Forecasting Out of Sample: Combination Forecast Methods

Jack K. Strauss

Miller Chair of Applied Economics

University of Denver

Bank of Indonesia Presentation

October 8, 2015

Challenges

- Why Combine?
  Model Uncertainty
  Difficult to identify a single best model.
  Diversification gains occur if models are not collinear
- When to Combine?
  Individual models are misspecified
  Instability is possible
  Short track record
- What to Combine?
  Forecasts using different information sets
  Forecasts based on different modelling approaches
  Surveys with time series with financial variables

Applications

Forecast combinations have been successfully applied in many areas of forecasting (Timmermann)

- GDP
- Inflation, interest rates
- Stock returns
- Employment growth,
- Regional variables such as housing prices or jobs
- Exchange rates and currency volatility
- political risk
- outcomes of football games, city populations, meteorological data

Why Do?

"The results have been virtually unanimous: combining multiple forecasts leads to increased forecast accuracy" (Clemens, 1989).
Makridakis and Hibon (2000) shows in the M3-competition, which involves forecasting 3003 time series concluded: "The accuracy of the combinations of various methods outperforms, on average, the specific methods being combined and does well in comparison to other methods."
Stock and Watson (2001,2004) undertook an extensive study of 150 variables and find that pooling outperforms predictions from the single best model. They show pooling particularly helps forecast inflaiton. Mercellino (2004) shows pooling also works with European data.

Challenges

Forecast combination:
Weighted average of individual model forecasts
Since the seminal work of Bates and Granger (1969), forecast combinations have come to be viewed as a simple and effective way to improve and robustify the forecasting performance.
There is increasing interest in forecasting methods that use large databases. Forecast combinations are now in widespread use in central banks, among private sector forecasters and in academic studies.

- What types of forecasts benefit most from combination,
- Why combination schemes are optimal in a given forecast situation
- When to expect the greatest advantage from forecast combination
- How to forecast using combination methods.

We will use an out of sample forecasting approach. OOS avoids look ahead bias (avoids data mining), and resembles forecaster in real time. OOS approach is more robust to unmodeled instability (Clark and McCracken, 2005, Giacomini and White, 2006, Giacomini and Rossi, 2009, 2010) or to overfit (McCracken, 1998, Clark, 2004).
Divide sample into an insample period that estimates a simple ARDL, training period to obtain weights and an out of sample period to test the forecast.
Use MSFE - mean square forecast error (Actual - forecast)*(Actual - forecast)
Horse race between the model and AR (or historical avg) benchmark
Choose a long horizon period to emphasize repeated tests of the model, graph it to show consistency over time. Horizon period will equal 4 or 8.
Expanding window (recursive) compared to a rolling window.

Combinations important when

- Forecasting Out-of-Sample (OOS) with Combination methods important when:
  - Structural instability
    - Predictive ability can vary markedly over time
    - Difficult to anticipate structural changes, and more Difficult to estimate/predict timing of breaks
  - Model mis-specification
    - uncertainty about correct model
    - Plethora of potential predictors - International, Supply shocks, Monetary, Exchange rates, Indian Variables
  - Lack of theory Theoretical models are highly stylized and fail to guide model selection
  - Data difficulties -mis-measurement, reliance on a single variable poses problems.

  As a result, important to reduces forecast variance (Rapach, Strauss & Zhou, 2010) and combination forecasts are a Type of shrinkage forecast

Strategies with many potential predictors

- General ("kitchen sink") model
  - Include all potential predictors in single model
- Simple Combinations
  - Mean, Median, Trimmed Mean
  - Simple Combination Methods often work well
- Cluster Methods
  - Rank predictors into useful and nonuseful groups
- Principal Components Methods
- Mean Squared Forecast Errors (MSFE) Discount
- Factor/Beta model
  - State variable = f(small # of aggregate factors)

Intuition

- Models maybe incomplete
- Employ different information or databases, or mis-measured data
- Some maybe surveys
- others nonlinear
- Some models may produce biased estimates
- Can combine forecasts or combine information (factors models)

Intuition

- Over-reliance on one/small number of predictors is risky
- Incorporate information from large number of potential predictors to diversify risk
  - Similar to portfolio diversification (Timmermann, 2006)
  - Of course, theory should inform selection of potential predictors
- However, over-parameterized models usually perform very poorly
- So, we need to incorporate information efficiently, imposing various types of restrictions
- Strong relation to aggregate economy can warrant beta or factor structure

It works...in theory, empirical emplications and in Nepal

Stock and Watson (1996) undertake a systematic study of a wide variety of economic time series and find that the majority of these are subject to change.
Diebold and Pauly (1987), Clements and Hendry (1998, 1999, 2006), Pesaran and Timmermann (2005) and Timmermann (2006) view model instability as an important determinant of forecasting performance and a potential reason for combining models.
Stock and Watson show that Business Cycles determinants are not systematic; sometimes due to inventory buildup, too high interest rates, oil shocks, and now financial crises. So the variables that forecast in change overtime.

