

# Forecast combination and model averaging using predictive measures

Jana Eklund and Sune Karlsson  
Stockholm School of Economics

# 1 Introduction

- Combining forecasts robustifies and improves on individual forecasts (Bates & Granger (1969))
- Bayesian model averaging provides a theoretical motivation and performs well in practice (Min & Zellner (1993), Madigan & Raftery (1994), Jacobson & Karlsson (2004))
- BMA based on an in-sample measure of fit, the marginal likelihood
- We suggest the use of an out-of-sample, predictive measure of fit, the predictive likelihood

## 2 Forecast combination using Bayesian model averaging

- $\mathfrak{M} = \{\mathcal{M}_1, \dots, \mathcal{M}_M\}$  a set of possible models under consideration
  - Likelihood function  $L(\mathbf{y} | \theta_i, \mathcal{M}_i)$
  - Prior probability for each model,  $p(\mathcal{M}_i)$
  - Prior distribution of the parameters in each model,  $p(\theta_i | \mathcal{M}_i)$
- Posterior model probabilities

$$p(\mathcal{M}_i | \mathbf{y}) = \frac{m(\mathbf{y} | \mathcal{M}_i) p(\mathcal{M}_i)}{\sum_{j=1}^M m(\mathbf{y} | \mathcal{M}_j) p(\mathcal{M}_j)}$$
$$m(\mathbf{y} | \mathcal{M}_i) = \int L(\mathbf{y} | \theta_i, \mathcal{M}_i) p(\theta_i | \mathcal{M}_i) d\theta_i$$

with  $m(\mathbf{y} | \mathcal{M}_i)$  the *prior predictive density* or *marginal likelihood*

- Model averaged posterior

$$p(\phi|\mathbf{y}) = \sum_{j=1}^M p(\phi|\mathbf{y}, \mathcal{M}_j) p(\mathcal{M}_j|\mathbf{y})$$

for  $\phi$  some function of the parameters

- Accounts for model uncertainty
- In particular

$$\hat{y}_{T+h} = E(y_{T+h}|\mathbf{y}) = \sum_{j=1}^M E(y_{T+h}|\mathbf{y}, \mathcal{M}_j) p(\mathcal{M}_j|\mathbf{y})$$

- Choice of models

- Posterior model probabilities,  $p(\mathcal{M}_i|\mathbf{y})$
- Bayes factor

$$BF_{ij} = \frac{P(\mathcal{M}_i|\mathbf{y})}{P(\mathcal{M}_j|\mathbf{y})} \bigg/ \frac{P(\mathcal{M}_i)}{P(\mathcal{M}_j)} = \frac{m(\mathbf{y}|\mathcal{M}_i)}{m(\mathbf{y}|\mathcal{M}_j)}$$

### 3 The predictive likelihood

- Split the sample  $\mathbf{y} = (y_1, y_2, \dots, y_T)'$  into two parts with  $m$  and  $l$  observations, with  $T = m + l$ .

$$\mathbf{y}_{T \times 1} = \begin{bmatrix} \mathbf{y}_{m \times 1}^* \\ \tilde{\mathbf{y}}_{l \times 1} \end{bmatrix} \begin{array}{l} \text{training sample} \\ \text{hold-out sample} \end{array}$$

- The training sample  $\mathbf{y}^*$  is used to convert the prior into a posterior

$$p(\theta_i | \mathbf{y}^*, \mathcal{M}_i)$$

- Leads to *posterior predictive density* or *predictive likelihood* for the hold-out sample  $\tilde{\mathbf{y}}$

$$p(\tilde{\mathbf{y}} | \mathbf{y}^*, \mathcal{M}_i) = \int_{\theta_i} L(\tilde{\mathbf{y}} | \theta_i, \mathbf{y}^*, \mathcal{M}_i) p(\theta_i | \mathbf{y}^*, \mathcal{M}_i) d\theta_i$$

- *Partial* Bayes factors

$$PBF_{ij} = \frac{p(\tilde{\mathbf{y}} | \mathbf{y}^*, \mathcal{M}_i)}{p(\tilde{\mathbf{y}} | \mathbf{y}^*, \mathcal{M}_j)} = \frac{m(\mathbf{y} | \mathcal{M}_i)}{m(\mathbf{y} | \mathcal{M}_j)} \bigg/ \frac{m(\mathbf{y}^* | \mathcal{M}_i)}{m(\mathbf{y}^* | \mathcal{M}_j)}$$

