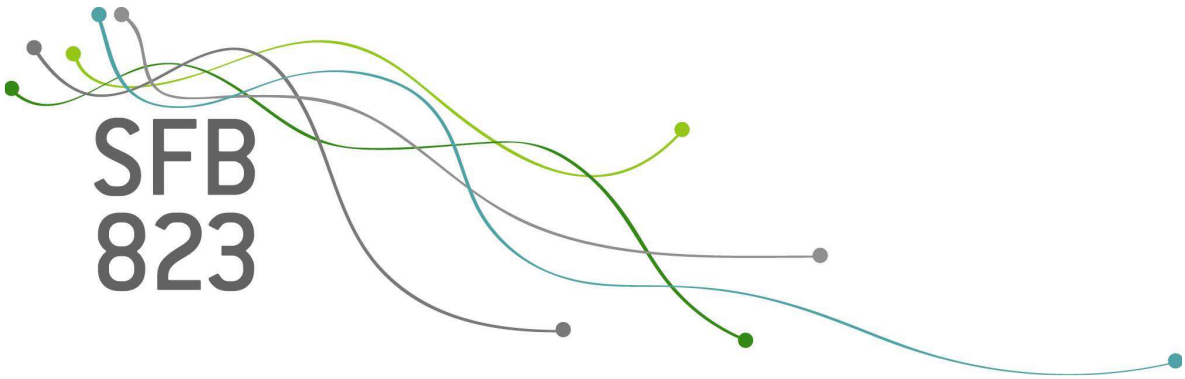
SFB 823

Discussion Paper

Model order selection for cascade autoregressive (CAR) models

Steffen Köhler

Nr. 3/2021



Model Order Selection for Cascade Autoregressive (CAR) Models

Steffen Köhler*

January 13, 2021

Abstract

In recent years, Cascade Autoregression (CAR) models enjoy increasing popularity in applied econometrics. This is due to the fact that they are able to approximate both short- and long-memory processes and are easy to implement. However, their model order, namely the timing of the steps, relies on ad-hoc decisions rather than being data-driven. In this paper, techniques for model order selection of CAR models in finite samples are presented. The approaches are evaluated in an extensive simulation study, as well as in an empirical application. The results suggest that model order selection may provide gains in both in- and out-of-sample performance.

Keywords: cascade models, HAR, model selection, long-memory, time-series modeling, forecasting, cross-validation

JEL: C01, C13, C22, C52, C58

Acknowledgements: The author is grateful for very constructive comments and suggestions by Vasyl Golosnoy and the participants of the 2nd UA Ruhr Econometrics Seminar in Essen. This work was partly supported by the Collaborative Research Center 'Statistical modeling of nonlinear dynamic processes' (SFB823, project A1) of the German Research Foundation (DFG).

1 Introduction

Over the last decades, it has become a goal to model an underlying time-series as parsimoneously as possible: In theory, any covariance-stationary process can be represented by an autoregressive (AR) or moving average (MA) model of infinite order. In practice however, estimation of AR or MA models of higher order entails the estimation of many unknown parameters. This in turn leads to high estimation risk and thus, decreases the forecasting performance of the AR and MA model. Therefore, modern model specification follows the major

*Faculty of Management and Economics, Ruhr-Universität Bochum, *E-mail:* steffen.koehler@rub.de

principles coined by Box and Jekins (1970). Parsimoneous specifications (i.e. including fewer past observations in the model) are preferred over less sparse ones. This typically involves determining the model order by the means of information criteria, or using mixtures of AR and MA models (ARMA). For a more recent review of model selection in time series, see e.g. Leeb and Pötscher (2009).

Despite their popularity, there are however drawbacks of ARMA models which are of theoretical and practical relevance. First, in contrast to pure AR models, ARMA models do not possess the Markov property, so they can not be used for applications that require this property. Second, they are used only for short-memory processes, that are characterized by an exponentially decaying autocovariance function (ACVF). Therefore, they are not able to capture the dynamics of long-memory processes with hyperbolically decaying ACVF.
One example of a long-memory process is the measure of realized volatility (RV), a proxy for the true volatility of e.g. daily stock returns. To approximate long-memory dynamics, the order of the ARMA model must be rather large, leading to the same problem of overfitting that simple AR or MA models suffer from (see Brockwell and Davis 1991).
To overcome this shortcoming, Granger and Joyeux (1980) proposed the autoregressive fractionally integrated moving average (ARFIMA) model. Although being a true long-memory model, estimation of ARFIMA models is not straightforward and prediction accuracy depends strongly on the underlying process (see Crato and Ray 1996, Baille et al. 2012 and the references therein). As an alternative to ARFIMA, cascade autoregressive (CAR) models gained in popularity, coined by the heterogeneous autoregressive model for realized volatility (HAR-RV) of Corsi (2009). It was specifically designed for RV time-series, but can be applied to other time-series as well. CAR models result from imposing linear restrictions on AR models of large order. Hence, an important advantage over conventional AR models is that CAR models are able to parsimoniously reproduce the hyperbolic (exponential) decay of the ACVF of long-memory (short-memory) processes. In theory, this allows the use of CAR models as an easy-to-implement approximation of many different processes.
Aside from being sparsely parameterized, CAR models offer additional benefits to practitioners: since they belong to the class of AR models, they possess the Markov property, can be easily estimated by Ordinary Least Squares (OLS) and thus, a broad range of model diagnostics can be used.

Because the HAR model of Corsi (2009) is the most known special case of CAR and is primarily applied to RV time-series, there exist different extensions of the model that are tailored to this purpose: Andersen et al. (2007) amended the formulation by including a jump component, resulting in the HAR-RV-J model. Corsi et al. (2010) also included a jump component into the model, but moreover, accounted for the fact that RV is an inconsistent estimate of integrated volatility in the presence of jumps. The LHAR-CJ model by Corsi and Renò

(2012) includes another factor - aside from the jump component - that allows for an asymmetric response of the variance to positive or negative shocks. Patton and Sheppard (2015) account for this asymmetry by using realized semivariances and present a panel version of the HAR model. A different modification was introduced by McAleer and Medeiros (2008), that allows for smooth transitioning between multiple regimes of the coefficients. For a literature overview, see Wen et al. (2016) and Corsi et al.. (2012).

All these extensions have in common, that the general order of the HAR model remains unchanged: In the standard HAR model, the weekly and monthly averages, together with today's observation determine tomorrow's value of the time series. This specification of the model order is an implicit assumption that is theoretically based only on the heterogeneous market hypothesis of Müller et al.. (1997). Also, this theoretical justification is valid for financial time series only. Despite this, in empirical finance applications (cf. Bekaert and Hoerova 2014, Golosnoy et al. 2019, Wilms et al.. 2020), the standard HAR order prevails.

However, using a specific order of a CAR model has a large impact on the model's dynamics. Hence, imposing one ad-hoc specification to fit all applications might not be appropriate.

To illustrate this, figure 1 shows the autocorrelation function (ACF) of two equally persistent CAR models of different order. Both processes have the same coefficient vector $(0.15, 0.2, 0.6)'$, but include different lagged

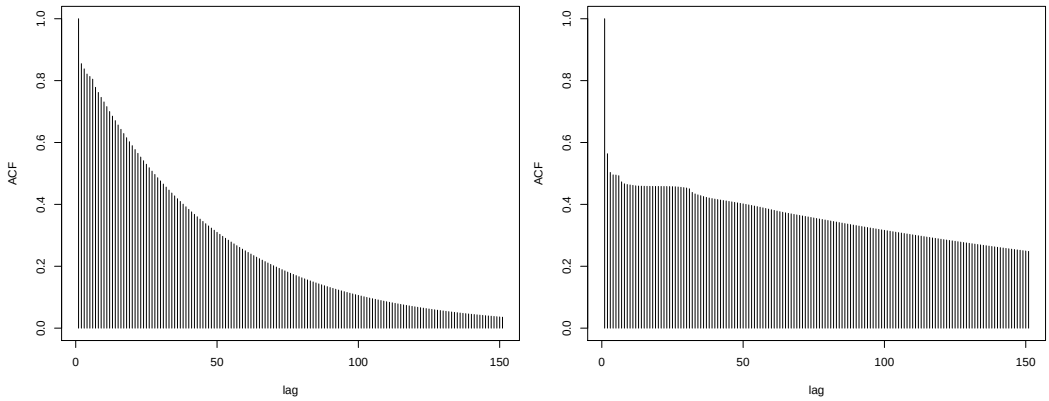


Figure 1: ACFs of CAR processes with exponential (left) and hyperbolical (right) decay.

means. The left process uses averages over the last two and five days, which results in an exponentially decaying ACF. The right, hyperbolically decaying ACF results from a CAR model that uses averages over the last five and 30 days. For the latter process, the cascade-like pattern with distinct steps can be observed clearly.

