Multiscale autoregression on adaptively detected timescales

Rafal Baranowski Piotr Fryzlewicz*

January 14, 2020

Abstract

We propose a multiscale approach to time series autoregression, in which linear regressors for the process in question include features of its own path that live on multiple timescales. We take these multiscale features to be the recent averages of the process over multiple timescales, whose number or spans are not known to the analyst and are estimated from the data via a change-point detection technique. The resulting construction, termed Adaptive Multiscale AutoRegression (AMAR) enables adaptive regularisation of linear autoregressions of large orders. The AMAR model permits the longest timescale to increase with the sample size, and is designed to offer simplicity and interpretability on the one hand, and modelling flexibility on the other. As a side result, we also provide an explicit bound on the tail probability of the ℓ_2 norm of the difference between the autoregressive coefficients and their OLS estimates in the AR(p) model with i.i.d. Gaussian noise when the order p potentially diverges with, and the autoregressive coefficients potentially depend on, the sample size. The R package **amar** provides an efficient implementation of the AMAR modelling, estimation and forecasting framework.

Key words: multiscale modelling, long time series, structural breaks, breakpoints, regularised autoregression, piecewise-constant approximation.

*Author for correspondence. Department of Statistics, London School of Economics, Houghton Street, London WC2A 2AE, UK. Email: p.fryzlewicz@lse.ac.uk. Work supported by the Engineering and Physical Sciences Research Council grant no. EP/L014246/1.

1 Introduction

Autoregression in time series modelling is arguably the most frequently used device to characterise temporal dependence in data, and has only recently been used in areas as diverse as the modelling of human face aging (Wang et al., 2019), reliability and measurement error in psychological measurement (Schuurman and Hamaker, 2019), modelling and forecasting regional tourism demand (Assaf et al., 2019) and the modelling of human kinematics (Zeng et al., 2019). The methodological and theoretical foundations of autoregressive modelling in time series are extensively covered in many excellent monographs, including Kitagawa (2010), Shumway and Stoffer (2010) and Brockwell and Davis (2016).

The classical linear autoregressive model of order p (AR(p)) for univariate time series X_t assumes that X_t is a linear but otherwise unconstrained function of its own past values $X_{t-1}, \ldots, X_{t-p}$, plus white-noise-like innovation ε_t, that is

$$X_t = \beta_1 X_{t-1} + \ldots + \beta_p X_{t-p} + \varepsilon_t, \quad t = 1, \ldots, T. \tag{1}$$

However, in situations in which the application of model (1) yields a large p, either in absolute terms or relative to T (perhaps in an attempt to reflect long-range dependence in X_t), it may be tempting to consider instead an alternative approach, in which X_t is regressed explicitly on some other features of its own past, rather than on the individual variables $X_{t-1}, \ldots, X_{t-p}$. As a conceptual example, consider the problem of modelling mid- and high-frequency financial returns, where X_t represents a fine-scale, e.g. one-minute, return on a financial instrument. In the hope of improving the predictive power, the analyst may wish to model X_t as depending not only on the past few one-minute returns, but also perhaps on past returns on lower frequencies, such as one hour or one day. Representing this in an unconstrained way as in (1) with a large value of p would lead to obvious over-parameterisation.

Motivated by this observation, this paper proposes what we call a multiscale approach to time series autoregression, in which we include as linear regressors for X_t features of the path $X_1, \ldots, X_{t-1}$ that live on multiple timescales. To fix ideas, in this article, we take

these multiscale features to be the recent averages of X_t over multiple time spans, which are not necessarily known to the analyst a priori and need to be estimated from the data. This leads to the following Adaptive Multiscale Autoregressive model of order q (AMAR(q)) for X_t:

$$X_t = \alpha_1 \frac{1}{\tau_1}(X_{t-1} + \ldots + X_{t-\tau_1}) + \ldots + \alpha_q \frac{1}{\tau_q}(X_{t-1} + \ldots + X_{t-\tau_q}) + \varepsilon_t, \quad t = 1, \ldots, T, \quad (2)$$

where the timescales $1 \leq \tau_1 < \tau_2 < \ldots < \tau_q$ and the scale coefficients $\alpha_1, \ldots, \alpha_q \in \mathbb{R}$ are unknown, the number of scales q is possibly much smaller than the largest timescale τ_q, and ε_t is a white-noise-like innovation process. Here, we use the term "adaptive" to reflect the fact that the timescales in the AMAR model automatically adapt to the data in the sense of being selected in a data-driven way, rather than being known a priori. The AMAR(q) model is a particular, multiscale, sparsely parameterised, version of the AR(τ_q) process. It permits the longest timescale τ_q to be large, perhaps of the order of tens or even hundreds. We propose to estimate the unknown number of scales q and the scale parameters $\tau_1, \ldots \tau_q$ via a change-point detection technique.

The concept of the AMAR model, but no details of its modus operandi or its estimation procedure, appears in the Oberwolfach report Fryzlewicz (2013). The current work makes precise several aspects of the AMAR framework and makes it operable. We now provide an overview of other related literature.

In the multivariate context, Reinsel (1983) consider a model in which the current time series variable depends linearly on a small number of index variables which are linear combinations of its own past values; in contrast to our setting, these index variables are assumed to be known a priori. Reduced-rank time series multivariate autoregression, which provides a way of reducing the parameterisation for multivariate time series via the use of automatically chosen index variables, is considered in Velu et al. (1986) and Ahn and Reinsel (1988), but this approach is not explicitly designed to be multiscale or to able to cope with autoregressions of large orders.

Ferreira et al. (2006) introduce a class of bi-scale univariate time series models that consist

of two main building blocks: Y_t, $t = 1, \ldots, Tm$, the fine-level process, where $m > 1$ is known, and the coarse-level aggregate process $X_t = m^{-1} \sum_{j=1}^{m} Y_{tm-j} + \varepsilon_t$, where the noise term $\varepsilon_t \sim \mathcal{N}(0, \sigma^2)$ is i.i.d. and independent of Y_t. Ferreira et al. (2006) recommend choosing a simple model for Y_t, e.g. AR(1), and show with this choice, X_t can emulate long-memory behaviour. An MCMC-based procedure is used for estimation. For reviews of this approach, see Ferreira and Lee (2007) and Ferreira et al. (2010). In contrast to this framework, AMAR assumes that the timescales are not known a priori, and uses coarse-level information for fine-level modelling, rather than vice versa.

Ghysels et al. (2004) propose MIDAS (Mixed Data Sampling) regression, in which time series observed at finer scales are used to model one observed at a lower frequency. In the notation of the previous paragraph, the MIDAS model is defined as $X_t = \beta_0 + \sum_{i=1}^{p} b_i(Y_{tm-i}; \boldsymbol{\beta}) + \varepsilon_t$, where $b_1(\cdot; \boldsymbol{\beta}), \ldots, b_p(\cdot; \boldsymbol{\beta})$ are given functions of the lagged observations recorded at a higher frequency and of a low-dimensional vector of unknown parameters $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_q)'$, and ε_t is random noise. For each recorded observation of X_t, m values of Y_i are sampled. MIDAS models have found multiple applications in the finance and macroeconomics literature, e.g. in the forecasting of quarterly GDP growth by using monthly business cycle indicators (Clements and Galvão, 2009; Bai et al., 2013), or of daily financial data (Andreou et al., 2013). Depending on the specification of $b_i(\cdot; \boldsymbol{\beta})$, estimation is typically done using either ordinary least squares or nonlinear least squares; for details and examples see Ghysels et al. (2007). Here we mention one particular form of $b_i(\cdot; \boldsymbol{\beta})$ from Forsberg and Ghysels (2007): $X_t = \beta_0 + \sum_{j=1}^{q} \beta_j \sum_{i=1}^{\tau_j} Y_{tm-i} + \varepsilon_t$, where $1 \leq \tau_1 < \ldots \leq \tau_q$ are known integers; the authors refer to the regressors as "step functions". One important difference with the AMAR framework is that $\tau_1, \ldots, \tau_q$ in our model are unknown. Ghysels et al. (2007) review and describe some possible extensions of the MIDAS framework.

In heterogeneous autoregressive (HAR) modelling (Corsi, 2009), the quantity of interest (in this specific work, daily latent volatility) is regressed on its past realised averages over given known multiple timescales (here: realised daily, weekly and monthly volatilities). The authors show that the model is able to imitate long-memory behaviour without, in fact, possessing the long-memory property. Numerous extensions and applications of the

HAR approach have been considered; see e.g. Hwang and Shin (2014) for infinite-order, long-memory HAR models; Hwang and Shin (2013) for testing for structural breaks in a long-memory HAR model; Wen et al. (2016) for the use of the HAR model with structural breaks in forecasting the volatility of crude oil futures; McAleer and Medeiros (2008) for a multiple regime, smooth transition HAR model; and Cubadda et al. (2017) for a multivariate extension of the HAR framework. Corsi et al. (2012) provide a review of HAR modelling of realised volatility. Müller et al. (1997) introduce a heterogeneous ARCH (HARCH) model, in which the squared volatility is a linear function of squared returns at different (again, known) time horizons. Raviv et al. (2015) compare and combine several models, including HAR but also and primarily multivariate models, to forecast day-ahead (coarse-scale) electricity prices by incorporating hourly (fine-scale) prices as regressors.

Multiscale modelling in statistics is often carried out through the formalism offered by wavelets, multiscale systems of (typically) compactly supported, oscillatory functions. For reviews of wavelet applications in statistics, see e.g. Vidakovic (1999) or Nason (2010). In particular, wavelets have also been used in the multiscale modelling of time series, see e.g. Nason et al. (2000) for the use of wavelets as building blocks in a multiscale, moving-average-like time series model and Schroeder and Fryzlewicz (2013) for the use of adaptive (unbalanced) Haar wavelets in the modelling and estimation of the time varying mean of a time series. The estimation technique introduced in the latter work involves change-point detection, but we emphasise that this is in the context of detecting changes in the first-order structure of a time series, rather than in the autoregressive parameter vector as in AMAR. Percival and Walden (2006) review the use of wavelets in time series analysis. Maeng and Fryzlewicz (2019) introduce bi-scale autoregression, in which the more remote autoregressive coefficients are assumed to be sampled from a smooth function; this is done to regularise the estimation problem and thus facilitate estimation of the coefficients if the autoregression order is large. The rough and smooth regions of the AR coefficient space are identified through a technique akin to change-point detection. The approach is different from AMAR in that only two scales are present (while in AMAR the number of scales is unknown a priori and is chosen adaptively from the data), and the scales are defined by the

degree of coefficient smoothness rather than by their spans as in AMAR.

The AMAR model is a particular instance of a linear regression model in which the coefficients have been grouped into (unknown) regions of constancy. The group lasso approach (Yuan and Lin, 2006) assumes that the groups are known and it therefore would not be suitable for AMAR. The fused lasso approach (Tibshirani et al., 2005), which uses a total-variation penalty on the vector of regressors, could in principle be used for the fitting of a piecewise-constant approximation to the estimated vector of AR coefficients, but consistent detection of scales in the AMAR model is effectively a multiple change-point detection problem, and it is known (see e.g. Cho and Fryzlewicz (2011)) that the total variation penalty is not optimal for this task.

The Long Short-Term Memory (LSTM) model of the recurrent neural network (Hochreiter and Schmidhuber, 1997) uses a bi-scale modelling approach whereby the new hidden state at each time point combines (in a particular way that has been learned from the data) long-range "cell state" information with more recent information originating from the previous hidden state and instantaneous input. LSTM models feature prominently in many important applications, such as handwriting recognition, speech recognition and machine translation. Their use in time series forecasting is less well explored and the theoretical understanding of their behaviour in the context of time series modelling is extremely limited, but see Petnehazi (2019) for a recent review. The complexity of LSTM models means that long samples are typically required to train them.

Finally, we note that our notion of "multiscale autoregression" is different from that in, for example, Basseville et al. (1992) or Daoudi et al. (1999), who consider statistical modelling on dyadic trees, motivated by the wavelet decomposition of data. In contrast, we are interested in the explicit multiscale modelling of the time evolution of the original process X_t (i.e. there is no prior multiscale transformation to speak of).

Against the background of the existing literature, the unique contributions of this work can be summarised as follows. Unlike the existing multiscale and index-based approaches to autoregression described above, the scales $\tau_1, \ldots, \tau_q$ in the AMAR model are not assumed to be known by the analyst and are estimable from the data; so is their number q. The AMAR

model is able to accommodate autoregressions of large order: the largest-scale parameter τ_q is permitted to increase with the sample size T at a rate close to $T^{1/2}$. The consistent estimation of the number of scales q and their spans $\tau_1, \ldots, \tau_q$ is achieved by a "narrowest-over-threshold"-type change-point detection algorithm (Baranowski et al., 2019) adapted to the AMAR context, and this paper both justifies this choice and shows how to overcome the significant methodological and theoretical challenges that arise in this adaptation. Being only based on the past averages of the process but enabling data-driven selection of their number and spans, the AMAR framework is designed to offer simplicity and interpretability on the one hand, and modelling flexibility on the other. As a side result, we also provide an explicit bound on the tail probability of the ℓ_2 norm of the difference between the autoregressive coefficients and their OLS estimates in the AR(p) model with i.i.d. Gaussian noise. The bound can be used to study consistency of the OLS estimators when the order p potentially diverges with, and the autoregressive coefficients potentially depend on, the sample size T.

The paper is organised as follows. Section 2 describes the AMAR framework and the narrowest-over-threshold procedure for estimating the number and the spans of the timescales. The relevant consistency theorem is also formulated, as is the side result mentioned in the previous paragraph. Section 3 describes some practical aspects of AMAR modelling, discusses the computational complexity of the AMAR algorithm and illustrates its finite-sample performance in a simulation study. Section 4 shows that AMAR models offer good predictive power in terms of out-of-sample forecasting of high- and mid- frequency financial returns, in an application to stock price series for a number of companies listed on New York Stock Exchange (NYSE). Section 5 concludes with a brief discussion, and the two appendices contain the proofs of our results.

The R package **amar** (Baranowski and Fryzlewicz, 2016a) provides an efficient implementation of our proposal. The R code used in all numerical examples reported in this paper is available from our GitHub repository (Baranowski and Fryzlewicz, 2016b).

