IMF INSTITUT
FOR CAPACITY DEVELO

Forecast Combinations

Joint Vienna Institute/ IMF ICD

Macro-econometric Forecasting and Analysis

JV16.12, L09, Vienna, Austria, May 24, 2016

Presenter
Mikhail Pranovich

Lecture Objectives

- Introduce the idea and rationale for forecast averaging
- Identify forecast averaging implementation issues
- Become familiar with a number of forecast averaging schemes

Introduction

- Usually, multiple forecasts are available to decision makers
- Differences in forecasts reflect:
  - differences in subjective priors
  - differences in modeling approaches
  - differences in private information
- It is hard to indentify the true DGP
  - should we use a single forecast or an “average” of forecasts?

Introduction

- Disadvantages of using a single forecasting model:
  - may contain misspecifications of an unknown form
    - e.g., some variables are missing
  - one statistical model is unlikely to dominate all its rivals at all points of the forecast horizon
- Combining separate forecasts offers :
  - a simple way of building a complex, more flexible forecasting model to explain the data
  - some insurance against “breaks” or other non-stationarities that may occur in the future

Outline of the lecture

1. What is a combination of forecasts?
2. The theoretical problem and implementation issues
3. Methods to assign weights
4. Improving the estimates of the theoretical model performance
5. Conclusion – Key Takeaways

Part I. What is a combination of forecasts?

- General framework and notation
- The forecast combination problem
- Issues and clarifications

General framework

- Today (at time *T*) we want to forecast the value of y_{T+h} (at *T+h*)
- We have *M* different forecasts:
    - model-based (econometric model, or DSGE), or judgmental (consensus forecasts)
    - the model(s) or judgment(s) are our own or of others
    - some models or information sets might be unknown: only the end product – forecasts – are available
- How to combine *M* forecasts into one forecast?
- Is there any advantage in combining vs. selecting the “best” among the *M* forecasts?

Notation

- y_T is the value of *Y* at time *t* (today is *T*)
- *h* is the forecasting horizon
- $\hat{y}_{T+h,m}$ is an unbiased (point) forecast of y_{T+h} at time *T*
- *m*= 1,…,*M* the indices of the available forecasts/models
- $e_{T+h,m} = y_{T+h} - \hat{y}_{T+h,m}$ is the forecast error of model *m*
- $\sigma_{T+h,m} = Var(e_{T+h,m})$ is the forecast error variance
- $\sigma_{T+h,i,j} = Cov(e_{T+h,i}, e_{T+h,j})$ covariance of forecast errors
- $w_{T,h} = (w_{T,h,1},\ldots, w_{T,h,M})'$ is a vector of weights
- $L(e_{t+h})$ is the loss from making a forecast error
- $E\{L(e_{t+h})\}$ is the risk associated with a forecast

Interpretation of loss function *L(e)*

- Squared error loss (mean squared forecasting error: MSFE)

$$L(e_{T+h}) = (e_{T+h})^2$$

  - equal loss from over/under prediction
  - loss increases quadratically with the error size

- Absolute error loss (mean absolute forecasting error: MAFE)

$$L(e_{T+h}) = | e_{T+h} |$$

  - equal loss from over/under prediction
  - proportional to the error size

- Linex loss (γ>0 controls the aversion against positive errors, γ<0 controls the aversion against negative errors)

$$L(e_{T+h}) = \exp(\gamma e_{T+h}) - \gamma e_{T+h} + 1$$

The forecast combination problem

- A combined forecast is a weighted average of *M* forecasts:

$$\hat{y}^{c}_{T+h} = \sum_{m=1}^{M} w_{T,h,m}\hat{y}_{T+h,m}$$

- The forecast combination problem can be formally stated as:

Problem 1: *Choose weights* $w_{T,h,i}$ *to minimize the loss function* $E[(e^{c}_{T+h})^2]$ *subject to* $\sum_{m=1}^{M} w_{T,h,m} = 1$

See Appendix 1 for generalization

- *<u>Note</u>: Here we assume MSFE-loss, but it could be any other*

Clarification: combining forecasting errors

- Notice that since $\sum_{i=1}^{M} w_{T,h,i} = 1$ then

$$e^{c}_{T+h} = y_{T+h} - \hat{y}^{c}_{T+h} = y_{T+h} \sum_{m=1}^{M} w_{T,h,m} - \sum_{m=1}^{M} w_{T,h,m} \hat{y}_{T+h,m} =$$

$$= \sum_{m=1}^{M} w_{T,h,m} \left(y_{T+h} - \hat{y}_{T+h,m} \right) =$$

$$= \sum_{m=1}^{M} w_{T,h,m} e_{T+h,m}$$

- Hence, if weights sum to one, then the expected loss from the combined forecast error is

$$E\left[L\left(e^{c}_{T+h} \right) \right] = E\left[L\left(\sum_{i=1}^{M} w_{T,h,i} e_{T+h,i} \right) \right]$$

Summary: what is the problem all about? (II)

Problem 1: *Choose weights* $w_{T,h,i}$ *to minimize the loss function* $E[(e^c_{T+h})^2]$ *subject to* $\sum_{i=1}^{M} w_{T,h,i} = 1$

- We want to find optimal weights (the theoretical solution to **Problem 1**)
    - How can we estimate <u>optimal</u> weights from a sample of data?
    - Are these estimates good?

