

Optimizing forecasts for inflation and interest rates by time-series model averaging

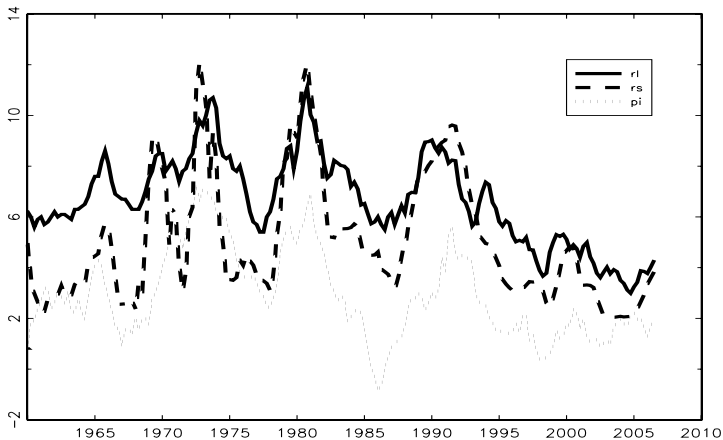
Adusei Jumah and Robert M. Kunst

Institute for Advanced Studies Vienna and University of Vienna

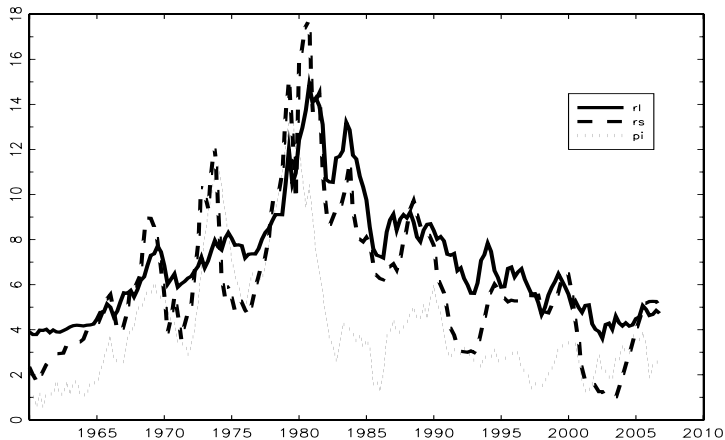
Presented at the ISF'2008, Nice

- 1 Introduction
- 2 The rival prediction models
- 3 Prediction horse race
- 4 Parametric bootstrap validation
- 5 Model averaging
- 6 Summary and conclusion

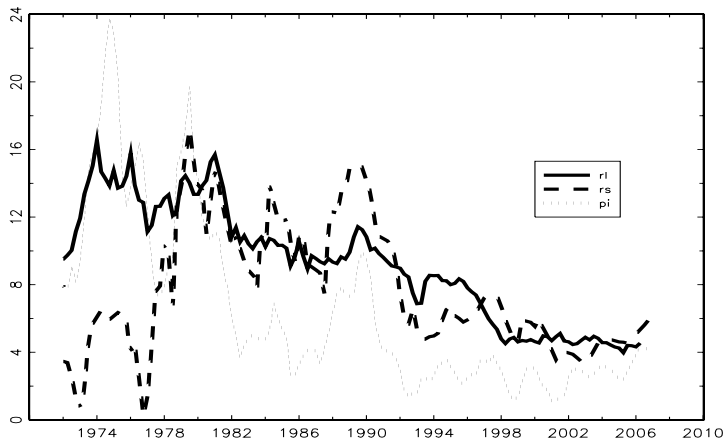
German data



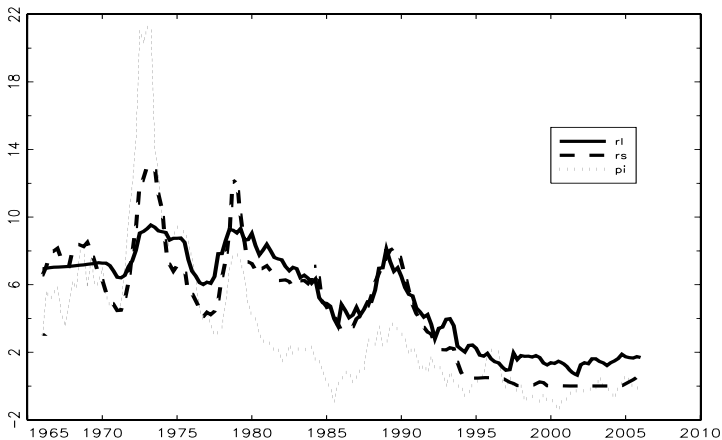
U.S. data



U.K. data



Japanese data



Stylized facts

Bivariate systems of interest rates at different maturity and of systems of rates of inflation plus interest rate have some recognizable or supported properties:

