Cross validation of prediction models for seasonal time series by parametric bootstrapping

Robert M. Kunst
robert.kunst@univie.ac.at

University of Vienna
and
Institute for Advanced Studies Vienna

Prepared for the Statistiktage in Linz, Austria
September 18–20, 2007
The forecaster’s problem

Does (or to what degree does) in-sample evidence yield a reliable recommendation on relative out-of-sample prediction performance of candidate model classes?

Asymptotically, the model selected by information criteria defines the best prediction model (Shibata, 1980). Prediction performance itself defines a consistent information criterion (Wei, 1992). Inoue and Kilian (2006) aim at optimizing prediction and argue that model selection by information criteria may be preferable to direct optimization of prediction but their entire setup is asymptotical. In a small sample, anything can happen.
An artificial example highlights the issues

Data are generated by a fourth-order autoregression

\[ X_t = 0.5X_{t-1} + 0.25X_{t-2} + 0.125X_{t-3} + 0.0625X_{t-4} + \varepsilon_t, \]

with i.i.d. N(0,1) errors \( \varepsilon_t \). Similar models were studied by McQuarrie and Tsai (1998).
Some tools for studying prediction performance

1. Out-of-sample prediction evaluation for the end of the sample (OOS)
2. Significance tests on forecasting performance
3. Cross validation
4. Simulating the models (parametric bootstrapping)
OOS for the last ten observations

![Graph showing root MSE vs. number of observations for different AR models.](image-url)
Cross validation for the artificial data

Cross validation of prediction models for seasonal time series by parametric bootstrapping

Robert M. Kunst  robert.kunst@univie.ac.at
University of Vienna and Institute for Advanced Studies Vienna

Cross validation of prediction models for seasonal time series by parametric bootstrapping
Parametric bootstrapping of artificial data + OOS

Cross validation of prediction models for seasonal time series by parametric bootstrapping
How are these curves constructed?

1. Trajectories of length \( n = 50 \) are drawn from the AR(4) model.
2. AR\((p)\) models are fitted to these trajectories.
3. Trajectories of length \( n = 50 + 10 \) from the fitted models are simulated and predicted in an OOS style. Last ten observations are kept for the OOS evaluation.
What do we learn from OOS and CV experiments?

- 1 lag is best for very small samples, 2 lags are best for $n < 80$, 3 lags dominate until $n = 200$, when the true specification takes over.
- OOS and CV yield similar conclusions, which may not be true for non-stationary generation processes, when OOS may be preferable, as it emphasizes the final part of the sample.
What do we learn from the bootstrapping experiments?

For the sample size of $n = 50$, parsimonious structures dominate at least at larger step sizes, even if higher-order models are true. Models do not necessarily win their own contest.
A real-world data example

A ten-country panel on quarterly barley prices for several decades. Four countries are shown. There are long random-walk swings and there is substantial seasonal variation.
Specification of candidate prediction models

Series correspond to first-order integrated models with added seasonal variation. Obvious candidate is a vector autoregression on differences.

- Degrees of freedom rule out autoregression orders larger than one.
- Cointegration fails, possibly due to an ongoing convergence process across Europe.
- Exogenous variables (temperature etc.) may improve prediction but are not in focus here.
- Long swings may be unpredictable. Focus on short-run prediction, in particular on the seasonal portion.
The two candidate models

Model A is a first-order vector autoregression with a deterministic seasonal cycle

\[ Y_t = \mu + \mathbf{A} \mathbf{D}_t + \Phi Y_{t-1} + \varepsilon_t, \]

where \( \mathbf{D}_t \) contains the seasonal waves \( \cos \pi t, \cos \pi t/2, \sin \pi t/2 \).

