AUSTRIAN JOURNAL OF STATISTICS
Volume 37 (2008), Number 3&4, 271–284

Cross Validation of Prediction Models for Seasonal Time Series by Parametric Bootstrapping

Robert M. Kunst
University of Vienna and Institute for Advanced Studies Vienna, Austria

Abstract: Out-of-sample prediction for the final portion of a sample is a popular tool for model selection in model-based forecasting. We suggest to add a simulation step to this exercise, where pseudo-samples are generated (parametrically bootstrapped), conditional on the observed data and on any of the candidate models, and these pseudo-samples are predicted using any of the candidate models. The technique is demonstrated by an artificial univariate time-series specification that highlights the main features, and also by a real-life multivariate application to agricultural price data.

In the exemplary data set on quarterly European barley prices, strong seasonal variation is obvious and represents a crucial feature in constructing good models for short-run prediction. Following some preliminary statistical testing, we restrict focus to vector autoregressions with deterministic seasonal cycles. We also consider a restricted specification that imposes a common seasonal cycle on all countries. While the restriction is formally rejected by hypothesis tests, it assists in reducing prediction errors. The parametric bootstrap experiments show that this improvement by using an invalid restriction is systematic.

Zusammenfassung: Ex-ante Prognosen für den zeitlich jüngsten Teil der Stichprobe sind ein populäres Werkzeug der Modellwahl in modellbasierter Prognose. Wir schlagen vor, einen Simulationsschritt zu diesem Werkzeug hinzuzufügen, in welchem Pseudo-Stichproben generiert werden (parametrisches Bootstrapping), die den beobachteten Daten und jedem der Kandidatenmodelle entsprechen. Hierauf werden die Pseudo-Stichproben durch jedes der Kandidatenmodelle prognostiziert. Die Technik wird sowohl an Hand eines künstlichen univariaten Zeitreihenmodells demonstriert, wie auch an einer realen Anwendung auf landwirtschaftliche Preise.

In dem Beispiel von Quartalsdaten europäischer Gerstenpreise ist saisonale Variation deutlich erkennbar. Diese stellt eine wichtige Charakteristik dar, welche zur Erstellung guter Modelle für kurzfristige Prognosen entscheidend ist. Auf Grund einiger statistischer Voruntersuchungen konzentrieren wir uns auf Vektorautoregressionen mit deterministischen Saisonzyklen. Wir erwägen auch eine eingeschränkte Spezifikation, die einen gemeinsamen Saisonzyklus für alle Länder annimmt. Obwohl diese Restriktion formell von den Daten abgelehnt wird, ist sie doch zur Optimierung der Prognose geeignet. Das parametrische Bootstrap-Experiment zeigt, dass die Verbesserung der Prognosegüte systematisch ist.

1 Introduction

In model-based forecasting, researchers customarily first specify a small set of candidate model classes and then select a member of this set according to a criterion. Some base their choice directly on information criteria (IC), while others prefer the model that dominates its rivals with regard to out-of-sample prediction (OOS), i.e. the ultimate purpose of the selected model, over a portion of the available sample.

The correspondence of the true model and the best forecasting model is not trivial. Under the restrictive assumption of correct specification—that is, the true data-generating model is contained in one of the candidate model classes—a consistent estimator yields the correct parameter value in large samples, and the well known textbook theorem that conditional expectation minimizes the squared forecasting error (MSE) guarantees that the true model also optimizes prediction. For the sample sizes that are typical of economics data, the comfort given by this basic fact is limited at best. A simple model that restricts some small and poorly identified parameters at zero will dominate its correctly specified rival in comparatively large samples. We demonstrate this feature by a small artificial example for autoregressive processes.

It is surprising that particularly the econometric literature so often recommends subjecting the forecaster's preliminary model choice to a battery of specification tests, thus implicitly equating mis-specification aspects and deficiencies of the forecasting model. While this step tends to increase the prediction model's sophistication unduely, a recent emphasis on testing the significance of differences in forecasting performance relative to a benchmark model tends to impose an excessive penalty on complexity. In our view, neither approach helps in identifying the optimum prediction model, unless there are additional costs or benefits involved in using sophistication, which cannot be captured by usual loss criteria such as MSE.

There is a rich literature on the relationships between the two main selection paradigms, i.e. out-of-sample prediction over a test sample (OOS) and information criteria. Originally, information criteria were inspired by the forecaster's problem—hence, the name of the FPE criterion, 'final prediction error', which is asymptotically equivalent to Akaike's AIC. We just mention Shibata (1980) who shows that lag-order selection via AIC optimizes asymptotic predictive properties among autoregressions, and, more recently, Wei (1992) who establishes that OOS optimization defines a valid information criterion for quite general selection problems.