This example shows that specifying the order of a CAR model involves two decisions; First, the number of cascades (averages) included in the model need to be set. Second, one needs to decide how long these steps should be, i.e. over which time-horizon the averages should be calculated.

In this paper, ways how to select the order of a CAR model in a data-driven way are presented. This ap-

proach has two implications: First, it makes the CAR model applicable outside of empirical finance, because the structure is no longer driven by empirical finance considerations. Second, even in financial applications, the standard HAR specification may not fit all realized volatility time-series. Applying model selection allows us to discover those divergences. Audrino and Knaus (2016) and Audrino et al. (2018) also presented a procedure for CAR model selection based on shrinkage. However, estimation is computationally expensive and thus, deprives the CAR model of its main advantages – its simplicity. We therefore restrict ourselves to selection techniques which are already known from autoregressive models. Furthermore, they do not require distributional assumptions, because in practice, the time-series' true distribution is typically unknown.

The remainder of the paper is organized as follows. In section 2, we present the class of linear CAR models and analyze their stochastic properties. In section 3 we present approaches for selecting the CAR model order which are based on either least squares (LS), general method of moments (GMM), or cross-validation. The performance of the model selection procedures is illustrated in an extensive Monte Carlo study in section 4, where we investigate the ability of our procedures to find the correct CAR model order. We further evaluate the approaches in an empirical application in section 5. In section 6 we will summarize the findings and give a brief outlook over the implications of the results and possible extensions.

2 Linear CAR models

Let y_t be a covariance-stationary process observed at time $t = 1, ..., T$. If y_t follows a CAR model, then it is defined as the weighted sum of q past averages plus an intercept. The averages are constructed over windows of different lengths. The length of window i is denoted by s_i, so the corresponding average is defined as

$$\bar{y}_{t-1,s_i} = \frac{1}{s_i} \sum_{j=1}^{s_i} y_{t-j}, \qquad i = 1, ..., q. \tag{1}$$

The number of averages q is also referred to as the number of 'steps', while s_i is the 'width' of the i-th step. Without loss of generality, we assume that $s_1 < s_2 < ... < s_q$.

The order of a CAR model is therefore given by the parameter q and the $q \times 1$ dimensional vector $\mathbf{s} = (s_1, ..., s_q)'$. Then, the CAR$(q, \mathbf{s})$ model reads as

$$y_t = \mu + \delta_1 \bar{y}_{t-1,s_1} + ... + \delta_q \bar{y}_{t-1,s_q} + u_t, \qquad u_t \sim \text{iid}(0, \sigma^2), \tag{2}$$

$$y_t = \mu + \boldsymbol{\delta}' \bar{\mathbf{y}}_{t-1} + u_t, \qquad \text{with} \qquad \bar{\mathbf{y}}_{t-1} = (\bar{y}_{t-1,s_1}, ..., \bar{y}_{t-1,s_q})'. \tag{3}$$

The coefficient vector $\delta = (\delta_1, ..., \delta_q)'$ contains the weights of the corresponding averages. A prominent example of a CAR model is the standard HAR of Corsi (2009) which is specified by $q = 3$ and $\mathbf{s} = (1, 5, 22)'$.

As mentioned earlier, CAR$(q, \mathbf{s})$ could be seen as a restricted AR(p) model with $p = s_q$.

The AR(s_q) representation with coefficient vector $\boldsymbol{\phi}$ can be easily obtained through (4):

$$\mathbf{R}'\boldsymbol{\delta} = \boldsymbol{\phi} = (\phi_1, ..., \phi_{s_q})', \tag{4}$$

where $\mathbf{R}$ is a $q \times s_q$ matrix with the i-th row given by $(\iota'_{s_i}, \mathbf{0}'_{(s_q - s_i)})/s_i$, with ι_{s_i} ($\mathbf{0}_{(s_q - s_i)}$) being a column vector of length s_i ($s_q - s_i$) containing only ones (zeros).

For example, in the standard HAR model, the matrix $\mathbf{R}$ is equal to

$$\begin{pmatrix} 1 & 0 & \cdots & & & & \cdots & 0 \\ 1/5 & 1/5 & 1/5 & 1/5 & 1/5 & 0 & \cdots & 0 \\ 1/22 & \cdots & & & & & \cdots & 1/22 \end{pmatrix}.$$

Properties of linear CAR models

Using the transformation from (4), stationarity can be easily checked; using $\boldsymbol{\phi}$, we can construct the characteristic polynomial

$$\boldsymbol{\phi}(z) = 1 - \phi_1 z - ... - \phi_{s_q} z^{s_q}.$$

If the roots of $\boldsymbol{\phi}(z) = 0$ all lie outside the unit-circle, the CAR model is stationary.

If a CAR model is found to be stationary, its first and second moments can also be calculated via the AR(s_q) representation. Then, the expectation of y_t is given as

$$E(y_t) \quad = \quad \mu/(1 - \boldsymbol{\phi}'\iota_{s_q}).$$

For presentation convenience and without loss of generality, we set $\mu = 0$ for the remainder of this paper. This imposes no restriction, since this property can always be ensured by previously de-meaning the time-series. Similarly, the autocovariance function (ACVF) of y_t at lag ℓ, $\gamma_\ell = E(y_t y_{t-\ell})$ can be obtained. It is given by the Yule-Walker equations:

$$\gamma_\ell = \boldsymbol{\gamma}'_{\ell-1}\mathbf{R}'\boldsymbol{\delta}, \qquad \text{with} \qquad \boldsymbol{\gamma}_{\ell-1} = (\gamma_{\ell-1}, ..., \gamma_{\ell-s_q})' \quad \text{for } \ell = 0, 1, ... \tag{5}$$

They allow to determine the complete ACVF and autocorrelation function (ACF) of a CAR model, conditional on $\boldsymbol{s}$ and $\boldsymbol{\delta}$.

One advantage of CAR models is that, depending on the model order, they are able to approximate exponential and hyperbolic decay of the ACVF. For example, setting $s_q = 1$ and $q = 1$ results in a AR(1) model with exponentially decaying ACVF. On the other hand, if s_q increases, the decay of the ACVF becomes more hyperbolic.

3 Selecting CAR model order

CAR models are parameterized through q, $\mathbf{s} = (s_1, ..., s_q)'$ and $\delta = (\delta_1, ..., \delta_q)'$. Conditional on $\mathbf{s}$ and q, estimation of $\boldsymbol{\delta}$ is easily carried out via an ordinary least squares (OLS) regression of (3). Therefore, we focus on data-driven ways to select $\mathbf{s}$ and q. This selection can be divided into three steps:

First, s_q is of special interest, because it determines the order of the underlying restricted AR model. Section 3.1 presents techniques to estimate s_q. Having estimated s_q, we need to determine in a second step the remaining paremeters of $\boldsymbol{s}$, conditional on a pre-specified value q. The selection procedures for $\mathbf{s}$ are presented in section 3.2. Thirdly, q needs to be estimated from the data as well, by repeating step two for different values of q and deciding, which q fits the data best. This is presented in section 3.3.

3.1 Selection of s_q

Although it is of high importance for CAR model selection, there is little theoretical guidance, how to estimate s_q. One asymptotically valid criterion for choosing s_q is that for an AR(s_q) model, the partial autocorrelation function (PACF) drops to zero at lags larger than s_q. Unfortunately, this approach is not feasible in finite samples, because it is highly unlikely that the PACF is exactly zero at some lags. In a similar manner, Audrino et al.. (2015) presented a test for the coefficients of an unrestricted AR model of large order; after lag s_q, the coefficients should drop to zero. Unfortunately, they show that this test is too conservative for most applications. However, as figure 1 indicates, in order to approximate long memory behavior, it is recommendable to select s_q not too small. Therefore, possible ad-hoc values for s_q exist.

For example, the HAR model fixes $s_q = 22$ so it refers to a monthly component, if data were sampled daily. Another candidate for s_q could be the criterion of Schwert (1989), that sets $s_q = 12\lfloor \sqrt[4]{T/100} \rfloor$, with $\lfloor x \rfloor$ denoting the integer part of x. Although it is primarily used in unit-root tests, it may be of interest in the setting of CAR models, too.

