A Comparison of Model Selection Procedures for Predicting Turning Points in Financial Time Series

Thorsten Poddig and Claus Huber

University of Bremen, Chair of Finance
FB 7, Hochschulring 4
D-28359 Bremen, Germany
poddig@uni-bremen.de
huberc@uni-bremen.de

Abstract. The aim of this paper is to compare the influence of different model selection criteria on the performance of ARMA- and VAR-models to predict turning points in nine financial time series. As the true data generating process (DGP) in general is unknown, so is the model that mimics the DGP. In order to find the model which fits the data best, we conduct data mining by estimating a multitude of models and selecting the best one optimizing a well-defined model selection criterion. In the focus of interest are two simple in-sample criteria (AIC, SIC) and a more complicated out-of-sample model selection procedure. We apply Analysis of Variance to assess which selection criterion produces the best forecasts. Our results indicate that there are no differences in the predictive quality when alternative model selection criteria are used.

1 Introduction

Forecasting turning points (TP) in financial time series is one of the most fascinating (and possibly rewarding) aspects in finance. In this paper, we implement a Monte-Carlo-based regression approach introduced by Wecker[1] and enhanced by Kling[2] to produce probabilistic statements for near-by TPs in monthly financial time series. This method needs forecasts of future values of the time series. Those can be predicted by an econometric model, which is assumed to mimic the true data generating process (DGP). The performance of forecasting models can be judged in two different ways. *In-sample* predictive accuracy is measured with the data already used for model development and estimation of the coefficients. In contrast, *out-of-sample* assessment focuses on the ability of the model to predict unknown datapoints. Hence forecasting models are usually needed to predict unknown datapoints, in this paper their out-of-sample predictions are in the focus of interest. To evaluate whether the out-of-sample forecasts from the models are reliable, backtesting is performed: Using a simulation period of 68 months, out-of-sample predictions are produced for each month, using for model estimation only the data available until this month. Unfortunately, in

J.M. Żytkow and J. Rauch (Eds.): PKDD'99, LNAI 1704, pp. 492–497, 1999.

most applications the true DGP is unknown, and so is the specification of the model. The problem is that a multitude of models exists, and one out of these models is able to fit the DGP best. In order to find this model that produces the best out-of-sample predictions in the backtesting period, two model selection procedures can be applied. In-sample model selection can be implemented easily, using e.g. the Akaike or Schwartz information criterion (AIC, SIC). More complicated is out-of-sample model selection, which can be conducted by forming a separate cross-validation subset (CV) from the training data. The CV is not used to estimate the coefficients but only to validate the model on unknown data. The aim of this paper is to compare the performance of models selected by two classical in-sample model selection criteria (AIC, SIC) with the performance of an out-of-sample validation procedure. To evaluate the performance of the models in the backtesting period, we take the view of a participant in the financial markets. Here one is not interested in optimizing statistical criteria, like Mean Squared Error etc., but in obtaining an acceptable profit. A performance criterion in this spirit is the Cumulative Wealth (CW). We predict TPs in nine financial time series of monthly periodicity with ARMA- and VAR-models using rolling regressions. In order to obtain statistically significant results, we examine the TP predictions from ARMA- resp. VAR-models created by the different model selection methods using Analysis of Variance (ANOVA).

Due to space limitations, this paper had to be shortened considerably. The full version is available from the internet[3].

2 The Detection of Turning Points in Financial Time Series

As a first step to obtain a probabilistic statement about a near-by turning point one has to define a rule when a TP in the time series is detected. The turning point indicator

$$z_t^P = \begin{cases} 1, \text{ if } x_t > x_{t+i}, i = -\tau, -\tau+1, \ldots, -1, 1, \ldots, \tau-1, \tau \\ 0, \text{ otherwise} \end{cases} \quad (1)$$

is defined as a local extreme value of τ preceding and succeeding datapoints. The trough indicator z_t^T is defined in an analogous way. As we investigate monthly time series, we define τ=2. At time t the economist knows only the current and past datapoints $x_t, x_{t-1}, \ldots, x_{t-\tau+1}, x_{t-\tau}$. The future values $x_{t+1}, \ldots, x_{t+\tau-1}$, $x_{t+\tau}$ have to be estimated using e.g. ARMA- and VAR-models. With those estimates we applied a Monte-Carlo based procedure developed by Wecker [1] and Kling [2] to obtain probabilistic statements about near-by TPs. A TP is detected if the probability reaches or exceeds a certain threshold θ, e.g. θ=.5.

