Forecast Combinations

Oxford, July-August 2013

Allan Timmermann[1]

[1]UC San Diego, CEPR, CREATES

Disclaimer: The material covered in these slides represents work from a book on economic forecasting jointly authored by Graham Elliott and Allan Timmermann (UCSD)

1. Introduction: When, why and what to combine?
2. Optimal Forecast Combinations: Theory
   - Optimal Combinations under MSE loss
3. Estimating Forecast Combination Weights
   - Weighting schemes under MSE loss
   - Forecast Combination Puzzle
   - Rapach, Strauss and Zhou, RFS 2010
   - Elliott, Gargano, and Timmermann, JoE Forthcoming
   - Time-varying combination weights
4. Model Combination
5. Density Combination
   - Geweke and Amisano (2011)
6. Bayesian Model Averaging
   - Avramov, JFE 2002
7. Conclusion

Key issues in forecast combination

- **Why combine?**
  - Many models or forecasts with 'similar' predictive accuracy
    - Difficult to identify a single best forecast
    - State-dependent performance
  - Diversification gains
- **When to combine?**
  - Individual forecasts are misspecified
  - Unstable forecasting environment (past track record unreliable)
  - Short track record; use 1-over-N weights?
- **What to combine?**
  - Forecasts using different information sets
  - Forecasts based on different modeling approaches (linear/nonlinear)
  - Surveys, econometric model forecasts

Essentials of forecast combination

- **Dimensionality reduction:** Combination reduces the information in a vector of forecasts to a single summary measure using a set of combination weights
- **Optimal combination** chooses weights to minimize the expected loss of the combined forecast
  - More accurate forecasts tend to get larger weights
  - Combination weights also reflect correlations across forecasts
  - Estimation error is important to combination weights
- **Irrelevance Proposition**: In a world with no model misspecification, infinite data samples (no estimation error) and complete access to the information sets underlying the individual forecasts, there is no need for forecast combination.

When to combine?

- Combined forecast $f(\hat{f}_1, \hat{f}_2)$ dominates individual forecasts $\hat{f}_1$ and $\hat{f}_2$ if

$$E\left[\mathcal{L}(\hat{f}_i, y_{T+h})\right] > \min_{f(.)} E\left[\mathcal{L}(f(\hat{f}_1, \hat{f}_2), y_{T+h})\right], \quad \text{for } i = 1, 2$$

- $\mathcal{L}$: loss function, e.g., MSE loss $(y - f)^2$
- y_{T+h} : outcome h periods ahead
- h : forecast horizon
- Forecast combination is essentially a model selection and parameter estimation problem with special constraints on the estimation problem

Applications of forecast combinations

- Forecast combinations have been successfully applied in several areas of forecasting:
  - Gross National Product
  - currency market volatility and exchange rates
  - inflation, interest rates, money supply
  - stock returns
  - meteorological data
  - city populations
  - outcomes of football games
  - wilderness area use
  - check volume
  - political risks
- Estimation of GDP
- Averaging across values of unknown parameters

Two types of forecast combinations

1. Data underlying the forecasts are not observed:
   - Treat individual forecasts like any other conditioning information (data) and estimate the best possible mapping from the forecasts to the outcome
2. Data underlying the model forecasts is observed: 'model combination'

- Using a middle step of first constructing forecasts limits the flexibility of the final forecasting model. Why not directly map the underlying data to the forecasts?
  - Estimation error plays a key role in the risk of any given method. Model combination yields a risk function which, through parsimonious use of the data, could result in an attractive risk function
  - Combined forecast can be viewed simply as a different estimator of the final model

Combinations of forecasts: theory

- Restrict conditioning information to a set of m forecasts $z_t = \{\hat{f}_{1t+h|t}, ..., \hat{f}_{mt+h|t}\}$
- The optimal combination is the function of the forecasts $f(\hat{f}_{1t+h|t}, \hat{f}_{2t+h|t}, ..., \hat{f}_{mt+h|t})$ that solves

$$\min_{f(.)} E\left[\mathcal{L}(f(\hat{f}_{1t+h|t}, \hat{f}_{2t+h|t}, ..., \hat{f}_{mt+h|t}), y_{t+h})|Z_t\right]$$

- Optimality of the combined forecast is conditional on observing the forecasts $\{\hat{f}_{1t+h|t}, \hat{f}_{2t+h|t}, ..., \hat{f}_{mt+h|t}\}$ rather than the underlying information sets used to construct the forecasts
- If the model $f(.)$ is a linear index, the combination is a linear combination with weights $\omega_1, ..., \omega_m$

Combinations of forecasts: theory

- Specialized concepts in optimal forecast combination arise from additional restrictions placed on the search for combination models
- Because the underlying 'data' are forecasts, they can be expected to obtain non-negative weights that sum to unity,

$$0 \leq \omega_i \leq 1, \qquad i = 1, ..., m$$

- Such constraints can be used to reduce the relevant parameter space for the combination weights and offer a more attractive risk function
- No need to constrain z_t to include only the set of observed forecasts $\{\hat{f}_{1t+h|t}, ..., \hat{f}_{mt+h|t}\}$. This information could be augmented to include other observed variables, z_t :

$$\min_{f(.)} E\left[\mathcal{L}(f(\hat{f}_{1t+h|t}, \hat{f}_{2t+h|t}, ..., \hat{f}_{mt+h|t}, z_t), y_{t+h})\right]$$

Combinations of two forecasts

- Two individual forecasts f_1, f_2 with forecast errors $e_1 = y - f_1$, $e_2 = y - f_2$
- Both forecasts assumed to be unbiased: $E[e_i] = 0$
- Variances of forecast errors: σ_i^2, $i = 1, 2$. Covariance is σ_{12}
- Combined forecasts will also be unbiased if the weights add up to one:

$$f = \omega f_1 + (1 - \omega) f_2$$

- Combined forecast error is a weighted average of the individual forecast errors:

$$\begin{aligned} e(\omega) &= y - \omega f_1 - (1 - \omega) f_2 = \omega e_1 + (1 - \omega) e_2 \\ E[e(\omega)] &= 0 \\ Var(e(\omega)) &= \omega^2 \sigma_1^2 + (1 - \omega)^2 \sigma_2^2 + 2\omega(1 - \omega)\sigma_{12} \end{aligned}$$

Combinations of two forecasts: optimal weights

- Solving for the MSE-optimal combination weights,

$$\begin{aligned} \omega^* &= \frac{\sigma_2^2 - \sigma_{12}}{\sigma_1^2 + \sigma_2^2 - 2\sigma_{12}} \\ 1 - \omega^* &= \frac{\sigma_1^2 - \sigma_{12}}{\sigma_1^2 + \sigma_2^2 - 2\sigma_{12}} \end{aligned}$$

- Combination weight can be negative if $\sigma_{12} > \sigma_1^2$ or $\sigma_{12} > \sigma_2^2$
- Negative weight on a forecast does not mean that it has no value - it means the forecast can be used to offset the prediction errors of other models
- Weakly correlated forecast errors: weights are the relative variance σ_2^2/σ_1^2 of the forecasts:

$$\omega^* = \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2} = \frac{\sigma_2^2/\sigma_1^2}{1 + \sigma_2^2/\sigma_1^2}$$

- Greater weight is assigned to more precise models (small σ_i^2)

Combinations of multiple unbiased forecasts

- $e = \iota_m y - f$: vector of forecast errors;

$$E[e] = 0, \quad \Sigma_e = E[ee']$$

- Minimizing MSE:

$$\begin{aligned} \omega^* &= \arg\min_{\omega} \omega' \Sigma_e \omega, \\ & \quad s.t.\ \omega' \iota_m = 1 \end{aligned}$$