In econometrics, Inoue and Kilian (2006) have contributed to this literature by obtaining the surprising result that information criteria dominate OOS searches. This result, however, focuses on asymptotic properties, while true-life forecasters may rather be interested in small-sample performance. Moreover, Inoue and Kilian (2006) assume that, even in large samples, OOS optimization is restricted to a portion of the sample, which leads to an obvious loss of information as compared to information criteria, to cross validation, or to Wei's assumptions.

The traditional separation of the sample into a larger training and a smaller test sample may limit the power of the OOS under correct specification. In real-life comparisons, however, it may also be beneficial, as it emphasizes the latest part of the sample that, in the advent of slowly changing structures, may be the most relevant part for approximating

the variables beyond the sample end. By contrast, IC selection and cross validation weight early and late sample portions uniformly.

A main drawback of OOS methods remains that they base the selection decision on just a few test-sample data points. We suggest to study the forecasting properties of each candidate model by a further simulation step. For each candidate, pseudo-samples are generated according to the parameter value estimated from the sample. These pseudo-samples are 'predicted' using any of the candidates according to the estimate from the pseudo-sample. Similar experiments have been reported by Clements and Smith (1999) but they are still not common in the literature. Such simulation experiments reveal quantitative as well as qualitative features that may assist the forecaster's model choice. For example, consider that model A data are forecasted best by using model A but much worse by using model B, while both prediction models are on a par for forecasting model B data. Then, the forecaster may tend to prefer to use model A, given the data, whether other criteria would support A over B or not.

Apart from the constructed autoregressive example, we apply the technique to a time-series panel of quarterly European barley prices that was also analyzed by Jumah and Kunst (2006). The strong seasonal variation in the data suggests that modelling seasonality is the key to its predictability. Jumah and Kunst (2006) find that seasonal cycles are mainly deterministic, and they apply the recently developed technique of monitoring seasonal convergence by rolling samples (Franses and Kunst, 2007). Whereas that procedure fails to support the existence of a common seasonal cycle across Europe, an OOS prediction comparison for the last part of the sample sees the statistically rejected common-cycle model in the lead. The parametric bootstrapping method demonstrates that the result is indeed systematic. The wrong model outperforms the correct one, as its incorrect rank restriction avoids the estimation of some poorly identified parameters.

The plan of this paper is as follows. Section 2 considers an artificial time-series problem for autoregressions of varying order, in the spirit of McQuarrie and Tsai (1998). Section 3 turns to the real-life data set of barley prices. Section 4 summarizes and concludes.

2 An Artificial Example

The conflict between the aims of optimizing finite-sample prediction and of finding the true data-generating model class is best seen in model structures with declining parameter values. In small samples, it is beneficial to suppress the parameters with small values, although these values differ from zero. We recall that the modeler's aim is not a scientific search for true structures but forecasting.

In detail, 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 ε_t. Comparable models were studied by McQuarrie and Tsai (1998). In this experiment and in all others reported in this paper, i.i.d. N(0,1) series are generated by repeated draws from the GAUSS routine RNDN. To study small-sample performance, we construct samples of length $n + 10$ for $n = 10k$ and $k = 1, \ldots, 10$.

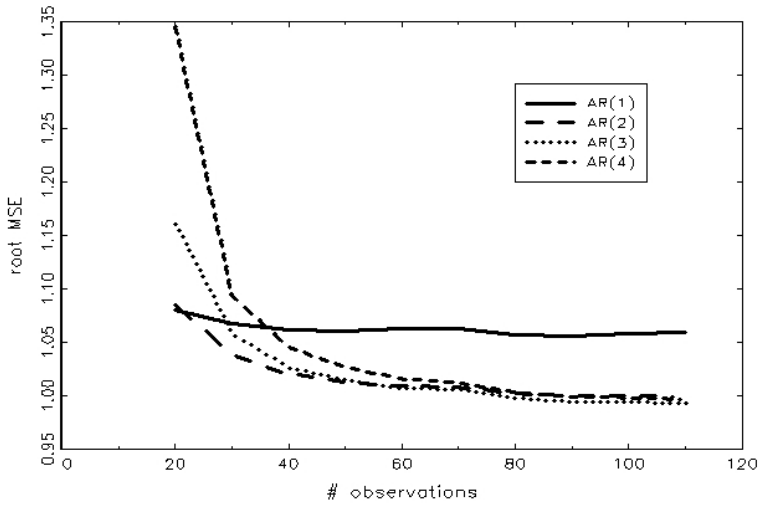

Figure 1: Forecasting performance of estimated autoregressive structures of varying lag order for data generated from AR(4) models, measured by squared prediction error for the last ten observations in samples of size n.