Model B restricts the rank of the coefficient matrix \( \mathbf{A} \) to one. This expresses the synchronicity of seasonal cycles across individuals (countries):

\[ Y_t = \mu + \mathbf{a} \mathbf{b}' \mathbf{D}_t + \Phi Y_{t-1} + \varepsilon_t, \]

with 10–vector \( \mathbf{a} \) and 3–vector \( \mathbf{b} \).
Model B is nested in model A, so standard hypothesis tests might be considered for in-sample model selection. The rank restriction is reflected in a $\chi^2$–distribution for the test statistic under model B. The implied $p$–value for rolling samples can be plotted against time. Forecasters may be interested in the right part of the plot.
In the early part, the rank restriction is accepted. In the right part (from 1992), the rank restriction is rejected. This result tends to discourage usage of the restricted model B for prediction.
The rejected model forecasts better

Table: Out-of-sample forecasting performance for the barley price series

<table>
<thead>
<tr>
<th></th>
<th>unrestricted model A</th>
<th>restricted model B</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>RMSE</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>core series</td>
<td>0.0446</td>
<td>0.0433</td>
</tr>
<tr>
<td>added series</td>
<td>0.0482</td>
<td>0.0421</td>
</tr>
<tr>
<td><strong>MAE</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>core series</td>
<td>0.0374</td>
<td>0.0365</td>
</tr>
<tr>
<td>added series</td>
<td>0.0385</td>
<td>0.0341</td>
</tr>
</tbody>
</table>

Out-of-sample one-step forecasts for last ten observations (40 data points added series, 44 data points core series after dropping 16 extrapolated data points). Core series are Belgium, Denmark, Spain, France, Netherlands, and the United Kingdom; added series are Austria, Germany, Finland, and Sweden. RMSE and MAE denote root mean squared errors and mean absolute errors.
Cross validation yields a less clear picture

Table: Cross validation for the barley price series

<table>
<thead>
<tr>
<th></th>
<th>unrestricted model A</th>
<th>restricted model B</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>RMSE</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>core series</td>
<td>0.0330</td>
<td>0.0366</td>
</tr>
<tr>
<td>added series</td>
<td>0.0314</td>
<td>0.0286</td>
</tr>
<tr>
<td><strong>MAE</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>core series</td>
<td>0.0248</td>
<td>0.0277</td>
</tr>
<tr>
<td>added series</td>
<td>0.0239</td>
<td>0.0217</td>
</tr>
</tbody>
</table>

Core series are Belgium, Denmark, Spain, France, Netherlands, and the United Kingdom; added series are Austria, Germany, Finland, and Sweden. RMSE and MAE denote root mean squared errors and mean absolute errors.
Simulations are started from actual observations at a time point $t$ toward the end of the sample, where all ten countries are observed. $t$ is varied over four quarters of a year.

For each $t$, 20,000 pseudo-samples are generated with sample size $n = 48, 100, 200, 500$. This yields 80,000 replications.

All pseudo-samples are predicted using model A and B.

The simulation is sensitive to the variance of $\varepsilon$ (relative to the seasonal cycle in $A$).
The restricted model B generates the data

The true model dominates for $h = 1$ and also at larger horizons but less so, as sample size increases. Long blue dashes for $n = 48$, curves above for $n = 100, 200, 500$. 
The unrestricted model A generates the data

The incorrect, restricted model dominates for most $h$ but dependence on the seasonal cycle is obvious. There is little change as sample size increases. Long blue dashes for $n = 48$, other curves for $n = 100, 200, 500$. 
What do we learn?

- The bootstrap answers the question what would happen if the data were really generated by models A/B and if I used models A/B for forecasting (4 cases).
- Given the data, model B should even be used when it is in fact incorrect.
- The experiment shows that the features are systematic and not artifacts of a few data points.
- By focusing on the last part of the sample, forecasting experiments may be more informative than traditional model selection, in the presence of non-stationarity.
- Any simulation of prediction models helps to identify instability or infeasibility of the stochastic processes.

Robert M. Kunst  robert.kunst@univie.ac.at
University of Vienna and Institute for Advanced Studies Vienna
Cross validation of prediction models for seasonal time series by parametric bootstrapping
The aim of all modelling is to decompose the data information into systematic features and noise.

Forecasting experiments are tuned at isolating the part of systematic information that cannot be used for prediction (in the examples, higher-order AR coefficients or part of the $A$ matrix).
Thank you for your attention