- Optimal combination weights:

$$\begin{aligned} \omega^* &= (\iota_m' \Sigma_e^{-1} \iota_m)^{-1} \Sigma_e^{-1} \iota_m, \\ MSE(\omega^*) &= \omega^{*\prime} \Sigma_e \omega^* = (\iota_m' \Sigma_e^{-1} \iota_m)^{-1} \end{aligned}$$

Optimality of equal weights

- Equal weights (EW) play a special role in forecast combination
- EW are optimal in population when the individual forecast errors have identical variance, σ^2, and identical pair-wise correlations ρ

$$\begin{aligned} \Sigma_e^{-1} \iota_m &= \frac{\iota_m}{\sigma^2(1+(m-1)\rho)} \\ (\iota_m' \Sigma_e^{-1} \iota_m)^{-1} &= \frac{\sigma^2(1+(m-1)\rho)}{m} \end{aligned}$$

- This situation holds to a close approximation when all models are based on similar data and perform roughly the same
- More generally, EW are the optimal combination weights when the unit vector lies in the eigen space of Σ_e.
  - Both a sufficient and a necessary condition for equal weights to be optimal

Estimating combination weights

- In practice, combination weights need to be estimated using past data
  - Once we use estimated parameters, the population-optimal weights no longer have any optimality properties in a 'risk' sense
  - For any forecast combination problem, there is typically no single optimal forecast method with estimated parameters
  - Risk functions for different estimation methods will typically depend on the data generating process
    - we prefer one method for some processes and different methods for other data generating processes

Estimating combination weights

- Treating the forecasts as data means that all issues related to how to estimate forecast models from data are relevant
- In the case of forecast combination, the "data" is not the outcome of a random draw but can be regarded as unbiased (if not precise) forecasts of the outcome
- This suggests imposing special restrictions on the combination schemes
  - Under MSE loss, linear combination schemes might impose

$$\sum_i \omega_i = 1, \qquad \omega_i \in [0, 1]$$

- Simple combination schemes such as EW satisfy these constraints and do not require estimation of any parameters
  - EW can be viewed as a reasonable prior when no data has been observed

Estimating combination weights

Existence of many estimation methods boils down to a number of standard issues in constructing forecasts:

- role of estimation error
- lack of a single optimal estimation scheme
- simple methods are difficult to beat in practice
- common baseline is to use a simple EW average of forecasts:

$$f^{ew}_{t+h|t} = \frac{1}{m}\sum_{i=1}^{m} f_{i,t+h|t}$$

- no estimation error here since the combination weights are imposed rather than estimated

Simple combination methods

- Equal-weighted forecast

$$f^{ew}_{t+h|t} = \frac{1}{m}\sum_{i=1}^{m} f_{i,t+h|t}$$

- Median forecast

$$f^{median}_{t+h|t} = median\{f_{i,t+h|t}\}_{i=1}^{m}$$

- Trimmed mean. Order forecasts $\{f_{1,t+h|t} \leq f_{2,t+h|t} \leq ... \leq f_{m-1,t+h|t} \leq f_{m,t+h|t}\}$. Trim top/bottom $\lambda\%$

$$f^{trim}_{t+h|t} = \frac{1}{m(1-2\lambda)} \sum_{i=\lfloor \lambda m+1 \rfloor}^{\lfloor (1-\lambda)m \rfloor} f_{i,t+h|t}$$

Bates-Granger restricted least squares

- Bates and Granger (1969): use plug-in weights in the optimal solution based on the estimated variance-covariance matrix
- This is numerically identical to restricted least squares estimator of the weights from a regression of the outcome on the vector of forecasts $f_{t+h|t}$ and no intercept subject to the restriction that the coefficients sum to one:

$$f^{BG}_{t+h|t} = \hat{\omega}'_{OLS} f_{t+h|t} = (\iota' \hat{\Sigma}_e^{-1} \iota)^{-1} \iota' \hat{\Sigma}_e^{-1} f_{t+h|t}$$

- $\hat{\Sigma}_\varepsilon = (T-h)^{-1} \sum_{t=1}^{T-h} e_{t+h|t} e'_{t+h|t}$: sample estimator of error covariance matrix

Diebold and Pauly (1987) shrinkage estimator

- Forecast combination weights formed as a weighted average of the prior $\omega_p = \iota_m / m$ and the least squares estimates $\hat{\omega}_{OLS}$:

$$\hat{\omega}_B = \tilde{A}\hat{\omega}_{OLS} + m^{-1}(I - \tilde{A})\iota_m$$

- Empirical Bayes approach sets $\tilde{A} = I(1 - \hat{\sigma}^2 / \hat{\tau}^2)$
- $\hat{\sigma}^2$: MLE for variance of the residuals from the OLS combination regression
- $\hat{\tau}^2 = (\hat{\omega}_{ols} - \omega_p)'(\hat{\omega}_{ols} - \omega_p) / tr[(Z_f' Z_f)]^{-1}$
- Z_f : matrix of regressors (ignoring the constant)
- $Z_f' Z_f$ is an unscaled estimate of the variance-covariance matrix of the forecasts

Weights inversely proportional to MSE or rankings

- Ignore correlations across forecast errors and set weights proportional to the inverse of the models' MSE-values:

$$\omega_i = \frac{MSE_i^{-1}}{\sum_{i=1}^{m} MSE_i^{-1}}$$

- Aiolfi and Timmermann (2006) propose a robust weighting scheme that weights forecast models inversely to their rank, $Rank_{it+h|t}$

$$\hat{\omega}_{it+h|t} = \frac{Rank_{it+h|t}^{-1}}{\sum_{i=1}^{m} Rank_{it+h|t}^{-1}}$$

- Best model gets a rank of 1, second best model a rank of 2, etc.

Forecast combination puzzle

- Empirical studies often find that simple equal-weighted forecast combinations perform very well compared with more sophisticated combination schemes that rely on estimated combination weights
  - Smith and Wallis (2009): "*Why is it that, in comparisons of combinations of point forecasts based on mean-squared forecast errors ..., a simple average with equal weights, often outperforms more complicated weighting schemes.*"
- Errors introduced by estimation of the combination weights could overwhelm any gains from setting the weights to their optimal values over using equal weights
- Explanations of the puzzle based on estimation error must show that
  - estimation error is large and/or
  - gains from setting the combination weights to their optimal values are small relative to using equal weights

Forecast combination puzzle - is it estimation error?

