

Extending the Scope of Automatic Time Series Model Selection: The Package *autots* for R

Dongik Jang^a, Hee-Seok Oh^a, Donghoh Kim^{1,b}

^aDepartment of Statistics, Seoul National University

^bDepartment of Applied Mathematics, Sejong University

Abstract

In this paper, we propose automatic procedures for the model selection of various univariate time series data. Automatic model selection is important, especially in data mining with large number of time series, for example, the number (in thousands) of signals accessing a web server during a specific time period. Several methods have been proposed for automatic model selection of time series. However, most existing methods focus on linear time series models such as exponential smoothing and autoregressive integrated moving average (ARIMA) models. The key feature that distinguishes the proposed procedures from previous approaches is that the former can be used for both linear time series models and nonlinear time series models such as threshold autoregressive (TAR) models and autoregressive moving average-generalized autoregressive conditional heteroscedasticity (ARMA-GARCH) models. The proposed methods select a model from among the various models in the prediction error sense. We also provide an R package *autots* that implements the proposed automatic model selection procedures. In this paper, we illustrate these algorithms with the artificial and real data, and describe the implementation of the *autots* package for R.

Keywords: Automatic model selection, ARIMA models, exponential smoothing, TAR models, ARMA-GARCH models, time series, R.

1. Introduction

Automatic model selection for a large number of time series is critical in many applications such as network-based services operations. For example, internet-based enterprises and organizations have received a large number of internet signals including possible cyber attacks that should be detected and resolved rapidly before any serious problems arise. In this case, an automatic system that identifies the proper model of signals is clearly more desirable than manual selection conducted explicitly by an analyst. Even when a time series is sufficiently important to merit the analysis of an expert, the result from automatic model selection might be useful as an initial model for manual selection.

The main objective of this paper is to propose an automatic system that selects the model that is best-fitted to a time series from among pre-specified classes of possible models without user supervision. Previously, there have been many automatic methods proposed by Gómez and Maravall (1998), Liu (1989), Mélard and Pasteels (2000), Hyndman and Khandakar (2008). Some commercial software also provides an automatic time series model selection. However, the main limitation of the existing approaches is that they consider only linear models, such as the autoregressive integrated

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2009-0086944).

¹ Corresponding author: Associate Professor, Department of Applied Mathematics, Sejong University, 98 Gunja-Dong, Gwangjin-Gu, Seoul 143-747, Korea. E-mail: donghohkim@sejong.ac.kr

moving average (ARIMA) and exponential smoothing models. Since a time series observed in the real world might have inherent nonlinear features such as heteroscedastic variance, the scope of the existing methods is restrictive.

In this paper, we propose two procedures, *best-model-based method* and *tree-based method*. These perform automatic time series model selection from among the classes of nonlinear and linear models. We consider the classical ARIMA, exponential smoothing linear models, TAR, and ARMA-GARCH nonlinear models. The proposed algorithms select the best model from among exponential smoothing, ARIMA, TAR and ARMA-GARCH models.

We provide an R package `autots` that implements the two proposed algorithms and explain the usage of the package `autots` through several examples so that the readers can utilize the proposed methods for model selection. The source code, binary, and user manual are available at <http://stat.snu.ac.kr/heeseok/autots>.

2. The Scope of Time Series Models

2.1. Linear time series models

The most popular classes of linear time series models are ARIMA models and exponential smoothing models. ARIMA models are often used to model linear dynamic structures and to depict linear relationships among lagged variables. Exponential smoothing models are popular in the analysis of economic and financial market data for smoothing and forecasting. The weighted moving average and Holt-Winters' model are specific forms of exponential smoothing models. We follow Hyndman and Khandakar (2008) when selecting a linear model.

2.2. Nonlinear time series models

There are many nonlinear time series forms. From the perspective of computing time and usefulness, we focus on two nonlinear parametric forms—GARCH for fluctuating volatility of financial data (Engle, 1982; Bollerslev, 1986) and TAR for biological and economic data (Tong, 1990; Tiao and Tsay, 1994).