Since we are interested in estimating the parameters from data, we focus on the selection of s_q via information criteria. Two popular information criteria for AR models are Akaike's (AIC, cf. Akaike 1973, 1974) or Bayes' (BIC, cf. Schwartz 1978): Several AR models of increasing order $p = 1, ..., p_{max}$ are fitted to the data[1] and s_q is equal to the value p that minimizes

$$IC(p) = \ln(\widehat{\sigma}_p^2) + \mathcal{C}_T \frac{p + \eta_T(p)}{T - p_{max}} \tag{6}$$

with $\widehat{\sigma}_p^2$ being the estimated residual variance of the fitted AR(p) model. Fixing $\eta_T(p) = 1$ and setting $\mathcal{C}_T = 2$ yields the AIC, while $\eta_T(p) = 1$ and $\mathcal{C}_T = \ln(T - p_{max})$ results in the BIC.

[1] For estimation, we use the sample from $p_{max} + 1$ to T and use the remaining values as starting values. This way, the sample size is constant for all AR(p) models.

However, in finite samples for processes with high persistency, Ng and Perron (2001) show that s_q is likely to be underestimated, so they introduced modified information criteria by allowing $\eta_T(p)$ to depend on sample statistics. These modified criteria are not calculated based on AR models. Instead, a Dickey-Fuller (cf. Dickey and Fuller, 1979 and Said and Dickey, 1984) autoregression is fitted to the data for increasing values of $p = 1, ..., p_{max}$:

$$\Delta y_t = \widehat{\beta}_0 y_{t-1} + \sum_{i=1}^{p} \widehat{\beta}_i \Delta y_{t-i} + \widehat{e}_{t,p}.$$

The parameters used in (6) are then given by

$$\begin{aligned} \widehat{\sigma}_p^2 &= \frac{1}{T - p_{max}} \sum_{t=p_{max}+1}^{T} \widehat{e}_{t,p}^2 \\ \eta_T(p) &= \widehat{\beta}_0^2 / \widehat{\sigma}_p^2 \cdot \sum_{t=p_{max}+1}^{T} y_t^2. \end{aligned}$$

Setting $\mathcal{C}_T = 2$ then yields the modified Akaike information criterion (MAIC) and $\mathcal{C}_T = \ln(T - p_{max})$ results in the modified BIC (MBIC). Ng and Perron (2001) showed that values of s_q found by minimizing the modified information criteria can be expected to be much larger than those found by AIC and BIC. Although the primary use of the estimates was unit-root testing, Ng and Perron (2001) highlighted the relevance of their findings for other applications, so we use them in the context of CAR models. Throughout the applications of this paper, we set the maximum number of lags to $p_{max} = 50$.

3.2 Selection of s

For a given number of steps $q \geq 3$ and s_q, $q - 1$ integer step widths need to be selected. It is reasonable to assume that the most recent observation has a distinctive influence on today's value of y_t. Therefore we fix $s_1 = 1$, so $q - 2$ unknown parameters remain.[2] These parameters are collected in the vector $\mathfrak{s}_q = (s_2, ..., s_{q-1})'$. To emphasize that $\mathfrak{s}_q$ controls the linear restrictions, the restriction matrix from (4) will be denoted $\mathbf{R}(\mathfrak{s}_q)$. The elements of $\mathfrak{s}_q$, s_i are integer-valued, limited from above by s_q and ordered, so it holds that $s_{i-1} < s_i < s_{i+1}$. Therefore, the amount of possible values of $\mathfrak{s}_q$ is limited. Denote the set of all possible combinations $\mathcal{S}_q$, with $\mathfrak{s}_q \in \mathcal{S}_q$, so that in general, there are $|\mathcal{S}_q| = \binom{s_q - 2}{q-2}$ possible combinations in total.

To illustrate this, consider the following example: for $q = 3$, only one width, s_2, needs to be determined and the set of all possible values of s_2 is $\mathcal{S}_q = \{2, 3, ..., s_q - 1\}$. In total, there are $|\mathcal{S}_q| = s_q - 2$ possibilities, whereas for $q = 4$ there are two steps (s_2, s_3) with $(s_q - 2)(s_q - 3)/2$ possibilities.

We contrast three possible methods for obtaining an estimate of the model order, $\widehat{\mathfrak{s}}_q$. First, we present a

[2] CAR models with $q = 1$ or $q = 2$ are trivial special cases, because they are completely determined by s_1 and s_q.

least squares (LS) approach which is based on minimizing the residual sum of squares (RSS). LS produces an estimate that maximizes the in-sample fit. Second, the step widths could be obtained from minimizing the distance between the empirical ACVF and the theoretical ACVF of the CAR, which ultimately results in GMM estimation. This approach is also similar to Yule-Walker estimation of unrestricted AR models. In applications with focus on predictions, selection based on Cross-Validation may be more appropriate, so we present the approach of Bergmeir et al.. (2018) as a third possible way to obtain an estimate of $\mathfrak{s}_q$.

Least Squares

Let $\mathbf{x}_t = (y_{t-1}, ..., y_{t-s_q})'$ contain lagged observations of y_t. For given q, s_q and conditional on each $\mathfrak{s}_q \in \mathcal{S}_q$, we are able to obtain $\widehat{\boldsymbol{\delta}}(\mathfrak{s}_q)$, the (conditional) OLS estimate of $\boldsymbol{\delta}$. The RSS conditional on $\mathfrak{s}_q$ is given by

$$RSS(\mathfrak{s}_q) = \sum_{t=1}^{T} (y_t - \widehat{\boldsymbol{\delta}}(\mathfrak{s}_q)' \bar{\mathbf{y}}_{t-1})^2. \tag{7}$$

The estimate $\widehat{\mathfrak{s}}_q$ is defined as the value that minimizes $RSS(\mathfrak{s}_q)$ among all possible combinations.

$$\widehat{\mathfrak{s}}_q = \arg \min_{\mathfrak{s}_q \in \mathcal{S}_q} RSS(\mathfrak{s}_q).$$

Selection Based on Wald Statistic

While LS is concerned with minimizing RSS, this approach minimizes the distance between the empirical ACVF and the restricted ACVF resulting from the CAR model.

Since y_t is a stationary process, the elements of the ACVF at lag ℓ can be consistently and non-parametically estimated by $\widehat{\gamma}_\ell = \frac{1}{T} \sum_{\ell+1}^{T} y_t y_{t-\ell}$. Let $\widehat{\boldsymbol{\gamma}} = (\widehat{\gamma}_1, ..., \widehat{\gamma}_{s_q})'$ denote y_t's estimated ACVF up to lag s_q. Moreover, let $\widehat{\boldsymbol{\gamma}}(\mathfrak{s}_q)$ be the ACVF up to lag s_q, resulting from an estimated CAR model with given $\mathfrak{s}_q$. If a CAR model is the true data-generating process (DGP), the two quantities should be approximately equal, i.e.

$$\widehat{\boldsymbol{\gamma}} - \widehat{\boldsymbol{\gamma}}(\mathfrak{s}_q) \approx \mathbf{0}_{s_q}. \tag{8}$$

Using the Yule-Walker equation from (5) in vector notation, conditional on $\mathfrak{s}_q$, the estimated ACVF of the CAR model up to lag s_q is defined by

$$\widehat{\boldsymbol{\gamma}}(\mathfrak{s}_q) = \widehat{\boldsymbol{\Gamma}} \mathbf{R}(\mathfrak{s}_q)' \widehat{\boldsymbol{\delta}}(\mathfrak{s}_q),$$

with $\widehat{\boldsymbol{\Gamma}}$ being a $s_q \times s_q$ symmetric Toeplitz matrix defined by the vector $(\widehat{\gamma}_0, ..., \widehat{\gamma}_{s_q-1})'$.