A participant in the financial markets usually is not interested in MSE, MAE, etc. but in economic performance. Since our models do not produce return forecasts but probabilities for TPs, we have to measure performance indirectly by generating trading signals from those probabilities: A short position is taken when a peak is detected (implying the market will fall, trading signal s=-1), a long

position in the case of a trough s=+1), and the position of the previous period is maintained if there is no TP. With the actual period-to-period return $r_{actual,t}$ we can calculate the return $r_{m,t}$ from a TP forecast of our model: $r_{m,t} = s \cdot r_{actual,t}$. In this paper we deal with log-differenced data, so the Cumulative Wealth can be computed by adding the returns over T periods: $CW = \sum_{t=1}^{T} r_{m,t}$. To test the ability of the ARMA and VAR models to predict TPs, we investigate nine financial time series, namely DMDOLLAR, YENDOLLAR, BD10Y (performance index for the 10 year German government benchmark bond), US10Y, JP10Y, MSWGR (performance index for the German stock market), MSUSA, MSJPA, and the CRB-Index. The data was available in monthly periodicity from 83.12 to 97.12, equalling 169 datapoints. To allow for the possibility of structural change in the data, we implemented rolling regressions: After estimating the models with the first 100 datapoints and forecasting the τ succeeding datapoints, the data-window of the fixed size of 100 datapoints was put forth for one period and the estimation procedure as well as the Monte-Carlo-simulations were repeated until the last turning point was predicted for 97.10. Thereby we obtained 68 out-of-sample turning point forecasts. We estimated a multitude of models for each model class: 15 ARMA-models from (1,0), (0,1), (1,1),..., to (3,3) and 3 VAR models VAR(1), (2), and (3) comprising all nine variables. We do not specify a model and estimate all rolling regressions with this model. Rather we specify a *class* of models (ARMA and VAR). Within a class the best model is selected for forecasting. As an extreme case, a different model specification could be chosen for every datapoint (within the ARMA class e.g. the ARMA(1,0) model for the first rolling regression, ARMA(2,2) for the second etc.).
Popular in-sample model selection criteria are AIC and SIC. Applying AIC and SIC for model selection within the first rolling regression, we estimated a multitude of e.g. ARMA-models with 100 datapoints and chose the model with the lowest AIC to forecast the τ future datapoints. In contrast to the simple implementation of AIC and SIC, the out-of-sample procedure for model selection is more complicated. Therefore we divided the training data in two subsequent, disjunct parts: an estimation (=training) subset (70 datapoints) and a validation subset (30 datapoints, see figure 1).

The first 70 datapoints from t-99 to t-30 were used to estimate the models, which were validated with respect to their abilities to predict TPs on the following 30 datapoints from t-29 to t. The decision which model is the "best" within the out-of-sample selection procedure was made with respect to CW: the model with the highest CW was selected. The specification of this model, e.g. ARMA(2,2), then was re-estimated with the 100 datapoints from t-99 to t to forecast the at time t unknown τ values of the time series which are necessary to decide whether there is a turning point at time t.

As a result of model selection with the two in-sample criteria AIC, SIC, and the out-of-sample procedure with regard to CW we obtain three sequences of TP forecasts each for ARMA- and VAR-models for the out-of-sample back-testing period of the 68 months. Two ARMA-sequences with a threshold θ=.5 could look like table 1. The first four columns refer to the number of the rol-

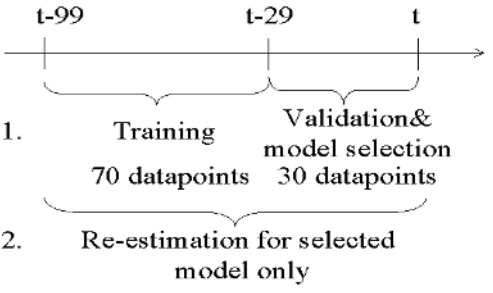

Fig. 1. Division of the database

ling regressions and the training, validation, and forecast period, respectively. For AIC and SIC model selection was performed on the 100 datapoints of the training and validation subset as a whole. The 5th (7th) column gives the specification of the ARMA-model selected by CW (AIC), the 6th (8th) column gives the corresponding CW- (AIC)-value.

Table 1. ARMA-sequence as an example for the rolling regressions

				CW		***AIC***	
RR	training	validation	forecast	Spec.	$CW-value$	Spec.	$AIC-value$
1	83.12-89.9	89.10-92.3	92.4-92.5	(2,2)	.179	(3,3)	-5.326
2	84.1-89.10	89.11-92.4	92.5-92.6	(1,0)	.253	(1,1)	-5.417
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
68	89.8-95.4	95.5-97.10	97.11-97.12	(3,0)	.815	(2,3)	-5.482

The first TP forecast was produced for 92.4 (with the unknown values of 92.5 and 92.6), the last for 97.10. The 68 out-of-sample forecasts of the model sequences generated this way are finally evaluated with respect to CW.

To judge whether the econometric models are valuable forecasting tools, one would like to test if the model class under consideration is able to outperform a simple benchmark in the backtesting period. When forecasting economic time series, a simple benchmark is the naive forecast. Using the last certain TP statement can be regarded as a benchmark in this sense. As τ=2, the last certain TP statement can be made for t-2, using the datapoints from t-4 to t. A valuable forecasting model should be able to outperform this Naive TP Forecast (NTPF) in the backtesting period.