- Asymptotically consistent model choice requires  $T/m \rightarrow \infty$
- *Predictive* weights for forecast combinations

$$p(\mathcal{M}_i | \tilde{\mathbf{y}}, \mathbf{y}^*) = \frac{p(\tilde{\mathbf{y}} | \mathbf{y}^*, \mathcal{M}_i) p(\mathcal{M}_i)}{\sum_{j=1}^M p(\tilde{\mathbf{y}} | \mathbf{y}^*, \mathcal{M}_j) p(\mathcal{M}_j)}$$

- Can use improper priors on parameters of the models
- Forecast combination is based on weights from predictive likelihood
- Model specific posteriors based on the full sample
- Additional complication: How to choose the size of the training sample,  $m$ , and the hold-out sample,  $l$ ?

### 3.1 Small sample results

- Linear model

$$\mathbf{y} = \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}$$

$$\mathbf{Z} = (\iota, \mathbf{X})$$

- Prior

$$\boldsymbol{\gamma} \sim N\left(\mathbf{0}, c\sigma^2 (\mathbf{Z}'\mathbf{Z})^{-1}\right)$$

$$p(\sigma^2) \propto 1/\sigma^2$$

- The predictive likelihood is given by

$$p(\tilde{\mathbf{y}}) \propto \left(\frac{S^*}{m}\right)^{-l/2} \frac{|\mathbf{M}^*|^{\frac{1}{2}}}{|\mathbf{M}^* + \tilde{\mathbf{Z}}'\tilde{\mathbf{Z}}|^{\frac{1}{2}}} \\ \times \left[ m + \frac{1}{(S^*/m)} (\tilde{\mathbf{y}} - \tilde{\mathbf{Z}}\boldsymbol{\gamma}_1)' \left(\mathbf{I} + \tilde{\mathbf{Z}}(\mathbf{M}^*)^{-1}\tilde{\mathbf{Z}}'\right)^{-1} (\tilde{\mathbf{y}} - \tilde{\mathbf{Z}}\boldsymbol{\gamma}_1) \right]^{-T/2}$$

$$S^* = \frac{c}{c+1} (\mathbf{y}^* - \mathbf{Z}^* \hat{\boldsymbol{\gamma}}^*)' (\mathbf{y}^* - \mathbf{Z}^* \hat{\boldsymbol{\gamma}}^*) + \frac{1}{c+1} \mathbf{y}^{*'} \mathbf{y}^*$$

$$\gamma_1 = \frac{c}{c+1} \hat{\boldsymbol{\gamma}}^*,$$

$$\mathbf{M}^* = \frac{c+1}{c} \mathbf{Z}^{*'} \mathbf{Z}^*$$

- Three components

- In sample fit,  $(S^*/m)^{-l/2}$

- Dimension of the model,  $|\mathbf{M}^*|^{\frac{1}{2}} / |\mathbf{M}^* + \tilde{\mathbf{Z}}' \tilde{\mathbf{Z}}|^{\frac{1}{2}}$

- Out of sample prediction,

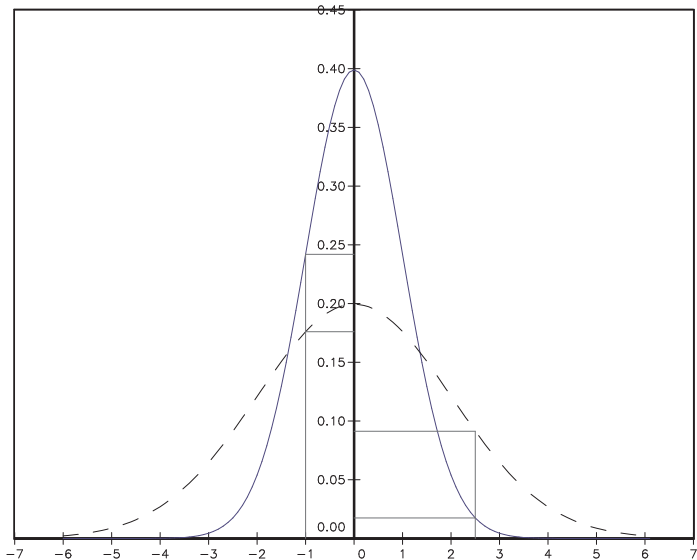
$$\left[ m + \frac{1}{(S^*/m)} (\tilde{\mathbf{y}} - \tilde{\mathbf{Z}} \boldsymbol{\gamma}^*)' (\mathbf{I} + \tilde{\mathbf{Z}} (\mathbf{M}^*)^{-1} \tilde{\mathbf{Z}}')^{-1} (\tilde{\mathbf{y}} - \tilde{\mathbf{Z}} \boldsymbol{\gamma}^*) \right]^{-T/2}$$