As stated earlier, the difference in (8) is not exactly zero, due to the restriction matrix. Moreover, the distance might not be equal over all lags of the ACVF. Hence, it would be optimal to have a scalar distance measure that accounts for both. To this end, we first need asymptotic (model-free) properties of the ACVF of y_t:

Using Bartlett's formula (cf. Schlittgen and Streitberg 2001), it is known that the empirical ACVF, $\widehat{\boldsymbol{\gamma}}$, is asymptotically normally distributed,

$$\widehat{\boldsymbol{\gamma}} \overset{a}{\sim} N(\boldsymbol{\gamma}, \Sigma). \tag{9}$$

The element in the i-th row and j-th column of the $s_q \times s_q$ covariance matrix Σ is defined by:

$$\Sigma_{i,j} = \frac{1}{T} \sum_{\ell=-\infty}^{\infty} \gamma_\ell \gamma_{\ell+i-j} + \gamma_{\ell-j}\gamma_{\ell+i} \tag{10}$$

In finite samples, it is not possible to compute the elements of Σ, because of the sum with limits $-\infty$ and ∞. However, since y_t is stationary, we know that $\gamma_{-\ell} = \gamma_\ell$ converges towards zero, as ℓ increases. Therefore, it is advisable to choose a value $\ell_{max} < \infty$, up to which the ACVF is calculated. ℓ_{max} should be chosen in a way that the ACVF is close to zero.[3] Using the empirical ACVF instead of the theoretical, $\Sigma_{i,j}$ can be estimated by

$$\widehat{\Sigma}_{i,j} = \frac{1}{T} \sum_{\ell=-\ell_{max}}^{\ell_{max}} \widehat{\gamma}_\ell \widehat{\gamma}_{\ell+i-j} + \widehat{\gamma}_{\ell-j}\widehat{\gamma}_{\ell+i}. \tag{11}$$

Using these results, the distance between $\widehat{\boldsymbol{\gamma}}$ and $\boldsymbol{\gamma}(\mathfrak{s}_q)$ is then measured by the Wald statistic, which is equal to the squared Mahalanobis distance

$$W(\mathfrak{s}_q) = \left[\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}(\mathfrak{s}_q)\right]' \Sigma^{-1} \left[\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}(\mathfrak{s}_q)\right] = \left[\widehat{\boldsymbol{\gamma}} - \boldsymbol{\Gamma}\mathbf{R}(\mathfrak{s}_q)'\boldsymbol{\delta}\right]' \Sigma^{-1} \left[\widehat{\boldsymbol{\gamma}} - \boldsymbol{\Gamma}\mathbf{R}(\mathfrak{s}_q)'\boldsymbol{\delta}\right]. \tag{12}$$

Note that, if a CAR model – with known parameters s_q, $\mathfrak{s}_q$ and $\boldsymbol{\delta}$ – is the true DGP, then $W(\mathfrak{s}_q)$ follows a χ^2-Distribution with q degrees of freedom.

$W(\mathfrak{s}_q)$ from (12) relies on true values which have to be replaced by estimates in practice. $\boldsymbol{\Gamma}$ and Σ are replaced by their empirical and unrestricted counterparts, $\widehat{\boldsymbol{\Gamma}}$ and $\widehat{\Sigma}$. Then, the distance measured by (12), depends only on $\mathbf{R}(\mathfrak{s}_q)$ and the conditional OLS estimate $\widehat{\boldsymbol{\delta}}(\mathfrak{s}_q)$, obtained from the regression in (3).

The feasible version of (12) is therefore given by

$$W(\mathfrak{s}_q) = \left[\widehat{\boldsymbol{\gamma}} - \widehat{\boldsymbol{\Gamma}}\mathbf{R}(\mathfrak{s}_q)'\widehat{\boldsymbol{\delta}}(\mathfrak{s}_q)\right]' \widehat{\Sigma}^{-1} \left[\widehat{\boldsymbol{\gamma}} - \widehat{\boldsymbol{\Gamma}}\mathbf{R}(\mathfrak{s}_q)'\widehat{\boldsymbol{\delta}}(\mathfrak{s}_q)\right]. \tag{13}$$

If only q and s_q are given, $W(\mathfrak{s}_q)$ can be calculated for all $\mathfrak{s}_q \in \mathcal{S}_q$. The minimum of $W(\mathfrak{s}_q)$ then defines the optimal value $\widehat{\mathfrak{s}}_q$:

$$\widehat{\mathfrak{s}}_q = \underset{\mathfrak{s}_q \in \mathcal{S}_q}{argmin}\ W(\mathfrak{s}_q) \tag{14}$$

In the following proposition, we will show that $\widehat{\mathfrak{s}}_q$ is a consistent estimate for its true counterpart.

Proposition 1. *Let y_t be a stationary, ergodic process, defined by* (3)*, $\boldsymbol{\delta}_0$ and $\mathfrak{s}_0 = (1, \mathfrak{s}_0, s_q)'$. Let the set of all possible values of $\mathfrak{s}_q$ be $\mathcal{S}_q$, with known q and s_q, such that $\mathcal{S}_q$ is compact and $\mathfrak{s}_0 \in \mathcal{S}_q$. Furthermore, let $\widehat{\mathfrak{s}}_q \in \mathcal{S}_q$ be some estimate for the true vector $\mathfrak{s}_0$. Then $W(\mathfrak{s}_q)$, defined by* (13)*, attains its minimum only at $\widehat{\mathfrak{s}} = \mathfrak{s}_0$.*

Proof. Define $g(\boldsymbol{\phi}) = \widehat{\boldsymbol{\gamma}} - \widehat{\boldsymbol{\Gamma}}\boldsymbol{\phi}$, where $\boldsymbol{\phi}$ is completely determined by $\mathfrak{s}_q$ and $\boldsymbol{\delta}$. Suppose $\widehat{\boldsymbol{\gamma}}$ is estimated by $\gamma_\ell = T^{-1}\sum_{t=\ell+1}^{T}(y_t - \overline{y})(y_{t-\ell} - \overline{y})$, so that $\widehat{\boldsymbol{\gamma}} \xrightarrow{p} \boldsymbol{\gamma}$, $\widehat{\boldsymbol{\Gamma}} \xrightarrow{p} \boldsymbol{\Gamma}$ and $\widehat{\Sigma} \xrightarrow{p} \Sigma$. Then it follows, that minimizing (13) with respect to $\boldsymbol{\delta}$ and $\mathfrak{s}_q \in \mathcal{S}_q$ is equivalent to the Generalized Method of Moments and thus, ensures that $plim\ \widehat{\boldsymbol{\delta}} = \boldsymbol{\delta}_0$ and $plim\ \widehat{\mathfrak{s}}_q = \mathfrak{s}_0$. □

[3] We set $\ell_{max} = 250$ for the simulation study and the empirical application.

Under the Nullhypothesis that y_t follows an CAR model defined by $\mathbf{s} = \mathbf{s}_0$, it holds that $W_s \overset{a}{\sim} \chi^2_{s_q - q}$, which can be used for testing specific CAR models:

$$W_q = W(\widehat{\mathfrak{s}}_q), \quad W_q \overset{a}{\sim} \chi^2_{s_q - q}. \tag{15}$$

Cross-Validation

While LS aims to maximize in-sample fit, cross-validation (CV) maximizes out-of-sample fit. Since y_t is autocorrelated and not iid, standard CV is not applicable, i.e. re-sampling and leaving out random observations would destroy the time-dependency of the series.

Recently, Bergmeir et al. (2018) derived a way of implementing CV for selecting the order of pure AR models. The procedure is similar to the leave-one-out CV and leaves the time structure intact. It can also be used in the context of CAR models, so we apply it to select $\mathfrak{s}_q$.

Although y_t itself is not iid, conditional on $\mathbf{x}_t = (y_{t-1}, ..., y_{t-s_q})'$ it is. Moreover, $y_t|\mathbf{x}_t$ has mean $\boldsymbol{\delta}(\mathfrak{s}_q)'\mathbf{R}(\mathfrak{s}_q)\mathbf{x}_t$ and variance σ^2. Hence, this allows for the removal of n pairs of information $(y_t, \mathbf{x}_t)$ that are chosen at random. These $j = 1, ..., n$ pairs $(y_j^o, \mathbf{x}_j^o)$ form the out-set that will be used for validation. Correspondingly, the remaining pairs are called the in-set. Over the in-set, we use OLS to obtain $\widetilde{\boldsymbol{\delta}}(\mathfrak{s}_q)$, conditional on $\mathfrak{s}_q$.[4] Then, over the out-set, the mean squared forecasting error (MSFE) is calculated by

$$MSFE(\mathfrak{s}_q) = n^{-1} \sum_{j=1}^{n} \left(y_j^o - \widetilde{\boldsymbol{\delta}}(\mathfrak{s}_q)'\mathbf{R}(\mathfrak{s}_q)\mathbf{x}_j^o \right)^2. \tag{16}$$

This procedure is repeated k times, also referred to as folds. For each fold, y_t is partitioned into in- and out-set. The out-set is chosen randomly, but it is ensured that, across all k folds, every pair $(y_t, \mathbf{x}_t)$ is exactly once in the out-set (for details, see Bergmeir et al. 2018). This automatically fixes $n = T/k$.

Over each fold, the MSFE is calculated, resulting in k values of MSFE for each $\mathfrak{s}_q \in \mathcal{S}_q$. Denote $MSFE(\mathfrak{s}_q, i)$ the MSFE of the i-th fold and the model defined by $\mathfrak{s}_q$. Then, $\widehat{\mathfrak{s}}_q$ is the value of $\mathfrak{s}_q$ that produces on average the lowest MSFE over all folds:

$$\widehat{\mathfrak{s}}_q = \underset{\mathfrak{s}_q \in \mathcal{S}_q}{\arg\min} \frac{1}{k} \sum_{i=1}^{k} MSFE(\mathfrak{s}_q, i)$$

Following Bergmeir et al. (2018), we use $k = 5$ in the simulation study and the empirical application.