The last portion of 10 observations constitutes the test sample for OOS evaluation and is therefore excluded from estimation within the given model class. The candidate model classes are specified as autoregressions of lag orders 1 to 4. Over the ten observations of the test sample, models are re-estimated adding one observation at a time, as in real-world forecasting. Thus, shown mean squared errors for n are actually averages over samples of sizes n to $n + 9$. Parameter estimation is conducted using least squares.

The outcome of this little experiment is given as Figure 1. The first-order model dominates for very short samples, while the second-order model remains in the lead until around $n = 50$, when third-order models take over. The true AR(4) structure is outperformed slightly even for $n = 100$, and it is not even competitive for very small samples.

While this clear ranking can be obtained in a lab situation by using 10,000 replications, the decision for a single trajectory relies on 10 data points only. The alternative technique of *cross validation* aims at predicting any data point $t \geq 5$ by using the remainder of the sample for parameter estimation. To obtain a realistic time-series design, we discard all observations from estimation that contain the predicted data point, either as the 'dependent' variable or as a lag. For example, for predicting X_5, coefficient parameters were fitted using observations X_t, $t \leq 4$ and $t \geq 10$. Details on cross validation of time-series models are found in McQuarrie and Tsai (1998), while the concept is originally due to Stone (1974).

The outcome of this cross-validation experiment is given as Figure 2. Qualitatively, it is very similar to the OOS evaluation in Figure 1, which is to be expected for stationary generation designs with time-constant parameters. Such similarities justify the current usage of the expression 'cross validation' for OOS evaluations, even though this may be at odds with the original concept.

Finally, Figure 3 introduces the parametric bootstrapping technique. To obtain these graphs, estimated parameter values were taken as the basis for drawing new trajectories with generating AR models of orders one to four, and these were again predicted using

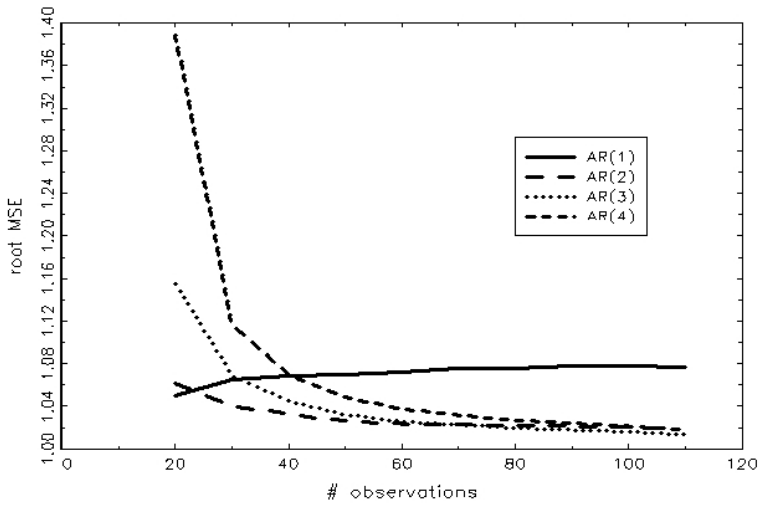

Figure 2: Cross validation of estimated autoregressive structures of varying lag order for data generated from AR(4) models, measured by squared out-of-sample fitting error for all single observations in samples of size n.

any of the four order specifications. Instead of rolling over the sample end and focusing on one-step errors, as before, we now evaluate multi-step predictions for step sizes $h = 1, \ldots, 10$. We kept the sample size constant at $n = 50$ and set the number of replications to 100,000. This number evolves from two simulation steps. 100 'true' trajectories are simulated for the given model (1), and for each data set we then generate 1000 parametric bootstraps according to the estimated parameter values.

This experiment shows that, excepting the order-four model with its difficult small coefficient, each fitted structure dominates its own trajectories at $h = 1$. As h increases, however, the ranking changes to the benefit of more parsimonious structures. The visual impression may support the usage of AR(2) as a forecasting device, as this specification dominates when it is correct and incurs small loss when it is actually incorrect. We recall that the true lag order is not assumed as known to the forecaster.

3 A Real-World Data Example

3.1 The Data

The data on quarterly barley prices is constructed from the Eurostat data base. Original Eurostat data is monthly but contains too many missing values. Furthermore, most time-series methods for seasonal data are tuned to the quarterly case—such as the monitoring of seasonal convergence by Franses and Kunst (2007)—or are considerably better developed and more powerful for quarterly observations—such as the HEGY test by Hylleberg, Engle, Granger, and Yoo (1990). For these reasons, monthly prices were aggregated to quarters by averaging over the three months that constitute a quarter.