---

**Figure 1** Predictive likelihood for models with small and large prediction error variance.

---



## 4 MCMC

- Impossible to include all models in the calculations
  - Reduce the number of models by restricting the maximum number of variables to  $k'$
  - Only consider "good" models
- Use reversible jump MCMC to identify good models
- Exact posterior probabilities calculated conditional on the set of visited models

---

**Algorithm 1** Reversible jump Markov chain Monte Carlo

---

Suppose that the Markov chains is at model  $\mathcal{M}$ , having parameters  $\theta_{\mathcal{M}}$ .

1. Propose a jump from model  $\mathcal{M}$  to a new model  $\mathcal{M}'$  with probability  $j(\mathcal{M}'|\mathcal{M})$ .
2. Accept the proposed model with probability

$$\alpha = \min \left\{ 1, \frac{p(\tilde{\mathbf{y}}|\mathbf{y}, \mathcal{M}') p(\mathcal{M}') j(\mathcal{M}|\mathcal{M}')}{p(\tilde{\mathbf{y}}|\mathbf{y}, \mathcal{M}) p(\mathcal{M}) j(\mathcal{M}'|\mathcal{M})} \right\}$$

3. Set  $\mathcal{M} = \mathcal{M}'$  if the move is accepted otherwise remain at the current model.
-

- Two types of moves
  1. Draw a variable at random and drop it if it is in the model or add it to the model (if  $k_{\mathcal{M}} < k'$ ). This step is attempted with probability  $p_A$ .
  2. Swap a randomly selected variable in the model for a randomly selected variable outside the model (if  $k_{\mathcal{M}} > 0$ ). This step is attempted with probability  $1 - p_A$ .

## 5 Simulation results

- Investigate the effect of the size of the hold-out sample
- Same design as Fernández, Ley & Steel (2001).

– 15 possible predictors,  $\mathbf{x}_1, \dots, \mathbf{x}_{10}$  generated as  $NID(0, 1)$  and

$$(\mathbf{x}_{11}, \dots, \mathbf{x}_{15}) = (\mathbf{x}_1, \dots, \mathbf{x}_5) (0.3, 0.5, 0.7, 0.9, 1.1)' (1, \dots, 1) + \mathbf{e},$$

where  $\mathbf{e}$  are  $NID(0, 1)$  errors.

– Dependent variable

$$y_t = 4 + 2x_{1,t} - x_{5,t} + 1.5x_{7,t} + x_{11,t} + 0.5x_{13,t} + \varepsilon_t,$$

with  $\varepsilon_t \sim N(0, 6.25)$

- $\mathfrak{M}$ -closed view, true model assumed to be part of the model set
- $\mathfrak{M}$ -open view, variables  $x_1$  and  $x_7$  excluded from set of possible predictors
- Three data sets, with last 20 observations set aside for forecast evaluation
  - $T = 120$  (30 years of quarterly data),
  - $T = 250$
  - $T = 400$
  - 100 samples of each sample size

- Prior on models

$$p(M_j) \propto \delta^{k_j} (1 - \delta)^{k' - k_j},$$

where  $k_j$  is the number of variables included in model  $j$ ,  $k' = 15$  and  $\delta = 0.2$ .

- g-prior with  $c = k'^3 = 3375$

- The Markov chain is run for 70 000 replicates, with the first 20 000 draws as burn-in
- Suggests that 70-80% of the data should be kept for the hold-out sample

---

**Table 1** RMSFE for simulated data sets

---

	min for $l$	PL	ML
small data set, $\mathfrak{M}$ -closed	83	2.6333	2.6406
medium data set, $\mathfrak{M}$ -closed	177	2.5064	2.5268
medium data set, $\mathfrak{M}$ -open	182	3.5919	3.6499
large data set, $\mathfrak{M}$ -closed	322	2.5308	2.5310
large data set, $\mathfrak{M}$ -open	302	3.3956	3.4605

---

$\mathfrak{M}$ -closed model :

$$y_t = 4 + 2x_{1,t} - x_{5,t} + 1.5x_{7,t} + x_{11,t} + 0.5x_{13,t} + \sigma\varepsilon_t, \quad (1)$$

with standard deviation 2.5

$\mathfrak{M}$ -open model:

$$y_t | x_{-1,-7} = -1.034x_{2,t} - 1.448x_{3,t} - 1.862x_{4,t} - 3.276x_{5,t} + 1.414x_{11,t} \quad (2) \\ + 0.414x_{12,t} + 0.914x_{13,t} + 0.414x_{14,t} + 0.414x_{15,t}$$