General problem of finding optimal forecast combination

- Let:
  - $\boldsymbol{u}$ an (M x 1) vector of 1's,
  - and $\boldsymbol{\Sigma}$ the (M x M) covariance matrix of the forecast errors $\Sigma = E\{ee'\}$

- It follows that

$$\sum_{m=1}^{M} w_{T,h,m} = u'w, \quad e_{T+h}^{c} = \sum_{m=1}^{M} w_{T,h,m} e_{T+h,m} = w'e, \quad (e_{T+h}^{c})^2 = w'ee'w$$

- For the MSFE loss, the optimal w's are the solution to the problem:

$$E\{(e_{T+h}^{c})^2\} = E\{w'ee'w\} = w'E\{ee'\}w = w'\Sigma w \Rightarrow \min_{w}$$

$$s.t.\ u'w = 1$$

- To find optimal weights it is therefore important to know (or have a "good" estimate) of $\boldsymbol{\Sigma}$

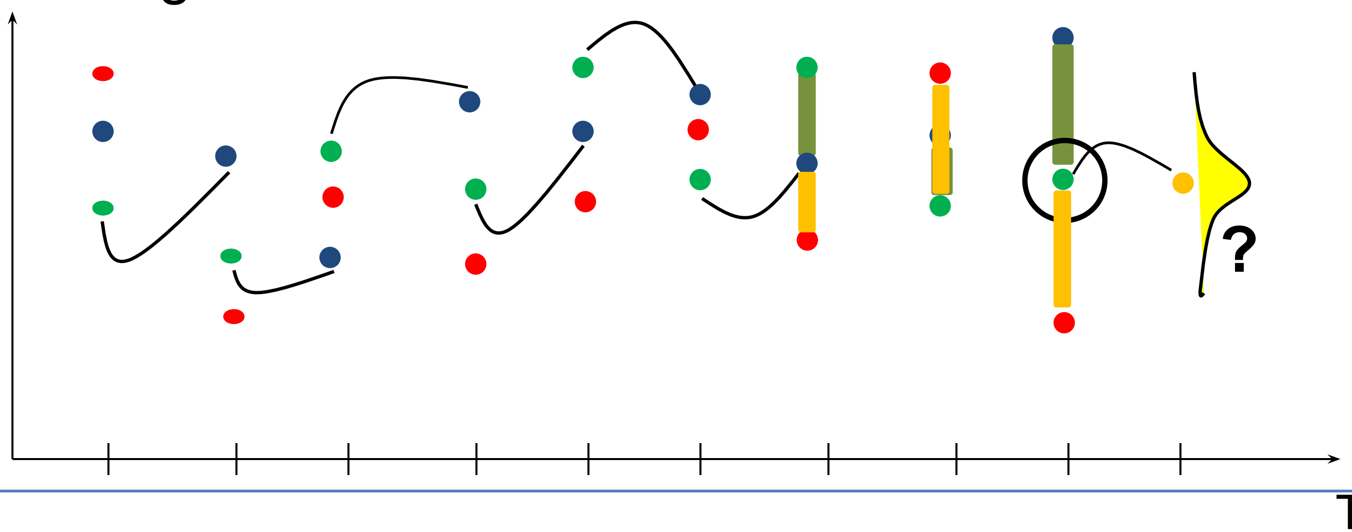
Issues and clarifications

- Do weights have to sum to one?
    - If forecasts are unbiased, this guarantees unbiased combination forecast
- Is there a difference between averaging across *forecasts* and across *forecasting models*?
    - If you *know* the models and the models are *linear* in parameters, there is no difference
- Is it better to combine forecasts rather than information sets?
    - Combining information sets is theoretically better*
    - practically difficult'/impossible: if sets are different, then the joint set may include so many variables that it will not be possible to construct a model that includes all of them

* Clemen (1987) shows that this depends on the extent to which information is common to forecasters

Summary: what is the problem all about? (I)

- Observations of a variable *Y*
- Forecast observations of *Y:*
    - *forecast 1*
    - *…*
    - *forecast M*
- Forecasting errors
- Question: how much weight to assign to each of forecasts, given past performance and knowing that there will be a forecasting error?



Part II. The theoretical problem and implementation issues

- A simple example with only 2 forecasts
- The general N forecast framework
- Issue 1: do weights sum to 1?
- Issue 2: are weights constant over time?
- Issue 3: are estimates of weights good?

