stationary data is that parameters β, $\mathbf{\Sigma}$ and σ in model (1) can change as the business context changes. A structural change is simulated as follows. Two different cases are randomly selected from those candidates listed in Table 1. A set of parameters is created for the first case, and demand and individual forecast data for periods 1-30 are generated based on these parameters. From the second case we generate parameters and data for periods 31-54. Since the change occurs in time

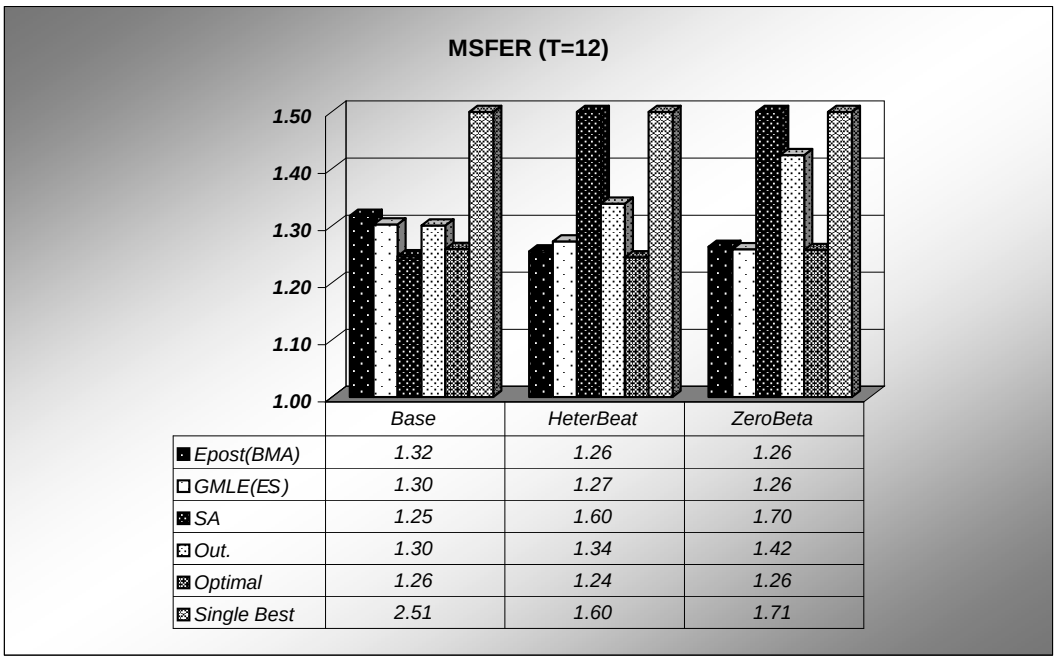


	Base	HeterCor	HighCor
Epost(BMA)	1.32	1.22	1.14
GMLE(ES)	1.30	1.23	1.19
SA	1.25	1.12	1.03
Out.	1.30	1.14	1.04
Optimal	1.26	1.26	1.25
Single Best	2.51	1.68	1.18