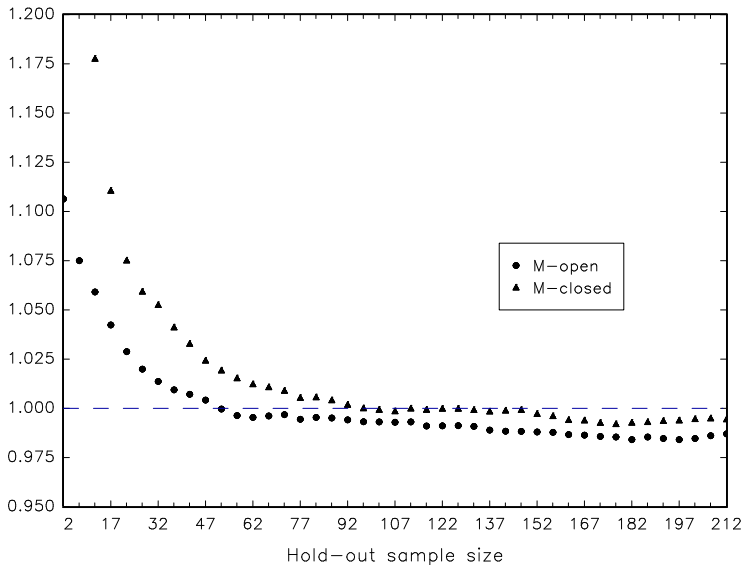
with standard deviation 3.355.



---

**Figure 2** Ratio of RMSFE for predictive likelihood and marginal likelihood as a function of  $l$  for the simulated medium data set.

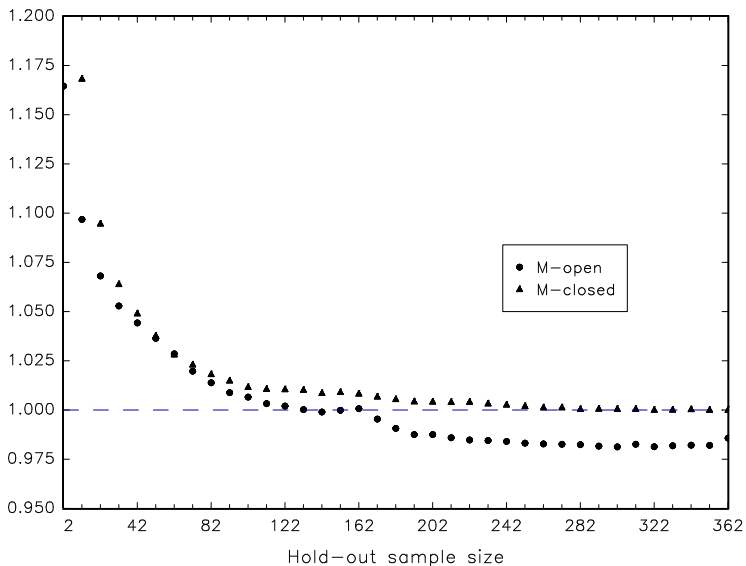
---



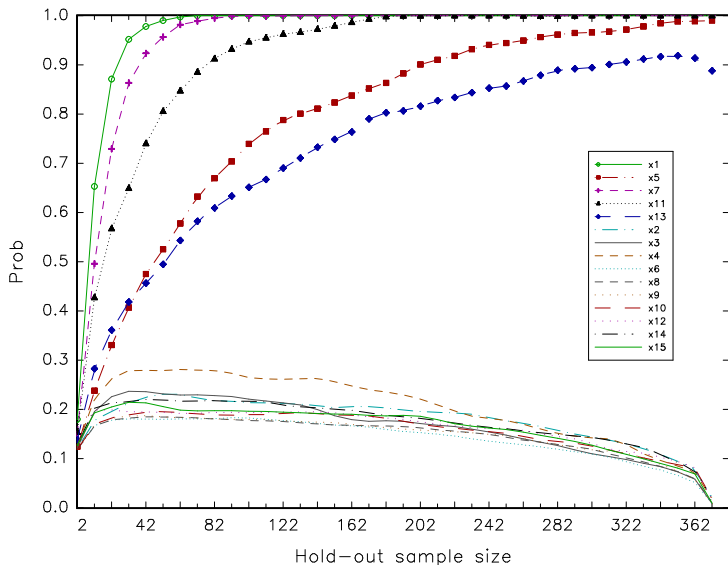
---

**Figure 3** Ratio of RMSFE for predictive likelihood and marginal likelihood as a function of  $l$  for the simulated large data set.