- In sufficiently large samples OLS estimation error should be of the order $(m-1)/T$
  - Unless equal weights are close to being optimal, estimation error is unlikely to be the full story, at least when m/T is small
- If poor forecasting methods get weeded out, most forecasts in any combination have similar forecast error variances, leading to a nearly constant diagonal of Σ_e
  - Differences across correlations would be required to cause deviations from EW
  - Large unpredictable component outside all of the forecasts pushes correlations towards positive numbers
- Explanations that aim to solve the forecast combination puzzle by means of large estimation errors require model misspecification or more complicated DGPs than is assumed when estimating the combination weights

Rapach-Strauss-Zhou (RFS, 2010)

- Quarterly data, 1947-2005
- 15 variables from Goyal and Welch (2008)
- Individual univariate prediction models:

$$\begin{aligned} r_{t+1} &= \alpha_i + \beta_i x_{it} + \varepsilon_{it+1} \\ \hat{r}_{t+1|t,i} &= \hat{\alpha}_{t,i} + \hat{\beta}_{t,i} x_{it} \end{aligned}$$

- Forecast combination:

$$\hat{r}^c_{t+1|t} = \sum_{i=1}^{N} \omega_{t,i} \hat{r}_{t+1|t,i}$$

Combination weights (Rapach et al.)

$$
\begin{aligned}
\hat{r}^{c}_{t+1|t} &= \sum_{i=1}^{N} \omega_{t,i} \hat{r}_{t+1|t,i} \\
\omega_{t,i} &= 1/N \\
\omega_{t,i} &= \frac{DMSPE_{t,i}^{-1}}{\sum_{j=1}^{N} DMSPE_{t,j}^{-1}} \\
DMSPE_{t,i} &= \sum_{s=T_0}^{t} \theta^{\tau-1-s} (r_{s+1} - \hat{r}_{s+1,i})^2
\end{aligned}
$$

- DMSPE is the discounted mean squared prediction error, using a discount factor, $\theta \leq 1$

Rapach-Strauss-Zhou (RFS, 2010): results

Table 1
Equity premium out-of-sample forecasting results for individual forecasts and combining methods

Individual predictive regression model forecasts						Combination forecasts		
Predictor	R^2_{OS} (%)	Δ (%)	Predictor	R^2_{OS} (%)	Δ (%)	Combining method	R^2_{OS} (%)	Δ (%)
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
			Panel A. 1965:1–2005:4 out-of-sample period					
D/P	0.34*	0.55	*LTY*	−3.09	2.29	Mean	3.58***	2.34
D/Y	0.25*	1.41	*LTR*	0.33	1.30	Median	3.04***	1.03
E/P	0.36	0.64	*TMS*	−2.96	5.14	Trimmed mean	3.51***	2.11
D/E	−1.42	0.58	*DFY*	−2.72	−0.83	DMSPE, θ = 1.0	3.54***	2.41
SVAR	−12.97	0.13	*DFR*	−1.10	0.57	DMSPE, θ = 0.9	3.49***	2.59
B/M	−2.60	−0.58	*INFL*	−0.84	1.39			
NTIS	−0.91	0.08	*I/K*	1.44**	2.80	Mean, CT	3.23***	1.25
TBL	−2.78	2.60						
			Panel B. 1976:1–2005:4 out-of-sample period					
D/P	−5.08	−0.70	*LTY*	−5.59	−0.89	Mean	1.19*	0.57
D/Y	−6.22	−0.54	*LTR*	−0.27	1.43	Median	1.51**	0.53
E/P	−1.70	0.75	*TMS*	−7.24	2.08	Trimmed mean	1.23*	0.59
D/E	−2.26	−1.65	*DFY*	−2.48	−1.18	DMSPE, θ = 1.0	1.11*	0.54
SVAR	−22.47	0.06	*DFR*	−2.14	−0.64	DMSPE, θ = 0.9	1.01*	0.46
B/M	−4.72	−1.27	*INFL*	−0.08	0.45			
NTIS	0.10	0.60	*I/K*	−3.47	−0.85	Mean, CT	1.20*	0.55
TBL	−7.31	−0.82						
			Panel C. 2000:1–2005:4 out-of-sample period					
D/P	10.32*	12.96	*LTY*	−0.32	0.24	Mean	3.04**	2.31
D/Y	10.40*	12.98	*LTR*	−1.72	2.57	Median	1.56*	0.28
E/P	8.02*	9.53	*TMS*	−4.98	4.23	Trimmed mean	2.98**	2.12
D/E	0.56	0.50	*DFY*	−0.53	−1.52	DMSPE, θ = 1.0	2.56**	1.65
SVAR	−5.62	−1.64	*DFR*	−2.10	1.76	DMSPE, θ = 0.9	2.66**	1.97
B/M	2.32	3.09	*INFL*	−1.42	0.57			
NTIS	−4.09	1.33	*I/K*	8.96**	9.13	Mean, CT	2.43**	1.32
TBL	−2.50	−0.20						

R^2_{OS} is the Campbell and Thompson (2008) out-of-sample R^2 statistic. Utility gain (Δ) is the portfolio management fee (in annualized percentage return) that an investor with mean-variance preferences and risk aversion coefficient of three would be willing to pay to have access to the forecasting model given in Column (1), (4), or (7) relative to the historical average benchmark forecasting model; the weight on stocks in the investor's portfolio is restricted to lie between zero and 1.5 (inclusive). Statistical significance for the R^2_{OS} statistic is based on the *p*-value for the Clark and West (2007) out-of-sample *MSPE-adjusted* statistic; the statistic corresponds to a one-sided test of the null hypothesis that the competing forecasting model given in Column (1), (4), or (7) has equal expected square prediction error relative to the historical average benchmark forecasting model against the alternative hypothesis that the competing forecasting model has a lower expected square prediction error than the historical average benchmark forecasting model. *, **, and *** indicate significance at the 10%, 5%, and 1% levels, respectively.