2 Methodology and theory

2.1 Motivation and notation

AMAR(q) is an instance of a sparsely parametrised autoregressive (AR) time series model, and therefore (2) can be rewritten as

$$X_t = \beta_1 X_{t-1} + \ldots \beta_p X_{t-p} + \varepsilon_t, \quad t = 1, \ldots, T, \tag{3}$$

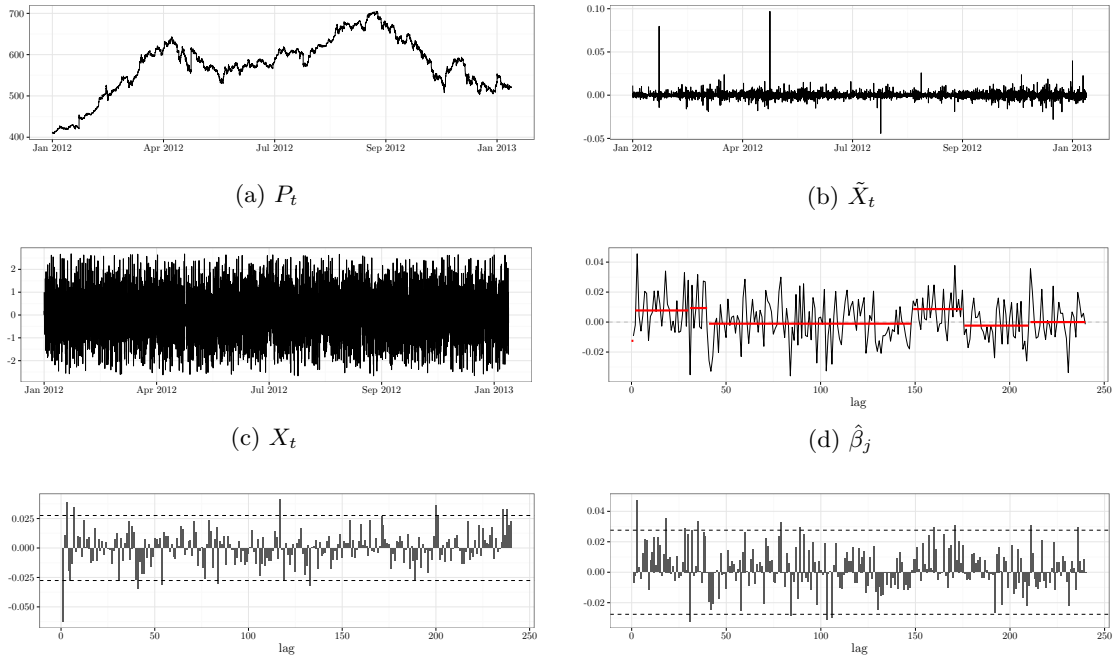
$$\beta_j = \sum_{k:\tau_k \geq j} \frac{\alpha_k}{\tau_k}, \quad j = 1, \ldots, p, \tag{4}$$

for any $p > \tau_q$. We refer to (3) and (4) as an AR(p) representation of the AMAR(q) process, noting that $\beta_j = 0$ for $j = \tau_q + 1, \ldots, p$. Let $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \ldots, \hat{\beta}_p)'$ be the Ordinary Least Squares (OLS) estimator of $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$. Then $\hat{\beta}_j$'s trivially decompose as

$$\hat{\beta}_j = \beta_j + (\hat{\beta}_j - \beta_j), \quad j = 1, \ldots, p. \tag{5}$$

The coefficients $\beta_1, \ldots, \beta_p$ form a piecewise-constant vector with change-points at the timescales $\tau_1, \ldots, \tau_q$, and thus the hope is that the timescales can be estimated consistently using a multiple change-point detection technique. This observation motivates the following estimation procedure for AMAR models. First, we choose a large p and find the OLS estimates of the autoregressive coefficients in the AR(p) representation of the AMAR(q) process. Then, we estimate the timescales by identifying the change-points in (5), using for this purpose an adaptation of the Narrowest-Over-Threshold (NOT) approach of Baranowski et al. (2019). Once the timescales are estimated, we estimate the scale coefficients via least squares. Figure 1d shows an example estimate of the piecewise-constant AMAR coefficient vector.

Our motivation for using the NOT approach as a change-point detector in this context is that it enjoys the following change-point isolation property: in each detection step, the NOT algorithm is guaranteed (with high probability) to be only selecting for consideration sections of the input data (here: the vector $(\hat{\beta}_1, \ldots, \hat{\beta}_p)$) that contain at most a single change-point each. This is a key fact that makes our version of the NOT method amenable



(a) P_t (b) $\tilde{X}_t$

(c) X_t (d) $\hat{\beta}_j$

(e) acf $\tilde{X}_t$ (f) acf X_t

Figure 1: Trades data for Apple Inc. from January 2012 to January 2013. 1a: trade price P_t sampled every 10 minutes. 1b: log-returns $\tilde{X}_t = \log(P_t/P_{t-1})$. 1c: normalised log-returns X_t (see Section 4.2 for details). 1d: the OLS estimates of the AR(p) coefficients with $p = 240$ (thin black) and the piecewise-constant AMAR estimate of the coefficients (red). 1e and 1f: sample acf for, respectively, $\tilde{X}_t$ and X_t.

to a theoretical analysis in the AMAR framework. Another example of a change-point detection method that has this isolation property is the IDetect technique of Anastasiou and Fryzlewicz (2019), which we envisage could also be adapted to the AMAR context.

In a typical application of the AMAR(q) model, we envisage that the number of timescales q will be small in comparison to the maximum timescale τ_q. In order to model this phenomenon, we work in a framework where the timescales $\tau_1, \ldots, \tau_q$ possibly diverge with, and the coefficients $\alpha_1, \ldots, \alpha_q$ depend on, the sample size T. However, for economy of notation we suppress the dependence of α_j, τ_j, q and X_t on T in the remainder of the paper. The following two quantities will together measure the difficulty of our change-point problem detection problem (with the convention $\tau_0 = 0$ and $\tau_{q+1} = p$):

$$\delta_T = \min_{j=1,\ldots,q+1} |\tau_j - \tau_{j-1}|, \tag{6}$$

$$\underline{\alpha}_T = \min_{j=1,\ldots,q} |\beta_{\tau_j+1} - \beta_{\tau_j}| = \min_{j=1,\ldots,q} |\alpha_j| \tau_j^{-1}. \tag{7}$$

Let $\mathbb{C}$ denote the complex plane. For any AR(p) process, we define its characteristic polynomial by

$$b(z) = 1 - \sum_{j=1}^{p} \beta_j z^j, \tag{8}$$

where $z \in \mathbb{C}$. Furthermore, the unit circle is denoted by $\mathbb{T} = \{z \in \mathbb{C} : |z| = 1\}$. Finally, for any vector $\mathbf{v} = (v_1, \ldots, v_k)' \in \mathbb{R}^k$ the Euclidean norm is denoted by $\|\mathbf{v}\| = \sqrt{\sum_{j=1}^{k} v_k^2}$.

We end this section by emphasising again that the purpose of change-point detection in our context is not to find change-point in the AMAR(q) process itself; indeed, this paper only studies stationary AMAR processes, which themselves contain no change-points. The aim of change-point detection in the AMAR context is to segment the possibly long vector of the estimated autoregressive coefficients into regions of piecewise constancy, and thereby estimate the unknown timescales $\tau_1, \ldots, \tau_q$.

The next section prepares the ground for the study of our estimation procedure by considering large deviations for estimated coefficients in AR(p) models under diverging p.

2.2 Large deviations for the OLS estimator in AR(p)

We obtain a tail probability bound on the Euclidean norm of the difference between the OLS estimator $\hat{\boldsymbol{\beta}}$ of the autoregressive parameters $\boldsymbol{\beta}$ in model (3), with all bounds explicitly depending on T, p and the other parameters of the AR(p) process. The following theorem holds.

Theorem 2.1 *Suppose X_t, $t = 1, \ldots, T$, follow the AR(p) model* (3) *and assume that the innovations $\varepsilon_1, \ldots, \varepsilon_T$ are i.i.d. $\mathcal{N}(0,1)$. Assume the initial conditions $X_t = 0$ a.s. for $t = 0, -1, \ldots, -p+1$, and that all roots of the characteristic polynomial $b(z)$ given by* (8) *lie outside the unit circle. Let $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \ldots, \hat{\beta}_p)'$ be the OLS estimate of the vector of the autoregressive coefficients $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$. Then there exist universal constants κ_1, κ_2, $\kappa_3 > 0$ not depending on T, p or $\boldsymbol{\beta}$ s.t. if $\sqrt{T} > \kappa_2 p \log(T)$, then we have*

$$\mathbb{P}\left(\left\| \hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \right\| \leq \kappa_1 (\underline{b}/\overline{b})^2 \left\| \boldsymbol{\beta} \right\| \frac{p \log(T) \sqrt{\log(T+p)}}{\sqrt{T} - \kappa_2 p \log(T)} \right) \geq 1 - \frac{\kappa_3}{T}, \tag{9}$$

where $\underline{b} = \min_{z \in \mathbb{T}} |b(z)|$ and $\overline{b} = \max_{z \in \mathbb{T}} |b(z)|$.

Theorem 2.1 implies that, with high probability, the differences $\hat{\beta}_j - \beta_j$ in (5) converge to zero with $T \to \infty$, provided that $\frac{p \log(T) \sqrt{\log(T+p)}}{\sqrt{T} - \kappa_2 p \log(T)} \to 0$.

In a setting in which both the order p and the autoregressive coefficients in model (3) do not depend on the sample size T, properties of the OLS estimators are well-established. Lai and Wei (1983) show that, without assumptions on the roots of the characteristic polynomial $b(z)$, the OLS estimators are strongly consistent if ε_t is a martingale difference sequence with conditional second moments bounded from below and above. Barabanov (1983) obtains similar results independently, under slightly stronger assumptions on the noise sequence. Bercu and Touati (2008) give an exponential inequality for the OLS estimators in the AR(1) model with i.i.d. Gaussian noise.

Algorithm 1 NOT algorithm for estimation of timescales in AMAR models

Input: Estimates $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \ldots, \hat{\beta}_p)'$; F_T^M = set of M intervals in $[1, p]$ whose start- and end-points have been drawn independently with replacement; $\mathcal{S} = \emptyset$.

Output: Set of estimated timescales $\mathcal{S} = \{\hat{\tau}_1, \ldots, \hat{\tau}_{\hat{q}}\} \subset \{1, \ldots, p\}$.

procedure NOT($\hat{\boldsymbol{\beta}}, s, e, \zeta_T$)
  if $e = s$ **then** STOP
  else
    $\mathcal{M}_{s,e} := \left\{m : [s_m, e_m] \in F_T^M, [s_m, e_m] \subset [s, e]\right\}$
    if $\mathcal{M}_{s,e} = \emptyset$ **then** STOP
    else
      $\mathcal{O}_{s,e} := \left\{m \in \mathcal{M}_{s,e} : \max_{s_m \le b < e_m} \mathcal{C}^b_{s_m, e_m}\left(\hat{\boldsymbol{\beta}}\right) > \zeta_T\right\}$
      if $\mathcal{O}_{s,e} = \emptyset$ **then** STOP
      else
        $m^* :\in \operatorname{argmin}_{m \in \mathcal{O}_{s,e}} |e_m - s_m + 1|$
        $b^* := \operatorname{argmax}_{s_{m^*} \le b < e_{m^*}} \mathcal{C}^b_{s_{m^*}, e_{m^*}}\left(\hat{\boldsymbol{\beta}}\right)$
        $\mathcal{S} := \mathcal{S} \cup \{b^*\}$
        NOT($\hat{\boldsymbol{\beta}}, s, b^*, \zeta_T$)
        NOT($\hat{\boldsymbol{\beta}}, b^* + 1, e, \zeta_T$)
      end if
    end if
  end if
end procedure

2.3 Timescale estimation via the Narrowest-Over-Threshold approach

To estimate the timescales $\tau_1, \ldots, \tau_q$, at which the change-points in model (5) are located, we adapt the Narrowest-Over-Threshold (NOT) approach of Baranowski et al. (2019), with the CUSUM contrast function $\mathcal{C}^b_{s,e}(\cdot)$ suitable for the piecewise-constant model, defined by

$$\mathcal{C}^b_{s,e}(\mathbf{v}) = \left| \sqrt{\frac{e-b}{(e-s+1)(b-s+1)}} \sum_{t=s}^{b} v_t - \sqrt{\frac{b-s+1}{(e-s+1)(e-b)}} \sum_{t=b+1}^{e} v_t \right|. \tag{10}$$

In Baranowski et al. (2019), NOT was shown to recover the number and locations of change-points (the latter at near-optimal rates) in the 'piecewise-constant signal + i.i.d. Gaussian noise' model. Although it is challenging to establish the corresponding consistency and near-optimal rates in problem (5) due to the complex dependence structure in the 'noise' $\hat{\beta}_j - \beta_j$, we show in Section (2.4) that NOT estimators in model (5) enjoy properties similar to those established in the i.i.d. Gaussian setting.

Let $\zeta_T > 0$ be a significance threshold with which to identify large CUSUM values. The NOT procedure for the estimation of the timescales in the AMAR(q) model is described in Algorithm 1. It is a key ingredient of the AMAR estimation algorithm, given in the next section.

2.4 AMAR estimation algorithm and its theoretical properties

We now introduce our proposed estimation procedure for the parameters of the AMAR model. We refer to it as the AMAR algorithm, and its steps are described in Algorithm 2. An efficient implementation of the procedure is available in the R package **amar** (Baranowski and Fryzlewicz, 2016a).

Algorithm 2 AMAR algorithm

Input: Data $X_1, \ldots, X_T$, threshold ζ_T and M, p.
Output: Estimates of the relevant scales $\hat{\tau}_1, \ldots, \hat{\tau}_{\hat{q}}$ and the corresponding AMAR coefficients $\hat{\alpha}_1, \ldots, \hat{\alpha}_{\hat{q}}$.

procedure AMAR($X_1, \ldots, X_T$, p, ζ_T)
Step 1 Find $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \ldots, \hat{\beta}_p)'$, the OLS estimates of the autoregressive coefficients in the AR(p) representation of AMAR(q).
Step 2 Call NOT($\hat{\boldsymbol{\beta}}, 1, p, \zeta_T$) from Algorithm 1 to find the estimates of the timescales; Sort them in increasing order to obtain $\hat{\tau}_1, \ldots, \hat{\tau}_{\hat{q}}$.
Step 3 With the timescales in (2) set to $\hat{\tau}_1, \ldots, \hat{\tau}_{\hat{q}}$, find $\hat{\alpha}_1, \ldots, \hat{\alpha}_{\hat{q}}$, the OLS estimates of the scale coefficients $\alpha_1, \ldots, \alpha_q$.
end procedure

To study the theoretical properties of the timescale estimators $\hat{\tau}_1, \ldots, \hat{\tau}_{\hat{q}}$, we make the following assumptions.