Insurance against breaks

Hendry and Clements (2004) argue that forecast combinations can provide insurance against extraneous (deterministic) structural breaks when individual forecasting models are misspecified. Their analysis provides supporting evidence that simple combinations can work well under a single end-of-sample break in the process governing the dynamics of the predictor variables. They consider a wide array of designs for the break and find that combinations work particularly well when the predictors are shifted in opposite directions and are positively correlated.

Stock and Watson (2001) support model instability and find that the performance of combined forecasts tends to be more stable than that of the individual constituent forecasts entering in the combinations.

It predicts GDP and inflation

Stock and Watson (2003): ”We undertake an empirical analysis of quarterly data on up to 38 candidate indicators (mainly asset prices) for seven OECD countries for a span of up to 41 years (1959 to 1999). The conclusions from the literature review and the empirical analysis are the same. Some asset prices predict either inflation or output growth in some countries in some periods. Which series predicts what, when and where is, however, itself difficult to predict good forecasting performance by an indicator in one period seems to be unrelated to whether it is a useful predictor in a later period. Intriguingly, forecasts produced by combining these unstable individual forecasts appear to improve reliably upon univariate benchmarks.”

Instability is the norm

” We conclude forecasts based on individual indicators are unstable. For example, in the U.S., recursive (i.e. simulated out of sample) forecasts of the four-quarter growth of industrial production using the term spread were substantially more accurate than a simple autoregressive benchmark from 1971 to 1984, but were substantially less accurate than the autoregressive benchmark from 1985 to 1999. More generally, finding an indicator that predicts well in one period is no guarantee that it will predict well in later periods; indeed, whether an indicator-based forecast outperforms an autoregressive benchmark in a subsequent period appears to be independent of whether it has done so in the past. This, along with evidence based on formal stability tests, suggests that instability of predictive relations based on asset prices (and most other candidate leading indicators) is the norm.”

Insample tests misleading

Stock and Watson: ”the most common method of identifying a potentially useful predictor is to rely on in-sample significance tests such as Granger causality tests, this turns out to provide no assurance that the identified predictive relation is stable. Indeed, the empirical results indicate that a significant Granger causality statistic contains little or no information about whether the indicator has been a reliable predictor.”
Massive literature document yield or credit spread, stock market (or other financial variables) as predictors of GDP or inflation, but the real-time evidence is weak as they work for one period, but then fail to predict a different period.
Reliance is fragile or not robust.

Reliance on single predictors risky

”In a similar vein, Cecchetti, Chu and Steindel (2000) performed a simulated out of sample forecasting experiment on various candidate leading indicators of inflation, from 1985 to 1998 in the U.S., including interest rates, term and default spreads, and several nonfinancial indicators. They concluded that none of these indicators, financial or nonfinancial, reliably predicts inflation in bivariate forecasting models, and that there are very few years in which financial variables outperform a simple autoregression. Because they assessed performance on a year by year basis, these findings have great sampling variability and it is difficult to know how much of this is due to true instability. Their findings are, however, consistent with Stock and Watson‘s (1996) results based on formal stability tests that time variation in these reduced form bivariate predictive relations is widespread in the U.S. data.
Our reading of this literature suggests that many of these forecasting relations are ephemeral. ”

Combinations work

Stock and Watson: The results are not entirely negative, however. Rather than focusing on individual asset prices, all of which have their deficiencies as leading indicators, these results suggest instead that combining information from a large number of asset prices can lead to reliable forecasts.
”Suitably combining the information in the various predictors appears to circumvent the worst of these instability problems. For example, the median of the forecasts of output growth based on individual asset prices produces a forecast that is reliably more accurate than the AR benchmark, even though the individual forecasts used to compute the median are not. Similarly, forecasts of inflation that combine information from measures of real activity and output gaps appear to be reliable and stable, even though the individual component forecasts are not.”

Diversify

- Multitude of variables furthers help diversify against model misspecification and data reporting issues, and relevant for Nepal.
- All models are simplifications of reality,
- Forecasting a priori dont know which model is correct, data is accurate or will be revised.
- Dont want to put eggs all in one basket so Diversify so rely on more than one forecast
- "Timmerman: ”A simple portfolio diversification argument motivates the idea of combining forecasts, Bates and Granger (1969). It is difficult to fully appreciate the strength of the diversification or hedging argument underlying forecast combination.”