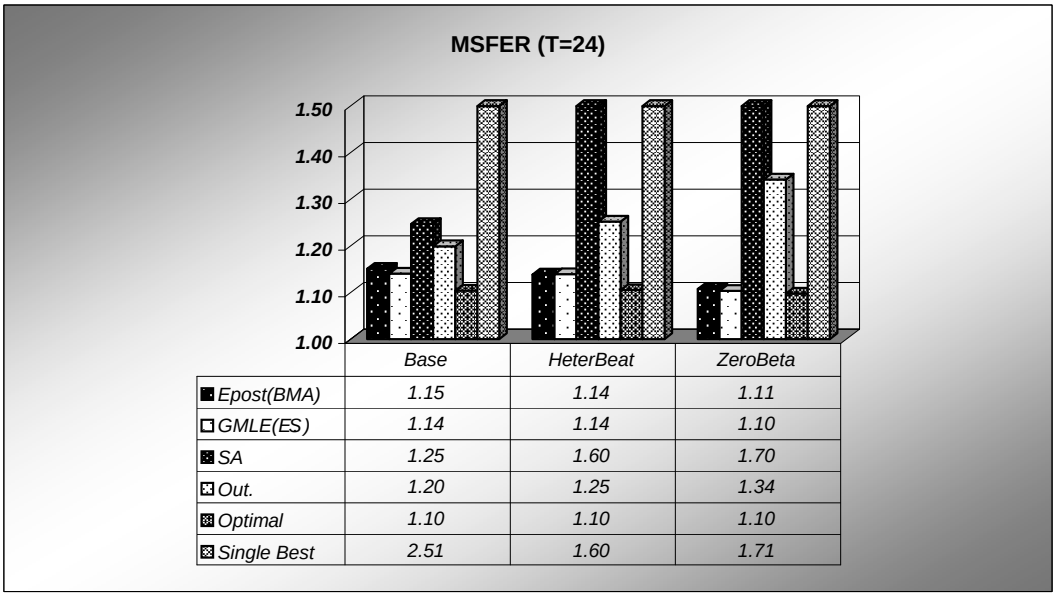


	Base	HeterCor	HighCor
Epost(BMA)	1.15	1.11	1.09
GMLE(ES)	1.14	1.12	1.11
SA	1.25	1.12	1.03
Out.	1.20	1.09	1.03
Optimal	1.10	1.10	1.09
Single Best	2.51	1.68	1.18

Figure 1: Impact of Correlation Among Individual Forecasts

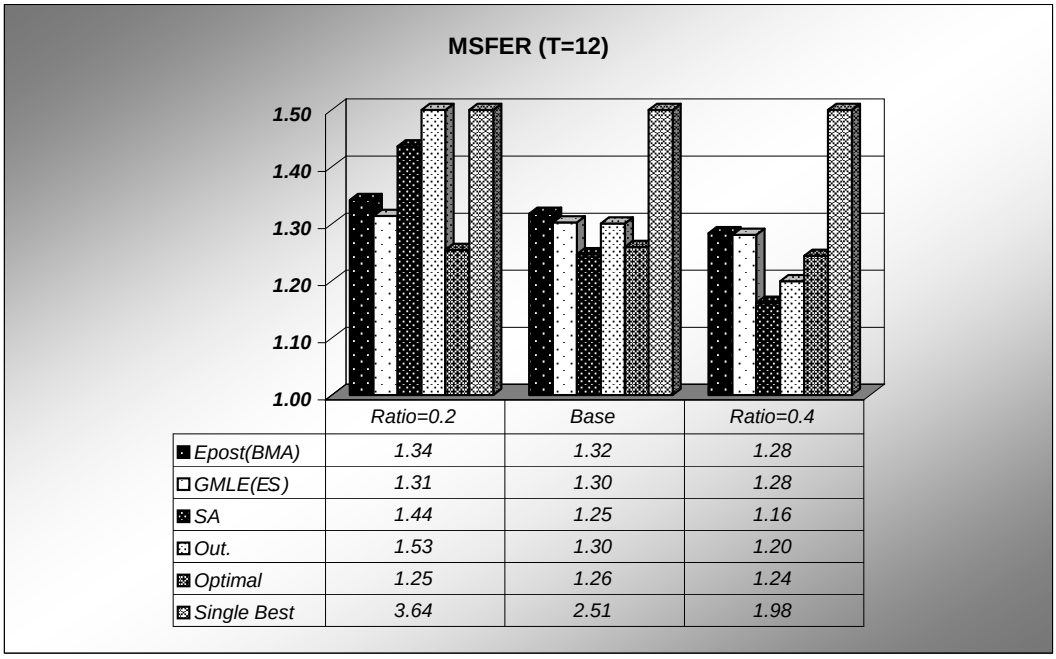


	Base	HeterBeat	ZeroBeta
Epost(BMA)	1.32	1.26	1.26
GMLE(ES)	1.30	1.27	1.26
SA	1.25	1.60	1.70
Out.	1.30	1.34	1.42
Optimal	1.26	1.24	1.26
Single Best	2.51	1.60	1.71