Optimal weights in population (M = 2)

- Assume we have 2 unbiased forecasts ($E(e_{T+h,m}) = 0$) and combine:

$$\hat{y}^c_{T+h} = w\hat{y}_{1,T+h} + (1-w)\hat{y}_{2,T+h}$$

$$E\{(e^c_{T+h})^2\} = E\{(we_{T+h,1} + (1-w)e_{T+h,2})^2\} =$$

$$= w^2\sigma^2_{T+h,1} + 2w(1-w)\sigma_{T+h,1,2} + (1-w)^2\sigma^2_{T+h,2} \Rightarrow \min_w$$

Result 1: The solution to **Problem 1** is

$$w = \frac{\sigma^2_{T+h,2} - \sigma_{T+h,1,2}}{\sigma^2_{T+h,1} + \sigma^2_{T+h,2} - 2\sigma_{T+h,1,2}}$$ weight of $\hat{y}_{1,T+h}$

$$1 - w = \frac{\sigma^2_{T+h,1} - \sigma_{T+h,1,2}}{\sigma^2_{T+h,1} + \sigma^2_{T+h,2} - 2\sigma_{T+h,1,2}}$$ weight of $\hat{y}_{2,T+h}$

Interpreting the optimal weights in population

- Consider the ratio of weights

$$\frac{w}{1-w}=\frac{\sigma^2_{T+h,2}-\sigma_{T+h,1,2}}{\sigma^2_{T+h,1}-\sigma_{T+h,1,2}}$$

- A larger weight is assigned to a more precise forecast
- If the covariance of the two forecasts increases, a greater weight goes to a more precise forecast
- The weights are the same (w = 0.5) if and only if $\sigma^2_{T+h,2}=\sigma^2_{T+h,1}$
- This is similar to building a minimum-variance-portfolio (finance)

See Appendix 2: a generalization to *M*>2

Result: Forecast combination reduces error variance

- Compute the expected MSFE with the optimal weights:

$$E[(e^c_{T+h}(w^*))^2] = \frac{\sigma^2_{T+h,1}\sigma^2_{T+h,2}(1-\rho^2_{T+h,1,2})}{\sigma^2_{T+h,1}+\sigma^2_{T+h,2}-2\rho_{T+h,1,2}\sigma_{T+h,1}\sigma_{T+h,2}}$$

$|\rho| \leq 1$ Is the correlation coefficient

- Suppose $\sigma^2_{T+h,1} < \sigma^2_{T+h,2}$ (forecast 1 is more precise), then:

$$E[(e^c_{T+h}(w^*))^2] = \sigma^2_{T+h,1}\frac{\sigma^2_{T+h,2}(1-\rho^2_{T+h,1,2})}{\sigma^2_{T+h,1}+\sigma^2_{T+h,2}-2\rho_{T+h,1,2}\sigma_{T+h,1}\sigma_{T+h,2}} \leq \sigma^2_{T+h,1}$$

(see Appendix 3)

Result 2: $E[(e^c_{T+h}(\hat{w}))^2] \leq \min\{\sigma^2_{T+h,1}, \sigma^2_{T+h,2}\}$

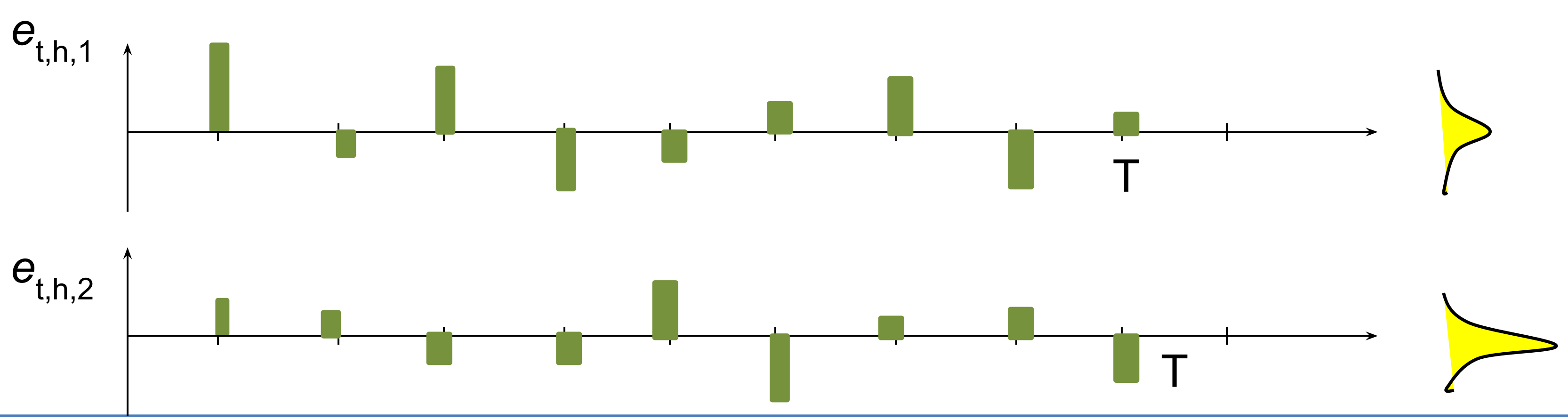
The combined forecast error variance is lower than the smallest of the forecasting error variances of any single model