3. The Proposed Methods

We propose two procedures for automatic time series model selection—*best-model-based method* and *tree-based method*. The main contribution of the proposed methods is that they extend the scope of model selection to nonlinear time series. Thus, it can be expected that the proposed selection methods have widespread applicability.

3.1. Best-model-based method

It is well known that Akaike Information Criteria (AIC) is a good measure to compare the models of the same class. However, using AIC to compare the models of different classes is difficult because the unification of the definitions of the likelihood functions may not be easy. One way to resolve this problem is to use a prediction error that measures the prediction performance of the models of different classes. Under the Gaussian assumption, AIC can be expressed as a term of prediction errors; this might be a rationale to use prediction errors instead of AIC.

We first select the models minimizing the AIC from among each class of models—exponential smoothing, ARIMA, TAR, and ARMA-GARCH. We can find the model that minimizes the prediction errors once selected from each class of models.

Suppose that we have a time series $X_1, \dots, X_m, X_{m+1}, \dots, X_{m+n}$. The first m series X_1, \dots, X_m are used to estimate the parameters and to find the models minimizing AIC within each class of models. The remaining series X_{m+1}, \dots, X_{m+n} are used to calculate the n -ahead prediction error:

$$\frac{1}{n} \sum_{i=m}^{m+n-1} \{X_{i+1} - \hat{X}_{i+1}(m)\}^2,$$

where $\hat{X}_j(m)$ denotes the predicted value at time j by a fitted model based on X_1, \dots, X_m . We consider the following iterative n -ahead prediction error as another prediction error:

$$\frac{1}{n} \sum_{i=m}^{m+n-1} \{X_{i+1} - \tilde{X}_{i+1}\}^2,$$

where \tilde{X}_{i+1} denotes the one-ahead prediction value computed with $X_1, \dots, X_m, \tilde{X}_{m+1}, \dots, \tilde{X}_i$.

Although prediction errors are a useful criterion for comparing the models, time series models having minimum prediction error may not necessarily satisfy model assumptions such as the normality of noise. Therefore, we check the model adequacy for the chosen model. If the residuals of the chosen model are non-normal, the second-best model that satisfies the normality condition is recommended.

The aforementioned procedure is summarized as follows:

- Step 1. Choose a model minimizing AIC from among each class of models—ARIMA, exponential smoothing model, TAR, and ARMA-GARCH.
- Step 2. Select the model minimizing prediction errors from among the four models chosen in Step 1.
- Step 3. Check the selected model for model adequacy. If the assumptions are not satisfied, then the second-best model is selected as the best model.

The best-model-based method is required to compare the time series models of linear and nonlinear model classes. Compared to the tree-based method described below, the best-model-based method is computationally more intensive. The major advantage of this approach is that this method might provide lower misspecification error than the tree-based method since all possible models including linear and nonlinear models are evaluated throughout the whole procedure without any screening (testing) steps.

3.2. Tree-based method

The tree-based method first chooses the models only from among the linear model classes (ARIMA and exponential smoothing model) and then decides whether or not the scope of the selection should be extended to the nonlinear model classes.

Drawing from our experiences, we use the Brock-Dechert-Sheinkman(BDS) test (Brock *et al.*, 1987) for the nonlinear test based on the residuals from the selected linear model. Note that the BDS test examines the spatial dependency of a time series. To compute the spatial dependency, the series is embedded in an m -space and the dependency is examined by counting the near points for which the distance is less than the specific value. Brock *et al.* (1996) showed that the BDS test statistic is asymptotically standard normal. For more details on the BDS test, we refer to Brock *et al.* (1996).

If the linearity of the residuals from the selected linear model is not satisfied, then we further extend the scope of selection to nonlinear model classes and choose the model that minimizes AIC and prediction errors. The tree-based method can be summarized as follows:

Step 1. Choose a model minimizing AIC from each linear model class—ARIMA and exponential smoothing model.

Step 2. Select the model minimizing prediction errors from among two models chosen in Step 1.

Step 3. Run a nonlinearity test(BDS test). If the residuals still hold nonlinear structures, then proceed to Step 4. Otherwise, select the linear model chosen in Step 2, and then, check it for model adequacy. If the normality assumption is satisfied, the selected model is chosen as the best model. Otherwise, the second-best nonlinear model is selected as the best model.