(A1) X_t follows the AMAR(q) model given in (2) with the innovations ε_t being i.i.d. $\mathcal{N}(0,1)$; the initial conditions satisfy $X_t = 0$ a.s. for $t < 0$.

(A2) $p > \tau_q$ and there exist constants $\theta < \frac{1}{2}$ and $c_1 > 0$ such that $p < c_1 T^\theta$ for all T.

(A3) The roots of the characteristic polynomial $b(z)$ given by (8) lie outside the unit circle. Furthermore, there exists a constant $c_2 > 0$ such that $c_2 \leq 1 - \sum_{j=1}^p |\beta_j| \leq \min_{z \in \mathbb{T}} |b(z)|$ uniformly in T (which also implies $\max_{z \in \mathbb{T}} |b(z)| \leq 1 + \sum_{j=1}^p |\beta_j| \leq 2$ uniformly in T).

(A4) $\lambda_T = c_3 T^{\theta-\frac{1}{2}}(\log(T))^{3/2}$, where θ is as in (A2) and $c_3 > 0$ is certain constant whose permitted range depends on c_1, c_2 given in (A2) and (A3). Also, $\delta_T^{1/2}\underline{\alpha}_T \geq \underline{c}\lambda_T$ for a sufficiently large $\underline{c} > 0$, where δ_T and $\underline{\alpha}_T$ are given by (6) and (7), respectively.

The Gaussianity assumption (A1) is made to simplify the theoretical arguments of the proof of Theorem 2.1, which is subsequently used to justify Theorem 2.2 below. As we argue in Section 2.2, Theorem 2.1 could possibly be extended to cover more complicated distributional scenarios for ε_t. (However, from a practical angle, the Gaussianity assumption appears to be reasonable from the point of view of the application of AMAR(q) in forecasting high-frequency returns in Section 4. In the applications, we first remove the volatility from the data and subsequently apply AMAR(q) modelling to the resulting residuals, an example of which can be seen in Figure 1c.)

Assumption (A2) imposes restrictions on both p and the maximum timescale τ_q, which are allowed to increase with $T \to \infty$, but at rates slower than $T^{1/2}$. A similar condition on p being the order of AR(p) approximations of an AR(∞) processes can be found in e.g. Ing and Wei (2005). Assumption (A3) implies that the AMAR(q) process X_t, $t = 1, \ldots, T$, is 'uniformly stationary' for all T: the requirement that $\min_{z\in\mathbb{T}} |b(z)|$ is bounded from below implies that the roots of the characteristic polynomial do not approach the unit circle $\mathbb{T}$ when $T \to \infty$, which in turn ensures that the X_t process is, heuristically speaking, uniformly sufficiently far from being unit-root.

Assumption (A4) controls both the minimum spacing between the timescales and the size of the jumps in (4). The quantity $\delta_T^{1/2}\underline{\alpha}_T$ used here is well-known in the change-point detection literature and characterises the difficulty of the multiple change-point detection problem. We have the following result.

Theorem 2.2 *Let assumptions (A1), (A2), (A3) and (A4) hold, and let $\hat{q}$ and $\hat{\tau}_1, \ldots, \hat{\tau}_{\hat{q}}$ denote, respectively, the number and the locations of the timescales estimated with Algorithm 2. There exist constants $C_1, C_2, C_3, C_4 > 0$ such that if $C_1\lambda_T \leq \zeta_T \leq C_2\delta_T^{1/2}\underline{\alpha}_T$, and*

$M \geq 36T\delta_T^{-2}\log(T\delta_T^{-1})$, *then for all sufficiently large* T *we have*

$$\mathbb{P}\left(\hat{q} = q, \max_{j=1,\ldots,q} |\hat{\tau}_j - \tau_j| \leq \epsilon_T\right) \geq 1 - C_4 T^{-1}, \tag{11}$$

with $\epsilon_T = C_3\lambda_T^2\underline{\alpha}_T^{-2}$.

The main conclusion of Theorem 2.2 is that Algorithm 2 estimates the number of the timescales correctly, while the corresponding locations of the estimates lie close to the true timescales, both with a high probability. Under certain circumstances, Algorithm 2 recovers the exact locations of the timescales. Consider e.g. the case when both the number of scales q and the scale coefficients $\alpha_1, \ldots, \alpha_q$ in (2) are fixed, while the timescales increase with T such that $\delta_T \sim p \sim T^\theta$ ('$\sim$' means that the quantities in question grow at the same rate with $T \to \infty$). This is a challenging setting, in which $\underline{\alpha}_T \sim T^{-\theta}$ and $\|\boldsymbol{\beta}\| \sim T^{-\theta/2}$, where the coordinates of $\boldsymbol{\beta}$ are given by (4), so the signal strength decreases to 0 when $T \to \infty$. Here $\delta_T^{1/2}\underline{\alpha}_T \sim T^{-\theta/2}$, consequently (A4) can only be met if θ in (A2) satisfies the additional requirement $\theta \leq \frac{1}{3}$. The distance between the true timescales and their estimates is then not larger than $\epsilon_T \sim T^{4\theta-1}(\log(T))^3$, which tends to zero if $\theta < \frac{1}{4}$. In this case, (11) simplifies to $\mathbb{P}\left(\hat{q} = q,\, \hat{\tau}_j = \tau_j\ \forall j = 1, \ldots, q\right) \geq 1 - C_4 T^{-1}$, when T is sufficiently large.

3 Practicalities and simulated examples

3.1 Parameter choice

Threshold ζ_T. We test three different approaches to the practical choice of threshold ζ_T.

1. We use a threshold of the minimum rate of magnitude permitted by Theorem 2.2, that is $\zeta_T = CT^{\theta-1/2}(\log(T))^{3/2}$ with $\theta = 0$. In the simulation study of Section 3.3, we show results for $C = 0.25$ and $C = 0.5$.

2. For any $\zeta_T > 0$, denote by $\hat{X}_t(\zeta_T)$ the forecast of X_t obtained via Algorithm 2 and by $\hat{q}(\zeta_T)$ the number of the estimated timescales. We select the threshold that minimises

the Schwarz Information Criterion (SIC) defined as follows:

$$\mathrm{SIC}(\zeta_T) = T \log \left(\sum_{t=1}^{T} (X_t - \hat{X}_t(\zeta_T))^2 \right) + 2\hat{q}(\zeta_T) \log(T), \tag{12}$$

where (12) is minimised over ζ_T such that $\hat{q}(\zeta_T) \leq q_{max} = 10$.

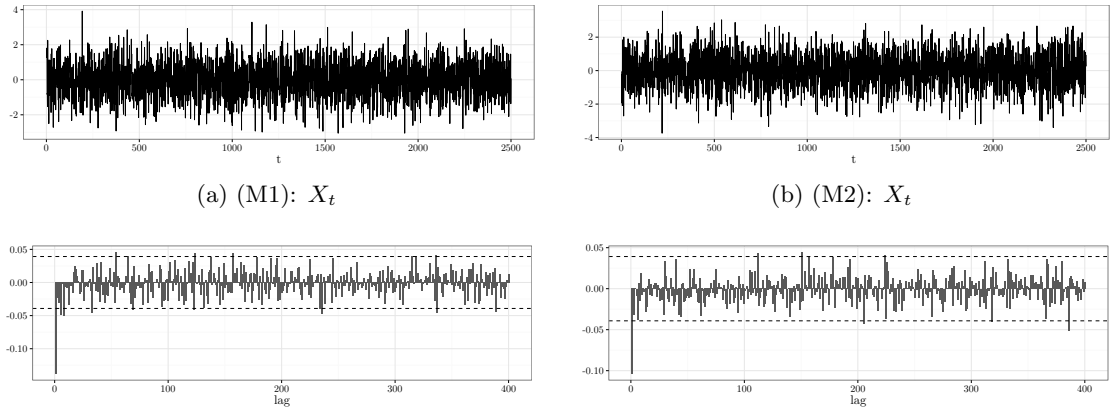
3. Another, application-specific way of choosing ζ_T is discussed in Section 4, where we apply AMAR(q) modelling to the forecasting of high-frequency financial returns.

Number M of random intervals. In line with the recommendation in Baranowski et al. (2019), we set $M = 10000$.

The autoregressive order p. We refrain from giving a universal recipe for the choice of p. In the real-data examples reported later, we choose the p that corresponds to a large "natural" time span. For example, if X_t represents 5-minute returns on a financial instrument, we can take p equal to the length of a trading week or trading month expressed in the number of 5-minute intervals. In principle, the SIC criterion (12) can be minimised with respect to both ζ_T and p, but this would increase the computational burden, and we do not pursue this direction in this work.

3.2 Computational complexity of the AMAR algorithm

The calculation of the OLS estimates in Steps 1 and 3 of Algorithm 2 takes $O(Tp^2)$ operations. The values of $\mathcal{C}_{s,e}^{b}(\mathbf{v})$ can be computed for all b in $O(e-s)$ operations, hence the complexity of Step 2 is $O(Mp)$. This term is typically dominated by $O(Tp^2)$, and therefore the usual computational complexity of the AMAR algorithm is $O(Tp^2)$. The **amar** package uses an efficient implementation of OLS estimation available from the R package **RcppEigen** (Bates and Eddelbuettel, 2013).



(a) (M1): X_t (b) (M2): X_t

(c) (M1): acf (d) (M2): acf

Figure 2: Examples of data generated according to (2) with parameters specified by (M1) and (M2) and $T = 2500$ observations.

3.3 Simulation study

To illustrate the finite sample behaviour and performance of our proposal, we apply Algorithm 2 to simulated data. All computations are performed in the **amar** package. The R code used is available from the GitHub repository (Baranowski and Fryzlewicz, 2016b). The data are simulated from (2) for the following two scenarios.

(M1) Three timescales $\tau_1 = 1$, $\tau_2 = \lfloor 20 \log(T) \rfloor$, $\tau_3 = \lfloor 40 \log(T) \rfloor$ with the corresponding coefficients $\alpha_1 = -0.115$, $\alpha_2 = -2.15$ and $\alpha_3 = -15$, and i.i.d. $\mathcal{N}(0, 1)$ noise ε_t.

(M2) Four timescales $\tau_1 = 1$, $\tau_2 = \lfloor 20 \log(T) \rfloor$, $\tau_3 = \lfloor 10 \log(T)^2 \rfloor$, $\tau_4 = \lfloor 20(\log(T))^2 \rfloor$ with the corresponding coefficients $\alpha_1 = -0.115$, $\alpha_2 = -3.15$, $\alpha_3 = -15$, $\alpha_4 = 10$, and i.i.d. $\mathcal{N}(0, 1)$ noise ε_t.

Figure 2 shows typical sample paths and sample autocorrelation functions for both scenarios with $T = 2500$ observations. Here we observe that, apart from lag 1, the sample autocorrelation function fails to pick up the serial dependence in the data, which illustrates the interesting ability of AMAR models to "mask" as low-order AR processes. This in turn may point to the inappropriateness of the sample autocorrelation function as a tool for analysing AMAR processes.

We consider two aspects of the estimators obtained with Algorithm 2. In order to assess

Method	Model	T	p	FP_0	TP_0	FN_0	$\text{FP}_{\log(T)}$	$\text{TP}_{\log(T)}$	$\text{FN}_{\log(T)}$	RPE
SIC	(M1)	2500	412	2.0	1.8	1.1	1.2	2.6	0.4	0.119
		5000	440	1.6	2.1	0.9	0.8	2.9	0.1	0.047
		10000	468	1.1	2.4	0.6	0.5	3.0	0.0	0.018
		25000	505	0.4	2.8	0.2	0.2	3.0	0.0	0.006
		50000	532	0.3	2.9	0.1	0.2	3.0	0.0	0.002
THR $C = 0.25$	(M1)	2500	412	1.4	1.8	1.2	0.7	2.5	0.5	0.106
		5000	440	0.8	1.9	1.1	0.3	2.5	0.5	0.074
		10000	468	0.3	2.0	1.0	0.0	2.3	0.7	0.054
		25000	505	0.2	2.1	0.9	0.0	2.3	0.7	0.049
		50000	532	0.1	2.0	1.0	0.0	2.1	0.9	0.057
THR $C = 0.5$	(M1)	2500	412	0.4	1.3	1.7	0.1	1.6	1.4	0.411
		5000	440	0.3	1.5	1.5	0.1	1.7	1.3	0.289
		10000	468	0.2	1.8	1.2	0.0	2.0	1.0	0.169
		25000	505	0.1	2.0	1.0	0.0	2.1	0.9	0.066
		50000	532	0.0	2.0	1.0	0.0	2.0	1.0	0.061
SIC	(M2)	2500	1324	2.5	1.3	2.7	1.8	2.0	2.0	0.259
		5000	1550	2.7	1.4	2.6	1.5	2.5	1.5	0.161
		10000	1796	2.8	1.9	2.1	1.6	3.0	1.0	0.080
		25000	2150	2.3	2.2	1.8	1.1	3.5	0.6	0.038
		50000	2441	1.5	2.7	1.3	0.6	3.6	0.4	0.018
THR $C = 0.25$	(M2)	2500	1324	10.0	1.2	2.8	9.1	2.1	1.9	0.475
		5000	1550	3.6	1.5	2.5	2.7	2.4	1.6	0.177
		10000	1796	2.6	1.6	2.4	1.4	2.8	1.1	0.095
		25000	2150	1.5	2.2	1.8	0.6	3.1	0.9	0.048
		50000	2441	0.8	2.5	1.5	0.3	3.1	0.9	0.043
THR $C = 0.5$	(M2)	2500	1324	1.4	1.1	2.9	0.9	1.6	2.5	0.366
		5000	1550	0.9	1.2	2.8	0.4	1.7	2.3	0.292
		10000	1796	0.6	1.3	2.7	0.1	1.8	2.2	0.253
		25000	2150	0.5	1.7	2.3	0.0	2.2	1.8	0.140
		50000	2441	0.2	2.0	2.0	0.0	2.2	1.8	0.129

Table 1: Simulation results for AMAR models with parameters given in Section 3.3 for growing sample sizes T, with the Relative Prediction Error, TP_η, FN_η and FP_η given by, respectively, (13), (14), (15) and (16), averaged over 100 simulations.

the performance of the method in terms of (in-sample) forecasting accuracy, we consider the Relative Prediction Error (RPE), defined as

$$\text{RPE} = \frac{\sum_{t=1}^{T}(\hat{X}_t - \mu_t)^2}{\sum_{t=1}^{T}\mu_t^2}, \tag{13}$$

where $\hat{X}_t = \hat{\alpha}_1 \frac{1}{\hat{\tau}_1}(X_{t-1}+\ldots+X_{t-\hat{\tau}_1})+\ldots+\hat{\alpha}_{\hat{q}} \frac{1}{\hat{\tau}_{\hat{q}}}(X_{t-1}+\ldots+X_{t-\hat{\tau}_{\hat{q}}})$ is the AMAR estimate of the conditional mean $\mu_t = \alpha_1 \frac{1}{\tau_1}(X_{t-1}+\ldots+X_{t-\tau_1})+\ldots+\alpha_q \frac{1}{\tau_q}(X_{t-1}+\ldots+X_{t-\tau_q}) = X_t - \varepsilon_t$. We also investigate the accuracy of Algorithm 2 in terms of the estimation of the timescales $\tau_1, \ldots, \tau_q$. To this end, we consider the following three measures:

$$\text{TP}_\eta = \left|\{j : \exists_k |\hat{\tau}_k - \tau_j| \leq \eta\}\right|, \tag{14}$$

$$\text{FN}_\eta = \left|\{j : \nexists_k |\hat{\tau}_k - \tau_j| \leq \eta\}\right|, \tag{15}$$

$$\text{FP}_\eta = \hat{q} - \text{TP}_\eta, \tag{16}$$

with $\eta = 0$ and $\eta = \log(T)$. For $\eta = 0$, TP_η, FN_η and FP_η are the number of, respectively, true positives, false negatives and false positives.