3.3 Selection of q

Using the methods presented, we are able obtain the estimate $\widehat{\mathbf{s}} = (1, \widehat{\mathfrak{s}}_q, \widehat{s}_q)'$ conditional on a pre-specified value of q. It is worth noting that an increase or decrease in q does not change all elements of $\widehat{\mathbf{s}}$; per construction, the first and last step parameters remain the same, so changes in q only affect $\widehat{\mathfrak{s}}_q$.

[4] In general, $\widetilde{\boldsymbol{\delta}}(\mathfrak{s}_q)$ is different from $\widehat{\boldsymbol{\delta}}(\mathfrak{s}_q)$, because it does not use all the data.

In most applications, $q = 3$ or $q = 4$ should suffice, because one of the main advantages of CAR models is that they are sparse. Increasing q therefore gradually nullifies this advantage, because more parameters, i.e. step width and the corresponding weight, need to be estimated from the data. This in turn leads to lower predictive power of the model. For example, Hwang and Shin (2014) showed that $q = 3$ may already be optimal regarding the trade-off between better approximation and higher estimation risk.

However, in theory, the choice of q is only limited by the relationship $2 \leq q \leq s_q$.[5] Moreover, q can be selected data-driven, which might be desirable for some applications:

The in sample fit of a model with q steps can be measured by RSS_q. Increasing the number of steps decreases on average the residual sum of squares, i.e. $RSS_{q+1} \leq RSS_q$, because the CAR model becomes more flexible. Hence, it may be tempting to include more steps into the model to achieve better in sample results. On the other hand, maximizing in sample fit usually leads to overfitting and thus, deteriorated out of sample forecasting performance.

To obtain a parsimoniously parameterized model that accounts for this trade-off a model selection criterion $\mathcal{P}(q, \widehat{\mathfrak{s}}_q)$ can be used, that penalizes the inclusion of additional steps

$$(\widehat{q}, \widehat{\mathfrak{s}}_{\widehat{q}}) = \arg\min_q \mathcal{P}(q, \widehat{\mathfrak{s}}_q).$$

$(\widehat{q}, \widehat{\mathfrak{s}}_{\widehat{q}})$ are then the CAR model order estimates.

For $\mathcal{P}(q, \widehat{\mathfrak{s}}_q)$, we consider AIC and BIC, which are the most prominent model selection criteria in the context of time-series models. They are given as

$$AIC = \ln\left(\frac{RSS_q}{T - s_q}\right) + 2\frac{q+2}{T - s_q}, \tag{17}$$

$$BIC = \ln\left(\frac{RSS_q}{T - s_q}\right) + \ln(T - s_q) \cdot \frac{q+2}{T - s_q}. \tag{18}$$

The appropriateness of a CAR model with q steps can be tested by using the test statistic given in (15), because if the number of steps is correct, W_q follows a $\chi^2_{s_q - q}$ distribution.

Another way to test the adequacy of the estimated model is to use the standard F-test, which implicitly relies on the assumption that an unrestricted AR(s_q) is the true DGP; denote RSS_q the RSS of the estimated Cascade model and by RSS_{s_q} the RSS of the unrestricted AR(s_q) model. For stationary time-series, the F statistic is then given by

$$F = \frac{RSS_q - RSS_{s_q}}{RSS_{s_q}} \frac{T - 2s_q}{s_q - q} \xrightarrow[T\to\infty]{H_0} F_{s_q - q, T - 2s_q}.$$

Lastly, the adequacy of the model can be verified by the means of residual diagnostics, e.g. the Ljung-Box test (cf. Ljung and Box 1978).

[5] If $q = s_q$, the CAR model is equal to the unrestricted AR(s_q)

4 Monte Carlo study

In this section, we compare the methods for order selection of CAR models, ie. the selection of s_q and $\mathfrak{s}$, in two simulation studies. Throughout the simulations, we use different CAR models following

$$y_t = \delta_1 y_{t-1} + \delta_2 \overline{y}_{t-1,s_2} + ... + \delta_q \overline{y}_{t-1,s_q} + u_t, \quad u_t \overset{iid}{\sim} N(0,1). \tag{19}$$

Hence, each model is completely determined by the vectors $\mathbf{s} = (1, s_2, ..., s_q)'$ and $\boldsymbol{\delta} = (\delta_1, ..., \delta_q)'$. $\boldsymbol{\delta}$ is constructed in a way that it only depends on a persistency parameter $\xi = \sum_{i=1}^{q} \delta_i$. To model long-memory behavior, ξ needs to be close to, but smaller than one[6]. Based on ξ, we define $\boldsymbol{\delta}$:

$$\boldsymbol{\delta} = \xi \cdot \begin{cases} (0.1, 0.45, 0.45)' & \text{for } q = 3 \\ (0.1, 0.3, 0.3, 0.3)' & \text{for } q = 4. \end{cases} \tag{20}$$

4.1 Simulation study for selection procedures for s_q

To compare the performance of the selection of s_q according to the four information criteria (AIC, BIC, MAIC, and MBIC), we simulate the CAR models according to (19) and (20) with two adjustments: first, we fix $q = 3$, because the effect of the number of steps on $\widehat{s}_q$ is assumed to be negligible. The same applies to the value of s_2, so we fix it to $s_2 = \lfloor \frac{s_q}{2} \rfloor$, with $\lfloor x \rfloor$ denoting the integer part of x. Hence, the DGP only depends on ξ and s_q. The different parameterizations are given in table 1.

We draw $B = 10,000$ samples, each of length $T = 1000$, from twelve different CAR models. For each sample, we estimate $\widehat{s}_q$ according to the minimum of the information criteria and obtain the estimation error, i.e. the distance between $\widehat{s}_q$ and s_q. Table 1 reports the Mean-Squared-Error (MSE) of the estimates across all replications.

On average, the MSE increases in s_q for all four information criteria, especially when the processes are more persistent. Very large values of s_q are therefore harder to estimate precisely than smaller values. This holds regardless of the chosen selection criterion. However, the selection approaches perform quite differently: overall, BIC does not seem to be a good choice for estimating s_q in this setting, because it produces by far the largest MSE for larger s_q. In those cases, the MSE associated with AIC estimates is also larger than those of MAIC and MBIC. This indicates that AIC and BIC are not suited for model selection of an high order AR process. Interestingly, even if the order is rather low, i.e. $s_q = 10$, the conventional information criteria are outperformed by MBIC, if the persistency is high. Therefore, the modified information criteria appear to be preferable over the conventional ones. While the MAIC estimates of s_q exhibit lower MSE if the true order and persistency are high, the MBIC produces on average the lowest MSE among the four information criteria. Therefore, without prior knowledge of the true underlying process, the MBIC should be preferred.

[6] $\xi = 1$ results in a non-stationary CAR model.

	Parametrization		MSE			
	s_q	ξ	MAIC	MBIC	AIC	BIC
1.	10	0.80	141.413	52.727	26.425	**21.837**
2.	10	0.90	33.100	**6.259**	25.740	18.969
3.	10	0.95	12.170	**5.077**	25.965	17.091
4.	15	0.80	114.052	49.179	**40.724**	70.717
5.	15	0.90	33.835	**12.755**	33.676	62.077
6.	15	0.95	16.592	**19.584**	31.943	59.215
7.	20	0.80	92.904	**46.828**	70.939	169.224
8.	20	0.90	40.117	**29.680**	59.722	128.071
9.	20	0.95	**31.353**	50.976	56.239	117.179
10.	25	0.80	80.007	**54.201**	134.433	357.475
11.	25	0.90	**49.547**	57.086	112.306	263.666
12.	25	0.95	**53.260**	97.362	103.149	231.752

Note: Bold values indicate the lowest MSE for a given specification.

Table 1: Results of the simulation study for s_q

4.2 Simulation study for selection procedures for $\mathfrak{s}_q$

We are interested in evaluating the three different selection procedures for $\mathfrak{s}_q$, conditional on s_q. To this end, we assume $s_q = 22$ to be known. We also simulate the process given by (19) and (20) for different choices of $\mathbf{s}$, i.e. with $q = 3$ or $q = 4$ steps. For $q = 3$ ($q = 4$), this entails the estimation of s_2 (s_2 and s_3) by the approaches presented in section 3.