- 1 In the short run, they often behave like random walks or martingales, erratic movements;
- 2 In the longer run, they refute the random-walk model, as they appear bounded;
- 3 There is some evidence on common movement, supported by the Fisher effect and by ideas of constant term premia (CAMPBELL AND SHILLER 1987, HALL, ANDERSON, AND GRANGER 1992);
- 4 In the presented samples, episodes of negative real rates and of inverted term structure are recognizable. It will be difficult to exploit EC equilibria for prediction.

The aim of this study

We want to find out which simple bivariate closed-loop time-series model is best suited for predicting inflation and interest rates in our data sets.

Restricting attention to closed-loop time-series models excludes all potential exogenous influences, dummy variables etc. The framework may permit trivariate structure in a later stage but we keep to bivariate models for simplicity of the designs.

The emphasis is on prediction. We do not intend to establish model validity or to give policy recommendations.

In provocative brevity: Fisher effect and term premium are wonderful economic concepts but: **Are they useful for forecasting?**

The unrestricted vector autoregression

For $y = (r_s, r_l)'$, $y = (\pi, r_j)'$, $j = s, l$, the simple linear dynamic structure **VAR**

$$y_t = \mu + \sum_{j=1}^p \Phi_j y_{t-j} + \varepsilon_t,$$

with a backdrop *n.i.d* $\{\varepsilon_t\}$, has a well-known time-series estimation bias toward stability. It generates mean reversion in both variables.

The differenced vector autoregression

The model **dVAR**

$$\Delta y_t = \mu + \sum_{j=1}^p \Phi_j \Delta y_{t-j} + \varepsilon_t$$

with ‘two unit roots’ matches the short-run I(1) behavior in both components but imposes no equilibrium among them.

The error-correction VAR

The model **EC-VAR**

$$\Delta y_t = \mu + \alpha_1 \beta_1' y_{t-1} + \sum_{j=1}^p \Phi_j \Delta y_{t-j} + \varepsilon_t$$

with 'one unit root' matches the short-run $I(1)$ behavior and it also imposes cointegration. We only consider pre-defined vectors of differences $\beta_1' = (1, -1)$ (Fisher effect or stable yield spread).

The threshold error-correction VAR

The model **th-VAR**

$$\Delta y_t = \mu + \alpha_1 \beta_1' y_{t-1} + \alpha_2 \beta_2' y_{t-1} I(|\beta_2' y_{t-1} - \eta| > c) + \sum_{j=1}^p \Phi_j \Delta y_{t-j} + \varepsilon_t$$

is globally stable but behaves like the EC-VAR for ‘regular’ values of $\beta_2' y$. We consider $\beta_2 = (1, 0)'$ and $(0, 1)'$ only.

Procedures for comparing forecast accuracy

- 1 Procedures not using a test sample
 - (a) Procedures not tuned to the task of forecasting
 - Restriction tests (nested searches)
 - Specification tests (search valid models)
 - BIC etc. (search true models)
 - (b) Procedures tuned to the task of forecasting
 - AIC etc. searches
 - **Model averaging weights**
- 2 Procedures using a test sample
 - **Horse races**
 - **Parametric bootstrap validation**

Rival models are non-nested

$$dVAR \subset EC - VAR \subset VAR \\ \cap \\ th - VAR$$

Nested pairs can be tested by hypothesis tests (multivariate unit-root tests by JOHANSEN for the sequence in the first line). Univariate unit-root tests on variables and on $r_j - \pi$, $j = s, l$, $r_l - r_s$ can also be used. In general, statistical evidence shows a preference for unit roots, although many statistics are close to the significance boundary. Not decisive, heterogeneous across countries, of doubtful value for predictability.

A typical horse race design

- Samples are split into a training and a test part, the share of the test part not too large. We use 40 out of around 190 as a maximum;
- Single-step and multi-step predictions based on each of the candidate models for observations $s = n - 39, \dots, n$. Models fitted on $t = 1, \dots, s - 1$;
- Summarize performance by statistics such as MSE or MAE.