We apply Algorithm 2 with the thresholds $\zeta_T = CT^{-1/2}(\log(T))^{3/2}$ with $C = 0.25$ and $C = 0.5$ (the corresponding methods are termed 'THR $C = 0.25$' and 'THR $C = 0.5$', respectively), and with the threshold chosen using the SIC criterion given by (12) (termed 'SIC'). The order of the AR(p) representation is set to $p = \lfloor 40 \log(T) \rfloor + 100$ in (M1) and $p = \lfloor 20(\log(T))^2 \rfloor + 100$ in (M2). We set $M = 10000$. Table 1 shows the results. We observe that for all methods, the average RPE decreases with T. SIC performs the best in this respect, achieving the lowest RPE in almost all cases. In terms of the estimation of the timescales, we also observe that the performance of all methods improves with T, with SIC yielding the best results. For example, in (M1) and with $T \geq 10000$, SIC always identifies three timescales close to the true ones, as $\text{TP}_{\log(T)} = 3$ in those cases. For $T \geq 25000$, SIC also very often recovers the exact locations of the timescales, as the average TP_0 is close to 3.

4 Application to high-frequency data from NYSE TAQ database

4.1 Data

We use the AMAR modelling and forecasting approach to the analysis of 5-minute and 10-minute returns on a number of stocks traded on NYSE or Nasdaq. The selected companies are shown in Table 2 and represent a variety of industries. The trade data are obtained from the NYSE Trades and Quotes database (through Wharton Research Data Services), for the time span covering 10 years from January 2004 to December 2013. The R code used to obtain the results is available from the GitHub repository Baranowski and Fryzlewicz (2016b).

Ticker	Company	Industry
AAPL	Apple Inc.	Computer Hardware
BAC	Bank of America Corp	Banks
CVX	Chevron Corp.	Oil & Gas Exploration & Production
CSCO	Cisco Systems	Networking Equipment
F	Ford Motor	Automobile Manufacturers
GE	General Electric	Industrial Conglomerates
GOOG	Alphabet Inc.	Internet Software & Services
MSFT	Microsoft Corp.	Systems Software
T	AT&T Inc.	Telecommunications

Table 2: Ticker symbols and the industries for the companies analysed in Section 4.

4.2 Data preprocessing

The data are preprocessed in the three steps. First, data cleaning is performed using the methodology of Brownlees and Gallo (2006), implemented in the R package **TAQMNGR** (Calvori et al., 2015). Second, to obtain the price series observed at the required frequency, we divide the trading day into time intervals of equal length (we consider 5-minute and 10-minute intervals). For each interval, the price process P_t is defined as the price of the last trade observed in the relevant 'bin'. When there are no trades in a given interval, P_t is set to the price of the latest available trade. Computations for this step are also performed in the **TAQMNGR** package.

Third, we remove the volatility from the log-returns $\tilde{X}_t = \log(P_t/P_{t-1})$ using the NoVaS transformation approach (Politis, 2003, 2007). The NoVaS estimate of the (squared) volatility is defined as

$$\hat{\sigma}_t^2(\lambda) = (1-\lambda)\hat{\sigma}_{t-1}^2(\lambda) + \lambda \tilde{X}_t^2, \tag{17}$$

with the initial value $\hat{\sigma}_0^2(\lambda) = 1$ and $\lambda \in (0,1)$ being the tuning parameter. The NoVaS transformation is similar to the ordinary exponential smoothing (ES, Gardner (1985); Taylor (2004)), where σ_t^2 is estimated as the weighted average of the squared returns with exponentially decaying weights. However, the ES estimator depends only on observations prior to time t, while (17) also involves the current observation.

Politis and Thomakos (2013) recommend to choose λ such that the resulting residuals $\frac{\tilde{X}_t}{\hat{\sigma}_t(\lambda)}$ match a desired distribution. As our theoretical results are given under Gaussianity, we attempt to normalise $\tilde{X}_t$ by minimising the Jarque-Bera test statistic (Jarque and Bera, 1980), defined as

$$\mathrm{JB}(\lambda) = \frac{n}{6}\left(\hat{\gamma}(\lambda)^2 + \frac{1}{4}(\hat{\kappa}(\lambda)-3)^2\right), \tag{18}$$

where $\hat{\gamma}(\lambda)$ and $\hat{\kappa}(\lambda)$ denote, respectively, the sample skewness and the sample kurtosis computed for the residuals $\frac{\tilde{X}_t}{\hat{\sigma}_t(\lambda)}$, computed on the validation set (of generic length n) defined in the next section.

4.3 Rolling window analysis

We conduct a rolling window analysis in which we compare forecasts obtained with AMAR(q) models to those obtained with the standard AR(p) model. A detailed description of the procedure applied for a single window is given in Algorithm 3. The window size is set to 252 days, which is approximately the number of trading days in a calendar year. For each window, the data are split into three parts. The first half (approximately 6 months) is used as the training set, on which we estimate the parameters for the analysed candidate models. The subsequent 3 months are used as the validation set, on which we select the model yielding the best forecasts (in terms of the R^2 statistic, defined below). The last

Algorithm 3 AMAR training algorithm

Input: Price series P_t observed at a chosen frequency; the maximum number q_{max} of AMAR timescales; the maximum order p of the AR(p) approximation; the number of subsamples M.

Output: The estimated returns $\hat{X}_t$.

procedure TRAINAMAR($P_1, \ldots, P_T$, p, q_{max})

Step 1 Set $\mathcal{S}_{train} = \{1, \ldots, \lfloor 0.5T \rfloor\}$, $\mathcal{S}_{validate} = \{\lfloor 0.5T \rfloor + 1, \ldots, \lfloor 0.75T \rfloor + 1\}$, and $\mathcal{S}_{test} = \{\lfloor 0.75T \rfloor + 1, \ldots, T\}$.

Step 2 Set $\tilde{X}_t = \log(P_t/P_{t-1})$ for $t = 1, \ldots, T$.

Step 3 Find $\lambda^* = \operatorname{argmin}_{\lambda \in (0,1)} \mathrm{JB}(\lambda)$, where $\mathrm{JB}(\lambda)$ given by (18) is calculated using $\tilde{X}_t$ s.t. $t \in \mathcal{S}_{train}$. Set $X_t = \frac{\tilde{X}_t}{\hat{\sigma}(\lambda^*)}$, $t = 1, \ldots, T$.

Step 4 Using $\tilde{X}_t$ for $t \in \mathcal{S}_{train}$, find $\hat{\beta}_1, \ldots, \hat{\beta}_p$, the OLS estimates of the autoregressive coefficients for the AR(p) model.

Step 5 Apply $\mathrm{NOT}(\hat{\boldsymbol{\beta}}, 1, p, \zeta_T^{(k)})$ for all thresholds $\zeta_T^{(k)}$ such there are at most q_{max} timescales. Denote by $\mathcal{T}_1, \ldots, \mathcal{T}_N$ the resulting sets of timescales.

Step 6 For each $\mathcal{T}_k$, find the OLS estimates of $\alpha_1, \ldots, \alpha_q$, using X_t, $t \in \mathcal{S}_{train}$. Using those estimates, construct predictions $X_t^{(k)}$ for $t \in \mathcal{S}_{validate}$. Find $k^* = \operatorname{argmax}_{k=1,\ldots,N} \mathrm{R}^2_{validate}(k)$, where $\mathrm{R}^2_{validate}(k)$ is given by (19) computed for $X_t^{(k)}$ for $t \in \mathcal{S}_{validate}$.

Step 7 Find the OLS estimates of the AMAR coefficients for the timescales $\mathcal{T}_{k^*}$ using X_t such that $t \in \mathcal{S}_{validate}$.

Step 8 Using the model obtained in the previous step, find predictions $\hat{X}_t$ for $t \in \mathcal{S}_{test}$. Record R^2_{test} and HR_{test}.

end procedure

three months serve as the test set, on which we use the model selected on the validation set to construct out-of-sample forecasts for the normalised returns X_t. Once the forecasts are calculated, the window is moved so that the old test set becomes the new validation set.

Let $\hat{X}_t$ be a forecast of X_t for $t = 1, \ldots, T$. The main criterion we use to assess the predictions is defined as follows:

$$\mathrm{R}^2 = 1 - \frac{\sum_{t=1}^{T}(X_t - \hat{X}_t)^2}{\sum_{t=1}^{T} X_t^2}. \tag{19}$$

Naturally, $\mathrm{R}^2 = 0$ for $\hat{X}_t \equiv 0$, therefore $\mathrm{R}^2 > 0$ implies that the given forecast beats the 'zeros only' benchmark. We also investigate how often the sign of the forecast agrees with the sign of the observed return. To this end, we consider the hit rate defined as

$$\mathrm{HR} = \frac{|\{t = 1, \ldots, T : \mathrm{sgn}(\hat{X}_t) = \mathrm{sgn}(X_t), X_t \neq 0\}|}{|\{t = 1, \ldots, T : X_t \neq 0\}|}. \tag{20}$$

4.4 Results

Tables 3–6 show the results, with the p in Algorithm 3 set to 6 and 12 days. We observe that both AR and AMAR achieve positive average R^2 in the majority of cases, which means that they typically beat the 'zeros only' benchmark. In the majority of cases, AMAR is better than AR in terms of R^2 and it is always better in terms of the average hit rate.

5 Discussion

The AMAR estimation algorithm can also be used in large-order autoregressions in which the AR coefficients may not necessarily be piecewise constant, but possess a different type of regularity (e.g. be a piecewise polynomial of a higher degree). In such cases, the AMAR estimator would provide a piecewise constant approximation to the AR coefficient vector.

It would be of interest to investigate whether the AMAR philosophy could be extended to multiscale predictive features other than the averages of the process. An example of

Ticker	Method	R^2	HR − 50%	pct. zeroes
AAPL	AMAR	**0.00042**	**1.6**	2.07
	AR	-0.00036	1.2	
BAC	AMAR	**0.00231**	**2.2**	9.36
	AR	0.00186	0.9	
CVX	AMAR	**-0.00002**	**0.7**	4.24
	AR	-0.00026	0.5	
CSCO	AMAR	**0.00273**	**2.5**	9.32
	AR	0.00235	1.8	
F	AMAR	0.00586	**3.8**	16.75
	AR	**0.00601**	1.8	
GE	AMAR	**0.00208**	**2.2**	9.63
	AR	0.00197	2.1	
GOOG	AMAR	**0.00111**	**1.9**	1.09
	AR	0.00066	1.9	
MSFT	AMAR	**0.00393**	**2.9**	8.79
	AR	0.00386	2.2	
T	AMAR	0.00321	**2.2**	9.68
	AR	**0.00358**	0.7	

Table 3: Averages of the measures introduced in Section 4 evaluating the out-of sample performance of the forecasts obtained with the AMAR methodology and the standard AR model. The returns X_t are observed every 5 minutes, while the maximum time-scale and the maximum order for AMAR and AR are both set to $p = 480$ (6 trading days expressed in 5-minute intervals). For each measure, bold font indicates the better method.

Ticker	Method	R^2	$HR - 50\%$	pct. zeroes
AAPL	AMAR	**0.00038**	**1.7**	2.07
	AR	-0.00038	1.4	
BAC	AMAR	**0.00227**	**2.3**	9.35
	AR	0.00204	1.5	
CVX	AMAR	**0.00017**	**0.8**	4.23
	AR	-0.00034	0.5	
CSCO	AMAR	0.00287	**2.5**	9.36
	AR	**0.00302**	1.9	
F	AMAR	**0.00602**	**3.8**	16.73
	AR	0.00595	1.6	
GE	AMAR	**0.00206**	**2.2**	9.61
	AR	0.00184	2.1	
GOOG	AMAR	**0.00088**	**1.8**	1.08
	AR	0.00064	1.8	
MSFT	AMAR	**0.00375**	**2.8**	8.83
	AR	0.00347	1.8	
T	AMAR	0.00315	**2.2**	9.69
	AR	**0.00350**	0.9	

Table 4: Averages of the measures introduced in Section 4 evaluating the out-of sample performance of the forecasts obtained with the AMAR methodology and the standard AR model. The returns X_t are observed every 5 minutes, while the maximum time-scale and the maximum order for AMAR and AR are both set to $p = 960$ (12 trading days expressed in 5-minute intervals). For each measure, bold font indicates the better method.

Ticker	Method	R^2	HR − 50%	pct. zeroes
AAPL	AMAR	**0.00043**	**1.9**	1.41
	AR	-0.00034	1.5	
BAC	AMAR	**0.00063**	**1.9**	6.83
	AR	0.00011	0.7	
CVX	AMAR	**-0.00065**	**0.5**	3.12
	AR	-0.00075	0.2	
CSCO	AMAR	**0.00181**	**1.8**	6.45
	AR	0.00124	0.5	
F	AMAR	**0.00277**	**2.7**	12.52
	AR	0.00235	0.3	
GE	AMAR	**0.00202**	**1.8**	6.95
	AR	0.00177	1.4	
GOOG	AMAR	**0.00095**	**1.7**	0.72
	AR	0.00072	1.6	
MSFT	AMAR	0.00208	**2.1**	6.12
	AR	**0.00241**	1.2	
T	AMAR	**0.00309**	**1.9**	7.40
	AR	0.00260	0.7	

Table 5: Averages of the measures introduced in Section 4 evaluating the out-of sample performance of the forecasts obtained with the AMAR methodology and the standard AR model. The returns X_t are observed every 10 minutes, while the maximum time-scale and the maximum order for AMAR and AR are both set to $p = 240$ (6 trading days expressed in 10-minute intervals). For each measure, bold font indicates the better method.