Rapach-Strauss-Zhou (RFS, 2010): results

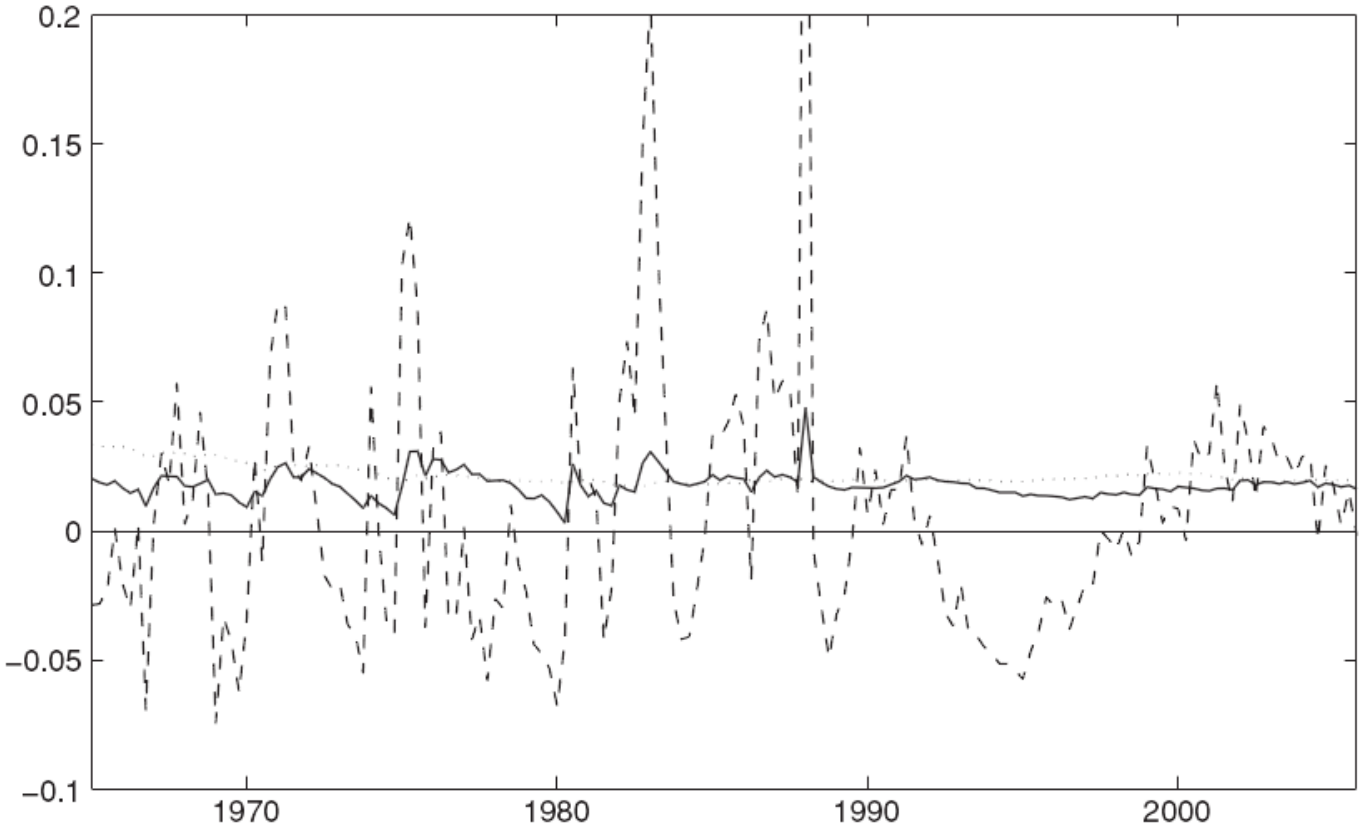


Figure 5
Equity premium forecasts for the mean combining method, historical average, and kitchen sink model, 1965:1–2005:4
The solid (dotted, dashed) line corresponds to the mean combining method (historical average, kitchen sink model) forecast.

Empirical Results (Rapach, Strauss and Zhou, 2010)

Table 5
R^2_{OS} statistics for out-of-sample equity premium combination forecasts during good, normal, and bad growth periods, 1965:1–2005:4

	Forecast horizon: one quarter				Forecast horizon: four quarters			
Combining method	Overall	Good	Normal	Bad	Overall	Good	Normal	Bad
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
				Panel A. Sorting on real GDP growth				
Mean	3.58***	1.82	1.71	6.17***	8.19***	3.07	3.63*	11.58***
Median	3.04***	2.67**	0.39	5.02***	6.99***	12.74***	6.35**	5.23***
Trimmed mean	3.51***	2.25*	1.24	5.94***	8.13***	5.41*	4.01*	10.63***
DMSPE, $\theta = 1.0$	3.54***	1.71	1.56	6.26***	7.87***	2.32	3.15	11.46***
DMSPE, $\theta = 0.9$	3.49***	1.60	1.36	6.33***	5.96***	4.71*	0.27	8.27***
				Panel B. Sorting on real profit growth				
Mean	3.58***	2.87*	−1.03	7.94***	8.19***	0.93	4.89*	14.72***
Median	3.04***	2.56**	0.21	5.74***	6.99***	1.14	8.00**	10.18***
Trimmed mean	3.51***	2.85*	−0.67	7.47***	8.13***	1.74	5.83**	13.55***
DMSPE, $\theta = 1.0$	3.54***	2.74*	−1.21	8.08***	7.87***	0.16	4.41	14.78***
DMSPE, $\theta = 0.9$	3.49***	2.51	−1.56	8.40***	5.96***	−4.28	2.00	14.70***
				Panel C. Sorting on real net cash flow growth				
Mean	3.58***	5.44**	2.17*	4.63**	8.19***	3.29*	8.81***	11.42***
Median	3.04***	4.12***	1.80**	4.25**	6.99***	4.99***	6.17**	9.48***
Trimmed mean	3.51***	5.01**	2.36**	4.47**	8.13***	4.39**	9.13***	10.04***
DMSPE, $\theta = 1.0$	3.54***	5.51**	2.13*	4.52**	7.87***	2.97*	8.50***	11.09***
DMSPE, $\theta = 0.9$	3.49***	5.88**	1.84*	4.15*	5.96***	0.53	6.66**	9.56***

This table reports the Campbell and Thompson (2008) R^2_{OS} statistic. The R^2_{OS} statistics are computed for the entire 1965:1–2005:4 forecast evaluation period (Overall) and three subperiods corresponding to the top third (Good), the middle third (Normal), and the bottom third (Bad) of observations sorted on the macroeconomic variable given in the panel heading. Statistical significance for the R^2_{OS} statistic is based on the p-value for the Clark and West (2007) out-of-sample *MSPE-adjusted* statistic; the statistic corresponds to a one-sided test of the null hypothesis that the combination forecast given in Column (1) has equal expected square prediction error relative to the historical average benchmark forecast against the alternative hypothesis that the combination forecast has a lower expected square prediction error than the historical average benchmark forecast. *, **, and *** indicate significance at the 10%, 5%, and 1% levels, respectively.

Rapach, Strauss, Zhou: main results

- Forecast combination methods dominate individual prediction models for stock returns out-of-sample
- Forecast combination reduces forecast variance
- Combined return forecasts are closely related to the economic cycle (NBER indicator)
- "Our evidence suggests that the usefulness of forecast combining methods ultimately stems from the highly uncertain, complex, and constantly evolving data-generating process underlying expected equity returns, which are related to a similar process in the real economy."

Elliott, Gargano, and Timmermann (JoE, forthcoming):

- Generalizes combination of m univariate models

$$f_{t+1|t} = \frac{1}{m}\sum_{i=1}^{m} x_t \hat{\beta}_i$$

to consider all $n_{k,K}$ $k-$variate models (out of a total of K possible predictors)

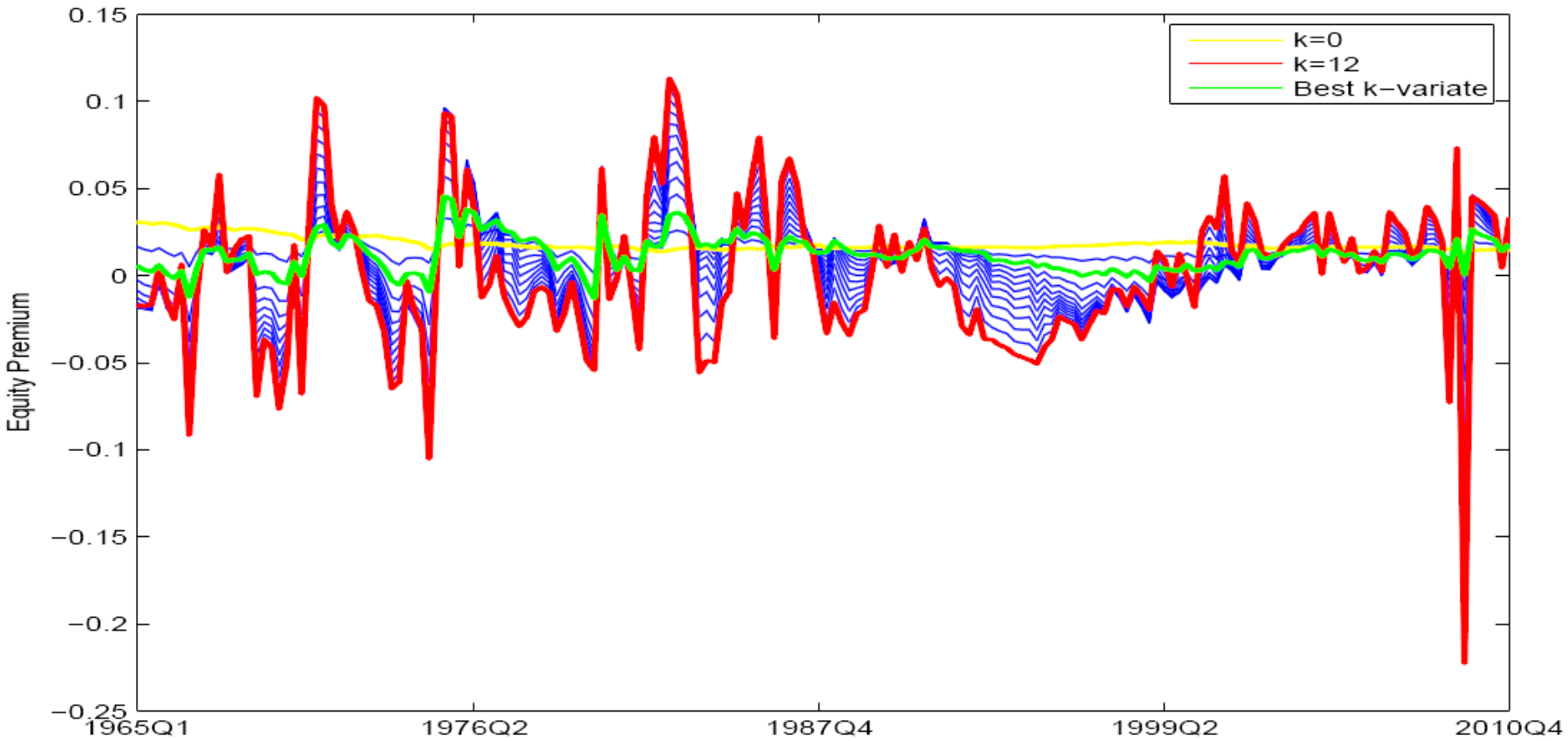
- For fixed K, the estimator for the complete subset regression, $\hat{\beta}_{k,K}$, can be written as