In order to produce a statistically significant result when comparing the model sequences generated by the different model selection criteria, we apply Analysis of Variance (ANOVA). The forecasts for ARMA-models, θ=.5, with respect to the evaluation criterion CW can be exhibited as in table 2 (the last column

contains the NTPF-results). The entry -.115 in the 3rd column of row 3 means that ARMA-models selected by AIC produced a CW of -.115 in the backtesting period when predicting turning points for MSWGR. Looking to the last row, column 3 reveals that the mean CW over all nine time series from the ARMA forecasts is -.192.

Table 2. Example for the exhibition of the results from the ARMA turning point forecasts

	Selection criteria			
	CW	AIC	SIC	NTPF
MSWGR	-.262	-.115	-.020	-.029
BD10Y	-.856	-.515	-.898	-.291
⋮	⋮	⋮	⋮	⋮
mean:	-.232	-.192	-.264	-.196

The block experiment of ANOVA can be used to test if the means of the columns (here the means from the TP predictions) and the means of the rows (the means from TP predictions for one of the time series) are identical. Thereby it is possible to compare the performance of the different model selection criteria. Additionally, the NTPF is included in the test to make sure that the models outperform the benchmark. The basic model of ANOVA is: $y_{ij} = \mu + \alpha_i + \beta_j + e_{ij}$, where y_{ij} represents the element in row i and column j of table 2, μ is the common mean of all y_{ij}, α_i is the block effect due to the analysis of different time series in the r rows of table 2, β_j the treatment effect of the p selection criteria (incl. NTPF) in the columns of table 2, and e_{ij} an iid, $N(0; \sigma^2)$ random factor. We want to test whether the treatment effects β_j are zero: $\beta_1 = \beta_2 = ... = \beta_p = 0$. In other words, we want to test the null hypothesis that there are no statistically significant effects due to the use of different model selection criteria on the TP forecasts from ARMA- and VAR-models. An F-test statistic is based on the idea that the total variation SST of the elements in table 2 can be decomposed in the variation between the blocks SSA, the variation between the selection criteria SSB, and the random variation SSE: $SST = SSA + SSB + SSE$. Estimators for SST, SSA, SSB and SSE can be computed as shown in [3]. Then an F-statistic can be computed (see [3]). The null is rejected, if F exceeds its critical value. The next section presents empirical results.

3 Empirical Results and Conclusion

The following table 3 exhibits the empirical results from the TP forecasts with ARMA- and VAR-models. The 2nd and 3rd column show the value for the F-statistic and its corresponding p-value. The 4th to 7th column contain the means

of the model sequence created by the selection criterion under consideration. E.g. the entry ".32" in row 2, column 2 gives the F-statistic for the null that the mean CW of ARMA-models, created by the use of the selection criteria AIC, SIC, CW, and the NTPF are all the same. The p-value of .8087 indicates that the null cannot be rejected at the usual levels of signficance (e.g. .10). Thus we have to conclude that there are no differences between TP forecasts from ARMA-models generated by different model selection criteria. Moreover, the ARMA forecasts do not differ significantly from the NTPF. Columns 4 to 7 exhibit the mean CW over all nine time series. The ARMA models selected by e.g. AIC managed to produce an average CW of .059 in the simulation period. This is only marginally higher than the mean CW from NTPF (.057).

Table 3. Empirical ANOVA results from the TP predictions

Model	F	p	AIC	SIC	CW	NTPF
ARMA	.32	.8087	.059	-.022	-.010	.057
VAR	.16	.9242	-.002	-.002	.015	.057

In general, the results indicate that there are no statistically significant differences between TP predictions from ARMA- and VAR-models (p-values .8087 and .9242). With concern to ARMA-models, AIC seems to be the best selection criterion with respect to CW (mean CW=0.059). This is only slightly better than the benchmark NTPF (mean CW=0.057) and cannot be considered as a reliable result. The other selection criteria even led to underperformance vs. NTPF. Results are even worse for VAR-models. All VARs underperformed the NTPF. Thus it must be doubted that ARMA- and VAR-models are valuable tools for predicting TPs in financial time series. If they are employed despite of the results achieved here, it might be a good choice to make use of in-sample selection criteria AIC and SIC. They led to comparable results as the out-of-sample validation procedure suggested in this paper and are less expensive to implement. If those results hold for other forecasting problems, evaluation criteria, and selection procedures as well has to be investigated by further research.

References

1. Wecker, W. (1979): Predicting the turning points of a time series; in: Journal of Business, Vol. 52, No. 1, 35-50
2. Kling, J.L. (1987): Predicting the turning points of business and economic time series; in: Journal of Business, Vol. 60, No. 2, 201-238
3. Poddig, T.; Huber, C. (1999): A Comparison of Model Selection Procedures for Predicting Turning Points in Financial Time Series - Full Version, Discussion Papers in Finance No. 3, University of Bremen, available at: www1.uni-bremen.de/~fiwi/