Ticker	Method	R^2	$HR - 50\%$	pct. zeroes
AAPL	AMAR	**0.00039**	**1.6**	1.41
	AR	-0.00050	1.5	
BAC	AMAR	**0.00074**	**1.9**	6.82
	AR	0.00033	0.7	
CVX	AMAR	**0.00001**	**0.8**	3.10
	AR	-0.00076	0	
CSCO	AMAR	**0.00154**	**1.8**	6.45
	AR	0.00138	0.6	
F	AMAR	**0.00288**	**3**	12.50
	AR	0.00242	-0.3	
GE	AMAR	0.00230	**2.3**	6.92
	AR	**0.00240**	1.5	
GOOG	AMAR	0.00060	1.5	0.71
	AR	**0.00093**	**1.6**	
MSFT	AMAR	**0.00233**	**2.2**	6.16
	AR	0.00220	1.5	
T	AMAR	**0.00321**	**2**	7.40
	AR	0.00278	0.6	

Table 6: Averages of the measures introduced in Section 4 evaluating the out-of sample performance of the forecasts obtained with the AMAR methodology and the standard AR model. The returns X_t are observed every 10 minutes, while the maximum time-scale and the maximum order for AMAR and AR are both set to $p = 480$ (12 trading days expressed in 10-minute intervals). For each measure, bold font indicates the better method.

such features could be the recent maxima, minima or other empirical quantiles of the process taken over a range of rolling windows. An "intelligent" version of such a generalised AMAR could be designed not just to choose the most relevant scales within a single family of features, but also to select the most relevant types of features from a suitably large dictionary.

A Proof of Theorem 2.1

We write the AR(p) model as

$$\mathbf{Y}_t = \mathbf{B}\mathbf{Y}_{t-1} + \varepsilon_t \mathbf{u}, \quad t = 1, \ldots, T, \tag{21}$$

where $\mathbf{Y}_t = (X_t, X_{t-1}, ..., X_{t-p+1})'$, the matrix of the coefficients

$$\mathbf{B} = \begin{pmatrix} \beta_1 & \beta_2 & \cdots & \beta_p \\ & \mathbf{I}_{p-1} & & 0 \end{pmatrix} \tag{22}$$

and $\mathbf{u} = (1, 0, \ldots, 0)' \in \mathbb{R}^p$. We start with a few auxiliary results.

Theorem A.1 (Parseval's identity, Theorem 1.9 in Duoandikoetxea (2001)) *For any complex-valued sequence $\{f_k\}_{k\in\mathbb{Z}}$ such that $\sum_{k\in\mathbb{Z}} |f_k|^2 < \infty$, the following identity holds*

$$\sum_{k\in\mathbb{Z}} |f_k|^2 = \int_{\mathbb{T}} |f(z)|^2 dm(z), \tag{23}$$

where $a(z) = \sum_{k\in\mathbb{Z}} a_k z^k$, $\mathbb{T} = \{z \in \mathbb{C} : \ |z| = 1\}$, $dm(z) = \frac{d|z|}{2\pi}$.

Lemma A.1 (Cauchy's integral formula) *Let $\mathbf{M} \in \mathbb{R}^{p\times p}$ be a real- or complex- valued matrix. Then for any curve Γ enclosing all eigenvalues of $\mathbf{M}$ and any $j \in \mathbb{N}$ the following holds*

$$\mathbf{M}^j = \frac{1}{2\pi i} \int_\Gamma z^j (z\mathbf{I}_p - \mathbf{M})^{-1} dz = \frac{1}{2\pi i} \int_\Gamma z^{j-1} (\mathbf{I}_p - z^{-1}\mathbf{M})^{-1} dz. \tag{24}$$

Lemma A.2 *Let* $\mathbf{B}$ *given by* (22) *be the matrix of coefficients of a stationary AR(p) process and let* $\mathbf{v} \in \mathbb{R}^p$. *For all* $z \in \mathbb{C}$ *such that* $\sum_{i=0}^{\infty} |\langle \mathbf{v}, \mathbf{B}^i \mathbf{u} \rangle| |z^i| < \infty$, *we have*

$$b(z) \sum_{i=0}^{\infty} \langle \mathbf{v}, \mathbf{B}^i \mathbf{u} \rangle z^i = b(z) \langle \mathbf{v}, (\mathbf{I}_p - z\mathbf{B})^{-1} \mathbf{u} \rangle = v(z), \tag{25}$$

where $v(z) = v_1 + v_2 z + \ldots + v_p z^{p-1}$, $b(z)$ *is given by* (8).

Proof. As $\sum_{i=0}^{\infty} |\langle \mathbf{v}, \mathbf{B}^i \mathbf{u} \rangle| |z^i| < 0$, we can change the order of summation in the left-hand side of (25)

$$(1 - \beta_1 z - \ldots - \beta_p z^p) \sum_{i=0}^{\infty} \langle \mathbf{v}, \mathbf{B}^i \mathbf{u} \rangle z^i = \left\langle \mathbf{v}, \left(\sum_{i=0}^{\infty} (1 - \beta_1 z - \ldots - \beta_p z^p) z^i \mathbf{B}^i \right) \mathbf{u} \right\rangle.$$

Define $\beta_0 = -1$, $\beta_k = 0$ for $k > p$. By direct algebra

$$\sum_{i=0}^{\infty} (1 - \beta_1 z - \ldots - \beta_p z^p) z^i \mathbf{B}^i = -\sum_{i=0}^{\infty} \left(\sum_{k=0}^{i} \beta_k \mathbf{B}^{i-k} \right) z^i := -\sum_{i=0}^{\infty} \mathbf{D}_i z^i.$$

The characteristic polynomial of $\mathbf{B}$ is given by $\phi(z) = (-1)^{p+1} \sum_{k=0}^{p} \beta_k z^{p-k}$. From the Cayley-Hamilton theorem, $\mathbf{B}$ is a root of ϕ, and, consequently for $i \geq p$,

$$\mathbf{D}_i = \mathbf{B}^{i-p} \sum_{k=0}^{i} \beta_k \mathbf{B}^{p-k} = \mathbf{B}^{i-p} \sum_{k=0}^{p} \beta_k \mathbf{B}^{p-k} = 0.$$

It remains to demonstrate that $\langle \mathbf{v}, \mathbf{D}_i \mathbf{u} \rangle = -v_{i+1}$ for $i = 0, \ldots, p-1$, which we show by induction. For $i = 0$, $\langle \mathbf{v}, \mathbf{D}_i \mathbf{u} \rangle = \beta_0 \langle \mathbf{v}, \mathbf{u} \rangle = -v_1$. When $i \geq 1$, matrices $\mathbf{D}_i$ satisfy $\mathbf{D}_i = \mathbf{B}\mathbf{D}_{i-1} + \beta_i \mathbf{I}_p$, therefore

$$\begin{aligned}
\langle \mathbf{v}, \mathbf{D}_i \mathbf{u} \rangle &= \langle \mathbf{v}, \mathbf{B}\mathbf{D}_{i-1} \mathbf{u} \rangle + \beta_i \langle \mathbf{v}, \mathbf{u} \rangle = \langle \mathbf{B}'\mathbf{v}, \mathbf{D}_{i-1} \mathbf{u} \rangle + \beta_i \langle \mathbf{v}, \mathbf{u} \rangle \\
&= \langle v_1(\beta_1, \ldots, \beta_p)' + (v_2, \ldots, v_p, 0)', \mathbf{D}_{i-1} \mathbf{u} \rangle + \beta_i \langle \mathbf{v}, \mathbf{u} \rangle = -v_1 \beta_i - v_{i+1} + v_1 \beta_i \\
&= -v_{i+1},
\end{aligned}$$

which completes the proof. □

Lemma A.3 *Let* $Z_1, Z_2, \ldots$ *be a sequence of i.i.d.* $\mathcal{N}(0,1)$ *random variables. Then for any*

integers $l \neq 0$ and $k > 0$, the following exponential probability bound holds

$$\mathbb{P}\left(\left|\sum_{t=1}^{k} Z_t Z_{t+l}\right| > kx\right) \leq 2\exp\left(-\frac{1}{8}\frac{kx^2}{6+x}\right). \tag{26}$$

For brevity, we will only show $\mathbb{P}\left(\sum_{t=1}^{k} Z_t Z_{t+l} > kx\right) \leq \exp\left(-\frac{1}{8}\frac{kx^2}{6+x}\right)$. The proof of $\mathbb{P}\left(\sum_{t=1}^{k} Z_t Z_{t+l} < kx\right) \leq \exp\left(-\frac{1}{4}kx\right)$ is similar and, combined with the above inequality, implies (26). By Markov's inequality, for any $x > 0$ and $\lambda > 0$ it holds that

$$\mathbb{P}\left(\sum_{t=1}^{k} Z_t Z_{t+l} > kx\right) \leq \exp\left(-kx\lambda\right) \mathrm{E}\exp\left(\lambda \sum_{t=1}^{k} Z_t Z_{t+l}\right).$$

By the convexity of $y \mapsto \exp(\lambda y)$ for any $\lambda > 0$, Theorem 1 in Vershynin (2011) implies

$$\mathrm{E}\exp\left(\lambda \sum_{t=1}^{k} Z_t Z_{t+l}\right) \leq \mathrm{E}\exp\left(4\lambda \sum_{t=1}^{k} Z_t \tilde{Z}_t\right),$$

where $\tilde{Z}_1, \ldots, \tilde{Z}_k$ are independent copies of $Z_1, \ldots, Z_k$. Using the independence by direct computation we get

$$\mathrm{E}\exp\left(4\lambda \sum_{t=1}^{k} Z_t \tilde{Z}_t\right) = \left(\mathrm{E}\exp\left(4\lambda Z_1 \tilde{Z}_1\right)\right)^k = \left(\mathrm{E}\exp\left(8\lambda^2 \tilde{Z}_1^2\right)\right)^k = \left(1 - 16\lambda^2\right)^{-\frac{1}{2}k}$$

provided that $0 < \lambda < \frac{1}{4}$, therefore $\mathbb{P}\left(\sum_{t=1}^{k} Z_t Z_{t+l} > kx\right) \leq \exp\left(-kx\lambda - \frac{k}{2}\log\left(1 - 16\lambda^2\right)\right)$. Taking $\lambda = \frac{-2+\sqrt{4+x^2}}{4x}$ minimises the right-hand side of this inequality. With this value of λ and using $\log(x) \leq x - 1$, we have

$$\begin{aligned}
\mathbb{P}\left(\sum_{t=1}^{k} Z_t Z_{t+l} > kx\right) &\leq \exp\left(\frac{k}{4}\left(2 - \sqrt{x^2+4} + 2\log\left(\frac{1}{4}\left(\sqrt{x^2+4}+2\right)\right)\right)\right) \\
&\leq \exp\left(\frac{k}{4}\left(2 - \sqrt{x^2+4} + \frac{1}{2}\left(\sqrt{x^2+4}+2\right) - 2\right)\right) \\
&= \exp\left(\frac{k}{8}\left(2 - \sqrt{x^2+4}\right)\right) = \exp\left(-\frac{1}{8}\frac{kx^2}{2+\sqrt{x^2+4}}\right) \\
&\leq \exp\left(-\frac{1}{8}\frac{kx^2}{6+x}\right),
\end{aligned}$$

which completes the proof. □

Lemma A.4 (Lemma 1 in Laurent and Massart (2000)) *Let $Z_1, Z_2, \ldots$ be a sequence of i.i.d. $\mathcal{N}(0,1)$ random variables. For any integer $k > 0$ and $x \in \mathbb{R}$ s.t. $x > 0$, the following exponential probability bounds hold*

$$\mathbb{P}\left(\sum_{t=1}^{k} Z_t^2 \geq k + 2\sqrt{kx} + 2x\right) \leq \exp(-x), \tag{27}$$

$$\mathbb{P}\left(\sum_{t=1}^{k} Z_t^2 \leq k - 2\sqrt{kx}\right) \leq \exp(-x). \tag{28}$$

Proof of Theorem 2.1. For $\mathbf{C}_T = \sum_{t=1}^{T-1} \mathbf{Y}_t \mathbf{Y}_t'$ and $\mathbf{A}_T = \sum_{t=1}^{T-1} \varepsilon_{t+1} \mathbf{Y}_t$, we have $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = \mathbf{C}_T^{-1} \mathbf{A}_T$. Consequently,

$$\left\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right\| \leq \lambda_{max}(\mathbf{C}_T^{-1}) \|\mathbf{A}_T\| = \lambda_{min}^{-1}(\mathbf{C}_T) \|\mathbf{A}_T\|, \tag{29}$$

where $\lambda_{min}(\mathbf{M})$ and $\lambda_{max}(\mathbf{M})$ denote, respectively, the smallest and the largest eigenvalues of a symmetric matrix $\mathbf{M}$. To provide an upper bound on $\left\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right\|$ given in Theorem 2.1, we will bound $\lambda_{min}(\mathbf{C}_T)$ from below and $\|\mathbf{A}_T\|$ from above, working on a set whose probability is large. Here we will show result more specific than (9), i.e.

$$\|\mathbf{A}_T\| \leq \left(32\underline{b}^{-2}\sqrt{1 + \|\boldsymbol{\beta}\|^2}\right) p \log(T) \sqrt{(1 + \log(T+p))T}, \tag{30}$$

$$\lambda_{min}(\mathbf{C}_T) \geq \bar{b}^{-2}\left(T - p(1 + 32\log(T)\sqrt{T})\right), \tag{31}$$

on the event

$$\mathcal{E}_T = \mathcal{E}_T^{(1)} \cap \mathcal{E}_T^{(2)} \cap \mathcal{E}_T^{(3)}, \tag{32}$$

where

$$\mathcal{E}_T^{(1)} = \bigcap_{1\le i<j\le p} \left\{ \left| \sum_{t=1}^{T-\max(i,j)} \varepsilon_t \varepsilon_{t+|i-j|} \right| < 32\log(T)\sqrt{T-\max(i,j)} \right\},$$
$$\mathcal{E}_T^{(2)} = \bigcap_{j=1}^{T} \left\{ \left| \sum_{t=1}^{T-j} \varepsilon_t \varepsilon_{t+j} \right| < 32\log(T)\sqrt{T-j} \right\},$$
$$\mathcal{E}_T^{(3)} = \left\{ \sum_{t=1}^{T-p} \varepsilon_t^2 > T-p-2\sqrt{\log(T)(T-p)} \right\}.$$