Estimating Σ

- The key ingredient for finding the optimal weights is the forecast error covariance matrix, e.g. for M=2:

$$\Sigma(e_{T+h,1}, e_{T+h,2}) = \begin{pmatrix} \sigma^2_{T+h,1} & \sigma_{T+h,1,2} \\ \sigma_{T+h,1,2} & \sigma^2_{T+h,2} \end{pmatrix}$$

- In reality, we do <u>not know the exact Σ</u>:
  - we can only estimate $\hat{\Sigma}$ (and then the weights) using past record of forecasting errors



Issues with estimating Σ

- Is the estimate of $\hat{\Sigma}$ based on the past forecasting errors "good"?
- If *forecasting history is short*, then $\hat{\Sigma}$ may be biased
- $\hat{\Sigma}$ may or may not depend on *t* (e.g., a model/forecaster *m* may become better than others over time – smaller $\sigma^2_{T+h,m}$)
  - If not, $\hat{\Sigma}$ converges to Σ as forecasting record lengthens
  - If it does, different issues: heteroskedasticity of any sort, serial correlation, etc.
- If such issues are there, the seemingly "optimal" forecast based on the estimated $\hat{\Sigma}$ might become inferior to other (simpler) combination schemes…

Optimality of equal weights

- The simplest possible averaging scheme uses equal weights
- The equal weights are also optimal weights if:
  - the variances of the forecast errors are the same
  - the pair-wise covariances of forecast errors are the same and equal to zero for $M > 2$
  - the loss function is symmetric, e.g. MSFE:
    - we are not concerned about the sign or the size of forecast errors

Empirical observation: Equal weights tend to perform better than many estimates of the optimal weights (Stock and Watson 2004, Smith and Wallis 2009)

Part III. Methods to estimate the weights: M is small relative to T (M<<T)

To combine or not to combine?

- Assess if one forecast encompasses information in other forecasts
- For MSFE loss, this involves using forecast encompassing tests
- Example: for 2 forecasts, estimate the regression

$$y_{t+h} = \beta_0 + \beta_1 \hat{y}_{t+h,1} + \beta_2 \hat{y}_{t+h,2} + \varepsilon_{t+h} \quad t = 1,2,\ldots T - h$$

- If you cannot reject…

$$H_0 : (\beta_0, \beta_1, \beta_2) = (0,1,0) \quad \rightarrow \text{forecast 1 encompasses 2}$$
$$H_0 : (\beta_0, \beta_1, \beta_2) = (0,0,1) \quad \rightarrow \text{forecast 2 encompasses 1}$$

- … there is no point in combining – use one of the models
- Rejection of H_0 implies that there is information in both forecasts that can be combined to get a better forecast

OLS estimates of the optimal weights

- Recall the general problem of estimating w_m for *m* forecasts (slide 12)
- We can use OLS to estimate the w_m's that minimize the MSFE (Granger and Ramanathan -1984):
  - we use history of past forecasts $\hat{y}_{t+h,m}$ over $t = 1,\ldots,T\!-\!h$ and $m=1,\ldots,M$ to estimate

$$y_{t+h} = w_1 \hat{y}_{t+h,1} + w_2 \hat{y}_{t+h,2} + \ldots + w_M \hat{y}_{t+h,M} + \varepsilon_{t+h}$$

or $$y_{t+h} = w_0 + w_1 \hat{y}_{t+h,1} + w_2 \hat{y}_{t+h,2} + \ldots + w_M \hat{y}_{t+h,M} + \varepsilon_{t+h} \quad s.t. \sum_{m=1}^{M} w_m = 1$$

  - including intercept w_0 takes care of a bias of individual forecasts

Reducing the dependency on sampling errors

- Assume that estimate $\hat{\Sigma}$ is affected by a sampling error (e.g., is biased due to a short forecast record)
- It makes sense to reduce the dependence of the weights on such a (biased) estimate $\hat{\Sigma}$
- Can achieve this by "shrinking" the optimal weights w's towards equal weights $1/M$ (Stock and Watson 2004)

$$w^s_{T,h,i} = \psi w_{T,h,i} + (1-\psi)\frac{1}{M}$$

$$\psi = \max\{0, 1 - kM/(T-h-M-1)\}$$

<u>Notice</u>:

- the parameter *k* determines the strength of the shrinkage
- as *T* increases relative to *M*, the estimated (e.g., OLS) weights become more important: $w^s_{T,h,i} \rightarrow w_{T,h,i}$
- Can you explain why?