The requirement of continuous price series of reasonable length restricts the analysis to ten countries: Austria, Belgium, Germany, Denmark, Spain, Finland, France, Netherlands, Sweden, and the United Kingdom. Sample ranges vary considerably across coun-

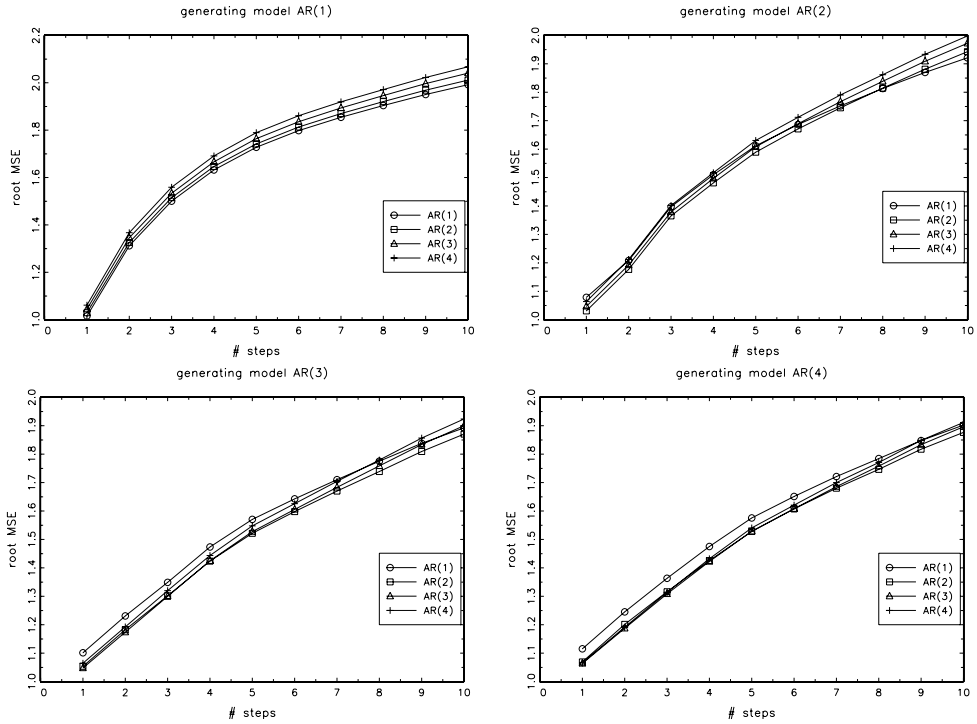

Figure 3: Forecasting performance of AR(1) to AR(4) models evaluated by parametric bootstrapping. 100 draws of the original AR(4) generating model and 1000 draws for the bootstrap at each lag order.

tries, and very few countries provide observations for the full range of 1970 to 2005. To enable the application of time-series analytic methods, we interpolated missing values and we extrapolated series with early endings. Because we aim at out-of-sample prediction evaluations, we used causal multivariate interpolation and extrapolation methods. All extrapolated observations are marked such that the prediction evaluation concerns forecasts of existing observations exclusively.

In order to stabilize variances and to enable the interpretation of first differences as inflation, all series were transformed by logarithms. Figure 4 shows time-series graphs for five countries with the longest samples and demonstrates the seasonal nature of the data.

3.2 Monitoring Seasonal Convergence

Franses and Kunst (2007) introduce a monitoring device for the study of convergence of deterministic seasonal cycles in panels of quarterly data. Their idea is as follows.

In a panel of N quarterly series with deterministic seasonal variation, the seasonal cycle is determined by three constants for each country. For $N \geq 3$, the $N \times 3$ coefficient matrix will typically have a rank of three. Rank deficiencies point to the occurrence of joint seasonal cycles across individuals.

A simple Fisher test statistic for the hypothesis of a rank of one in the coefficient matrix is calculated on rolling samples. If values of the statistic increase over time, this indicates a diverging tendency, whereas decreasing values indicate a converging tendency.