Step 4. Choose a model minimizing AIC from each nonlinear model class-TAR and ARMA-GARCH. Then, select the model minimizing prediction errors, and check it for model adequacy. If the normality assumption is satisfied, the selected model is chosen as the best model. Otherwise, the second-best nonlinear model is selected as the best model.

4. Numerical Experiments

4.1. Implementation of the autots package

We describe the implementation of the proposed automatic model selection algorithms in the `autots` package for R, and illustrate the proposed algorithms with the artificial data. The main function of the `autots` package is `Ats`, which executes the time series model selection automatically. The arguments of the `Ats` function are as follows:

```
> args(Ats)
function (x, method = c("TB", "BM"), model.set = c("all", "linear", "nonlinear"),
  one.step = TRUE, n.ahead = 1, verbose = FALSE)
```

The input arguments are as follows:

- `x` : A numeric vector or a univariate time series.
- `method` : Specifies the model selection method. If "TB", the best model is selected by the tree-based method. If "BM", the best model is selected by best-model-based method. The default is "TB".
- `model.set` : Specifies the model classes when best-model-based method is used. When `model.set` is set as "linear", "nonlinear" or "all", the candidate classes considered for model selection are linear models, nonlinear models, or both, respectively. The default is "all".
- `one.step` : If TRUE, `n.ahead` prediction error is computed based on iterative one-ahead prediction error.
- `n.ahead` : The number of time steps ahead for which prediction is required.
- `verbose` : If TRUE, the list of time series models considered and the supplementary information are printed on screen. The default is FALSE.

The output arguments are as follows:

- `call` : Matched call.

- `bestmodel` : Description of the model selected from among the linear and/or nonlinear model classes.
- `order` : Order of the selected model, when available.
- `coeff` : A vector of coefficients, when available.
- `fitted.values` : A vector of fitted values of the selected model.
- `cond.sdevi` : Conditional standard deviation, when available.
- `residuals` : A vector of residuals from the selected model.
- `s.residuals` : A vector of standardized residuals from the selected model.
- `sigma2` : Variance estimate.
- `frequency` : Frequency of input time series, when available.
- `series` : Object name of input time series.
- `x` : A vector of input time series.
- `bestfit` : List containing information related to the selected model.

Here, we use three artificial examples to assess the practical performance of the proposed automatic model selection methods and to demonstrate how the `autots` package is used. To that end, we generate three time series sets, one each, from the ARIMA, TAR, and ARMA-GARCH models, respectively.

For the first case, artificial data is generated from the ARIMA(2,0,2) model, and the tree-based method is performed as follows:

```
> set.seed(329)
> x <- arima.sim(n = 200, list(ar = c(0.8897, -0.4858),
+      ma = c(-0.2279, 0.2488)), sd = sqrt(0.1796))
> fit <- Ats(x, method="TB", verbose=TRUE)
Time series model is selected by tree-based method.
```

Linear time series model:

ARIMA(2,0,2) with zero mean is the best among ARIMA models.

ETS(A,N,N) is the best among exponential smoothing models.

Prediction errors:

Prediction errors of the best ARIMA and exponential smoothing models.

ARIMA: 0.0864492222882386, Exponential smoothing: 0.282615595316445

ARIMA(2,0,2) with zero mean is the best among linear time series models.

Nonlinearity test:

Nonlinearity test accepts null hypothesis.

Time series data seems to have linear structure.

Final model:

The best time series model is selected in ARIMA models.

```
> fit
```

Title:

Automatic time series model selection

Series:

x

Call:

Ats(x = x, method = "TB", verbose = TRUE)

Best model:

ARIMA(2,0,2) with zero mean

Autoregressive integrated moving average (ARIMA) model

Coefficient(s):

ar1	ar2	ma1	ma2
0.6477	-0.3447	0.0926	0.3549

The tree-based method selects ARIMA(2,0,2) as the best linear time series model, which is the true model. Since the nonlinearity test for the residual accepts the linearity hypothesis, the selection process stops.