Part IV. Methods to estimate the weights: when M is large relative to T

Premise: problems with OLS weights

- The problem with OLS weights:
    - If M is large relative to $T – h$ the OLS estimates loose precision and may not even be feasible (if $M > T – h$)
- Even if M is low relative to $T – h$, the OLS estimates of weights may be subject to a sampling error
    - the estimate $\hat{\Sigma}$ may depend on the sample used
- A number of other methods can be used when M is large relative to T

Relative performance weights

- An alternative to the of OLS weights:
  - ignore the covariance across forecast errors
  - compute weights based on past forecast performance
- For each forecast compute $MSFE_{T,h,m} = \frac{1}{T-h}\sum_{t=1}^{T-h} e_{t+h,m}^2$

MSFE weights (or relative performance weights)

$$w_{T,h,m} = \frac{\frac{1}{MSFE_{T,h,m}}}{\sum_{m=1}^{M} \frac{1}{MSFE_{T,h,m}}}$$

Emphasizing recent performance

- Compute:

$$MSFE_{T,h,m} = \frac{1}{\widetilde{T}} \sum_{t=1}^{T-h} e_{t+h,m}^{2} \delta(t)$$

where $\widetilde{T}$ is the number of periods with $\delta(t)>0$ and $\delta(t)$ can be either

$$\delta(t) = \begin{cases} 1 & \text{if } t \geq T-h-v \\ 0 & \text{if } t < T-h-v \end{cases}$$

Using only a part of forecasting history for forecast evaluation

or

$$\delta(t) = \delta^{T-h-t}$$

Discounted MSFE

Such MSFE weights emphasize the recent forecasting performance

Shrinking relative performance

- Consider instead

$$w_{T,h,m} = \frac{\left(\dfrac{1}{MSFE_{T,h,m}}\right)^k}{\sum_{m=1}^{M}\left(\dfrac{1}{MSFE_{T,h,m}}\right)^k}$$

- If k=1 we obtain standard MSFE weights
- If k=0 we obtain equal weights $1/M$

As parameter $k \to 0$ the relative performance of a particular model becomes less important

Performance weights with correlations

- MSFE weights ignore correlations between forecasting errors
- Ignoring it, when it is present decreases efficiency – larger forecasting variance from the combined forecast
- Consider instead

$$\hat{\Sigma}_{T,m,j} = \frac{1}{T-h}\sum_{t=1}^{T-h} e_{t+h,m} e_{t+h,j}$$

The relative performance weights adjusted for covariance:

$$w_{T,h,m} = \frac{\sum_{j \neq m} \hat{\Sigma}^{-1}_{T,m,j}}{\sum_{j,m} \hat{\Sigma}^{-1}_{T,m,j}}$$

- *<u>Note</u>:* this weighting scheme may be computationally intensive. For M models we need to calculate $M(M+1)/2$ different $\hat{\Sigma}_{T,m,j}$

Rank-based forecast combination

- Aiolfi and Timmerman (2006) allow the weights to be inversely related to the rank of the forecast
- The better is the forecast (e.g., according to MSFE) the higher is the rank r_m
- After all models are ranked form best to worst, the weights are:

$$w_{T,h,m} = \frac{\left(\frac{1}{r_{T,h,m}}\right)}{\sum_{m=1}^{M}\left(\frac{1}{r_{T,h,m}}\right)}$$

Trimming

- In forecast combination, it is often advantageous to discard models with the worst and best performance (i.e., trimming)
  - This is because simple averages are easily distorted by extreme forecasts/forecast errors
  - Trimming justifies the use of the median forecast
- Aiolfi and Favero (2003) recommend ranking the individual models by R^2 and discarding the bottom and top 10 percent.

Example

- Stock and Watson (2003): relative forecasting performance of various forecast combination schemes versus the AR (benchmark)

Table 3. MSFES of Combination Forecasts, Relative to Autoregression: Forecasts of 4-quarter Growth of Real GDP (h = 4)

	Canada	France	Germany	Italy	Japan	U.K.	U.S.
Forecast period	82:I – 98:IV	. .	82:I – 98:IV	82:I – 98:IV	82:I – 98:II	82:I – 98:IV	82:I – 98:IV
Univariate forecasts							
AR RMSE	0.025	.	0.018	0.019	0.023	0.018	0.016
random walk	0.99	.	1.05	1.31	2.97	0.96	1.04
Combination forecasts, full panel							
median	0.92	.	0.99	0.91	0.93	1.00	0.92
mean	0.88	.	1.00	0.82	0.88	0.98	0.90
trimmed mean	0.90	.	0.99	0.82	0.89	0.99	0.91
disc. mse(.9)	0.90	.	0.98	0.85	0.93	0.94	0.90
disc. mse(.95)	0.90	.	1.00	0.89	0.93	0.96	0.89
disc. mse(1)	0.91	.	1.00	0.91	0.92	0.97	0.87
recent best	0.85	.	1.26	0.71	0.97	0.80	1.67

Part V. Improving the Estimates of the Theoretical Model Performance: Knowing the parameters in the model

Question

- So far we assumed that we do not know models from which forecasts originate

- Would our estimates of the weights improve if we knew something about these models

  - e.g., if we knew the number of parameters?