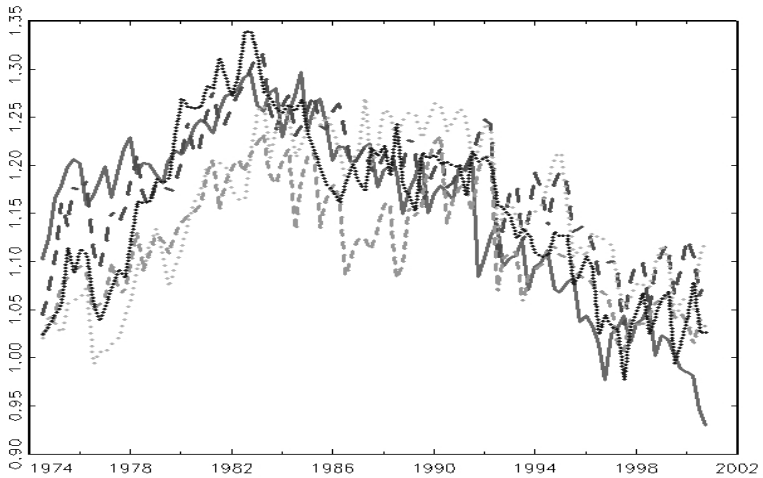

Figure 4: Time series plot of logarithmic barley prices.

Figure 5: Rolling test statistics (p-values) for a common seasonal cycle in barley prices. Countries in the sample are Belgium, Denmark, France, Spain, and the United Kingdom.

In Figure 5, the test statistic has been coded by nominal p-values, and the interpretation is reversed. An episode of 'convergence' is followed by 'divergence'. One might expect that out-of-sample prediction at the end of the sample should be influenced by the divergent period and that the unrestricted model should forecast better.

3.3 Forecasts

Out-of-sample mean predictions are based on vector autoregressions (VAR). Unit roots were found in all series, therefore VAR models are specified in first differences. By contrast, HEGY tests (Hylleberg et al., 1990) rejected seasonal unit roots, such that seasonal variation can be modelled successfully using deterministic dummy variables. Series with long samples (core set) are modelled as depending on the core set only, while other series (added set) are modelled as depending on all series. The distinction of core and added set is dictated by the lack of early data for the added set and is not to be seen as reflecting true causal directions.

Co-integration tests (Johansen, 1995) suggest that country series are not co-integrated. This is surprising from the viewpoint of economic theory and contradicts the 'law of one price'. We conjecture that convergence processes within the EU are mainly responsible for this feature, which is properly reflected in inferior forecasting performance of co-integrated VAR models (unreported control experiments). Hence, co-integrated models must be discarded for prediction, and interest should focus on VAR structures in pure first differences. Due to the limited sample size, higher-order VAR models cannot be used for prediction either, and we exclusively consider first-order VAR models.

In detail, we consider VAR systems of the form

$$Y_t = \mu + \mathbf{A}D_t + \Phi Y_{t-1} + \varepsilon_t\,, \tag{1}$$

where Y_t collects first differences of the original data ΔX_t in N-vectors, μ is an intercept vector, $\mathbf{A}$ is an $N \times 3$ matrix of seasonal coefficients, D_t collects three normalized deterministic seasonal cycles $\cos \pi t$, $\cos(\pi t/2)$, $\sin(\pi t/2)$, Φ is a block-triangular coefficient matrix of dimension $N \times N$, and ε_t is white-noise error with covariance matrix Σ. If the core variables inhabit the top N_1-portion of the vector Y_t, the matrix Φ has its north-east part of dimension $N_1 \times (N - N_1)$ restricted at zero.

Without further restrictions on $\mathbf{A}$, model (1) is equivalent to a model with four quarterly dummy variables and without an intercept. The form (1) is more convenient for describing seasonal features. For unrestricted matrix $\mathbf{A}$, the model corresponds to the maintained hypothesis of the convergence test of Figure 5. The null hypothesis corresponds to a matrix $\mathbf{A}$ with its rank restricted at one, such that all N countries have a common seasonal cycle. Then, the matrix can also be written as $\mathbf{A} = ab'$ with an N-vector a and a 3-vector b. We are mainly interested in the forecasting performance of the unrestricted model (1) and of the variant with the rank restriction. These two models will be called *model A* and *model B* in the following.

We note that model B does not imply the implausible feature that the climate is identical across Europe. Even in the presence of slight discrepancies among harvest months in different regions, proportionality in supplied quantities over quarters may occur. Moreover, the intense trade across Europe could prevent that price cycles match supply cycles exactly. Nonetheless, Figure 5 demonstrates that the rank restriction is rejected for most subsamples and, even more importantly, it is rejected toward the end of the sample.

It may come as a surprise, then, that Table 1 reflects a discrepancy between the in-sample statistical results (Figure 5) and the relative forecasting performance. Model B with restricted rank performs better than the formally supported general model. The evaluation considers one-step out-of-sample predictions for the last ten observations (two and a half years) but the ranking persists for larger horizons.