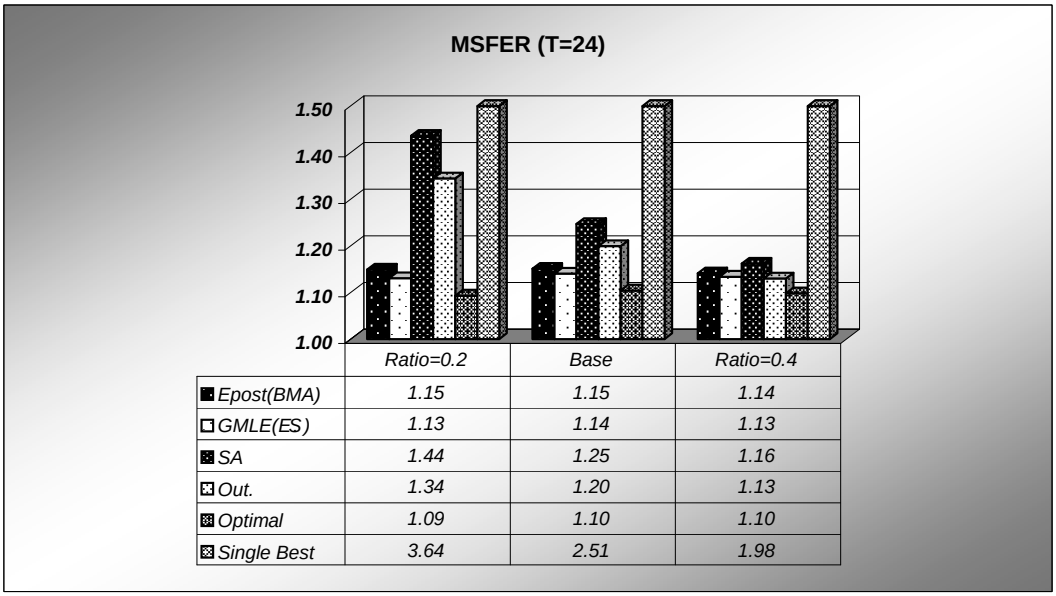


	Base	HeterBeat	ZeroBeta
Epost(BMA)	1.15	1.14	1.11
GMLE(ES)	1.14	1.14	1.10
SA	1.25	1.60	1.70
Out.	1.20	1.25	1.34
Optimal	1.10	1.10	1.10
Single Best	2.51	1.60	1.71

Figure 2: Impact of True Weights



	Ratio=0.2	Base	Ratio=0.4
Epost(BMA)	1.34	1.32	1.28
GMLE(ES)	1.31	1.30	1.28
SA	1.44	1.25	1.16
Out.	1.53	1.30	1.20
Optimal	1.25	1.26	1.24
Single Best	3.64	2.51	1.98



	Ratio=0.2	Base	Ratio=0.4
Epost(BMA)	1.15	1.15	1.14
GMLE(ES)	1.13	1.14	1.13
SA	1.44	1.25	1.16
Out.	1.34	1.20	1.13
Optimal	1.09	1.10	1.10
Single Best	3.64	2.51	1.98

Figure 3: Impact of the Residual Uncertainty Ratio

period 30 and the most recent 24 periods are used as historical, a mixture of pre-change and post-change data is used when one estimates combining weights for time periods from 31 through 53. As in the stationary data experiments, 2500 data sets are generated. After every 25 data sets, a new pair of cases is randomly picked.

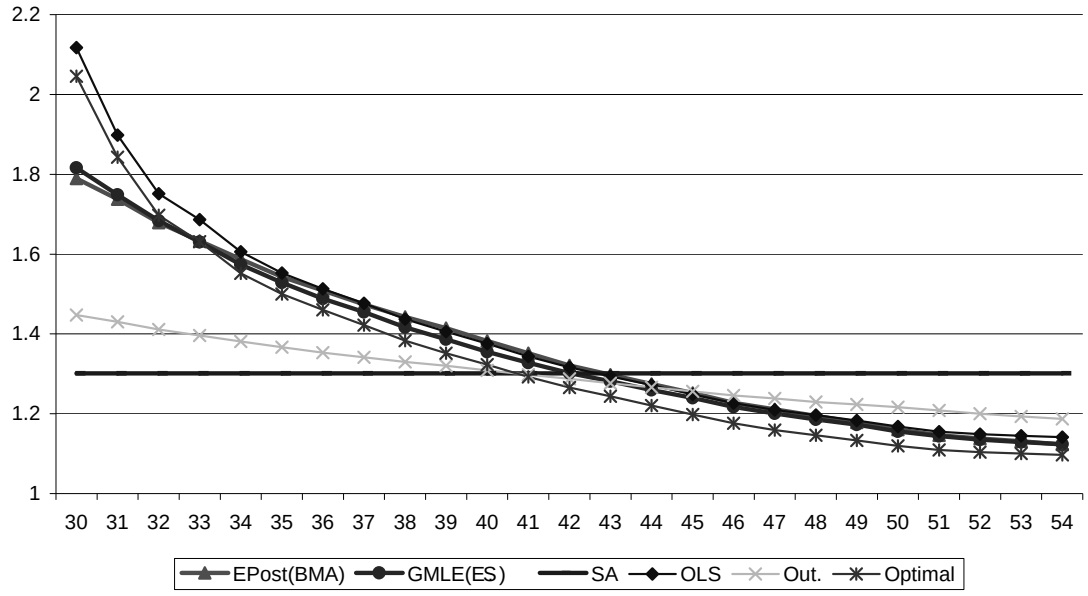


Figure 4: Evoution of SMFER over Time, for Non-Stationary Data

Figure 4[7] shows that how different combination methods respond after the structural change. An ideal method will converge quickly to its last value. As expected, both OLS and Optimal require more data to converge and produce relatively bad forecasts during the early transition periods. SA and Outperformance demonstrate good adaptability to structural change in early periods, but they converge to higher MSFER's. Epost(BMA) and GMLE(ES) have better performance than OLS and Optimal in early periods and converge to a low MSFER.