Hansen (2007) approach

- For a process y_t there may be an infinite number of potential explanatory variables $(x_{1t}, x_{2t}, \ldots)$
- In reality we deal with only a finite subset $(x_{1t}, x_{2t}, \ldots, x_{Nt})$
- Consider a sequence of linear forecasting models where model m uses the first k_m variables $(x_{1t}, x_{2t}, \ldots, x_{kt})$:

$$y_{t+h} = \sum_{j=1}^{k_m} \theta_j x_{jt} + b_{t,m} + \varepsilon_{t,m}$$

- with $b_{t,m}$ the approximation error of model *m:*

$$b_{t,m} = \sum_{j=k_m+1}^{\infty} \theta_j x_{jt}$$

- and the forecast given by

$$\hat{y}_{t+h} = \sum_{j=1}^{k_m} \theta_j x_{jt}$$

Hansen (2007) approach (2)

- Let $\hat{\varepsilon}_m$ be the vector of *T-h* (in-sample!) residuals of model *m*
- The {(*T-h*)x*M*} matrix collecting these residuals:

$$E = (\varepsilon_{t,1}, \ldots, \varepsilon_{t,M})$$

- *K* = (*k*$_1$, …, *k*$_M$) is an *M*x1 vector of the number of parameters in each model

- The Mallow criterion is minimized with respect to w

$$C_{T-h}(w) = w' EE' w + 2s^2 K' w \Rightarrow \min_w$$

where s^2 is the largest of all models sample error variance estimator

- The Mallow criterion is an unbiased approximation of the combined forecast MSFE:
  - Minimizing $C_{T\text{-}h}(w)$ delivers optimal weights w
- It is a quadratic optimization problem: numerical algorithms are available (e.g., in GAUSS, QPROG; in Excel, SOLVER)

Example of Hansen's approach (M = 2)

- We need to find w that minimizes the Mallow criterion:

$$(T-h)\left[\begin{pmatrix} w & 1-w \end{pmatrix}\begin{pmatrix} s_1^2 & s_{12} \\ s_{12} & s_2^2 \end{pmatrix}\begin{pmatrix} w \\ 1-w \end{pmatrix}\right] + 2s_2^2\begin{pmatrix} k_1 & k_2 \end{pmatrix}\begin{pmatrix} w \\ 1-w \end{pmatrix}$$

- Minimizing gives:

$$w = \frac{s_2^2((k_2-k_1)/(T-h)) - (s_2^2 - s_{12})}{s_1^2 + s_2^2 - 2s_{12}}$$

- The optimal weights
  - depend on the Var and Cov of residuals
  - penalize the larger model: the weight on the (first) smaller model increases with the size of the "larger" second model $k_2 > k_1$
- See appendix 7 for further details

Conclusions – Key Takeaways

- Combined forecasts imply diversification of risk (provided not all the models suffer from the same misspecification problem)
- Numerous schemes are available to formulate combined forecasts
- For a standard MSFE loss, the payoff from using covariances of errors to derive weights is small
- Simple combination schemes are difficult to beat

Thank You!

References

- Aiolfi, Capistran and Timmerman, 2010, "Forecast Combinations", in *Forecast Handbook*, Oxford, Edited by Michael Clements and David Hendry.
- Clemen, Robert, 1985, "Combining Forecasts: A Review and Annotated Bibliography," *International Journal of Forecasting*, Vol. 5, No. 4, pp. 559–583.
- Stock, James H., and Mark W. Watson, 2004, "Combination Forecasts of Output Growth in a Seven-Country Data Set," *Journal of Forecasting*, Vol. 23, No. 6, pp. 405–430.
- Timmermann, Allan, 2006. "Forecast Combinations," Handbook of Economic Forecasting, Elsevier.

Appendix

Appendix 1: generalization of problem 1

Let **w** be the (M x 1) vector of weights, **e** the (M x 1) vector of forecast errors, **u** an (M x 1) vector of 1s', and **Σ** the (M x M) variance covariance matrix of the errors

$$\mathbf{\Sigma} = E\left[\mathbf{ee}'\right]$$

It follows that

$$\sum_{i=1}^{M} w_{T,h,i} = \mathbf{u}'\mathbf{w} \qquad e_{T+h}^{c} = \sum_{i=1}^{M} w_{T,h,i} e_{T+h,i} = \mathbf{w}'\mathbf{e} \qquad \left(e_{T+h}^{c}\right)^{2} = \mathbf{w}'\mathbf{ee}'\mathbf{w}$$

$$E\left[\left(e_{T+h}^{c}\right)^{2}\right] = E\left[\mathbf{w}'\mathbf{ee}'\mathbf{w}\right] = \mathbf{w}'E\left[\mathbf{ee}'\right]\mathbf{w} = \mathbf{w}'$$

Problem 1: ***Choose w** to minimize **w'Σ w** subject to **u'w** = 1.*

Appendix 2: generalization of result 1

Result 1: Let **u** be an (M x 1) vector of 1s' and $\boldsymbol{\Sigma}_{T,h}$ the variance-covariance matrix of the forecast errors $e_{T,h,i}$. The vector of optimal weights ***w'*** with M forecasts is

$$\mathbf{w}' = \frac{\mathbf{u}'\boldsymbol{\Sigma}_{T,h}^{-1}}{\mathbf{u}'\boldsymbol{\Sigma}_{T,h}^{-1}\mathbf{u}}$$

For the proof and to see how this applies when M = 2 see Appendix 1

Appendix 2: generalization of result 1

Let e be the (M x 1) vector of the forecast errors. Problem 1: choose the vector w to minimize E[**w'ee'w]** subject to **u'w** = 1.