3.4 Cross Validation Literally

As a check for the validity of the results from the prediction experiment, we also cross-validated the models in the traditional sense of the word. That is, using all observations except for the observation at t, we fitted models and approximated the multivariate observation at t. In a time-series setting, even if only one observation is left out, more than one

Table 1: Forecasting performance for the barley price series.

		model A	model B
RMSE	core series	0.0446	0.0433*
	added series	0.0482	0.0421*
MAE	core series	0.0374	0.0365*
	added series	0.0385	0.0341*

Note: 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. Forecasting is evaluated out-of-sample single-step for the last ten time points of the sample. This yields 40 observations for the added series and 44 observations for the core series, after discarding 16 extrapolated observations.

Table 2: Cross validation for the barley price series.

		model A	model B
RMSE	core series	0.0330*	0.0366
	added series	0.0314	0.0286*
MAE	core series	0.0248*	0.0277
	added series	0.0239	0.0217*

Note: 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.

data point must be excluded for internal consistency. Here, where only first-order dynamics are permitted, two points at t and $t+1$ are omitted and t is forecast using estimates from $\{2,\ldots,t-1\} \cup \{t+2,\ldots,n\}$.

Table 2 shows that this cross-validation experiment is not entirely conclusive with regard to model choice. While the unrestricted model performs better for the core set, the restricted model is preferable for the added set. Interestingly, the improved performance for model A relative to the end-of-sample prediction of Table 1 does not match the monitoring result of Figure 5 that would support the restricted model B for earlier portions of the sample, which cross validation includes as criteria in contrast to the previous experiment.

Generally, measures improve relative to the out-of-sample prediction experiment. This may be due to the fact that more observations are used here on average but also to the inclusion of some re-constructed data, as a clear separation between true and constructed data is not tractable any more. One may assume that the prediction error for the interpolated data tends to be smaller, as these points have been predicted themselves.

3.5 Cross Validation by Parametric Bootstrapping

In order to get insights on whether the observed reversal of ranking between the in-sample inference stage and the OOS prediction stage is systematic, we again conduct a parametric

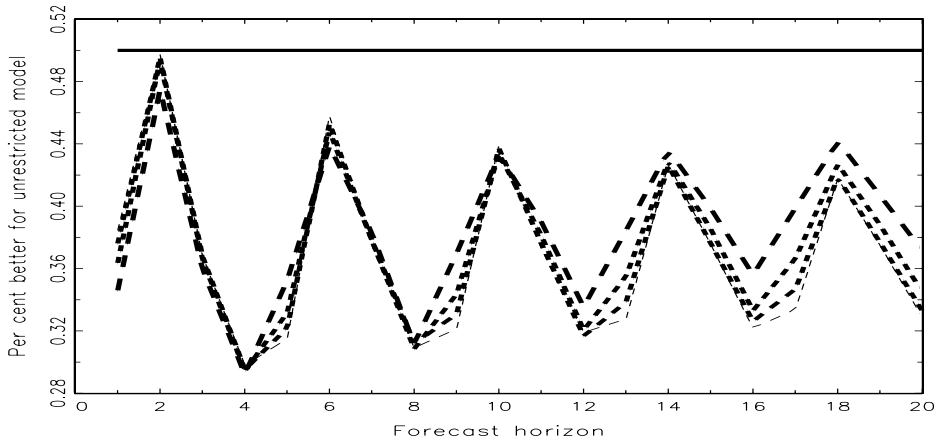

Figure 6: Generating model is the unrestricted model A. Forecast horizons on the abscissa, frequencies of lower MSE for prediction using model A on the ordinate. Long dashes for sample size $n = 48$, other curves for $n = 100, 200, 500$.

bootstrap experiment.

For this experiment, pseudo-data are generated from the fitted structures according to both rival models, with N-variate Gaussian errors—that is, repeated draws from the GAUSS random processor—and a variance matrix estimated from the sample. These pseudo-data are then predicted using both models. While the bootstrapped samples have constant parameter values, as fitted to the true observations, the prediction models are based on sample-specific parameter values, such as obtained from estimation. This design reflects the situation of a hypothetical forecaster properly.

The experiment has been conducted for pseudo-samples of sizes $n = 48, 100, 200, 500$, and the MSE is compared for predictions at horizons up to $h = 20$. Figures 6 and 7 show the frequency across the replications where the MSE for model A is lower than the MSE for model B. All frequencies are below 0.5 and thus favor model B. While the restricted model B outperforms the unrestricted model A more strongly when it is true, it also does so when it is in fact incorrect. The feature persists for larger n. As the sample size increases, the gains of using the restricted model gradually decline. The differences in performance among the core and added series are only minor and generally correspond to the summary graph.