For our second example, we generate data from TAR with two regimes. The parameters of the low and high regimes are $(\phi_{10}, \phi_{11}) = (0, 0.5)$ and $(\phi_{20}, \phi_{21}) = (0, -1.8)$, respectively. Since ARIMA(0,0,2), selected as the best linear model, does not satisfy linearity, the classes of the candidate models are extended to the nonlinear models and SETAR is chosen as the final model. Applying the tree-based method, we identify the true model and estimate the parameters.

```
> library(TSA)
```

```
> set.seed(777)
```

```
> y <- tar.sim(n=100, Phi1=c(0,0.5), Phi2=c(0,-1.8),
```

```
+ p=1, d=1, sigma1=1, thd=-1, sigma2=2)$y
```

```
> fit <- Ats(y, "TB", n.ahead=5, verbose=TRUE)
```

Time series model is selected by tree-based method.

Linear time series model:

ARIMA(0,0,2) with non-zero mean is the best among ARIMA models.

ETS(A,N,N) is the best among exponential smoothing models.

Prediction errors:

Prediction errors of the best ARIMA and exponential smoothing models.

ARIMA: 4.2885500007487, Exponential smoothing: 4.34413460239223

ARIMA(0,0,2) with non-zero mean is the best among linear time series models.

Nonlinearity test:

Nonlinearity test rejects null hypothesis.
Time series data seems to have nonlinear structure.

Nonlinear time series model:

SETAR with 2 regimes is the best among TAR models.
ARMA(0,2)+GARCH(1,2) is the best among ARMA-GARCH models.
Prediction errors:
Prediction errors of the best TAR and ARMA-GARCH model.
TAR: 1.86877566991823, ARMA-GARCH: 4.16333424703586
SETAR is the best among nonlinear time series models.

Final model:

The best time series model is selected in TAR models.

```
> fit
```

Title:

Automatic time series model selection

Series:

y

Call:

Ats(x = y, method = "TB", n.ahead = 5, verbose = TRUE)

Best model:

SETAR model with 2 regimes

Self exciting threshold autoregressive (SETAR) model

Coefficient(s):

Low regime:

const L phiL.1
1.122 0.800

High regime:

const H phiH.1 phiH.2
-0.352 -1.480 0.539

Threshold

Variable: $Z(t) = + (1) X(t) + (0) X(t-1)$

Value: 0.09068 (fixed)

Proportion of points in low regime: %

For our third example, we generate data from ARMA(1, 1)-GARCH(1, 1). As per the best-model-based method, LSTAR with 2 regimes gives the smallest prediction error. However, since normality

does not hold for LSTAR, the second-best model ARMA(1, 1)-GARCH(1, 1) is recommended.

```
> set.seed(326)
> spec <- garchSpec(model= list(omega=1.0e-5, alpha=0.2,
+   beta=0.7, ar=0.2, ma=.7))
> x <- garchSim(spec, n = 200)
> fit <- Ats(x, method="BM", model.set="all", n.ahead=20, verbose=TRUE)
Time series model is selected by best-model-based method.

Model set: Linear          | Non-Linear
           Exponential, ARIMA | TAR, ARMA-GARCH

Linear time series model:
ARIMA(1,0,1) with zero mean is the best among ARIMA models.
ETS(A,N,N) is the best among exponential smoothing models.
Prediction errors:
  Prediction errors of the best ARIMA and exponential smoothing models.
  ARIMA: 5.65567781050125e-05, Exponential smoothing: 0.000216145645675635
  ARIMA(1,0,1) with zero mean is the best among linear time series models.

Nonlinear time series model:
LSTAR with 2 regimes is the best among TAR models.
ARMA(1,1)+GARCH(1,1) is the best among ARMA-GARCH models.
Prediction errors:
  Prediction errors of the best TAR and ARMA-GARCH model.
  TAR: 5.09809491035092e-05, ARMA-GARCH: 5.64501607316534e-05
  LSTAR is the best among nonlinear time series models.

Normality test for residuals:
Standardize residuals don't seem to be normal distribution.
So we recommend another nonlinear time series models.
We recommend ARMA-GARCH models.
ARMA(1,1)+GARCH(1,1) is the best among nonlinear time series models.

Final model:
The best time series model is selected in ARMA-GARCH models.

> fit
Title:
Automatic time series model selection

Series:
x

Call:
Ats(x = x, method = "BM", model.set = "all", n.ahead = 20, verbose = TRUE)
```