By varying the parameters ξ and $\mathfrak{s}_q$, we simulate 24 different cascade models in total. The parametrizations are given in table 2; for example, parametrization 17 yields $\boldsymbol{\delta} = 0.9 \cdot (0.1, 0.3, 0.3, 0.3)' = (0.09, 0.27, 0.27, 0.27)'$ and $\mathbf{s} = (1, 2, 5, 22)'$, so that $\mathfrak{s}_q = (2, 5)'$ needs to be estimated.

Like in the previous simulations, we use $B = 10,000$ replications per model to calculate the performance measures. Throughout the simulation of each model, the number of observations T is set to 2000. The first 1000 observations are treated as in-sample whereas the remaining 1000 are used for forecasting.

In-sample, we employ the three approaches presented (LS, Wald and CV) to estimate the unknown vector $\mathfrak{s}_q$ and $\boldsymbol{\delta}$, resulting in $\widehat{\mathfrak{s}}_q(b, j)$ and $\widehat{\boldsymbol{\delta}}(b, j)$ for replication $b = 1, ..., B$ and model selection method j.

These estimates are then used to calculate measures of fit for the in-sample and out-of-sample period. In-sample,

we make use of the euclidean distance between $\mathfrak{s}_q$ and $\widehat{\mathfrak{s}}_q(b,j)$:

$$\vartheta(b,j) = ||\widehat{\mathfrak{s}}_q(b,j) - \mathfrak{s}_q||_2 = \sqrt{(\widehat{s}_2(b,j) - s_2)^2 + ... + (\widehat{s}_{q-1}(b,j) - s_{q-1})^2}, \quad \vartheta(b,j) \geq 0. \tag{21}$$

If $\widehat{\mathfrak{s}}_q(b,j)$ is equal to $\mathfrak{s}_q$, the euclidean distance is zero. Larger values indicate that $\widehat{\mathfrak{s}}_q(b,j)$ deviates substantially from its true value. For example, if $q = 3$, equation (21) reduces to $\vartheta(b,j) = \sqrt{(\widehat{s}_2(b,j) - s_2)^2}$.

Out-of-sample fit is evaluated based on the accuracy of the predictions: we first derive one-step ahead forecasts $\widehat{y}_{t|t-1}(b,j)$ for the out-of-sample period $1000 < t \leq T$, conditional on estimates $\widehat{\mathfrak{s}}_q(b,j)$ and $\widehat{\boldsymbol{\delta}}(b,j)$. The forecasts are then used to calculate the Root Mean Squared Forecasting Error (RMSFE):

$$RMSFE(b,j) = \sqrt{\frac{1}{1000}\sum_{t=1001}^{T}(\widehat{y}_{t|t-1}(b,j) - y_t)^2} \tag{22}$$

Upon completition of all B replications, we are able to compare the model selection approaches using the averages of the two performance measures given in (21) and (22), $\overline{\vartheta}(j)$ and $\overline{RMSFE}(j)$. Table 2 shows the results of the simulation for $q = 3$ (parametrizations 1 to 12) and $q = 4$ (parametrizations 13 to 24). We also provide results of three additional models; first, the true model with known $\mathbf{s}$. For this model, only $\boldsymbol{\delta}$ needs to be estimated and serves as a benchmark. Second, the standard HAR model and third, a cascade model that may useful in smaller samples and involves an ad-hoc model selection - denoted by 'fixed'. $\mathfrak{s}_q$ is not estimated from the data. Instead, $\mathbf{s}$ is completely determined by s_q and q:

$$\mathbf{s} = \begin{cases} \left(1, \lfloor\sqrt{s_q}\rfloor, s_q\right)' & \text{if } q = 3 \\ \left(1, \lfloor\sqrt{s_q}\rfloor, \lfloor\frac{s_q}{2}\rfloor, s_q\right)' & \text{if } q = 4, \end{cases}.$$

In-sample standard LS selection consistently outperforms Cross-Validation in terms of lower $\overline{\vartheta}(j)$. This is not suprising, since CV uses not the complete sample per each fold. Moreover, especially for $q = 3$, LS delivers estimates, that are on average the closest to the true value of $\mathfrak{s}_q$. For $q = 4$, the relative performance of the estimation procedure depends more strongly on the true value of $\mathfrak{s}_q$, rather than ξ; for processes with model order $\mathbf{s} = (1,2,5,22)'$ or $\mathbf{s} = (1,5,10,22)'$ Wald-based selection performs best in-sample.

Out-of-sample, all models produce RMSFE close to the theoretical value of one. Since the differences are small, we multiply the results by 100. This finding is independent from the choice of steps and persistency. Moreover, the average RMSFE are almost identical across the selection procedures, which stems from the fact that they often select the same $\widehat{\mathfrak{s}}_q$. However, the estimated models outperform the ad-hoc and HAR parametrization in many cases. In cases, where HAR produces the smallest average RMSFE, the true value of $\mathbf{s}$ is either very similar to or exactly equal to the HAR parametrization. Aside from those cases, Wald and LS produce on average the most accurate predictions.

	Parametrization		In-sample, $\overline{\vartheta}(j)$			Out-of-sample, $\overline{RMSFE}(j)$, scaled by factor 100					
$q=3$	ξ	**s**	LS	Wald	CV	LS	Wald	CV	True	fixed	HAR
1.	0.8	(1,2,22)'	0.084	**0.072**	0.101	**100.150**	100.153	100.154	100.115	100.983	101.176
2.	0.8	(1,5,22)'	0.671	**0.652**	0.746	100.265	100.269	100.275	100.148	100.386	**100.148**
3.	0.8	(1,10,22)'	**2.427**	2.435	2.645	100.320	**100.317**	100.328	100.177	100.483	100.443
4.	0.8	(1,15,22)'	**4.114**	4.486	4.328	100.278	100.271	100.274	100.135	100.273	**100.269**
5.	0.9	(1,2,22)'	**0.036**	0.042	0.044	**100.159**	100.166	100.162	100.138	101.201	101.446
6.	0.9	(1,5,22)'	**0.452**	0.478	0.498	100.277	100.287	100.284	100.174	100.476	**100.174**
7.	0.9	(1,10,22)'	**1.969**	2.075	2.173	**100.304**	100.313	100.316	100.165	100.555	100.505
8.	0.9	(1,15,22)'	**3.721**	4.265	3.947	100.282	100.281	**100.280**	100.139	100.311	100.307
9.	0.95	(1,2,22)'	**0.025**	0.027	0.027	**100.211**	100.215	100.214	100.194	101.360	101.632
10.	0.95	(1,5,22)'	**0.387**	0.423	0.429	100.259	100.274	100.267	100.156	100.492	**100.156**
11.	0.95	(1,10,22)'	**1.807**	2.010	1.977	**100.302**	100.318	100.313	100.161	100.599	100.536
12.	0.95	(1,15,22)'	**3.483**	4.295	3.755	**100.309**	100.314	100.312	100.167	100.363	100.355
$q=4$											
13.	0.8	(1,2,5,22)'	3.327	**2.558**	3.941	100.376	**100.335**	100.408	100.166	100.620	100.580
14.	0.8	(1,2,10,22)'	**3.796**	3.864	4.255	100.422	**100.418**	100.448	100.218	100.643	100.794
15.	0.8	(1,2,15,22)'	**4.785**	5.314	5.160	100.356	**100.351**	100.365	100.156	100.617	100.672
16.	0.8	(1,5,10,22)'	5.329	**5.102**	5.453	100.445	100.437	100.441	100.170	100.309	**100.253**
17.	0.9	(1,2,5,22)'	2.413	**1.825**	2.836	100.339	**100.303**	100.368	100.148	100.706	100.678
18.	0.9	(1,2,10,22)'	**3.065**	3.233	3.406	**100.372**	100.381	100.397	100.193	100.720	100.930
19.	0.9	(1,2,15,22)'	**4.233**	5.046	4.523	**100.367**	100.374	100.380	100.195	100.757	100.849
20.	0.9	(1,5,10,22)'	4.845	**4.643**	4.999	100.475	100.471	100.476	100.205	100.373	**100.324**
21.	0.95	(1,2,5,22)'	1.952	**1.411**	2.329	100.370	**100.338**	100.400	100.200	100.807	100.786
22.	0.95	(1,2,10,22)'	**2.716**	3.015	3.045	**100.389**	100.402	100.415	100.220	100.795	101.034
23.	0.95	(1,2,15,22)'	**4.008**	4.979	4.236	**100.342**	100.354	100.351	100.177	100.801	100.914
24.	0.95	(1,5,10,22)'	4.547	**4.487**	4.700	100.503	100.513	100.504	100.232	100.418	**100.372**

Table 2: Simulation results: Bold values indicate the lowest $\overline{\vartheta}$ or $\overline{RMSFE}$ (True model excluded).