Notice that E[**w'ee'w**] = **w'**E[**ee'**]**w** = **w'Σw**. The Lagrangean is

$$L = \mathbf{w'\Sigma w} - \lambda[\mathbf{u'w} - 1]$$

and the FOC is

$$\mathbf{\Sigma w} - \lambda\mathbf{u} = 0 \Rightarrow \mathbf{w}^* = \mathbf{\Sigma}^{-1}\mathbf{u}\lambda$$

Using **u'w** = 1 one can obtain λ

$$\mathbf{u'w}^* = \mathbf{u'\Sigma}^{-1}\mathbf{u}\lambda = 1 \Rightarrow \lambda = \left[\mathbf{u'\Sigma}^{-1}\mathbf{u}\right]^{-1}$$

Substituting λ back one gives

$$\mathbf{w}^* = \mathbf{\Sigma}^{-1}\mathbf{u}\left[\mathbf{u'\Sigma}^{-1}\mathbf{u}\right]^{-1}$$

Appendix 2: generalization of result 1 (M = 2)

Let $\Sigma_{t,h}$ be the variance-covariance matrix of the forecasting errors

$$\Sigma_{T,h} = \begin{pmatrix} \sigma^2_{T+h,1} & \sigma_{T+h,1,2} \\ \sigma_{T+h,1,2} & \sigma^2_{T+h,2} \end{pmatrix}$$

Consider the inverse of this matrix

$$\Sigma_{T,h}^{-1} = \frac{1}{\det|\Sigma_{T,h}|} \begin{pmatrix} \sigma^2_{T+h,2} & -\sigma_{T+h,1,2} \\ -\sigma_{T+h,1,2} & \sigma^2_{T+h,1} \end{pmatrix}$$

Let **u’** = [1, 1]. The two weights *w** and (*1* - *w**) can be written as

$$\begin{bmatrix} w^* & 1 - w^* \end{bmatrix} = \frac{\boldsymbol{\Sigma}_{\mathbf{T,h}}^{\mathbf{-1}}\mathbf{u}}{\mathbf{u}'\boldsymbol{\Sigma}_{\mathbf{T,h}}^{\mathbf{-1}}\mathbf{u}}$$

Optimal weights in population (M = 2)

- Assume we have 2 unbiased forecasts ($E(e_{T+h,m}) = 0$) and combine:

$$\hat{y}^c_{T+h} = w\hat{y}_{1,T+h} + (1-w)\,\hat{y}_{2,T+h}$$

$$E\{(e^c_{T+h})^2\} = E\{(we_{T+h,1} + (1-w)e_{T+h,2})^2\} =$$

$$= w^2\sigma^2_{T+h,1} + 2w(1-w)\sigma_{T+h,1,2} + (1-w)^2\sigma^2_{T+h,2} \Rightarrow \min_w$$

Result 1: The solution to **Problem 1** is

$$w = \frac{\sigma^2_{T+h,2} - \sigma_{T+h,1,2}}{\sigma^2_{T+h,1} + \sigma^2_{T+h,2} - 2\sigma_{T+h,1,2}} \quad \text{weight of } \hat{y}_{1,T+h}$$

$$1 - w = \frac{\sigma^2_{T+h,1} - \sigma_{T+h,1,2}}{\sigma^2_{T+h,1} + \sigma^2_{T+h,2} - 2\sigma_{T+h,1,2}} \quad \text{weight of } \hat{y}_{2,T+h}$$

Appendix 3

Notice that $\sigma^2_{T+h,1} + \sigma^2_{T+h,2} - 2\sigma_{T+h,1,2} = E\left[(e_{T+h,1} - e_{T+h,2})^2\right] \geq 0$

and that $\sigma^2_{T+h,1}(1 - \rho^2_{T+h,12}) \geq 0$

Need to show that the following inequality holds

$$\frac{\sigma^2_{T+h,2}(1-\rho^2_{T+h,1,2})}{\sigma^2_{T+h,1} + \sigma^2_{T+h,2} - 2\rho_{T+h,1,2}\sigma_{T+h,1}\sigma_{T+h,2}} \leq 1$$

Rearrange the above

$$\sigma^2_{T+h,2}(1-\rho^2_{T+h,1,2}) \leq \sigma^2_{T+h,1} + \sigma^2_{T+h,2} - 2\rho_{T+h,1,2}\sigma_{T+h,1}\sigma_{T+h,2}$$

$$0 \leq \sigma^2_{T+h,1} - 2\rho_{T+h,1,2}\sigma_{T+h,1}\sigma_{T+h,2} + \sigma^2_{T+h,2}\rho^2_{T+h,1,2}$$

$$0 \leq (\sigma_{T+h,1} - \sigma_{T+h,2}\rho_{T+h,12})^2$$

Appendix 4: trading-off bias vs. variance

- The MSFE loss function of *a forecast* has two components:
  - the squared bias of the forecast
  - the (ex-ante) forecast variance

$$E\{(y_{T+h}-\hat{y}_{T+h,m})^2\} = Bias^2_{T+h,m} + \sigma^2_y + Var(\hat{y}_{T+h,m})$$