It works in practice

Forecast combinations have been used successfully in empirical work in such diverse areas as forecasting Gross National Product, currency market volatility, inflation, money supply, stock prices, meteorological data, city populations, outcomes of football games, wilderness area use, check volume and political risks, c.f. Clemen (1989).
Summarizing the simulation and empirical evidence in the literature on forecast combinations, Clemen (1989) writes ”The results have been virtually unanimous: combining multiple forecasts leads to increased forecast accuracy.... in many cases one can make dramatic performance improvements.”

It works in practice

Makridakis and Hibon (2000) conducted the M3 competition which involved forecasting 3003 time series and concluded ”The accuracy of the combination of various methods outperforms, on average, the specific methods being combined and does well in comparison with other methods.”
Stock and Watson (2001, 2004) undertook an extensive study across 300 economic and financial variables using linear and nonlinear forecasting models and found that pooled forecasts outperform predictions from the single best model, confirming Clemens conclusion. The overall dominance of the combination forecasts holds at 1, 6, 12 month horizons.
Best combination methods combine forecasts across many different time-series models.
Marcellino (2004) extend it to a large European data set with essentially the same conclusions.

Timmermann: Choosing the single forecast with the best track record is often a bad idea

Many studies have found that combination dominates the best individual forecast in 00S forecasting experiments.
Makridakis et al (1982) reports simple AVG of 6 forecasting methods performed better than the underlying ind. forecasts.
In simulation experiments Gupta and Wilton (1987) also find combination superior to the single best forecast.
Makridakis and Winkler (1983) report large gains from simply AVG forecasts from ind. models over performance of best model.
Hendry and Clements (2002) explain the better performance of combination methods over the best individual model by misspecification of the models caused by deterministic shifts in the underlying DGP.
Naturally, the models cannot be misspecified in the same way with regard to this source of change, or else diversification gains = 0.

Trimming worst models often improves performance

Winkler and Makridakis (1983) find that including very poor models in an equal-weighted combination can substantially worsen forecasting performance.
Stock and Watson (2003) also find that the simplest forecast combination methods such as trimmed equal weights and slowly moving weights tend to perform well and that such combinations do better than forecasts from a dynamic factor model.

Shrinkage often improves performance.

A shrinkage estimator is an estimator that, either explicitly or implicitly, incorporates the effects of shrinkage. In loose terms this means that a naive or raw estimate is improved by combining it with other information.
Shrinkage is implicit in Bayesian inference and penalized likelihood inference, In contrast, simple types of Max-lld and ls procedures do not include shrinkage effects.
Shrinkage works with large number of explanatory variables. Diebold and Pauly (1990) and Stock and Watson report that shrinkage weights systematically improve forecasting .
Aiolfi and Timmermann (2004) first pre-select models into either quartiles or clusters based on past forecasting performance across models. Then pool forecasts within each cluster and estimate optimal combination weights that are shrunk towards equal weights. These conditional combination strategies outperform most strategies.

Keep the combinations weights simple

Stable, equal weights have so far been the workhorse of the combination literature and have set a benchmark that has proved surprisingly difficult to beat.
This is surprising since theory suggests no combination scheme to be dominant, since the various methods incorporate restrictions on the covariance matrix that are designed to trade off bias against reduced parameter estimation error.
The optimal bias can be expected to vary across applications, and the scheme that provides the best trade-off is expected to depend on:
the sample size,
the number of forecasting models involved,
the ratio of the variance of individual models‘ forecast errors as well as their correlations
and the degree of instability in the underlying DGP

In order to form simulated out-of-sample forecasts of $y^{\tau}_{t+\tau}$, we first divide the sample of T observations into in-sample and out-of-sample ortions, where the first I observations comprise the in-sample period and the last O observations make up the out-of-sample period ($T = I + O$). We compute the initial out-of-sample forecast, corresponding to $y^{\tau}_{I+\tau}$, based on the $predictor_{i,t}$ as

$$\widehat{y}^{\tau}_{i,I+\tau|I} = \widehat{\alpha}_I + \sum_{j=0}^{n_1-1} \widehat{\beta}_{j,I} y_{t-j} + \sum_{j=0}^{n_2-1} \widehat{\gamma}_{j,I} predictor_{i,t-j} \qquad (1)$$

where $\widehat{\alpha}_I$, $\widehat{\beta}_{j,I}$, and $\widehat{\gamma}_{j,I}$ are the OLS estimates of α, β_j, and γ_j, respectively.
Further, we may need a training period to obtain the weights to use for each ARDL model.