To sum up, in-sample gains can be considerably large, when estimating $\mathfrak{s}_q$ by the methods presented and LS seems to be the preferred choice. Out-of-sample, performing model selection also gives an advantage over fixed parametrizations, unless the true parameters are close to the assumed ones.

5 Empirical Application

In this section, we apply the procedures on time-series of realized variances (RV) and show the benefits of model selection, in contrast to using a fixed $\boldsymbol{s}$. The data is obtained from the Oxford-Man Institute's realized library by Heber et al. (2009) and consists of daily RV estimators of major stock indices, calculated using 5-minute intra-day returns. The indices used are DAX 30, EURO STOXX 50, FTSE 100, Nifty 50, Nikkei 225 and S&P 500. The data ranges from from 2005 to 2018 and consists of approximately 3550 observations each.[7]

The years 2005 to 2014 are treated as in-sample and used for estimation and selection of the model parameters. The remaining four years are used to construct out-of-sample one-step-ahead forecasts. Because of the presence of large outliers and skewness in the data, we use a log transform of the data. It is worth mentioning that our aim is to model the time-series of RV itself and not the underlying latent variance process.[8]

Using the methods presented, we estimate models with 3 and 4 steps and determine s_q by using MBIC.For further comparison, we also estimate AR(1) and ARMA(1,1) models.[9]

The estimated values for $\mathbf{s}$ are reported in table 3. For $q = 3$, the results are sometimes similar to the HAR specification, for example the OLS estimates for the DAX index are $(1, 5, 19)'$. However, the short-run dynamics are emphasized by the data, because s_2 is often smaller than 5 and s_q is always estimated to be smaller than 22. This becomes even more clear for $q = 4$, because s_2 and s_3 are almost exclusively estimated to be smaller or equal to 5.

In tables 4 and 5 it can be seen that these differences to the standard HAR formulation result in a better in-sample fit on average. Although the ARMA(1,1) model produces the lowest AIC and BIC for the S&P 500 index, the OLS approach provides the best in-sample fit for most of the RV time-series. AIC and BIC also often indicate, that the inclusion of the additional cascade step results in a better fit.

However, table 6 shows that the increased in-sample fit does not necessarily result in improved forecasts. For example, the Wald-based estimation and the EURO STOXX 50, the inclusion of an additional step decreases

[7] Differences in sample length are caused by, among other factors, different national holidays.

[8] RV is only a proxy for the latent variance process. Hence, a model that fits log(RV) best is not necessarily the best fitting model for the true volatility. For a discussion of comparing models fitted to variance proxies, see eg. Hansen and Lunde (2006) and Patton (2011)

[9] In contrast to the other models, we estimate the ARMA(1,1) using Maximum-Likelihood and assume normally distributed errors.

$q = 3$	LS	Wald	CV	fixed
DAX	(1, 5, 19)'	(1, 3, 19)'	(1, 5, 19)'	(1, 4, 19)'
STOXX	(1, 5, 19)'	(1, 3, 19)'	(1, 5, 19)'	(1, 4, 19)'
FTSE	(1, 4, 15)'	(1, 4, 15)'	(1, 4, 15)'	(1, 3, 15)'
Nifty 50	(1, 3, 16)'	(1, 3, 16)'	(1, 3, 16)'	(1, 4, 16)'
Nikkei	(1, 3, 16)'	(1, 3, 16)'	(1, 3, 16)'	(1, 4, 16)'
S&P	(1, 3, 18)'	(1, 2, 18)'	(1, 3, 18)'	(1, 4, 18)'
$q = 4$				
DAX	(1, 2, 5, 19)'	(1, 3, 5, 19)'	(1, 2, 5, 19)'	(1, 4, 9, 19)'
STOXX	(1, 3, 5, 19)'	(1, 3, 5, 19)'	(1, 3, 5, 19)'	(1, 4, 9, 19)'
FTSE	(1, 2, 5, 15)'	(1, 2, 5, 15)'	(1, 2, 5, 15)'	(1, 3, 7, 15)'
Nifty 50	(1, 3, 5, 16)'	(1, 3, 5, 16)'	(1, 2, 5, 16)'	(1, 4, 8, 16)'
Nikkei	(1, 2, 5, 16)'	(1, 2, 3, 16)'	(1, 3, 10, 16)'	(1, 4, 8, 16)'
S&P	(1, 2, 5, 18)'	(1, 2, 5, 18)'	(1, 2, 5, 18)'	(1, 4, 9, 18)'

Table 3: Estimation results, $\widehat{\mathbf{s}}$

the AIC (BIC) from $-2,520.29$ to $-2,534.04$ ($-2,502.78$ to $2,510.69$). But simultaneously, the scaled RMSFE increases from 66.57 to 66.63. However, the model selection procedures on average all produce more precise out-of-sample forecasts than the standard HAR approach and, aside from the FTSE 100, ARMA(1,1).

$q = 3$	LS	Wald	CV	fixed	HAR	ARMA(1,1)	AR(1)
DAX	**-3, 260.121**	-3, 259.414	**-3, 260.121**	-3, 246.814	-3, 256.078	-3, 250.953	-2, 836.476
STOXX	-2, 527.423	-2, 520.288	-2, 527.423	-2, 520.502	**-2, 528.036**	-2, 518.801	-2, 059.023
FTSE	**-2, 810.168**	**-2, 810.168**	**-2, 810.168**	-2, 797.053	-2, 797.379	-2, 804.616	-2, 256.185
Nifty 50	**-2, 933.014**	**-2, 933.014**	**-2, 933.014**	-2, 928.626	-2, 914.384	-2, 894.541	-2, 457.880
Nikkei	**-3, 195.107**	**-3, 195.107**	**-3, 195.107**	-3, 182.060	-3, 176.565	-3, 179.111	-2, 839.178
S&P	-2, 475.766	-2, 465.954	-2, 475.766	-2, 473.624	-2, 462.698	**-2, 500.127**	-1, 990.094
$q = 4$							
DAX	**-3, 269.263**	-3, 268.525	**-3, 269.263**	-3, 247.896	-3, 256.078	-3, 250.953	-2, 836.476
STOXX	**-2, 534.041**	**-2, 534.041**	**-2, 534.041**	-2, 522.856	-2, 528.036	-2, 518.801	-2, 059.023
FTSE	**-2, 817.542**	**-2, 817.542**	**-2, 817.542**	-2, 805.572	-2, 797.379	-2, 804.616	-2, 256.185
Nifty 50	**-2, 935.703**	**-2, 935.703**	-2, 935.338	-2, 926.865	-2, 914.384	-2, 894.541	-2, 457.880
Nikkei	**-3, 199.797**	-3, 197.708	-3, 195.428	-3, 181.256	-3, 176.565	-3, 179.111	-2, 839.178
S&P	-2, 492.811	-2, 492.811	-2, 492.811	-2, 471.769	-2, 462.698	**-2, 500.127**	-1, 990.094

Table 4: In-sample fit, AIC: Bold values indicate the lowest AIC.

$q = 3$	LS	Wald	CV	fixed	HAR	ARMA(1,1)	AR(1)
DAX	**-3, 242.621**	-3, 241.913	**-3, 242.621**	-3, 229.313	-3, 238.578	-3, 239.272	-2, 825.795
STOXX	-2, 509.910	-2, 502.775	-2, 509.910	-2, 502.989	**-2, 510.524**	-2, 507.112	-2, 048.334
FTSE	-2, 792.684	-2, 792.684	-2, 792.684	-2, 779.569	-2, 779.895	**-2, 792.949**	-2, 245.518
Nifty 50	**-2, 915.585**	**-2, 915.585**	**-2, 915.585**	-2, 911.197	-2, 896.955	-2, 882.911	-2, 447.249
Nikkei	**-3, 177.713**	**-3, 177.713**	**-3, 177.713**	-3, 164.666	-3, 159.171	-3, 167.503	-2, 828.569
S&P	-2, 458.297	-2, 448.484	-2, 458.297	-2, 456.154	-2, 445.228	**-2, 488.468**	-1, 979.435
$q = 4$							
DAX	**-3, 245.929**	-3, 245.190	**-3, 245.929**	-3, 224.562	-3, 238.578	-3, 239.272	-2, 825.795
STOXX	**-2, 510.691**	**-2, 510.691**	**-2, 510.691**	-2, 499.505	-2, 510.524	-2, 507.112	-2, 048.334
FTSE	**-2, 794.229**	**-2, 794.229**	**-2, 794.229**	-2, 782.260	-2, 779.895	-2, 792.949	-2, 245.518
Nifty 50	**-2, 912.464**	**-2, 912.464**	-2, 912.1000	-2, 903.627	-2, 896.955	-2, 882.911	-2, 447.249
Nikkei	**-3, 176.605**	-3, 174.516	-3, 172.236	-3, 158.064	-3, 159.171	-3, 167.503	-2, 828.569
S&P	-2, 469.518	-2, 469.518	-2, 469.518	-2, 448.476	-2, 445.228	**-2, 488.468**	-1, 979.435

Table 5: In-sample fit, BIC: Bold values indicate the lowest BIC.