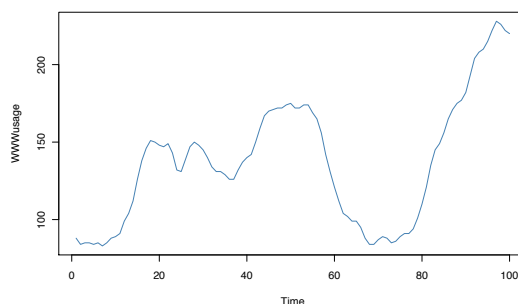



Figure 1: Time series plot for WWWusage data.

Best model:

ARMA(1,1)+GARCH(1,1)

Generalized autoregressive conditional heteroskedasticity (GARCH) model
with ARMA mean structure

Coefficient(s):

mu	ar1	ma1	omega	alpha1	beta1
7.981037e-05	3.033591e-01	8.029663e-01	1.145582e-10	8.464763e-01	4.450823e-01

4.2. Application to real data

In this section, through the analysis of two real data sets, the best models for which are known, we demonstrate the effectiveness of the proposed algorithms. The first real data, WWWusage data, is a time series of the number of users connected to the internet through a server every minute. This data was originally analyzed and modeled as ARIMA(3, 1, 0) by Makridakis *et al.* (1998). The other real data, djiclose data, is the closing values of the Dow Jones index on the first day of each month between October 1928 and December 2007. This data is available in R package expsmooth and its log yield was fitted as local-level GARCH with drift model by Hyndman *et al.* (2008).

The tree-based method is applied to choose the best model from among the linear time series model for WWWusage data displayed in Figure 1. We perform the function `Ats` as follows:

```
> fit <- Ats(WWWusage, model.set="linear", verbose=TRUE)
Time series model is selected by tree-based method.
```

Linear time series model:

ARIMA(3,1,0) is the best among ARIMA models.

ETS(A,Ad,N) is the best among exponential smoothing models.

Prediction errors:

Prediction errors of the best ARIMA and exponential smoothing models.

ARIMA: 0.643275512057917, Exponential smoothing: 4.00079892318515

ARIMA(3,1,0) is the best among linear time series models.

Nonlinearity test:

Nonlinearity test accepts null hypothesis.
Time series data seems to have linear structure.

Final model:

The best time series model is selected in ARIMA models.

ARIMA(3, 1, 0) and ETS(A,Ad,N) are selected from among the ARIMA and exponential smoothing models, respectively. As per the prediction errors, ARIMA(3, 1, 0) is finally selected as the best linear time series model: this result matches with that of Makridakis *et al.* (1998). The `Ats` object provides the information pertaining to the selected model.

```
> fit
```

Title:

Automatic time series model selection

Series:

WWWusage

Call:

```
Ats(x = WWWusage, model.set = "linear", verbose = TRUE)
```

Best model:

ARIMA(3,1,0)

Autoregressive integrated moving average (ARIMA) model

Coefficient(s):

ar1	ar2	ar3
1.1512	-0.6658	0.3459

Furthermore, the function `plot` provides various diagnostic plots such as ACF, PACF, and residual plot, so that the user can easily check the model adequacy and obtain more information about the features of the data. In addition, it is possible to draw plots of interests selectively. The following code depicts Figure 2, which shows the various time series plots of `WWWusage` data modeled by ARIMA(3, 1, 0):

```
> par(mfrow=c(3, 4)); plot(fit)
```

Make a plot selection (or 0 to exit):

1: Time Series	2: ACF of Observations
3: ACF of Squared Observations	4: Cross Correlation
5: Residuals	6: Standardized Residuals
7: ACF of Standardized Residuals	8: ACF of Squared Standardized Residuals
9: Cross Correlation between r^2 and r	10: QQ-Plot of Standardized Residuals
11: PACF of Observations	12: PACF of Standardized Residuals