[7]Only a subset of the combination methods is included.

4 Conclusions

The purpose of this paper is to explore the improvement of forecast accuracy by combining multiple forecasts. Based on a Bayesian combination framework, four new forecast combination methods are proposed. Computational experiments are conducted in both stationary and non-stationary environments. Combined forecasts are clearly superior to individual ones. The combination methods proposed by the authors exhibit the most stable and satisfactory performance in the scenarios tested, compared to other commonly used combination methods. OLS and Optimal require more historical data to calibrate, but they perform slightly better when they lots of representative data. Simple Average and Outperformance are very effective in certain settings, and perform very poorly in others (see the last paragraph of section 3.2). In Zhang et al (2004) we test these methods on a large data set from a major semiconductor manufacturer.

5 Acknowledgement

The authors wish to thank the Semiconductor Research Council, the National Science Foundation and IBM for financial support and advice, without which this research would not have been possible.

References

[1] Aksu C, Gunter SI. 1992. An empirical analysis of the accuracy of SA, OLS, ERLS and NRLS combination of forecasts. *International Journal of Forecasting* 8: 27-43.

[2] Bates JM, Granger CWJ. 1969. The combination of forecasts. *Operational Research Quarterly* 20: 451-468.

[3] Bordley RF. 1986. Technical note: Linear combination of forecasts with an intercept: A Bayesian approach. *Journal of Forecasting* 5: 243-249.

[4] Bunn DW. 1975. A Bayesian approach to the linear combination of forecasts. *Operational Research Quarterly* 26: 25-329.

[5] Chan CK, Kingsman BG, Wong H. 1999. A comparison of unconstrained and constrained OLS for the combination of demand forecasts: A case study of the ordering and stocking of bank printed forms. *Annals of Operations Research* 87: 129-140.

[6] Clemen RT. 1986. Linear constraints and the efficiency of combined forecasts. *Journal of Forecasting* 5: 31-38.

[7] Clemen RT. 1989. Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting* 5: 559-583.

[8] de Menezes LM, Bunn DW, Taylor JW. 2000. Review of guidelines for the use of combined forecasts. *European Journal of Operational Research* 120: 190-204.

[9] Dorfman JH, McIntosh CS. 2001. Imposing inequality restrictions: efficiency gains from economic theory. *Economics Letter* 71: 205-209.

[10] Fernandez C, Ley E, Steel MFJ. 2001. Benchmark priors for Bayesian model averaging. *Journal of Econometrics* 100: 381-427.

[11] Geweke J. 1986. Exact inference in inequality constrained normal linear regression model. *Journal of Applied Econometrics* 1: 127-141.

[12] Granger CWJ, Ramanathan R. 1984. Improved methods of forecasting. *Journal of Forecasting* 3: 197-204.

[13] Gunter SI. 1990. Theoretical justification of the efficiency of simple average combinations. Working Paper. Temple University, Philadelphia, PA.

[14] Gunter SI. 1992. Nonnegativity restricted least squares combinations. *International Journal of Forecasting* 8: 45-59.

[15] Holden K, Peel DA. 1989. Unbiased, efficiency and the combination of economic forecasts. *Journal of Forecasting* 8: 175-188.

[16] Leamer EE. 1978. *Specification Searches*. Wiley: New York.

[17] MacDonald R, Marsh IW. 1994. Combining exchange rate forecasts: What is the optimal consensus measure? *Journal of Forecasting* 13: 313-333.

[18] Min C, Zeller A. 1993. Bayesian and non-Bayesian methods for combining models and forecasts with application to forecasting international growth rates. *Journal of Econometrics* 56: 89-118.

[19] Pindyck RS, Rubinfeld DL. 1981. *Econometric models and economic forecasts*. New York: McGraw-Hill.

[20] Reid DJ. 1968. Combining three estimates of gross domestic product. *Economica* 35: 431-444.

[21] Roundy R. 2002. Report on practices related to demand forecasting for semiconductor products. *Technical Report* No. 1293, School of ORIE, Cornell University.