$q = 3$	LS	Wald	CV	fixed	HAR	ARMA(1,1)	AR(1)
DAX	53.907	**53.661**	53.907	53.813	53.898	53.911	59.974
STOXX	67.046	**66.565**	67.046	66.836	67.005	66.774	73.788
FTSE	64.471	64.471	64.471	64.451	64.527	**64.217**	72.701
Nifty 50	**53.674**	**53.674**	**53.674**	54.027	54.012	53.841	61.338
Nikkei	**62.164**	**62.164**	**62.164**	62.274	62.450	62.733	66.662
S&P	**62.711**	62.781	**62.711**	62.960	63.353	63.852	65.318
$q = 4$							
DAX	53.733	**53.636**	53.733	53.837	53.898	53.911	59.974
STOXX	**66.630**	**66.630**	**66.630**	66.842	67.005	66.774	73.788
FTSE	64.244	64.244	64.244	64.363	64.527	**64.217**	72.701
Nifty 50	**53.703**	**53.703**	53.711	54.008	54.012	53.841	61.338
Nikkei	62.215	62.171	**62.105**	62.259	62.450	62.733	66.662
S&P	62.998	62.998	62.998	**62.994**	63.353	63.852	65.318

Table 6: Out-of-sample fit, RMSFE (scaled by factor 100): Bold values indicate the lowest RMSFE.

6 Conclusion and possible extensions

In this paper, we presented three different ways for model selection in CAR models, namely the order of the underlying AR model, the number of steps and the width of each step. In a simulation study, we showed that the procedures are able to identify the model order in a satisfying way. This results in two major benefits over standard HAR procedures: First, by estimating the step pattern, one gain deeper insights into the dynamics of a time-series in-sample: we showed empirically, that the model order may deviate from the standard HAR formulation.

Second, we showed in a simulation study and empirically that, despite the additional estimation risk, out-of-sample forecasting performance is similar or even better than the standard HAR model.

Since we focused on univariate time-series, future research may concern model selection in a multivariate setting. Moreover, the CAR model results from imposing linear restrictions on a AR model. Hence, a natural next step would be considering non-linear restrictions.

References

[1] Akaike, H. (1973). Information theory and an extension of maximum likelihood principle. In *Proc. 2nd Int. Symposium on Information Theory*: 267–281.

[2] Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6): 716-723.

[3] Andersen, T. G., Bollerslev, T., Diebold, F. X., 2007. Roughing it up: Including jump components in the measurement, modeling, and forecasting of return volatility. *The Review of Economics and Statistics*, 89(4): 701–720.

[4] Audrino, F., Knaus, S. D. (2016). Lassoing the HAR model: A model selection perspective on realized volatility dynamics. *Econometric Reviews*, 35(8-10): 1485–1521.

[5] Audrino, F., Huang, C., Okhrin, O. (2018). Flexible HAR model for realized volatility. *Studies in Nonlinear Dynamics & Econometrics*, 23(3).

[6] Audrino, F., Camponovo, L., Roth, C. (2015). Testing the lag structure of assets' realized volatility dynamics. *Quantitative Finance and Economics*, 1(4): 363–387.

[7] Baillie, R. T., Kongcharoen, C., Kapetanios, G. (2012). Prediction from ARFIMA models: Comparisons between MLE and semiparametric estimation procedures. *International Journal of Forecasting*, 28(1): 46–53.

[8] Bekaert, G., Hoerova, M. (2014). The VIX, the variance premium and stock market volatility. *Journal of Econometrics*, 183(2): 181–192.

[9] Bergmeir, C., Hyndman, R. J., Koo, B. (2018). A note on the validity of cross-validation for evaluating autoregressive time series prediction. *Computational Statistics & Data Analysis*, 120: 70–83.

[10] Box, G., Jenkins, G. (1970). *Time Series Analysis: Forecasting and Control*. San Francisco: Holden-Day.

[11] Brockwell, P. J., Davis, R. A. (1991). *Time Series: Theory and Methods*. Springer Science & Business Media.

[12] Corsi, F. (2009). A simple approximative long-memory model of realized volatility. *Journal of Financial Econometrics*, 7: 174–196.

[13] Corsi, F., Pirino, D., Renò, R. (2010). Threshold bipower variation and the impact of jumps on volatility forecasting. *Journal of Econometrics*, 159(2): 276–288.

[14] Corsi, F., Renò, R. (2012). Discrete-time volatility forecasting with persistent leverage effect and the link with continuous-time volatility modeling. *Journal of Business & Economic Statistics*, 30(3): 368–380.

[15] Corsi, F., Audrino, F., Renò, R. (2012). HAR modeling for realized volatility forecasting. In *Handbook of Volatility Models and Their Applications* (vol. 3). John Wiley & Sons, Inc., New Jersey.

[16] Crato, N., Ray, B. K. (1996). Model selection and forecasting for long-range dependent processes. *Journal of Forecasting*, 15(2): 107–125.

[17] Dickey, D. A., Fuller, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American statistical association*, 74(366a): 427–431.

[18] Golosnoy, V., Hildebrandt, B., Köhler, S. (2019). Modeling and forecasting realized portfolio diversification benefits. *Journal of Risk and Financial Management*, 12(3): 116.

[19] Granger, C. W., Joyeux, R. (1980). An introduction to long-memory time series models and fractional differencing. *Journal of Time Series Analysis*, 1(1): 15–29.

[20] Hansen, P. R., and Lunde, A. (2006). Consistent ranking of volatility models. *Journal of Econometrics*, 131(1-2): 97–121.

[21] Hayashi, F. (2000). *Econometrics*. Princeton University Press: New Jersey.

[22] Heber, G., Lunde, A., Shephard, N., Sheppard, K. (2009). *Oxford-Man Institute's realized library (v0.3)*. Oxford-Man Institute, University of Oxford.

[23] Hwang, E., Shin, D. W. (2014). Infinite-order, long-memory heterogeneous autoregressive models. *Computational Statistics & Data Analysis*, 76: 339–358.

[24] Ljung, G. M., Box, G. E. (1978). On a measure of lack of fit in time series models. *Biometrika*, 65(2): 297–303.

[25] Leeb, H., Pötscher, B. M. (2009). Model Selection. In *Handbook of Financial Time Series*. Springer-Verlag Berlin Heidelberg

[26] McAleer, M., Medeiros, M. C. (2008). A multiple regime smooth transition heterogeneous autoregressive model for long memory and asymmetries. *Journal of Econometrics*, 147(1): 104–119.

[27] Müller, U. A., Dacorogna, M. M., Davé, R. D., Olsen, R. B., Pictet, O. V., Von Weizsäcker, J. E. (1997). Volatilities of different time resolutions—analyzing the dynamics of market components. *Journal of Empirical Finance*, 4(2-3): 213–239.

[28] Ng, S., Perron, P. (2001). Lag length selection and the construction of unit root tests with good size and power. *Econometrica*, 69(6): 1519–1554.

[29] Patton, A. J. (2011). Volatility forecast comparison using imperfect volatility proxies. *Journal of Econometrics*, 160(1): 246–256.

[30] Patton, A. J., Sheppard, K. (2015). Good volatility, bad volatility: Signed jumps and the persistence of volatility. *Review of Economics and Statistics*, 97(3): 683–697.

[31] Said, S. E., Dickey, D. A. (1984). Testing for unit roots in autoregressive-moving average models of unknown order. *Biometrika*, 71(3): 599–607.

[32] Schlittgen, R., Streitberg, B. H. J. (2001). *Zeitreihenanalyse*. Oldenburg.

[33] Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2): 461–464.

[34] Schwert, G. W. (1989). Tests for Unit Roots: A Monte Carlo Investigation. *Journal of Business & Economic Statistics*, 7: 147–160.

[35] Wen, F., Gong, X., Cai, S. (2016). Forecasting the volatility of crude oil futures using HAR-type models with structural breaks. *Energy Economics* 59: 400–413.

[36] Wilms, I., Rombouts, J., Croux, C. (2020). Multivariate volatility forecasts for stock market indices. *International Journal of Forecasting*. Advance online publication. doi:10.1016/j.ijforecast.2020.06.012