$$\begin{aligned} \hat{\beta}_{k,K} &= \Lambda_{k,K}\hat{\beta}_{OLS} + o_p(1) \\ \Lambda_{k,K} &\equiv \frac{1}{n_{k,K}}\sum_{i=1}^{n_{k,K}} \left(S_i'\Sigma_X S_i\right)^{-} \left(S_i'\Sigma_X\right). \end{aligned}$$

- $S_i : K \times K$ matrix with zeros everywhere except for ones in the diagonal cells corresponding to included variables, zeros for the excluded variables

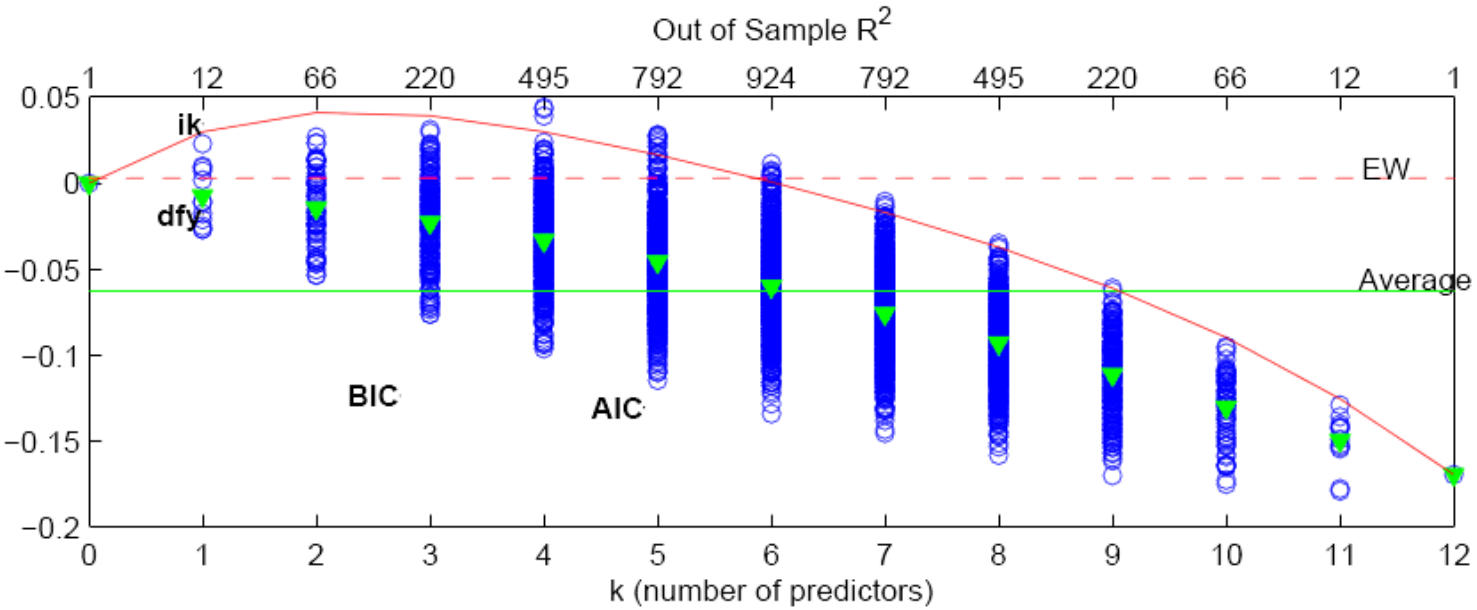
Elliott, Gargano, and Timmermann (JoE, forthcoming)

Figure 5: Out-of-sample forecasts of monthly stock returns for different k-variate subset combinations



Elliott, Gargano, and Timmermann (JoE, forthcoming)

Figure 7: Out-of-sample forecast performance. Each circle represents a single regression model, grouped according to the number of predictors the model contains. For a given value of k, the number of possible k-variate models, $\binom{12}{k} = \frac{12!}{k!(12-k)!}$, is reported on the upper x-axis at the top of the diagram. Triangles represent average values computed across all models with a given number of predictors, k. The horizontal line marked 'Average' shows the performance averaged across all 4096 models while the dotted horizontal line marked 'EW' refers to the performance of the equal-weighted forecast combination based on all models. The full curved line tracks the subset combination of the k-variate models. The best and worst univariate models are displayed as text strings above $k = 1$; AIC and BIC refer to the models recursively selected by these information criteria.



Adaptive combination weights

- Bates and Granger (1969) propose several adaptive estimation schemes
- Rolling window of the forecast models' relative performance over the most recent *win* observations:

$$\hat{\omega}_{i,t|t-h} = \frac{\left(\sum_{\tau=t-win+1}^{t} e_{i,\tau|\tau-h}^2\right)^{-1}}{\sum_{j=1}^{m}\left(\sum_{\tau=t-win+1}^{t} e_{j,\tau|\tau-h}^2\right)^{-1}}$$

- Adaptive updating scheme discounts older performance, $\lambda \in (0;1)$:

$$\hat{\omega}_{i,t|t-h} = \lambda\hat{\omega}_{i,t-1|t-h-1} + (1-\lambda)\frac{\left(\sum_{\tau=t-win+1}^{t} e_{i,\tau|\tau-h}^2\right)^{-1}}{\sum_{j=1}^{m}\left(\sum_{\tau=t-win+1}^{t} e_{j,\tau|\tau-h}^2\right)^{-1}}$$

- The closer to unity is λ, the smoother the combination weights

Time-varying combination weights

- Time-varying parameter (Kalman filter):

$$\begin{aligned} y_{t+1} &= \omega_t' \hat{f}_{t+1|t} + \varepsilon_{t+1} \\ \omega_t &= \omega_{t-1} + u_t, \quad cov(u_t, \varepsilon_{t+1}) = 0 \end{aligned}$$

- Discrete (observed) state switching (Deutsch et al., 1994):

$$y_{t+1} = I_{e_t \in A}(\omega_{01} + \omega_1' \hat{f}_{t+1|t}) + (1 - I_{e_t \in A})(\omega_{02} + \omega_2' \hat{f}_{t+1|t}) + \varepsilon_{t+1}$$