Finally, we will demonstrate that $\mathcal{E}_T$ satisfies

$$\mathbb{P}\left(\mathcal{E}_T\right) \ge 1 - \frac{5}{T}. \tag{33}$$

(29), (30), (31) and (33) combined together imply the statement of Theorem 2.1. The remaining part of the proof is split into three parts, in which we show (30), (31) and (33) in turn. In the calculations below, we will repeatedly use the following representation of $\mathbf{Y}_t$, which follows from applying (21) recursively:

$$\mathbf{Y}_t = \sum_{j=1}^{t} \varepsilon_j \mathbf{B}^{t-j}\mathbf{u} = \sum_{j=1}^{t} \varepsilon_{t-j+1}\mathbf{B}^{j-1}\mathbf{u}, \quad t = 1, 2, \ldots, T. \tag{34}$$

Upper bound for $\|\mathbf{A}_T\|$. The Euclidean norm satisfies $\|\mathbf{A}_T\| = \sup_{\mathbf{v}:\in\mathbb{R}^p \|\mathbf{v}\|=1} |\langle \mathbf{v}, \mathbf{A}_T\rangle|$, therefore we consider inner products $\langle \mathbf{v}, \mathbf{A}_T\rangle$ where $\mathbf{v} \in \mathbb{R}^p$ is any unit vector. By (34),

$$\langle \mathbf{v}, \mathbf{A}_T\rangle = \sum_{t=1}^{T-1} \langle \mathbf{v}, \mathbf{Y}_t\rangle \varepsilon_{t+1} = \sum_{t=1}^{T-1}\sum_{j=1}^{T-1} \left\langle \mathbf{v}, \mathbf{B}^{j-1}\mathbf{u}\right\rangle \varepsilon_{t-j+1}\varepsilon_{t+1} = \sum_{j=1}^{T-1} \left\langle \mathbf{v}, \mathbf{B}^{j-1}\mathbf{u}\right\rangle a_j,$$

where $a_j = \sum_{t=1}^{T-1} \varepsilon_{t-j+1}\varepsilon_{t+1} = \sum_{t=1}^{T-j} \varepsilon_t\varepsilon_{t+j}$. Lemma A.1 and Lemma A.2 applied to the

equation above yield

$$\begin{aligned}\sum_{j=1}^{T-1}\left\langle \mathbf{v}, \mathbf{B}^{j-1}\mathbf{u}\right\rangle a_j &= \frac{1}{2\pi i}\int_{\mathbb{T}}\left(\sum_{j=1}^{T-1} z^{j-1}a_j\right)\left\langle \mathbf{v}, (z\mathbf{I}_p - \mathbf{B})^{-1}\mathbf{u}\right\rangle dz \\ &= \frac{1}{2\pi i}\int_{\mathbb{T}}\left(\sum_{j=1}^{T-1} z^{j-1}a_j\right)\left(\sum_{j=1}^{p} z^{p-j}v_j\right) q(z)dz \\ &= \frac{1}{2\pi i}\int_{\mathbb{T}}\left(\sum_{j=0}^{T+p-1} z^{j}c_j\right) q(z)dz,\end{aligned}$$

where $q(z) = (z^p b(z^{-1}))^{-1}$ and $c_j = \sum_{i=0}^{j} a_{i+1}v_{p-j+i}$. Integrating by parts, we get

$$\frac{1}{2\pi i}\int_{\mathbb{T}}\left(\sum_{j=0}^{T+p-1} z^{j}c_j\right) q(z)dz = -\frac{1}{2\pi i}\int_{\mathbb{T}}\left(\sum_{j=0}^{T+p-1} z^{j+1}\frac{c_j}{j+1}\right) q'(z)dz.$$

Combining the calculations above and Cauchy's inequality we obtain the following bound.

$$\left\langle \mathbf{v}, \mathbf{A}_T\right\rangle \leq \sqrt{\sum_{j=0}^{T+p-1}\left(\frac{c_j}{j+1}\right)^2}\sqrt{\int_{\mathbb{T}}|q'(z)|^2 dm(z)}. \tag{35}$$

To further bound the first term on the right-hand side of (35), we recall that on the event $\mathcal{E}_T$ coefficients $|a_j| \leq 32\log(T)\sqrt{T}$, hence

$$\begin{aligned}\sqrt{\sum_{j=0}^{T+p-1}\left(\frac{c_j}{j+1}\right)^2} &= \sqrt{\sum_{j=0}^{T+p-1}\frac{1}{(j+1)^2}\left(\sum_{i=0}^{j} a_{i+1}v_{p-j+i}\right)^2} \\ &\leq \max_{j=0,\ldots,T+p-1}|a_j|\sqrt{\sum_{j=0}^{T+p-1}\frac{1}{(j+1)^2}\left(\sum_{i=0}^{j}|v_{p-j+i}|\right)^2} \\ &\leq 32\log(T)\sqrt{T}\sqrt{\sum_{j=0}^{T+p-1}\frac{\max(j+1,p)}{(j+1)^2}} \\ &\leq 32\log(T)\sqrt{(1+\log(T+p))T}.\end{aligned}$$

For the second term in (35), we calculate the derivative $q'(z) = -\frac{pz^{p-1}-\sum_{j=1}^{p}(p-j)\beta_j z^{p-j-1}}{(z^p b(z^{-p}))^2}$

and bound

$$\sqrt{\int_{\mathbb{T}} |q'(z)|^2 dm(z)} = \sqrt{\int_{\mathbb{T}} \left| \frac{pz^{p-1} - \sum_{j=1}^{p}(p-j)\beta_j z^{p-j}}{(z^p b(z^{-p}))^2} \right|^2 dm(z)}$$

$$\leq \frac{\sqrt{\int_{\mathbb{T}} \left| pz^{p-1} - \sum_{j=1}^{p}(p-j-1)\beta_j z^{p-j-1} \right|^2 dm(z)}}{\min_{|z|=1} |(z^p b(z^{-p}))|^2} =$$

$$= \underline{b}^{-2} \sqrt{\left(p^2 + \sum_{j=1}^{p}(p-j)^2 \beta_j^2 \right)} \leq \underline{b}^{-2} p \sqrt{1 + \|\boldsymbol{\beta}\|^2}.$$

Combining the bounds on the two terms, we obtain

$$\langle \mathbf{v}, \mathbf{A}_T \rangle \leq \left(32 \underline{b}^{-2} \sqrt{1 + \|\boldsymbol{\beta}\|^2} \right) p \log(T) \sqrt{(1 + \log(T+p))T}.$$

Taking supremum over $\mathbf{v} \in \mathbb{R}^p$ such that $\|\mathbf{v}\| = 1$ proves (30).

Lower bound for $\lambda_{min}(\mathbf{C}_T)$. Let $\mathbf{v} = (v_1, \ldots, v_p)'$ be a unit vector in $\mathbb{R}^p$. We begin the proof by establishing the following inequality

$$\langle \mathbf{v}, \mathbf{C}_T \mathbf{v} \rangle \geq \overline{b}^{-2} \sum_{i,j=1}^{p} v_i v_j \sum_{t=1}^{T-1} \varepsilon_{t-j+1} \varepsilon_{t-i+1}, \tag{36}$$

where $\varepsilon_t = 0$ for $t \leq 0$ and $\overline{b} = \max_{z \in \mathbb{T}} |b(z)|$. By Theorem A.1 and (34), we rewrite the quadratic form on the left-hand side of (36) to

$$\langle \mathbf{v}, \mathbf{C}_T \mathbf{v} \rangle = \sum_{t=1}^{T-1} \langle \mathbf{v}, \mathbf{Y}_t \rangle^2 = \int_{\mathbb{T}} \left| \sum_{t=1}^{T-1} \left\langle \mathbf{v}, \sum_{j=1}^{t} \varepsilon_j \mathbf{B}^{t-j} \mathbf{u} \right\rangle z^t \right|^2 dm(z) \tag{37}$$

$$= \int_{\mathbb{T}} \left| \sum_{t=1}^{T-1} \sum_{j=1}^{T-1} \varepsilon_j \omega_{t-j} z^t \right|^2 dm(z) \tag{38}$$

where $\omega_j = \langle \mathbf{v}, \mathbf{B}^j \mathbf{u} \rangle$ for $j \geq 0$, $\omega_j = 0$ for $j < 0$. Changing the order of summation and by

a simple substitution we get

$$\sum_{t=1}^{T-1}\sum_{j=1}^{T-1}\varepsilon_j\omega_{t-j}z^t=\sum_{j=1}^{T-1}\varepsilon_j z^j\sum_{t=1}^{T-1}\omega_{t-j}z^{t-j}=\sum_{j=1}^{T-1}\varepsilon_j z^j\sum_{t=0}^{T-j-1}\omega_t z^t. \tag{39}$$

Using the definition of ω_j, the fact that all eigenvalues of $\mathbf{B}$ have modulus strictly lower than one and Lemma A.2, (39) simplifies to

$$\begin{aligned}\sum_{j=1}^{T-1}\varepsilon_j z^j\sum_{t=0}^{T-j-1}\omega_t z^t&=\sum_{j=1}^{T-1}\varepsilon_j z^j\left\langle\mathbf{v},(\mathbf{I}_p-(\mathbf{B}z)^{T-j})(\mathbf{I}_p-\mathbf{B}z)^{-1}\mathbf{u}\right\rangle\\&=\sum_{j=1}^{T-1}\varepsilon_j\left(z^j\left\langle\mathbf{v},(\mathbf{I}_p-\mathbf{B}z)^{-1}\mathbf{u}\right\rangle-z^T\left\langle\mathbf{B}^{T-j}\mathbf{v},(\mathbf{I}_p-\mathbf{B}z)^{-1}\mathbf{u}\right\rangle\right)\\&=b(z)^{-1}\sum_{j=1}^{T-1}\varepsilon_j\left(z^j v(z)-z^T w_j(z)\right),\end{aligned}$$

where $v(z)=\sum_{k=1}^{p}v_k z_{k-1}$ and $w_j(z)=\sum_{k=1}^{p}(\mathbf{B}^{T-j}v)_k z^{k-1}$ for $j=0,\ldots,T-1$. The equation above, (37) and (39) combined together imply the following inequality

$$\begin{aligned}\langle\mathbf{v},\mathbf{C}_T\mathbf{v}\rangle&=\int_{\mathbb{T}}\left|b(z)^{-1}\sum_{j=1}^{T-1}\varepsilon_j\left(z^j v(z)-z^T w_j(z)\right)\right|^2 dm(z)\\&\geq\overline{b}^{-2}\int_{\mathbb{T}}\left|\sum_{j=1}^{T-1}\varepsilon_j\left(z^j v(z)-z^T w_j(z)\right)\right|^2 dm(z).\end{aligned}$$

Observe that $\sum_{j=1}^{T-1}\varepsilon_j\left(z^j v(z)-z^T w_j(z)\right)=\sum_{j=1}^{T-1}\varepsilon_j\left(z^j v(z)-z^T w_j(z)\right)=\sum_{t=1}^{T+p-1}c_t z^t$ is a trigonometric polynomial, therefore by Theorem A.1 and simple algebra

$$\begin{aligned}\int_{\mathbb{T}}\left|\sum_{j=1}^{T-1}\varepsilon_j\left(z^j v(z)-z^T w_j(z)\right)\right|dm(z)&=\sum_{t=1}^{T+p-1}|c_t|^2\geq\sum_{t=1}^{T-1}|c_t|^2=\sum_{t=1}^{T-1}\left(\sum_{j=1}^{p}v_j\varepsilon_{t-j+1}\right)^2=\\&=\sum_{i,j=1}^{p}v_j v_i\sum_{t=1}^{T-1}\varepsilon_{t-j+1}\varepsilon_{t-i+1},\end{aligned}$$

which proves (36).

We are now in a position to bound $\langle \mathbf{v}, \mathbf{C}_T \mathbf{v} \rangle$ from below. Rearranging terms in (36) yields

$$\begin{aligned}
\langle \mathbf{v}, \mathbf{C}_T \mathbf{v} \rangle &\geq \overline{b}^{-2} \left(\sum_{i=1}^{p} v_i^2 \sum_{t=1}^{n-i} \varepsilon_t^2 + \sum_{1 \leq i < j \leq p} v_i v_j \sum_{t=1}^{T-\max(i,j)} \varepsilon_t \varepsilon_{t+|j-i|} \right) \\
&\geq \overline{b}^{-2} \left(\sum_{t=1}^{T-p} \varepsilon_t^2 \sum_{i=1}^{p} v_i^2 - \max_{1 \leq i < j \leq p} \left| \sum_{t=1}^{T-\max(i,j)} \varepsilon_t \varepsilon_{t+|j-i|} \right| \left(\left(\sum_{i=1}^{p} |v_i| \right)^2 - \sum_{i=1}^{p} v_i^2 \right) \right) \\
&\geq \overline{b}^{-2} \left(\sum_{t=1}^{T-p} \varepsilon_t^2 - (p-1) \max_{1 \leq i < j \leq p} \left| \sum_{t=1}^{T-\max(i,j)} \varepsilon_t \varepsilon_{t+|j-i|} \right| \right).
\end{aligned}$$

Recalling the definition of $\mathcal{E}_T$, we conclude that on this event

$$\begin{aligned}
\langle \mathbf{v}, \mathbf{C}_T \mathbf{v} \rangle &\geq \overline{b}^{-2} \left(T - p - 2\sqrt{\log(T)(T-p)} - (p-1)32\log(T)\sqrt{T} \right) \\
&\geq \overline{b}^{-2} \left(T - p(1 + 32\log(T)\sqrt{T}) \right).
\end{aligned}$$

Taking infimum over $\mathbf{v} \in \mathbb{R}^p$ such that $\|\mathbf{v}\| = 1$ in the inequality above proves (31).