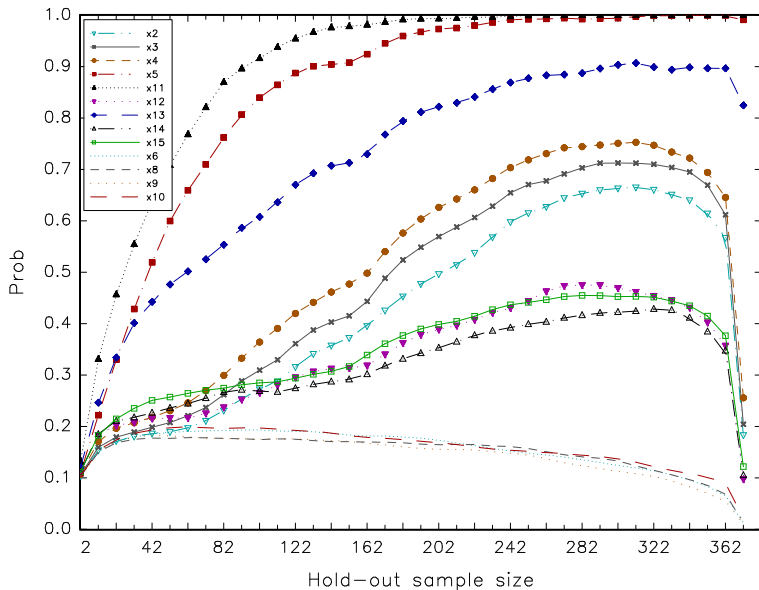
---



**Figure 4** Variable inclusion probabilities (average) for large data set,  $\mathfrak{M}$ -closed view



**Figure 5** Variable inclusion probabilities (average) for large data set,  $\mathfrak{M}$ -open view



## 6 Swedish inflation

- Simple regression model of the form

$$y_{t+h} = \alpha + \omega d_{t+h} + \mathbf{x}_t \beta + \varepsilon_t,$$

- Constant term and a dummy variable,  $d_t$ , for the low inflation regime starting in 1992Q1 always included
- Quarterly data for the period 1983Q1 to 2003Q4 on 77 predictor variables
- Dynamics
  - A preliminary run is used to select,  $\mathbf{x}_t^*$ , the 20 most promising predictors
  - The final run is based on these with one additional lag

$$y_{t+h} = \alpha + \omega d_{t+h} + \mathbf{x}_t^* \beta_1 + \mathbf{x}_{t-1}^* \beta_2 + \varepsilon_t,$$

- 4 quarter ahead forecasts for the period 1999Q1 to 2003Q4

- Maximum of 15 predictors, ( $k' = 15$ ),  $\delta = 0.1$
- $T = 64$ ,  $l = 44$  for the hold-out sample
- 5 000 000 replicates

---

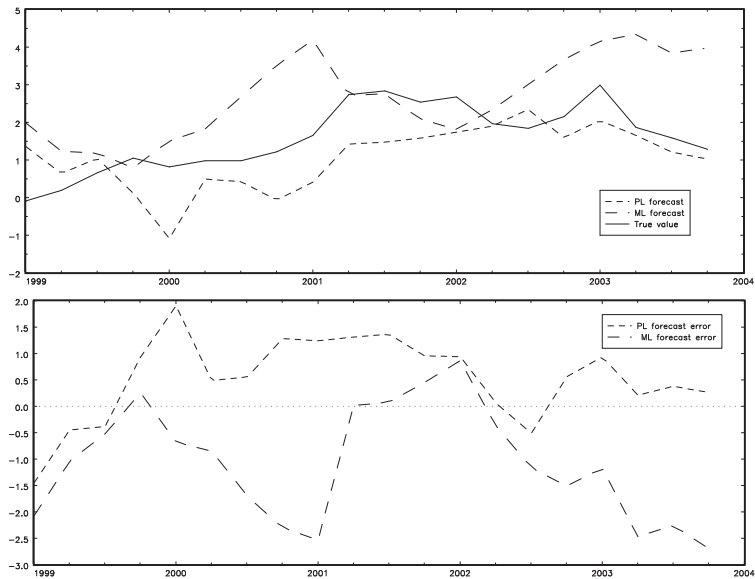
**Table 2** RMSFE of the Swedish inflation 4 quarters ahead forecast, for  $l = 44$ .