- Regime switching weights (Elliott and Timmermann, 2005):

$$\begin{aligned} y_{t+1} &= \omega_{0s_{t+1}} + \omega_{s_{t+1}}' \hat{f}_{t+1|t} + \varepsilon_{t+1} \\ pr(S_{t+1} &= s_{t+1} | S_t = s_t) = p_{s_{t+1} s_t} \end{aligned}$$

Combinations as a hedge against instability

- Forecast combinations can work well empirically because they provide insurance against model instability
  - Empirically, Elliott and Timmermann (2005) allow for regime switching in combinations of forecasts from surveys and time-series models and find strong evidence that the relative performance of the underlying forecasts changes over time
  - Performance of combined forecasts tends to be more stable than that of individual forecasts used in the empirical combination study of Stock and Watson (2004)
  - Combination methods that attempt to explicitly model time-variations in the combination weights often fail to perform well, suggesting that regime switching or model 'breakdown' can be difficult to predict or even to track through time

Model combination

- When the data underlying the individual forecasts is observed, we can construct forecasts from many different models and average over the resulting forecasts
- For linear combinations, the model average forecast is

$$f^c_{t+h|t} = \sum_{i=1}^{m} \omega_i f_{it+h|t}$$

- $f_{it+h|t}$: individual forecast that depends on some underlying data, z_t
- Same issues as when only the forecasts are observed - but new possibilities like BMA (Bayesian Model Averaging) arise

Classical approach to density combination

- Problem: we do not directly observe the outcome density—we only observe a draw from this—and so cannot directly choose the weights to minimize the loss between this object and the combined density
- Kullback Leibler (KL) loss for a linear combination of densities $\sum_{i=1}^{m} \omega_i p_{it}(y)$ relative to some unknown true density $p(y)$ is given by

$$\begin{aligned} KL &= \int p(y) \ln \left(\frac{p(y)}{\sum_{i=1}^{m} \omega_i p_i(y)} \right) dy \\ &= \int p(y) \ln \left(p(y) \right) dy - \int p(y) \ln \left(\sum_{i=1}^{m} \omega_i p_i(y) \right) dy \\ &= C - E \ln \left(\sum_{i=1}^{m} \omega_i p_i(y) \right) \end{aligned}$$

- C is constant for all choices of the weights ω_i
- Minimizing the KL distance is the same as maximizing the log score in expectation

Classical approach to density combination

- Use of the log score to evaluate the density combination is popular in the literature
- Geweke and Amisano (2011) use this approach to combine GARCH and stochastic volatility models for predicting the density of daily stock returns
- Under the log score criterion, estimation of the combination weights becomes equivalent to maximizing the log likelihood. Given a sequence of observed outcomes $\{y_t\}_{t=1}^{T}$, the sample analog is to maximize

$$\begin{aligned} \hat{\omega} &= \arg\max_{\omega} T^{-1} \sum_{t=1}^{T} \ln \left(\sum_{i=1}^{m} \omega_i p_{it}(y_t) \right) \\ s.t. \;\; \omega_i &\geq 0, \quad \sum_{i=1}^{m} \omega_i = 1 \quad \text{for all } i \end{aligned}$$

Prediction pools with two models (Geweke-Amisano, 2011)

- With two models, M_1, M_2, we have a predictive density

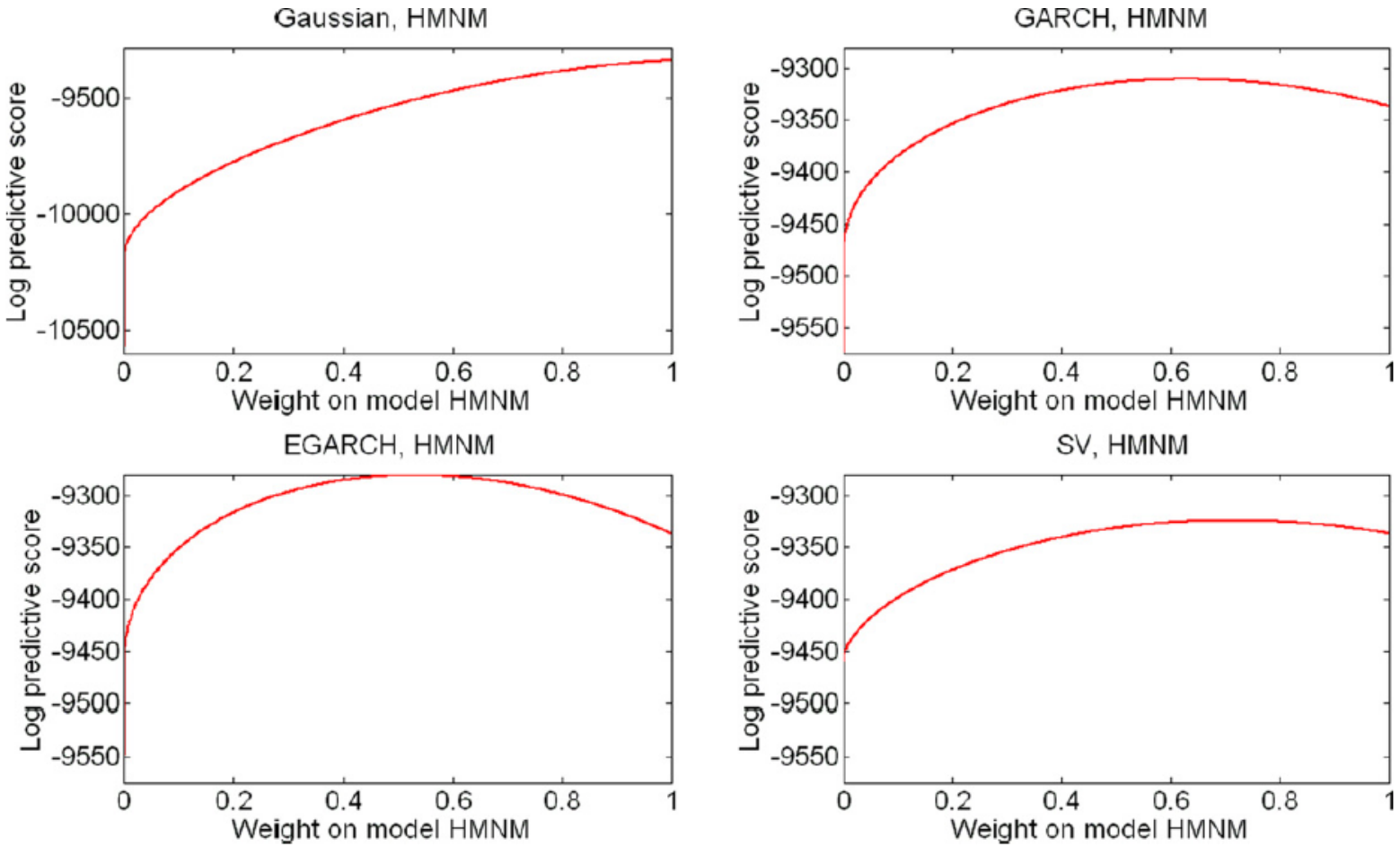
$$p(y_t|Y_{t-1}, M) = \omega p(y_t|Y_{t-1}, M_1) + (1-\omega)p(y_t|Y_{t-1}, M_2)$$

  and a predictive log score

$$\sum_{t=1}^{T} \log\left[wp(y_t|Y_{t-1}, M_1) + (1-w)p(y_t|Y_{t-1}, M_2)\right], \quad \omega \in [0,1]$$

- Empirical example: Combine GARCH and stochastic volatility models for predicting the density of daily stock returns

Log predictive score as a function of model weight, S&P500, 1976-2005 (Geweke-Amisano, 2011)



Weights in pools of multiple models, S&P500, 1976-2005 (Geweke-Amisano, 2011)

Table 3
Optimal pools of 6 and 5 models.