The results for one-step forecasts

		dVAR	VAR	EC-VAR	th-VAR
Two interest rates					
Germany	r_l	0.1041	0.2393	0.0962	0.1001
	r_s	0.0677	0.1492	0.0878	0.1186
United States	r_l	0.1588	0.2303	0.1218	0.1314
	r_s	0.3024	0.5997	0.2216	0.2018
United Kingdom	r_l	0.0592	0.2355	0.0625	0.0844
	r_s	0.1860	0.4132	0.1607	0.2138
Japan	r_l	0.0604	0.1063	0.0605	0.0580
	r_s	0.0143	0.0283	0.0462	0.0584
Inflation and short rate					
Germany	π	0.5633	0.1900	0.4205	0.4130
	r_s	0.1420	0.2507	0.1598	0.2050
United States	π	0.4711	0.3124	0.9236	0.9045
	r_s	0.2500	0.6492	0.2074	0.2274
United Kingdom	π	1.9613	0.6464	0.6960	0.5944
	r_s	0.3430	0.6324	0.0770	0.4296
Japan	π	1.3802	0.3205	0.4396	0.3941
	r_s	0.0192	0.0572	0.0406	0.0490
Inflation and long rate					
Germany	π	0.5520	0.1800	0.4055	0.3990
	r_l	0.1102	0.2570	0.1275	0.1597
United States	π	0.4406	0.3702	0.8764	0.8108
	r_l	0.1213	0.2425	0.1242	0.1332
United Kingdom	π	1.9387	0.3300	0.5751	0.5568
	r_l	0.0634	0.2210	0.0860	0.0866
Japan	π	1.2986	0.3505	0.4265	0.3887
	r_l	0.0684	0.1182	0.0715	0.0674

Summary of the horse races

- VAR, dVAR, and th-VAR have comparable performance for one-step forecasts. EC-VAR comes out worse;
- For longer prediction horizons, dVAR and EC-VAR are gaining ground;
- th-VAR has problems for larger horizons, due to its nonlinearity. It requires stochastic forecasts and averaging.

The idea of the technique

- 1 Assuming a specific model class as the DGP, estimate the corresponding parameter vector for the given sample;
- 2 Generate artificial samples of comparable length, with random errors;
- 3 Use any of the considered model classes to predict the pseudo-samples;
- 4 Repeat this exercise for each considered model as the DGP.

See also JUMAH AND KUNST (forthcoming, *Journal of Forecasting*)

What we may learn from the bootstrap validation

- Prediction models that predict well for any given DGP may be preferred;
- Prediction models that do not even dominate their own DGP are not likely to be good prediction models in practice;
- If the overall impression of the simulated evaluation and the data evaluation are similar, the assumed DGP may be a plausible candidate for the true model.

A summary of the results

DGP	Prediction model			
	VAR	dVAR	EC-VAR	th-VAR
VAR	0.296	0.277	0.236	0.190
dVAR	0.280	0.279	0.261	0.180
EC-VAR	0.309	0.321	0.289	0.072
th-VAR	0.292	0.291	0.251	0.165

Empirical frequency of the respective prediction model delivering the best forecast. Average over all countries and series, 96,000 cases in total for each DGP.

Interpretation of the results

- Models do not generally dominate their own home grounds;
- VAR and dVAR show the best performance;
- Except for the EC-VAR DGPs, relative performance of prediction models is largely independent of the DGP;
- EC-VAR DGPs tend to become unstable in some situations while th-VAR DGPs are OK. Fisher effects etc. tend to be violated in the tails.

The idea of model averaging

- 1 Instead of restricting attention to binary model selection, weighted averages of model forecasts may be considered that may outperform pure strategies;
- 2 Large weights on specific models may indicate that this model class is preferred if the researchers wishes to use pure strategies only.

Model averaging based on HANSEN, 2007

The method was developed for regression models with dependent variable Y . Minimize in $W = (w_1, \dots, w_m)'$

$$C_n(W) = (Y - X\hat{\Theta})'(Y - X\hat{\Theta}) + 2\hat{\sigma}^2 k(W),$$

where the $n \times K$ -matrix X comprises the K explanatory variables, and $\hat{\Theta}$ is the weighted average with weights w_m on model $m = 1, \dots, M$ of least-squares estimates of the coefficient Θ under model m . $k(W) = \sum_{m=1}^M w_m k_m$ is the essential dimension of the considered combination of models with parameter dimension k_m . Here, $M = 4$.

Technicalities and caveats

- 1 In contrast to familiar automatized weighting of candidate models, C_n must be minimized. We use a grid search;
- 2 HANSEN shows that the Mallows criterion outperforms other information criteria in his simulations;
- 3 The procedure has not been shown to work for non-nested candidates nor for $M \neq K$ nor time-series models and it is unlikely to be optimal (or consistent) for unit-root decisions.