---

	<i>Predictive likelihood</i>	<i>Marginal likelihood</i>
Forecast combination	0.9429	1.5177
Top 1	1.0323	1.5376
Top 2	0.9036	1.7574
Top 3	0.9523	1.6438
Top 4	1.0336	1.4828
Top 5	0.9870	2.0382
Top 6	0.9661	1.6441
Top 7	1.0534	1.5755
Top 8	1.1758	1.2905
Top 9	1.0983	1.8356
Top 10	1.0999	1.7202
Random walk	1.0251	1.0251

---

**Figure 6** Swedish data, 4 quarters ahead inflation forecast,  $l = 44$ .





---

**Table 3** Variables with highest posterior inclusion probabilities (average).

---

<i>Predictive likelihood</i>		<i>Marginal likelihood</i>	
Variable	Post. prob.	Variable	Post. prob.
1. Infla	0.5528	Pp1664	0.9994
2. InfRel	0.4493	Pp1529	0.9896
3. U314W	0.3271	InfHWg	0.9456
4. REPO	0.2871	AFGX	0.8104
5. IndProd	0.2459	PpTot	0.4996
6. ExpInf	0.2392	PrvEmp	0.4804
7. R5Y	0.1947	InfCns	0.4513
8. InfFl	0.1749	InfPrd	0.4105
9. MO	0.1533	R3M	0.4048
10. InfUnd	0.1473	Pp75+	0.3927
11. LabFrc	0.1409	ExpInf	0.3829
12. NewHouse	0.1245	InfFor	0.3786
13. InfImpP	0.1225	MO	0.1793
14. PrvEmp	0.1219	POilSEK	0.1702
15. PPP	0.1134	USD	0.1170

**Table 4** Posterior model probabilities, 4 quarters ahead forecast for 1999Q1 using predictive likelihood with  $l = 44$ .

Variable	Model				
	1	2	3	4	5
InfRel	×	×	×		×
InfRel <sub>-1</sub>				×	×
ExpInf	×	×	×	×	×
R5Y	×	×	×		×
InfF1	×	×		×	×
InfF1 <sub>-1</sub>			×		
InfUnd	×	×	×	×	×
USD	×	×	×		×
GDPTCW		×			×
GDPTCW <sub>-1</sub>				×	
Post. Prob	0.0538	0.0301	0.0218	0.0187	0.0184

**Table 5** Posterior model probabilities, 4 quarters ahead forecast for 1999Q1 using marginal likelihood.

Variable	Model				
	1	2	3	4	5
Pp1664	×	×	×	×	×
Pp1529	×	×	×	×	×
InfHWg	×	×	×	×	×
AFGX <sub>-1</sub>	×	×	×		
PpTot	×	×	×		×
PpTot <sub>-1</sub>				×	
R3M <sub>-1</sub>	×	×	×	×	×
InfFor	×				
InfFor <sub>-1</sub>		×			
POilSEK			×		
NewJob <sub>-1</sub>	×	×			
PP2534	×	×			
Post. Prob	0.1316	0.0405	0.0347	0.0264	0.0259

## 7 Conclusions

- The Bayesian approach to forecast combination works well
- The predictive likelihood improves on standard Bayesian model averaging based on the marginal likelihood
- The forecast weights based on predictive likelihood have good large and small sample properties
- Significant improvement when the true model or DGP not included in the set of considered models

## References

- Bates, J. & Granger, C. (1969), ‘The combination of forecasts’, *Operational Research Quarterly* **20**, 451–468.
- Fernández, C., Ley, E. & Steel, M. F. (2001), ‘Benchmark priors for Bayesian model averaging’, *Journal of Econometrics* **100**(2), 381–427.
- Jacobson, T. & Karlsson, S. (2004), ‘Finding good predictors for inflation: A Bayesian model averaging approach’, *Journal of Forecasting* **23**(7), 479–496.
- Madigan, D. & Raftery, A. E. (1994), ‘Model selection and accounting for model uncertainty in graphical models using Occam’s window’, *Journal of the American Statistical Association* **89**(428), 1535–1546.
- Min, C.-K. & Zellner, A. (1993), ‘Bayesian and non-bayesian methods for combining models and forecasts with applications to forecasting international growth rates’, *Journal of Econometrics* **56**(1-2), 89–118.