Gaussian	GARCH	EGARCH	*t*-GARCH	SV	HMNM	log score
0.000	0.000	0.319	0.417	0.000	0.264	−9264.83
0.000	0.060	X	0.653	0.000	0.286	−9284.30
0.000	0.000	0.471	X	0.000	0.529	−9280.34
0.000	0.000	0.323	0.677	0.000	X	−9296.08

The first six columns provide the weights for the optimal pools and the last column indicates the log score of the optimal pool. "X" indicates that a model was not included in the pool.

Bayesian Model Averaging (BMA)

$$p^c(y) = \sum_{i=1}^{m} \omega_i p(y|M_i)$$

- m models: $M_1, ..., M_m$
- BMA weights predictive densities by the posteriors of the models, M_i
- BMA is a model averaging procedure rather than a predictive density combination procedure per se
- BMA assumes the availability of both the data underlying each of the densities, $p_i(y) = p(y|M_i)$, and knowledge of how that data is employed to obtain a predictive density
- $p(M_i)$: prior probability that model i is the true model

Bayesian Model Averaging (BMA)

- Posterior probability for model i, given the data, Z, is

$$p(M_i|Z) = \frac{p(Z|M_i)p(M_i)}{\sum_{j=1}^{m} p(Z|M_j)p(M_j)}$$

- The combined model average is then

$$p^c(y) = \sum_{i=1}^{m} p(y|M_i)p(M_i|Z)$$

- Marginal likelihood of model i is

$$P(Z|M_i) = \int P(Z|\theta_i, M_i)P(\theta_i|M_i)d\theta_i$$

- $p(\theta_i|M_i)$: prior density of model i's parameters
- $p(Z|\theta_i, M_i)$: likelihood of the data given the parameters and the model

Constructing BMA estimates

Requirements:

- List of models $M_1, ..., M_m$
- Computation of $p(M_i|Z)$ requires computation of the marginal likelihood $p(Z|M_i)$ which can be time consuming
- Prior model probabilities $p(M_1), ..., p(M_m)$
- Priors for the model parameters $P(\theta_1|M_1), ..., P(\theta_m|M_m)$

Alternative BMA schemes

- Raftery, Madigan and Hoeting (1997) MC^3
- If the models' marginal likelihoods are difficult to compute, one can use a simple approximation based on BIC:

$$\omega_i = P(M_i|Z) \approx \frac{\exp(-0.5BIC_i)}{\sum_{i=1}^{m} \exp(-0.5BIC_i)}$$

- Remove models that appear not to be very good
  - Madigan and Raftery (1994) suggest removing models for which $p(M_i|Z)$ is much smaller than the posterior probability of the best model

Avramov (JFE, 2002)

- $m = 14$ different predictors
- $2^{14} =$ models
- monthly and quarterly stock returns, 1953-1998
- six Fama-French portfolios: size (S,B)$\times$(LMH) book to market

Avramov (JFE, 2002): Results

Table 2
Posterior probabilities of forecasting models based on a prior sample weighted against predictability
The first rows display cumulative posterior probabilities computed as $\mathscr{A}'\mathscr{P}$, where $\mathscr{A}$ is a $2^{14} \times 14$ matrix representing all forecasting models by their unique combinations of zeros and ones and $\mathscr{P}$ is a $2^{14} \times 1$ vector including posterior probabilities for all models. The second rows denote the highest-posterior-probability compositions represented by a combination of zeros and ones designating exclusions and inclusions of predictive variables, respectively. The stock universe comprises six portfolios identified by two letters designating increasing values of size (S,B) and book-to-market (L,M,H). Following are the predictors spanning the information set: dividend yield (Div); book-to-market (BM); earnings yield (EY); the momentum portfolio (WML); the difference in annualized yields of Moody's Baa and Aaa rated bonds (Def); the monthly rate of a three-month Treasury bill (Tbill); excess return on the value-weighted index (RET); the difference between the return on long-term corporate bonds and the return on long-term government bond (DEF); the difference between the monthly return on long-term government bond and the one-month Treasury bill rate (TERM); January Dummy (Jan); inflation rate (Inf); size premium (SMB); value premium (HML); and the difference in annualized yield of ten-year and one-year Treasury bills (Term). Figures displayed below are computed when investors encounter a hypothetical sample weighted against predictability.

Portfolio	Predictive variables													
	Div	BM	EY	WML	Def	Tbill	RET	DEF	TERM	Jan	Inf	SMB	HML	Term
SL	0.20	0.08	0.38	0.02	0.14	0.28	0.48	0.04	0.12	0.21	0.31	0.16	0.05	0.08
	0	0	**1**	0	0	0	0	0	0	0	**1**	0	0	0
SM	0.12	0.06	0.16	0.02	0.10	0.09	0.40	0.06	0.54	0.77	0.15	0.12	0.02	0.19
	0	0	0	0	0	0	0	0	**1**	**1**	0	0	0	0
SH	0.06	0.05	0.07	0.03	0.06	0.04	0.49	0.03	0.35	1.00	0.06	0.08	0.02	0.22
	0	0	0	0	0	0	**1**	0	0	**1**	0	0	0	0
BL	0.12	0.05	0.14	0.04	0.14	0.13	0.05	0.07	0.20	0.03	0.69	0.05	0.05	0.15
	0	0	0	0	0	0	0	0	0	0	**1**	0	0	0
BM	0.15	0.06	0.15	0.03	0.20	0.34	0.03	0.09	0.54	0.07	0.23	0.04	0.03	0.25
	0	0	0	0	0	0	0	0	**1**	0	0	0	0	0
BH	0.07	0.06	0.06	0.03	0.09	0.09	0.02	0.03	0.17	0.92	0.21	0.02	0.02	0.47
	0	0	0	0	0	0	0	0	0	**1**	0	0	0	**1**

Avramov (JFE, 2002): Results

Table 6
Bayesian model averaging: external validity
The table displays several statistics examining the properties of out-of-sample monthly forecast errors generated by several return-generating processes and the weighted forecasting model. The former set includes the all-inclusive model (All), the iid model (iid), and five models selected by adjusted R^2, AIC, SIC, FIC, and PIC, all of which are described by Bossaerts and Hillion (1999). We examine three prior specifications corresponding to a hypothetical sample size equal to 50, 100, and 25 observations per parameter. MPE is the mean forecast error. Efficiency stands for the estimated slope in the regression of forecast errors on predicted one-period-ahead returns. Serial correlation expresses the estimated slope in the regression of current on lagged forecast errors. The quantities t-statistic's are the corresponding statistics testing the equality of the forecast errors, of the correlation between forecast errors and future predicted returns (efficiency), and of serial correlations to zero. MSE is the mean squared error in percent. We adopt two different schemes having distinct asymptotic properties. The *rolling* scheme fixes the estimation window size and drops distant observations as recent ones are added. The *recursive* scheme uses all available data. The bottom part of the table displays mean squared errors for the quarterly sample corresponding to three prior scenarios, in which the number of hypothetical observations is equal to one third of the monthly counterparts, i.e., $T_0 = 17, 33$, and 8.