Kitchen sink model

- Consider N potential predictors, where N is "large"
- Let $\Delta y_{t+h}^{h} = (1/h) \sum_{j=1}^{h} \Delta y_{k,t+j}$
- $\Delta y_t = y_t - y_{t-1}$
- h denotes horizon, lag length chosen by AIC
- KS: $\Delta y_{t+h}^{h} = a + b\Delta y_t + \sum_{i=1}^{N} c_i x_{i,t} + e_{t+h}^{h}$
- Drawbacks
  - In-sample over-fitting
  - Typically delivers very poor out-of-sample forecasts
  - Estimation may be infeasible if N is large relative to in-sample period
- Eg, kitchen sink model performs very poorly for forecasting U.S. equity premium (Goyal & Welch, 2008; Rapach, Strauss)

Simple Combinations

- Simple Mean, Median, Trimmed Mean (remove best and worst ARDL performer)
- Individual ARDL: $\Delta y^h_{k,t+h} = a + b\Delta y_{k,t} + c_i x_{i,t} + e^h_{k,t+h}$
  - Individual model forecast based on predictor $x_{i,t}$
- Combination forecast: $\Delta \hat{y}^{h,c}_{k,t+h} = \sum_{i=1}^{N} w_i \Delta \hat{y}^{h,i}_{k,t+h}$
  - $\sum_{i=1}^{N} w_i = 1$
  - Weighted average of individual model forecasts
  - Various combining methods available
  - Simple (eg, $w_i = 1/N$)
  - Adv: Easy, Empirically works well
  - Use when large number of predictors, small number of observations (no training period req.) and there is substantial noise and breaks (e.g., stock returns)
  - Disadv: includes poor predictors and doesn't allow a training period to sort out potentially poor predictors

Discount MSFE combining method

Stock and Watson (2004) consider a combination method where the weights in equation depend inversely on the recent historical forecasting performance of the individual ARDL models. Their discount MSFE (DMSFE) combination method uses the following weights

$$w_{i,t} = \varepsilon_{i,t}^{-1} / \sum_{j=1}^{n} \varepsilon_{j,t}^{-1} \tag{2}$$

$$\varepsilon_{i,t} = \sum_{s=l}^{t-\tau} \delta^{t-\tau-s} (y_{s+h}^{\tau} - \widehat{y}_{i,s+\tau|s}^{\tau})^2 \tag{3}$$

where δ is a discount factor. We consider δ values of 1.0 and 0.9.[1]

[1] $\delta = 1$ implies there is no discounting, and $\delta < 1$ implies greater importance is attached to the recent forecasting performance of the individual models.

Cluster

We follow Aiolfi and Timmermann's (2006) $C(K, PB)$ algorithm, in which the combination forecast is the average of the individual forecasts generated by a cluster of individual models that have performed best(lowest MSFE).
To form the initial combination forecast, we first compute the MSFE for the individual forecasts over the initial holdout period (from $t = I + \tau$ to $t = I + H$), and group the individual models into K equal-sized clusters, where the 1st (2nd) cluster contains the individual models with the lowest (2nd) MSFE values.
1st combination forecast is the avg of the individual forecasts generated by the ARDL models included in the first cluster.
Following Aiolfi and Timmermann (2006), we consider $K = 2$ and $K = 3$ in our study.

Principal component combining methods

In this method, the combination forecast is based on the first m principal components of the individual forecasts. Let $PC^{\tau}_{1,s+\tau|s}, \ldots, PC^{\tau}_{m,s+\tau|s}$, $s = 1, \ldots, t$, represent the first m principal components of the uncentered second-moment matrix of the individual forecasts. To form a combination forecast, we estimate the following regression model:

$$y^{\tau}_{t+\tau} = \lambda_1 PC^{\tau}_{1,s+\tau|s} + \ldots + \lambda_m PC^{\tau}_{m,s+\tau|s} + \nu^{\tau}_{s+\tau} \tag{4}$$

where $s = 1, \ldots, t - \tau$. The combination forecast is given by $\widehat{y}^{\tau}_{t+\tau} = \widehat{\lambda}_1 PC^{\tau}_{1,s+\tau|s} + \ldots + \widehat{\lambda}_m PC^{\tau}_{m,s+\tau|s}$, where $\widehat{\lambda}_1, \ldots, \widehat{\lambda}_m$ are the OLS estimates of $\lambda_1, \ldots, \lambda_m$, respectively, in equation (4). We consider $m = 1$, $m = 2$, $m = 3$.

Predictive Likelihood Model Averaging
Baynesian Model averaging is widely used in academic literature as well as in practice. Critical to obtain weights during the forecast evaluation period.
One way is to use analogue of Baynesian for frequentist statistics. Akaike suggests uses the AIC - it is an asymptotically unbiased of minus twice the likelihood function.