Selection:

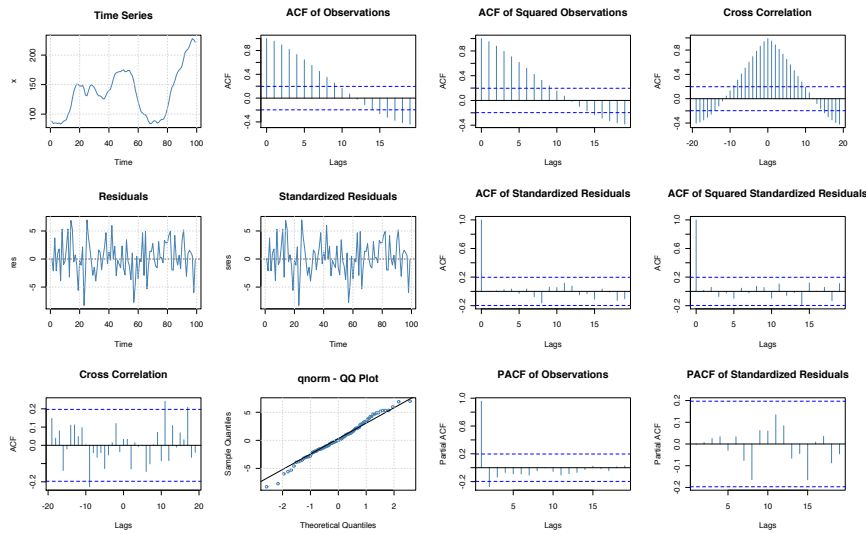


Figure 2: Various plots for the *WWWusage* data modeled by $ARIMA(3, 1, 0)$.

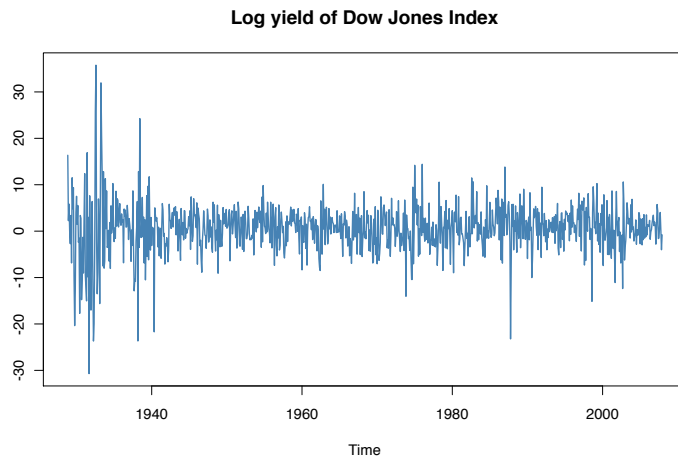


Figure 3: Log yield of the closing values of the Dow Jones index on the first day of each month between October 1928 and December 2007.

Next, we analyze the log yield of the closing values of the Dow Jones index on the first day of each month. The ARCH or GARCH models might fit well since Figure 3 shows that the log yield of *djiclose* does not have equal variances over time. We run a tree-based model selection by following codes that produce the description of the selection process and fitting results.

```
> library(expsmooth)
> data(djiclose)
> fit <- Ats(djiclose[-1,2], method="TB", n.ahead=20, verbose=TRUE)
Time series model is selected by tree-based method.
```

Linear time series model:

ARIMA(3,0,2) with non-zero mean is the best among ARIMA models.

ETS(A,N,N) is the best among exponential smoothing models.

Prediction errors:

Prediction errors of the best ARIMA and exponential smoothing models.

ARIMA: 5.918854131169, Exponential smoothing: 5.95056377136318

ARIMA(3,0,2) with non-zero mean is the best among linear time series models.

Nonlinearity test:

Nonlinearity test rejects null hypothesis.

Time series data seems to have nonlinear structure.

Nonlinear time series model:

LSTAR with 2 regimes is the best among TAR models.