In short, while barley prices do not really share a common seasonal cycle across Europe, assuming such a cycle helps in predicting the variables. The hypothesis that the additional parameters in model A are zero may be rejected but estimating these parameters does not aid the forecaster. In other words, the forecaster would need a stronger penalty for model complexity in this direction than is prescribed by hypothesis testing at usual significance levels. We note that the nested nature of the selection problem implies that AIC decision corresponds to testing at loose significance levels and that AIC likewise prefers the ‘bad’ model A. Thus, customary information criteria cannot replace the OOS experiments here.

The periodic fluctuation of Figure 6 points at a technical difficulty in setting up the bootstrap. Because of the non-stationary generating mechanism—a VAR in differences—

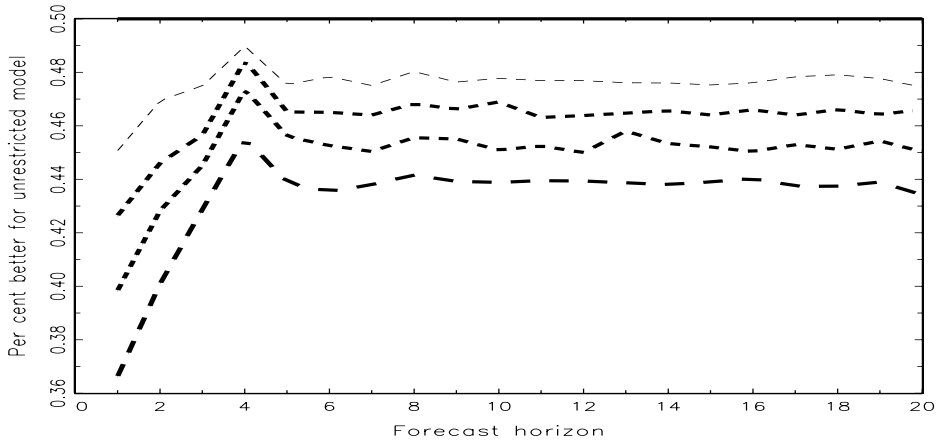

Figure 7: Generating model is the restricted model B. Forecast horizons on the abscissa, frequencies of lower MSE for prediction using model A on the ordinate. Long dashes for sample size $n = 48$, curves with shorter dashes for $n = 100, 200, 500$.

performance may depend on the quarter and on the starting values for the bootstrapped trajectories. We chose to start the bootstrapped samples from the last year where all series are available, we varied starting values over the four quarters of that year, and then we averaged over all quarters. This yields four times 10,000 replications instead of just 10,000. The averaging technique removed signs of periodicity from the model B bootstrap (see Figure 7) but not from the model A bootstrap.

While the shown figures are based on maximum-likelihood estimates for the covariance matrix Σ from the original samples, we found that the shapes are extremely sensitive to modifications with regard to this aspect. Larger residual variances, for example using degrees-of-freedom corrections, tend to enhance the benefits of model B. The share of variation due to the seasonal cycle decreases, and restricting seasonality improves prediction.

3.6 A Review of the Procedure

In this subsection, we summarize the parametric bootstrap technique that we used in the previous subsection.

After a thorough preliminary analysis, a phase during which many inappropriate models are considered and discarded, the forecaster is left with a small set of candidate models that may be useful for prediction. A traditional OOS 'horse race' can serve as a further guideline.

Then, we suggest the following steps:

1. Each candidate model is estimated over the full sample, typically by maximum likelihood or a convenient approximation.

2. For each candidate, the identified parameter values are used for generating R trajectories of a length that is comparable to the original sample size, maybe slightly longer such that there is room for manoeuvre for some OOS. If the utilized time-

series structures are non-stationary, starting values should be best obtained from a late part of the original sample.

3. Each candidate model is fitted to the pseudo-samples, omitting some final observations as test samples.
4. The test sample portions of the pseudo-samples are predicted. Averages of MSE across the R replications or comparable summary statistics can be reported.

According to the concept of parametric bootstrapping, the pseudo-samples for step #2 rely on draws from a normal random processor, with variances taken from sample estimates. It is straight forward to modify this step by utilizing a different distributional assumption.

As $n \to \infty$, the step #2 estimates will converge to true values for the correctly specified class, given that such a class exists. In strictly non-nested selection problems, other candidates will yield clearly worse predictions. For the other generating models, estimates converge to some pseudo-true values, prediction by the corresponding model will be best, although absolutely much less precise than for the true-true match.

If—as in our empirical example—models are nested, there are two cases for correct specification. Firstly, if the restricted model is correct, estimates are consistent for both models. The restricted model will yield better forecasts, due to efficiency, and graphs will be almost identical for both generating models. Second, if the restricted model is invalid but the general model is correct, the graphs will be quite different, as the pseudo-true and the true parameter values do not coincide. One could argue that the strong discrepancies between Figures 7 and 6 point to some evidence on the invalidity of the restricted model B. However, note that the outlined properties are asymptotic and that our sample are rather small, particularly for a 10-variate model.