Lower bound for $\mathbb{P}(\mathcal{E}_T)$. Recalling (32) and using a simple Bonferroni bound, we get

$$\begin{aligned}
\mathbb{P}(\mathcal{E}_T^c) &\leq p^2 \max_{1 \leq i < j \leq p} \mathbb{P} \left(\left| \sum_{t=1}^{T-\max(i,j)} \varepsilon_t \varepsilon_{t+|i-j|} \right| \geq 32\log(T)\sqrt{T - \max(i,j)} \right) \\
&\quad + T \max_{1 \leq j \leq T} \mathbb{P} \left(\left| \sum_{t=1}^{T-j} \varepsilon_t \varepsilon_{t+j} \right| < 32\log(T)\sqrt{T-j} \right) \\
&\quad + \mathbb{P} \left(\sum_{t=1}^{T-p} \varepsilon_t^2 > T - p - 2\sqrt{\log(T)(T-p)} \right) \\
&:= p^2 \max_{1 \leq i < j \leq p} P_{i,j}^{(1)} + T \max_{1 \leq j \leq T} P_j^{(2)} + P^{(3)}.
\end{aligned}$$

Lemma A.3 implies that

$$\begin{aligned}
P_{i,j}^{(1)} &\leq 2\exp\left(-\frac{1}{8} \frac{(32\log(T))^2}{6 + (\sqrt{T - \max(i,j)})^{-1} 32\log(T)} \right) \leq 2\exp\left(-2\log(T)\right) = \frac{2}{T^2}, \\
P_j^{(2)} &\leq 2\exp\left(-\frac{1}{8} \frac{(32\log(T))^2}{6 + (\sqrt{T-j})^{-1} 32\log(T)} \right) \leq 2\exp\left(-2\log(T)\right) = \frac{2}{T^2}.
\end{aligned}$$

Moreover, by Lemma A.4, $P^{(3)} \leq \exp(-\log(T)) = \frac{1}{T}$, hence, given that $p^2 < T$, we have

$\mathbb{P}(\mathcal{E}_T^c) \leq \frac{5}{T}$, which completes the proof. $\square$

B Proof of Theorem 2.2

The proof is split into four steps.

Step 1. Consider the event $\left\{\left\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right\| \leq \kappa_1(\underline{b}/\bar{b})^2 \|\boldsymbol{\beta}\| \frac{p\log(T)\sqrt{\log(T+p)}}{\sqrt{T}-\kappa_2 p\log(T)}\right\}$ where κ_1, κ_2 are as in Theorem 2.1. Assumption (A3) implies that $\underline{b}/\bar{b}$ and $\|\boldsymbol{\beta}\|$ are bounded from above by constants. Furthermore, by (A2), $p \leq C_1 T^{\theta}$, which implies that

$$\kappa_1(\underline{b}/\bar{b})^2 \|\boldsymbol{\beta}\| \frac{p\log(T)\sqrt{\log(T+p)}}{\sqrt{T}-\kappa_2 p\log(T)} \leq CT^{a-1/2}(\log(T))^{3/2} =: \lambda_T$$

for some constant $C > 0$ and a sufficiently large T. Define now

$$A_T = \left\{\left\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right\| \leq \lambda_T\right\} \tag{40}$$

By Theorem 2.1,

$$\mathbb{P}(A_T) \geq \mathbb{P}\left(\left\|\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}\right\| \leq \kappa_1(\underline{b}/\bar{b})^2 \|\boldsymbol{\beta}\| \frac{p\log(T)\sqrt{\log(T+p)}}{\sqrt{T}-\kappa_2 p\log(T)}\right) \geq 1 - \kappa_3 T^{-1}, \tag{41}$$

for some constant $\kappa_3 > 0$.

Step 2. For $j = 1, \ldots, q$, define the intervals

$$\mathcal{I}_j^L = (\tau_j - \delta_T/3, \tau_j - \delta_T/6) \tag{42}$$

$$\mathcal{I}_j^R = (\tau_j + \delta_T/6, \tau_j + \delta_T/3) \tag{43}$$

Recall that F_T^M is the set of M randomly drawn intervals with endpoints in $\{1, \ldots, p\}$. Denote by $[s_1, e_1], \ldots, [s_M, e_M]$ the elements of F_T^M and let

$$D_T^M = \left\{\forall j = 1, \ldots, q, \exists k \in \{1, \ldots, M\}, \text{ s.t. } s_k \times e_k \in \mathcal{I}_j^L \times \mathcal{I}_j^R\right\}. \tag{44}$$

We have that

$$\mathbb{P}\left((D_T^M)^c\right) \leq \sum_{j=1}^{q} \Pi_{m=1}^{M}\Big(1-\mathbb{P}\left(s_m \times e_m \in \mathcal{I}_j^L \times \mathcal{I}_j^R\right)\Big)$$
$$\leq q\left(1-\frac{\delta_T^2}{6^2p^2}\right)^M \leq \frac{p}{\delta_T}\left(1-\frac{\delta_T^2}{36p^2}\right)^M.$$

Therefore, $\mathbb{P}\left(A_T \cap D_T^M\right) \geq 1-\kappa_3 T^{-1} - T\delta_T^{-1}(1-\delta_T^2 p^{-2}/36)^M$. In the remainder of the proof, assume that A_T and D_T^M all hold. We specify the constants

$$C_1 = 2\sqrt{C_3}+1, \quad C_2 = \frac{1}{\sqrt{6}} - \frac{2\sqrt{2}}{\underline{C}}, \quad C_3 = (4\sqrt{2}+6).$$

(We need to ensure $\underline{C}C_2 > C_1$, and thus $C_2\delta_T^{1/2}\underline{f}_T > C_1\sqrt{\log(T)}$, i.e., we can select $\zeta_T \in [C_1\sqrt{\log(T)}, C_2\delta_T^{1/2}\underline{f}_T)$. This is indeed the case because $\underline{C}$ is sufficiently large.)

Step 3. We focus on a generic interval $[s,e]$ such that

$$\exists j \in \{1,\ldots,q\},\ \exists k \in \{1,\ldots,M\},\ \text{s.t. } [s_k,e_k] \subset [s,e] \text{ and } s_k \times e_k \in \mathcal{I}_j^L \times \mathcal{I}_j^R. \tag{45}$$

Fix such an interval $[s,e]$ and let $j \in \{1,\ldots,q\}$ and $k \in \{1,\ldots,M\}$ be such that (45) is satisfied. Let $b_k^* = \text{argmax}_{s_k \leq b \leq e_k} \mathcal{C}_{s_k,e_k}^b\left(\hat{\boldsymbol{\beta}}\right)$. By construction, $[s_k,e_k]$ satisfies $\tau_j - s_k + 1 \geq \delta_T/6$ and $e_k - \tau_j > \delta_T/6$. Let

$$\mathcal{M}_{s,e} = \left\{m : [s_m,e_m] \in F_T^M, [s_m,e_m] \subset [s,e]\right\},$$
$$\mathcal{O}_{s,e} = \{m \in \mathcal{M}_{s,e} : \max_{s_m \leq b < e_m} \mathcal{C}_{s_m,e_m}^b\left(\hat{\boldsymbol{\beta}}\right) > \zeta_T\}.$$

Our first aim is to show that $\mathcal{O}_{s,e}$ is non-empty. This follows from Lemma 2 in Baranowski et al. (2019) and the calculation below.

$$\begin{aligned}
\mathcal{C}^{b_k^*}_{s_k,e_k}\left(\hat{\boldsymbol{\beta}}\right) &\geq \mathcal{C}^{\tau_j}_{s_k,e_k}\left(\hat{\boldsymbol{\beta}}\right) \\
&\geq \mathcal{C}^{b_k^*}_{s_k,e_k}\left(\boldsymbol{\beta}\right) - \lambda_T \geq \left(\frac{\delta_T}{6}\right)^{1/2} |\alpha_j \tau_j^{-1}| - \lambda_T \geq \left(\frac{\delta_T}{6}\right)^{1/2} \underline{\alpha}_T - \lambda_T \\
&= \left(\frac{1}{\sqrt{6}} - \frac{\lambda_T}{\delta_T^{1/2}\underline{\alpha}_T}\right) \delta_T^{1/2}\underline{\alpha}_T \geq \left(\frac{1}{\sqrt{6}} - \frac{2\sqrt{2}}{\underline{C}}\right) \delta_T^{1/2}\underline{\alpha}_T = C_2 \delta_T^{1/2}\underline{\alpha}_T > \zeta_T.
\end{aligned}$$

Let $m^* = \operatorname{argmin}_{m\in\mathcal{O}_{s,e}}(e_m - s_m + 1)$ and $b^* = \operatorname{argmax}_{s_{m^*}\leq b<e_{m^*}} \mathcal{C}^b_{s_{m^*},e_{m^*}}\left(\hat{\boldsymbol{\beta}}\right)$. Observe that $[s_{m^*}, e_{m^*})$ must contain at least one change-point. Indeed, if this were not the case, we would have $\mathcal{C}^b_{s_{m^*},e_{m^*}}\left(\boldsymbol{\beta}\right) = 0$ and

$$\mathcal{C}^{b^*}_{s_{m^*},e_{m^*}}\left(\hat{\boldsymbol{\beta}}\right) = |\mathcal{C}^{b^*}_{s_{m^*},e_{m^*}}\left(\hat{\boldsymbol{\beta}}\right) - \mathcal{C}^{b^*}_{s_{m^*},e_{m^*}}\left(\boldsymbol{\beta}\right)| \leq \lambda_T < C_1\lambda_T \leq \zeta_T,$$

which contradicted $\mathcal{C}^{b^*}_{s_{m^*},e_{m^*}}\left(\hat{\boldsymbol{\beta}}\right) > \zeta_T$. On the other hand, $[s_{m^*}, e_{m^*})$ cannot contain more than one change-points, because $e_{m^*} - s_{m^*} + 1 \leq e_k - s_k + 1 \leq \delta_T$.

Without loss of generality, assume $\tau_j \in [s_{m^*}, e_{m^*}]$. Let $\eta_L = \tau_j - s_{m^*} + 1$, $\eta_R = e_{m^*} - \tau_j$ and $\eta_T = (C_1 - 1)^2\alpha_j^2\tau_j^{-2}\lambda_T^2$. We claim that $\min(\eta_L, \eta_R) > \eta_T$, because $\min(\eta_L, \eta_R) \leq \eta_T$ and Lemma 2 in Baranowski et al. (2019) would have resulted in

$$\begin{aligned}
\mathcal{C}^{b^*}_{s_{m^*},e_{m^*}}\left(\hat{\boldsymbol{\beta}}\right) &\leq \mathcal{C}^{b^*}_{s_{m^*},e_{m^*}}\left(\boldsymbol{\beta}\right) + \lambda_T \leq \mathcal{C}^{\tau_j}_{s_{m^*},e_{m^*}}\left(\boldsymbol{\beta}\right) + \lambda_T \leq \eta_T^{1/2}|\alpha_j\tau_j^{-1}| + \lambda_T \\
&= (C_1 - 1 + 1)\lambda_T = C_1\lambda_T \leq \zeta_T,
\end{aligned}$$

which contradicted $\mathcal{C}^{b^*}_{s_{m^*},e_{m^*}}\left(\hat{\boldsymbol{\beta}}\right) > \zeta_T$.

We are now in a position to prove $|b^* - \tau_j| \leq C_3\lambda_T\underline{\alpha}_T^{-2}$. Our aim is to find ϵ_T such that for any $b \in \{s_{m^*}, s_{m^*} + 1, \ldots, e_{m^*} - 1\}$ with $|b - \tau_j| > \epsilon_T$, we always have

$$(\mathcal{C}^{\tau_j}_{s_{m^*},e_{m^*}}\left(\hat{\boldsymbol{\beta}}\right))^2 - (\mathcal{C}^{b}_{s_{m^*},e_{m^*}}\left(\hat{\boldsymbol{\beta}}\right))^2 > 0. \tag{46}$$

This would then imply that $|b^* - \tau_j| \leq \varepsilon_T$. By expansion and rearranging the terms, we see

that (46) is equivalent to

$$\langle \boldsymbol{\beta}, \boldsymbol{\psi}^{\tau_j}_{s_{m^*},e_{m^*}} \rangle^2 - \langle \boldsymbol{\beta}, \boldsymbol{\psi}^{b}_{s_{m^*},e_{m^*}} \rangle^2 > \langle \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}, \boldsymbol{\psi}^{b}_{s_{m^*},e_{m^*}} \rangle^2 - \langle \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}, \boldsymbol{\psi}^{\tau_j}_{s_{m^*},e_{m^*}} \rangle^2$$
$$+ 2\Big\langle \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}, \boldsymbol{\psi}^{b}_{s_{m^*},e_{m^*}} \langle \boldsymbol{\beta}, \boldsymbol{\psi}^{b}_{s_{m^*},e_{m^*}} \rangle - \boldsymbol{\psi}^{\tau_j}_{s_{m^*},e_{m^*}} \langle \boldsymbol{\beta}, \boldsymbol{\psi}^{\tau_j}_{s_{m^*},e_{m^*}} \rangle \Big\rangle. \tag{47}$$

In the following, we assume that $b \geq \tau_j$. The case that $b < \tau_j$ can be handled in a similar fashion. By Lemma 4 in Baranowski et al. (2019), we have

$$\begin{aligned}\langle \boldsymbol{\beta}, \boldsymbol{\psi}^{\tau_j}_{s_{m^*},e_{m^*}} \rangle^2 - \langle \boldsymbol{\beta}, \boldsymbol{\psi}^{b}_{s_{m^*},e_{m^*}} \rangle^2 &= (\mathcal{C}^{\tau_j}_{s^*,e^*}(\boldsymbol{\beta}))^2 - (\mathcal{C}^{b}_{s_{m^*},e_{m^*}}(\boldsymbol{\beta}))^2 \\ &= \frac{|b-\tau_j|\eta_L}{|b-\tau_j|+\eta_L}(\alpha_j \tau_j^{-1})^2 =: \kappa.\end{aligned}$$

In addition, since we assume event A_T,

$$\begin{aligned}&\langle \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}, \boldsymbol{\psi}^{b}_{s_{m^*},e_{m^*}} \rangle^2 - \langle \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}, \boldsymbol{\psi}^{\tau_j}_{s_{m^*},e_{m^*}} \rangle^2 \leq \lambda_T^2, \\ &2\Big\langle \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}, \boldsymbol{\psi}^{b}_{s_{m^*},e_{m^*}} \langle \boldsymbol{\beta}, \boldsymbol{\psi}^{b}_{s_{m^*},e_{m^*}} \rangle - \boldsymbol{\psi}^{\tau_j}_{s_{m^*},e_{m^*}} \langle \boldsymbol{\beta}, \boldsymbol{\psi}^{\tau_j}_{s_{m^*},e_{m^*}} \rangle \Big\rangle \\ &\qquad \leq 2\|\boldsymbol{\psi}^{b}_{s_{m^*},e_{m^*}} \langle \boldsymbol{\beta}, \boldsymbol{\psi}^{b}_{s_{m^*},e_{m^*}} \rangle - \boldsymbol{\psi}^{\tau_j}_{s_{m^*},e_{m^*}} \langle \boldsymbol{\beta}, \boldsymbol{\psi}^{\tau_j}_{s_{m^*},e_{m^*}} \rangle\|_2 \lambda_T = 2\kappa^{1/2}\lambda_T,\end{aligned}$$

where the final equality is also implied by Lemma 4 in Baranowski et al. (2019). Consequently, (47) can be deducted from the stronger inequality $\kappa - 2\lambda_T\kappa^{1/2} - \lambda_T^2 > 0$. This quadratic inequality is implied by $\kappa > (\sqrt{2}+1)^2\lambda_T^2$, and could be restricted further to

$$\frac{2|b-\tau_j|\eta_L}{|b-\tau_j|+\eta_L} \geq \min(|b-\tau_j|, \eta_L) > (4\sqrt{2}+6)(\underline{\alpha}\tau_j^{-1})^{-2}\lambda_T^2 = C_3(\alpha_j\tau_j^{-1})^{-2}\lambda_T^2. \tag{48}$$

But since

$$\eta_L \geq \eta_T = (C_1-1)^2(\alpha_j\tau_j^{-1})^{-2}\lambda_T^2 = (2\sqrt{C_3})^2(\alpha_j\tau_j^{-1})^{-2}\lambda_T^2 > C_3(\alpha_j\tau_j^{-1})^{-2}\lambda_T^2,$$

we see that (48) is equivalent to $|b-\tau_j| > C_3(\alpha_j\tau_j^{-1})^{-2}\lambda_T^2$. To sum up, $|b^*-\tau_j| > C_3(\alpha_j\tau_j^{-1})^{-2}\lambda_T^2$ would result in (46), a contradiction. So we have proved that $|b^*-\tau_j| \leq C_3(\alpha_j\tau_j^{-1})^{-2}\lambda_T^2$.