ARMA(1,0)+GARCH(2,1) is the best among ARMA-GARCH models.

Prediction errors:

Prediction errors of the best TAR and ARMA-GARCH model.

TAR: 5.94608046772321, ARMA-GARCH: 5.92486371666828

ARMA(1,0)+GARCH(2,1) is the best among nonlinear time series models.

Final model:

The best time series model is selected in ARMA-GARCH models

```
> fit
```

Title:

Automatic time series model selection

Series:

djiclose[-1, 2]

Call:

Ats(x = djiclose[-1, 2], method = "TB", n.ahead = 20, verbose = TRUE)

Best model:

ARMA(1,0)+GARCH(2,1)

Generalized autoregressive conditional heteroskedasticity (GARCH) model
with ARMA mean structure

Coefficient(s):

mu	ar1	omega	alpha1	alpha2	beta1
6.385234e-01	-2.496118e-03	2.856601e-05	8.690347e-02	1.389224e-01	8.103082e-01

In the first stage, ARIMA(3,0,2), is selected with the smallest 20-ahead prediction error from among the linear model class. The nonlinearity test of residuals from ARIMA(3,0,2) indicates that a nonlinear model is required to explain the behavior of the log yield. In the second stage, the

ARMA(1,0)-GARCH(2,1) model is selected as the final model for the log yield of `djiclose` since the 20-ahead prediction errors of TAR and ARMA-GARCH are 5.946 and 5.925, respectively.

5. Conclusion

In this paper, we have proposed two automatic procedures to select the best time series model. The proposed procedures combine the classical model selection methods and newly proposed model selection algorithms. The results from the real data analyses and the case studies suggest that the proposed methods are promising. The main contribution of the proposed methods is that they can also be applied to nonlinear model classes. Therefore, the proposed algorithms have widespread acceptability and can be used for the model selection of various time series data. Furthermore, we provide an R package `autots` to implement the proposed algorithms. Further improvements can be made by including the various model selection criteria other than AIC and more nonlinear time series models; however, these topics are left for future research.

References

- Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity, *Journal of Econometrics*, **31**, 307–327.
- Brock, W. A., Dechert, W. D. and Sheinkman, J. A. (1987). *A Test of Independence Based on the Correlation Dimension*, Working paper no. 8702, Department of Economics, University of Wisconsin, Madison.
- Brock, W. A., Dechert, W. D., Sheinkman, J. A. and LeBaron, B. (1996). A test for independence based on the correlation dimension, *Econometric Reviews*, **15**, 197–235.
- Engle, R. F. (1982). Autoregressive conditional heteroskedasticity with estimates of the variance of United Kingdom inflation, *Econometrica*, **50**, 987–1007.
- Gómez, V. and Maravall, A. (1998). *Programs TRAMO and SEATS, Instructions for the Users*, Working paper 97001, Dirección General de Análisis y Programación Presupuestaria, Ministerio de Economía Hacienda.
- Hyndman, R. J. and Khandakar, Y. (2008). Automatic time series forecasting: The forecast package for R, *Journal of Statistical Software*, **27**, 1–22.
- Hyndman, R. J., Koehler, A. B., Ord, J. K. and Snyder, R. D. (2008). *Forecasting with Exponential Smoothing*, Springer, Berlin.
- Liu, L. M. (1989). Identification of seasonal ARIMA models using a filtering method, *Communications in Statistics, Part A Theory & Methods*, **18**, 2279–2288.
- Makridakis, S., Wheelwright, S. C. and Hyndman, R. J. (1998). *Forecasting: Methods and Applications*, John Wiley & Sons, New York.
- Mélard, G. and Pasteels, J. M. (2000). Automatic ARIMA modeling including intervention, using time series expert software, *International Journal of Forecasting*, **6**, 497–508.
- Tiao, G. C. and Tsay, R. S. (1994). Some advances in nonlinear and adaptive modeling in time series, *Journal of Forecasting*, **13**, 109–131.
- Tong, H. (1990). *Non-Linear Time Series: A Dynamical Systems Approach*, Oxford University Press, Oxford.