- Combining forecasts offers a tradeoff: increased overall bias vs. lower (ex-ante) forecast variance

$$E\{(y_{T+h}-\hat{y}_{T+h,m})^2\} = E\left\{\left(\sum_{m=1}^{M} w_{T+h,m} bias_{T+h,m} + \varepsilon_y + \sum_{m=1}^{M} w_{T+h,m}[\hat{y}_{T+h,m} - E\{\hat{y}_{T+h,m}\}]\right)^2\right\} =$$

$$= \sum_{m=1}^{M} w^2_{T+h,m} bias^2_{T+h,m} + \sigma^2_y + \sum_{m=1}^{M} w^2_{T+h,m} Var\{\hat{y}_{T+h,m}\}$$

Appendix 4

The MSFE loss function of *a forecast* has two components:

- the squared bias of the forecast
- the (ex-ante) forecast variance

$$E\{(y_{T+h} - \hat{y}_{T+h,m})^2\} = E\{(E(y_{T+h}) + \varepsilon_y - \hat{y}_{T+h,m})^2\} =$$

$$= E\{(E(y_{T+h}) + \varepsilon_y - E\{\hat{y}_{T+h,m}\} + E\{\hat{y}_{T+h,m}\} - \hat{y}_{T+h,m})^2\} =$$

$$= Bias_m^2 + \sigma_y^2 + Var\{\hat{y}_{T+h,m}\}$$

Appendix 5

Suppose that $x_{T,h,i} = \mathbf{Py}$ where **P** is an (m x T) matrix, **y** is a (T x 1) vector with all y_t , t = 1,…T. Consider:

$$\hat{\sigma}^2_{T,h,i} = \frac{1}{T-h}\sum_{t=1}^{T-h}\left(x_{t,h,i} - y_{t+h}\right)^2$$

$$= \frac{1}{T-h}\sum_{t=1}^{T-h}\left(x_{t,h,i} - E[y_{t+h}] - \varepsilon_{y,t+h}\right)^2$$

$$= \sigma^2_y + \frac{1}{T-h}\sum_{t=1}^{T-h}\left(x_{t,h,i} - E[y_{t+h}]\right)^2 - \frac{2}{T-h}\sum_{t=1}^{T-h}\varepsilon_{y,t+h}\left(x_{t,h,i} - E[y_{t+h}]\right)$$

Appendix 5

Consider:

$$\hat{\sigma}^2_{T,h,i} = \sigma^2_y + \frac{1}{T-h}\sum_{t=1}^{T-h}\left(x_{t,h,i} - E[y_{t+h}]\right)^2 - \frac{2}{T-h}\sum_{t=1}^{T-h}\varepsilon_{y,t+h}\left(x_{t,h,i} - E[y_{t+h}]\right)$$

$$= \ldots - \frac{2}{T-h}\boldsymbol{\varepsilon}'(\mathbf{Py} - \mathbf{E}[\mathbf{y}])$$

$$= \ldots - \frac{2}{T-h}\boldsymbol{\varepsilon}'(\mathbf{PE}[\mathbf{y}] + \mathbf{P}\boldsymbol{\varepsilon} - \mathbf{E}[\mathbf{y}])$$

$$= \ldots - \frac{2}{T-h}\boldsymbol{\varepsilon}'\mathbf{P}\boldsymbol{\varepsilon} - \boldsymbol{\varepsilon}'(\mathbf{I} - \mathbf{P})\mathbf{E}[\mathbf{y}]$$

$$\square\; MSPE_T - \frac{2}{T-h}E\left[\boldsymbol{\varepsilon}'\mathbf{P}\boldsymbol{\varepsilon}\right]$$

Appendix 6: Adaptive weights

- Relative performance weights may be sensitive to adding new forecast errors (may vary wildly)
- We can use an adaptive scheme that updates previous weights by the most recently computed weights
- E.g., for the MSFE weights (can use other weighting too):

$$w_{T,h,m} = \frac{\frac{1}{MSFE_{T,h,m}}}{\sum_{m=1}^{M} \frac{1}{MSFE_{T,h,m}}}$$

$$\breve{w}_{T,h,m} = \alpha \breve{w}_{T-1,h,m} + (1-\alpha) w_{T,h,m}, \quad 0 \leq \alpha \leq 1$$

- The update parameter *α* controls the degree of weights update from period *T*-1 to period *T*

Appendix 7: Example of Hansen’s approach (M = 2)

If the covariance term is zero the weight becomes

$$w = \frac{s_2^2((k_2 - k_1)/(T-h)) + s_2^2}{s_1^2 + s_2^2}$$

The Mallow criterion has a preference for smaller models, and models with smaller variance

- If $k_2 = k_1$, the criterion is equivalent to minimizing the variance of the combination fit