	$T_0 = 50$	$T_0 = 100$	$T_0 = 25$	All	iid	Adj R^2	AIC	SIC	FIC	PIC
The rolling scheme—monthly sample										
MPE	0.0006	0.0007	0.0003	−0.0006	0.0007	−0.0002	0.0000	−0.0023	0.0001	−0.0003
t-Statistic	0.4225	0.4944	0.2368	−0.3874	0.5126	−0.1551	0.0176	−1.5588	0.0365	−0.2117
Efficiency	−0.0563	−0.0287	−0.2335	−0.7874	−0.4371	−0.7642	−0.7919	−0.9454	−0.8709	−0.7926
t-Statistic	−0.1788	−0.0863	−0.8557	−7.8065	−1.3761	−7.4691	−6.9512	−7.9715	−7.4193	−7.2763
Serial correlation	0.0397	0.0499	0.0323	−0.0284	0.0684	−0.0043	−0.0051	0.0274	−0.0185	−0.0269
t-Statistic	0.6676	0.8326	0.5494	−0.4856	1.1288	−0.0738	−0.0895	0.5024	−0.3167	−0.4659
MSE	0.2137	0.2141	0.2139	0.2333	0.2155	0.2309	0.2298	0.2319	0.2339	0.2312
The recursive scheme—monthly sample										
MPE	−0.0003	−0.0004	−0.0003	0.0005	−0.0010	0.0007	0.0012	0.0013	0.0028	0.0020
t-Statistic	−0.1049	−0.1659	−0.1421	0.1847	−0.3924	0.3018	0.5103	0.5099	1.1455	0.8152
Efficiency	−0.2357	−0.1047	−0.4018	−0.6708	−0.4500	−0.5959	−0.5804	−0.7953	−0.7319	−0.7300
t-Statistic	−0.6675	−0.2572	−1.3455	−3.0407	−0.9504	−2.8158	−2.7292	−3.7994	−3.0175	−2.9242
Serial correlation	0.0401	0.0489	0.0372	0.0036	0.0706	0.0144	0.0143	0.0417	0.0144	0.0120
t-Statistic	0.6728	0.8079	0.6320	0.0655	1.1414	0.2597	0.2572	0.7386	0.2663	0.2194
MSE	0.2133	0.2133	0.2143	0.2231	0.2155	0.2197	0.2189	0.2260	0.2237	0.2239
MSEs for the quarterly sample										
Rolling	0.7546	0.7577	0.7651	0.9333	0.7777	0.9041	0.8629	0.8570	0.9286	0.9347
Recursive	0.7757	0.7678	0.7930	0.8312	0.7781	0.8163	0.8233	0.8952	0.8170	0.8337

Avramov (JFE, 2002): Results

Table 7
Variance decompositions
The table exhibits the marginal contribution of each source of uncertainty about predicted stock returns, i.e., model risk, estimation risk, and uncertainty attributed to forecast errors (denoted For. error), to the overall uncertainty about predicted returns. The variance components are given by

$$\text{var}\{R_{T+1}|D\} = \sum_{j=1}^{2^M} \text{P}(\mathcal{M}_j|D)[\mathbb{E}\{Y_j\} + \text{var}\{\lambda_j\} + (\tilde{\lambda} - \mathbb{E}\{\lambda_j\})(\tilde{\lambda} - \mathbb{E}\{\lambda_j\}')].$$

where R_{T+1} is the next-period excess return, $\text{P}(\mathcal{M}_j|D)$ is the posterior probability of model j, $\mathbb{E}\{Y_i\}$ and $\text{var}\{\lambda_j\}$ are the forecast error and parameter uncertainty components corresponding to model j, respectively. The model uncertainty component is given by $\sum_{j=1}^{2^M} \text{P}(\mathcal{M}_j|D)(\tilde{\lambda} - \mathbb{E}\{\lambda_j\})(\tilde{\lambda} - \mathbb{E}\{\lambda_j\})'$ where $\tilde{\lambda} = \sum_{j=1}^{2^M} \text{P}(\mathcal{M}_j|D)\mathbb{E}\{\lambda_j\}$ is the predicted mean of the next-period excess return that incorporates model uncertainty. The decompositions are performed separately for each of six equity portfolios formed as the intersection of two size (S,B) and three book-to-market (L,M,H) groups, and are presented for both monthly and quarterly samples. For each sample, we examine three specifications of the prior sample size T_0.

Portfolio	*Monthly observations*			*Quarterly observations*		
	Estimation risk	Model risk	For. error	Estimation risk	Model risk	For. Error
	$T_0 = 50$ observations per parameter			$T_0 = 17$ observations per parameter		
SL	0.02	0.05	0.93	0.06	0.09	0.85
SM	0.03	0.08	0.89	0.07	0.11	0.82
SH	0.04	0.02	0.94	0.06	0.10	0.84
BL	0.02	0.01	0.97	0.05	0.06	0.89
BM	0.02	0.02	0.96	0.06	0.10	0.84
BH	0.03	0.03	0.94	0.05	0.11	0.84

Avramov (JFE, 2002): Results

Table 8
Asset allocation and the economic loss of ignoring model uncertainty
The table exhibits asset allocations to six size book-to-market portfolios as percentages of the total invested wealth using three prior scenarios corresponding to a hypothetical prior sample size equal to 25, 50, and 100 observations per parameter. Asset allocations are derived for investment horizons of one, two, four, six, eight, and ten years, for relative risk-aversion coefficient (γ) equal to seven, and for current values of predictive variables (z_T) equal to actual-end-of-sample realizations. We also examine asset allocation when the current values are equal to historical means focusing on $T_0 = 50$. The table exhibits allocation to individual portfolios, total allocation to equities (Total), and a utility loss. Utility loss is computed as the loss in an annual certainty equivalent risk-free rate perceived by investors who are forced to ignore model uncertainty and, instead, allocate funds based upon several return-generating processes. The latter includes the all-inclusive model (All), and models selected by adjusted R^2, AIC, SIC, FIC, and PIC, all of which are described by Bossaerts and Hillion (1999).

Horizon	*Asset allocation*							*Utility loss*					
	SL	SM	SH	BL	BM	BH	Total	All	Adj R^2	AIC	SIC	FIC	PIC
$T_0 = 50$ observations per parameter, and z_T = end-of-sample realizations													
1	0.00	0.00	0.36	0.00	0.00	0.31	0.67	4.37	1.23	1.71	1.71	3.71	2.33
2	0.00	0.00	0.32	0.00	0.00	0.33	0.65	4.07	1.95	2.56	2.61	3.41	3.13
4	0.00	0.00	0.30	0.00	0.00	0.33	0.63	3.90	2.09	2.63	2.65	2.55	2.73
6	0.00	0.00	0.29	0.00	0.00	0.34	0.63	3.00	1.83	2.27	2.25	1.79	2.08
8	0.00	0.00	0.28	0.00	0.00	0.35	0.62	2.23	1.56	1.89	1.91	1.29	1.63
10	0.00	0.00	0.27	0.00	0.00	0.35	0.62	1.75	1.37	1.59	3.71	0.98	1.35

Avramov (JFE, 2002): main results

- BMA forecasts are more robust than individual forecasts, with unbiased and serially uncorrelated forecast errors
- Model uncertainty reduces the strength of the evidence on return predictability
- term and market risk premia appear to be the best predictors of stock returns

Conclusion

- Combinations of forecasts is motivated by
  - misspecified forecasting models due to, e.g., structural breaks
  - diversification across forecasts
  - private information used to compute individual forecasts (surveys)
- Simple, robust estimation schemes tend to work well
  - small samples (estimation errors in combination weights)
- Even if they do not always deliver the most precise forecasts, forecast combinations, particularly equal-weighted ones, generally do not deliver poor performance and so from a "risk" perspective represent a relatively safe choice
- Empirically, survey forecasts work well for many macroeconomic variables, but they tend to be biased and not very precise for stock returns