[22] Tiao GC, Zellner A. 1964. Bayes's theorem and the use of prior knowledge in regression analysis. *Biometrika* 51: 219-230.

[23] Zellner A. 1986. On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*, Goel PK, Zellner A (eds); North-Holland: Amsterdam.

[24] Zhang F., A. Heching, S. Hood, J. Hosking, Y. Leung, R. Roundy and J. Wong. 2004. Experiments on Combining Demand Forecasts with Semiconductor Data. *Technical Report* No. 1396, School of ORIE, Cornell University.

A Derivation of posterior distribution

The marginal posterior distribution of β is given by

$$p(\beta|\mathbf{y}) \propto \int_0^\infty (\sigma^{-2})^{\frac{v''+r+1}{2}} e^{-\frac{1}{2\sigma^2}[v''s''^2+(\beta-\hat{\mu})'\mathbf{W}(\beta-\hat{\mu})]} d\sigma$$

Let $z = \sigma^{-2}$. Then

$$p(\beta|\mathbf{y}) \propto \int_0^\infty \frac{1}{2} z^{\frac{v''+r-2}{2}} e^{-\frac{v''s''^2+(\beta-\hat{\mu})'\mathbf{W}(\beta-\hat{\mu})}{2}z} dz$$

Since $\int_0^\infty z^{\alpha-1}e^{-z/\gamma}dz = \gamma^\alpha\Gamma(\alpha)$,

$$p(\beta|\mathbf{y}) \propto \frac{1}{2}(\frac{2}{v''s''^2+(\beta-\hat{\mu})'\mathbf{W}(\beta-\hat{\mu})})^{\frac{v''+r-2}{2}}\Gamma(v''+r)$$

B Notes on Simplification and Computation

In this section of the appendix we present some of the algebraic simplifications that we found to be useful in developing computer code for these methods. Recall that in section 2.2.2 we assume that $p(\sigma) \propto \sigma^{-1}$, an improper distribution. In light of the definition of $p(\sigma)$ two lines before equation (2), this is tantamount to letting $v' \to 0$ and $v's'^2 \to 0$. Thus equation (2) simplifies to (4). We also note that in section 2.2.1 the following simplifications arise:

$$\begin{aligned} \mathbf{V}^{-1} &= g\mathbf{X}'\mathbf{X} \\ \mathbf{W} &= (g+1)\mathbf{X}'\mathbf{X} \\ \tilde{\mu} &= \frac{1}{1+g}\,[g\mu+\hat{\beta}] \\ v'' &= n \end{aligned}$$

Applying (4) we can now express (3) as

$$p(\beta,\sigma|\mathbf{y}) \propto (\sigma^{-2})^{\frac{1}{2}(2k+v+1)} e^{-\frac{1}{2\sigma^2}[(\beta-\mu)'\mathbf{V}^{-1}(\beta-\mu)+(\beta-\hat{\beta})'\mathbf{X}'\mathbf{X}(\beta-\hat{\beta})+vs^2]} \qquad (12)$$

which is the limit of the middle line of (3) as v' tends to zero. The term in brackets is key. The paper simplifies it as $v''s''^2 + (\beta-\tilde{\mu})'\mathbf{W}(\beta-\tilde{\mu})$. We claim that

$$v''s''^2 + (\beta-\tilde{\mu})'\mathbf{W}(\beta-\tilde{\mu}) = (\mathbf{y}-\mathbf{X}\beta)'(\mathbf{y}-\mathbf{X}\beta) + (\beta-\mu)'\mathbf{V}^{-1}(\beta-\mu) \quad (13)$$

We use (13) in computation. The proof of (13) is algebraic, starting from the exponent in (12) and applying the definition of s.

Consider the last equation of section 2.2.3, the basis for computing the mean of the posterior distribution of the weights via Monte-Carlo integration. In our code we only generate points for which $q(\beta_i) = 1$. Also, we use the uniform distribution, so $I(\beta_i)$ is independent of β_i. Both terms cancel in the ratio. The formula for $p(\beta_i|y)$ is the last equation of section 2.2.1. Note that the Gamma term and the $\frac{1}{2}$ in this equation are independent of β_i, so they will also cancel in the ratio. Finally, we apply (13), so

$$p(\beta|\mathbf{y}) \propto [(\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta) + (\beta - \mu)'\mathbf{V}^{-1}(\beta - \mu)]^{-\frac{n+k}{2}}$$

In the code, the i-th element of tt is this expression, evaluated for the the i-th randomly generated β vector. The i-th element of AA is the corresponding $(\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta)$ term, and the i-th element of BB is the corresponding $(\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta)$ term.