We feel, however, that such large-sample properties are potentially less interesting to the applied forecaster who is not really interested in the 'truth' of any candidate model. Rather, the graphs should be seen as tools for the selection of the optimal forecast model.

4 Summary and Conclusion

For a small artificial experiment and for a real-life data set, we demonstrated the application of traditional OOS (out-of-sample) forecasting, of cross validation, and of OOS via parametric bootstrap. Cross validation is not so often used in forecasting model selection but it may deserve attention if dynamic structures can be assumed as time-constant, as it exhausts the sample information fully, in contrast to OOS at the sample end.

While simulations similar to our bootstrap OOS can be found occasionally in the literature (e.g., Clements and Smith, 1999), usage of the technique is not common. Often, it is conducted asymmetrically, such that only one candidate model is simulated and all candidates are used to forecast the simulated data. This approach may be grounded in classical hypothesis testing that traditionally concentrates on properties 'under the null'. Simulating under the alternative at a fitted parameter value may then be interpreted as a form of confirmatory approach. Confirmatory approaches are often shunned in testing problems, last not least because of the tremendous effort involved in compiling critical values at distances from the null (e.g., see Dhrymes, 1998). Such arguments, however,

do not apply to forecasting model selection, and the programming and computer time involved in generating the graphs shown in this paper are reasonable.

The real-life example of agricultural prices depicts the typical situation of a forecaster. Most models can be discarded in an early stage of modelling, and then interest focuses on a handful of candidates that may form a nested, partially nested, or non-nested set. The outlined device can be used for any of these situations, and it clearly points to problematic aspects. For example, if a model class is supported by tests as well as short-run OOS but incurs enormous loss if it is by chance not really the generating model, the forecaster may avoid that class for prediction. This feature has been reported in conjunction with many non-linear time-series models that can offer excellent data description but rarely live up to their expectations if it comes to prediction. Bootstrap simulations can also be seen as answering to the warnings by Rissanen (2007) that traditional information criteria do not penalize model complexity sufficiently, as penalty terms depend on parameter dimension only. While Rissanen's code-length approach imposes a heavy burden on the user, the OOS bootstrap can be implemented at low cost.

It may be an obvious suggestion to explore even non-parametric bootstrap methods for the OOS evaluations. Our experience with drawing from residual distributions, however, is that this tends to blur the distinction among the—intrinsically parameterized—model classes. Model-based forecasting inevitably proceeds by simplifying the sample information, discarding a big portion of it as 'noise', and focusing on main features that have a good chance of surviving beyond the sample end.

References

Clements, M. P., and Smith, J. (1999). A Monte Carlo study of the forecasting performance of empirical SETAR models. *Journal of Applied Econometrics*, *14*, 123-141.

Dhrymes, P. (1998). *Time Series, Unit Roots and Cointegration.*

Franses, P. H. F., and Kunst, R. M. (2007). Analyzing a panel of seasonal time series: Does seasonality in industrial production converge across Europe? *Economic Modelling*, *24*, 954-968.

Hylleberg, S., Engle, R. F., Granger, C. W. J., and Yoo, B. S. (1990). Seasonal integration and cointegration. *Journal of Econometrics*, *44*, 215-238.

Inoue, A., and Kilian, L. (2006). On the selection of forecasting models. *Journal of Econometrics*, *130*, 273-306.

Johansen, S. (1995). *Likelihood-Based Inference in Cointegrated Vector Autoregressive Models*. Oxford University Press.

Jumah, A., and Kunst, R. M. (2006). Seasonal cycles in European agricultural commodity prices. *Economics Series*, *192*.

McQuarrie, A. D. R., and Tsai, C. (1998). *Regression & Time Series Model Selection*. World Scientific.

Rissanen, J. (2007). *Information and Complexity in Statistical Modeling*. Springer.

Shibata, R. (1980). Asymptotic efficient selection of the order of the model for estimating parameters of a linear process. *Annals of Statistics*, *8*, 147-164.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions (with discussion). *Journal of the Royal Statistical Society, Series B, 36*, 111-147.

Wei, C. Z. (1992). On predictive least squares principles. *Annals of Statistics, 20*, 1-42.

Author's Addresses:

Robert Kunst
University of Vienna
Department of Economics
BWZ
Brünner Straße 72
1210 Wien
Austria

and

Institute for Advanced Studies
Stumpergasse 56
1060 Wien
Austria

E-Mail: robert.kunst@univie.ac.at