The result of the in-sample Mallows weighting

		dVAR	VAR	EC-VAR	threshold VAR
Two interest rates					
Germany	r_l	0.49	0.49	0.02	0
	r_s	0.37	0.63	0	0
U.S.A.	r_l	0.32	0.62	0.06	0
	r_s	0	0.97	0.03	0
U.K.	r_l	0.67	0.33	0	0
	r_s	0.26	0.74	0	0
Japan	r_l	1.00	0	0	0
	r_s	0.35	0.65	0	0
Inflation and short interest rate					
Germany	π	0.10	0.90	0	0
	r_s	0.20	0.79	0.01	0
USA	π	0.47	0.53	0	0
	r_s	0.31	0.69	0	0
UK	π	0.62	0.38	0	0
	r_s	0.36	0.64	0	0
Japan	π	0.23	0.77	0	0
	r_s	0.22	0.77	0	0.01
Inflation and long interest rate					
Germany	π	1.00	0	0	0
	r_l	0.51	0.49	0	0
USA	π	0.59	0.40	0.01	0
	r_l	0.18	0.82	0	0
UK	π	0.33	0.67	0	0
	r_l	0.33	0.65	0	0.02
Japan	π	0.63	0.37	0	0
	r_l	0.63	0.37	0	0

Summary of the weighting results

- 1 The procedure often sets weights to zero, as negative weights are not allowed;
- 2 Often dVAR or VAR are selected exclusively;
- 3 EC-VAR and th-VAR get zero weights usually;
- 4 The weighting does not favor th-VAR due to its larger dimension.

Forecasting using the selected weights

		dVAR	VAR	EC-VAR	th-VAR	Mallows
Two interest rates						
Germany	r_f	0.1041	0.2393	0.0962	0.1186	0.2298
	r_s	0.0977	0.1492	0.0878	0.1186	0.1670
United States	r_f	0.1588	0.2303	0.1218	0.1174	0.2218
	r_s	0.3024	0.5997	0.2216	0.2018	0.5859
United Kingdom	r_f	0.0552	0.2355	0.0625	0.0844	0.1912
	r_s	0.1860	0.4132	0.1667	0.2138	0.4220
Japan	r_f	0.0604	0.1063	0.0605	0.0584	0.1099
	r_s	0.0148	0.0283	0.0462	0.0584	0.0219
Inflation and short rate						
Germany	π	0.5633	0.1900	0.4205	0.4130	0.1906
	r_s	0.1170	0.2507	0.1598	0.2050	0.2620
United States	π	0.4711	0.1424	0.9236	0.9045	0.3501
	r_s	0.2500	0.6492	0.2074	0.2274	0.6738
United Kingdom	π	1.9613	0.6464	0.6960	0.5944	0.6250
	r_s	0.3430	0.6324	0.2270	0.4296	0.6266
Japan	π	1.3802	0.2750	0.4396	0.3941	0.2764
	r_s	0.0170	0.0572	0.0406	0.0490	0.0607
Inflation and long rate						
Germany	π	0.5520	0.1100	0.4055	0.3990	0.1846
	r_f	0.1102	0.2570	0.1275	0.1597	0.2334
United States	π	0.4406	0.1792	0.8764	0.8108	0.3834
	r_f	0.1213	0.2425	0.1242	0.1170	0.2349
United Kingdom	π	1.9387	0.5320	0.5751	0.5568	0.5302
	r_f	0.0944	0.2210	0.0860	0.0866	0.2127
Japan	π	1.2986	0.3565	0.4265	0.3887	0.2174
	r_f	0.0684	0.1182	0.0715	0.0972	0.1185

Results of weighted forecasts

For most cases, the weighted forecasts are unable to outperform pure strategies. In-sample model selection remains inferior to selection based on test samples.

Summary of the candidates' performance

- The simple VAR is maybe the best predictor for one-step forecasts;
- The dVAR is almost as good—fails for inflation that calms down after inflationary episodes—and even improves for larger horizons, even if it is 'misspecified';
- The EC-VAR is a valuable economic model with a firm basis in the literature (Fisher effect etc.) but its prediction performance is poor. There may be stability problems in estimated structures (error-enhancing instead of error-correcting);
- The th-VAR is an interesting outsider that occasionally captures well the overall mean reversion in the variables in unusual regimes (remember inflation targets such as 2%). It loses ground for longer horizons.

Some tentative general conclusions

- It is once more confirmed that simple structures often outperform complex structures even if they are severely misspecified—when it comes to forecasting;
- Methods not using a test sample often deliver less adequate prediction models than methods using the training-test sample split;
- The tuning of model-averaging weights to multi-step forecasting may be an issue for further research.

Thank you for your attention