Step 4. With the arguments above valid on the event $A_T \cap B_T \cap D_T^M$, we can now proceed with the proof of the theorem. At the start of Algorithm 1, we have $s = 1$ and $e = T$ and, provided that $q \geq 1$, condition (45) is satisfied. Therefore the algorithm detects a change-point b^* in that interval such that $|b^* - \tau_j| \leq C_3(\alpha_j \tau_j^{-1})^{-2}\lambda_T^2$. By construction, we also have that $|b^* - \tau_j| < 2/3\delta_T$. This in turn implies that for all $l = 1, \ldots, q$ such that $\tau_l \in [s, e]$ and $l \neq j$ we have either $\mathcal{I}_l^L, \mathcal{I}_l^R \subset [s, b^*]$ or $\mathcal{I}_l^L, \mathcal{I}_l^R \subset [b^* + 1, e]$. Therefore (45) is satisfied within each segment containing at least one change-point. Note that before all q change points are detected, each change point will not be detected twice. To see this, we suppose that τ_j has already been detected by b, then for all intervals $[s_k, e_k] \subset [\tau_j - C_3(\alpha_j \tau_j^{-1})^{-2}\lambda_T^2 + 1, \tau_j - C_3(\alpha_j \tau_j^{-1})^{-2}\lambda_T^2 + 2/3\delta_T + 1] \cup [\tau_j + C_3(\alpha_j \tau_j^{-1})^{-2}\lambda_T^2 - 2/3\delta_T, \tau_j + C_3(\alpha_j \tau_j^{-1})^{-2}\lambda_T^2]$, Lemma 2 in Baranowski et al. (2019), together with the definition of A_T, guarantee that

$$\begin{aligned}
\max_{s_k \leq b < e} \mathcal{C}_{s_k, e_k}^b \left(\hat{\boldsymbol{\beta}}\right) &\leq \max_{s \leq b < e} \mathcal{C}_{s_k, e_k}^b (\boldsymbol{\beta}) + \lambda_T \\
&\leq \sqrt{C_3(\alpha_j \tau_j^{-1})^{-2}\lambda_T^2}\alpha_j \tau_j^{-1} + \sqrt{C_3(\alpha_{j+1}\tau_{j+1}^{-1})^{-2}\lambda_T^2}\alpha_{j+1}\tau_{j+1}^{-1} + \lambda_T \\
&< (2\sqrt{C_3} + 1)\lambda_T = C_1 \lambda_T \leq \zeta_T.
\end{aligned}$$

Once all the change-points have been detected, we then only need to consider $[s_k, e_k]$ such that

$$[s_k, e_k] \subset [\tau_j - C_3(\alpha_j \tau_j^{-1})^{-2} + 1, \tau_{j+1} + C_3(\alpha_{j+1}\tau_{j+1}^{-1})^{-2}]$$

for $j = 1, \ldots, q$. For such intervals, we have

$$\max_{s_k \leq b < e_k} \mathcal{C}_{s_k, e_k}^b \left(\hat{\boldsymbol{\beta}}\right) \leq \max_{s \leq b < e} \mathcal{C}_{s_k, e_k}^b (\boldsymbol{\beta}) + \lambda_T \leq \sqrt{C_3(\alpha_j \tau_j^{-1})^{-2}\lambda_T^2}\alpha_j \tau_j^{-1} + \lambda_T \leq C_1 \lambda_T \leq \zeta_T.$$

Hence the algorithm terminates and no further change-points are detected. □

References

S.K. Ahn and G. Reinsel. Nested reduced-rank autoregressive models for multiple time series. *Journal of the American Statistical Association*, 83:849–856, 1988.

A. Anastasiou and P. Fryzlewicz. Detecting multiple generalized change-points by isolating single ones. *Preprint*, 2019.

E. Andreou, E. Ghysels, and A. Kourtellos. Should macroeconomic forecasters use daily financial data and how? *Journal of Business and Economic Statistics*, 31:240–251, 2013.

A. Assaf, G. Li, H. Song, and M. Tsionas. Modeling and forecasting regional tourism demand using the Bayesian Global Vector Autoregressive (BGVAR) model. *Journal of Travel Research*, 58:383–397, 2019.

J. Bai, E. Ghysels, and J. Wright. State space models and MIDAS regressions. *Econometric Reviews*, 32:779–813, 2013.

A. Barabanov. On strong convergence of the method of least squares. *Avtomatika i Telemekhanika*, 44:119–127, 1983.

R. Baranowski and P. Fryzlewicz. **amar**: Adaptive Multiscale Autoregressive time series models, 2016a. URL `https://github.com/rbaranowski/amar`. R package version 1.00.

R. Baranowski and P. Fryzlewicz. Adaptive Multiscale Autoregressive time series models: simulation code, 2016b. URL `https://github.com/rbaranowski/amar-num-ex`.

R. Baranowski, Y. Chen, and P. Fryzlewicz. Narrowest-over-threshold detection of multiple change points and change-point-like features. *Journal of the Royal Statistical Society Series B*, 81:649–672, 2019.

M. Basseville, A. Benveniste, and A. Willsky. Multiscale autoregressive processes, part I: Schur-Levinson parametrizations. *IEEE Trans. Sig. Proc.*, 40:1915–1934, 1992.

D. Bates and D. Eddelbuettel. Fast and elegant numerical linear algebra using the **RcppEigen** package. *Journal of Statistical Software*, 52:1–24, 2013.

B. Bercu and A. Touati. Exponential inequalities for self-normalized martingales with applications. *The Annals of Applied Probability*, 18:1848–1869, 2008.

P. Brockwell and R. Davis. *Introduction to Time Series and Forecasting*. Springer, 3rd edition, 2016.

C. Brownlees and G. Gallo. Financial econometric analysis at ultra-high frequency: Data handling concerns. *Computational Statistics and Data Analysis*, 51:2232–2245, 2006.

F. Calvori, F. Cipollini, and G. Gallo. **TAQMNGR**: Manage tick-by-tick transaction data, 2015. URL `http://CRAN.R-project.org/package=TAQMNGR`. R package version 2015.2-1.

H. Cho and P. Fryzlewicz. Multiscale interpretation of taut string estimation and its connection to Unbalanced Haar wavelets. *Statistics and Computing*, 21:671–681, 2011.

M. Clements and A. Galvão. Forecasting US output growth using leading indicators: An appraisal using MIDAS models. *Journal of Applied Econometrics*, 24:1187–1206, 2009.

F. Corsi. A simple approximate long-memory model of realized volatility. *Journal of Financial Econometrics*, 7:174–196, 2009.

F. Corsi, F. Audrino, and R. Reno. HAR modeling for realized volatility forecasting. In *Handbook of Volatility Models and Their Applications*, pages 363–382. John Wiley & Sons, New Jersey, USA, 2012.

G. Cubadda, B. Guardabascio, and A. Hecq. A vector heterogeneous autoregressive index model for realized volatility measures. *International Journal of Forecasting*, 33:337–344, 2017.

K. Daoudi, A. B. Frakt, and A. S. Willsky. Multiscale autoregressive models and wavelets. *IEEE Transactions on Information Theory*, 45:828–845, 1999. doi: 10.1109/18.761321.

J. Duoandikoetxea. *Fourier Analysis*, volume 29 of *Graduate Studies in Mathematics*. American Mathematical Society, 2001.

M. Ferreira and H. Lee. *Multiscale Modeling: a Bayesian Perspective*. Springer, 2007.

M. Ferreira, M. West, H. Lee, and D. Higdon. Multi-scale and hidden resolution time series models. *Bayesian Analysis*, 1:947–967, 2006.

M. Ferreira, A. Bertolde, and S. Holan. Analysis of economic data with multiscale spatio-temporal models. In *Handbook of Applied Bayesian Analysis*, pages 295–318. Oxford University Press, 2010.

L. Forsberg and E. Ghysels. Why do absolute returns predict volatility so well? *Journal of Financial Econometrics*, 5:31–67, 2007.

P. Fryzlewicz. On multi-zoom autoregressive time series models. *Oberwolfach Reports*, 48/2013:21–24, 2013.

E. Gardner. Exponential smoothing: The state of the art. *Journal of Forecasting*, 4:1–28, 1985.

E. Ghysels, P. Santa-Clara, and R. Valkanov. The MIDAS touch: Mixed data sampling regression models. Technical report, University of North Carolina and UCLA, 2004.

E. Ghysels, A. Sinko, and R. Valkanov. MIDAS regressions: Further results and new directions. *Econometric Reviews*, 26:53–90, 2007.

S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 9: 1735–1780, 1997.

E. Hwang and D.W. Shin. A CUSUM test for a long memory heterogeneous autoregressive model. *Economics Letters*, 121:379–383, 2013.

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

C.K. Ing and C.Z. Wei. Order selection for same-realization predictions in autoregressive processes. *The Annals of Statistics*, 33:2423–2474, 2005.

C. Jarque and A. Bera. Efficient tests for normality, homoscedasticity and serial independence of regression residuals. *Economics Letters*, 6:255–259, 1980.

G. Kitagawa. *Introduction to Time Series Modeling.* Chapman & Hall/CRC, 2010.

T.L. Lai and C.Z. Wei. Asymptotic properties of general autoregressive models and strong consistency of least-squares estimates of their parameters. *Journal of Multivariate Analysis*, 13:1–23, 1983.

B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *The Annals of Statistics*, 28:1302–1338, 2000.

H. Maeng and P. Fryzlewicz. Regularized forecasting via smooth-rough partitioning of the regression coefficients. *Electronic Journal of Statistics,* to appear, 2019.

M. McAleer and M. Medeiros. A multiple regime smooth transition Heterogeneous Autoregressive model for long memory and asymmetries. *Journal of Econometrics*, 147:104–119, 2008.

U. Müller, M. Dacorogna, R. Davé, R. Olsen, O. Pictet, and J. von Weizsäcker. Volatilities of different time resolutions — Analyzing the dynamics of market components. *Journal of Empirical Finance*, 4:213–239, 1997.

G. Nason. *Wavelet Methods in Statistics with R.* Springer Science and Business Media, 2010.

G. Nason, R. von Sachs, and G. Kroisandt. Wavelet processes and adaptive estimation of the evolutionary wavelet spectrum. *Journal of the Royal Statistical Society: Series B*, 62: 271–292, 2000.

D. Percival and A. Walden. *Wavelet Methods for Time Series Analysis.* Cambridge University Press, 2006.

G. Petnehazi. Recurrent neural networks for time series forecasting. *Preprint*, 2019.

D. Politis. A normalizing and variance-stabilizing transformation for financial time series. In *Recent Advances and Trends in Nonparametric Statistics*, pages 335–347. Elsevier, 2003.

D. Politis. Model-free versus model-based volatility prediction. *Journal of Financial Econometrics*, 5:358–359, 2007.

D. Politis and D. Thomakos. NoVaS transformations: flexible inference for volatility forecasting. In *Recent Advances and Future Directions in Causality, Prediction, and Specification Analysis*, pages 489–525. Springer, 2013.

E. Raviv, K. Bouwman, and D. van Dijk. Forecasting day-ahead electricity prices: Utilizing hourly prices. *Energy Economics*, 50:227–239, 2015.

G. Reinsel. Some results on multivariate autoregressive index models. *Biometrika*, 70: 145–156, 1983.

A. Schroeder and P. Fryzlewicz. Adaptive trend estimation in financial time series via multiscale change-point-induced basis recovery. *Statistics and Its Interface*, 6:449–461, 2013.

N. Schuurman and E. Hamaker. Measurement error and person-specific reliability in multilevel autoregressive modeling. *Psychological Methods*, 24:70–91, 2019.

R. Shumway and D. Stoffer. *Time Series Analysis and Its Applications*. Springer, 4th edition, 2010.

J. Taylor. Volatility forecasting with smooth transition exponential smoothing. *International Journal of Forecasting*, 20:273–286, 2004.

R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight. Sparsity and smoothness via the fused lasso. *J. R. Statist. Soc. B*, 67:91–108, 2005.

R. Velu, G. Reinsel, and D. Wichern. Reduced rank models for multiple time series. *Biometrika*, 73:105–118, 1986.

R. Vershynin. A simple decoupling inequality in probability theory. Technical report, University of Michigan, 2011. URL `https://www.math.uci.edu/~rvershyn/papers/decoupling-simple.pdf`.

B. Vidakovic. *Statistical Modeling by Wavelets*. Wiley, 1999.

W. Wang, Y. Yan, Z. Cui, J. Feng, S. Yan, and N. Sebe. Recurrent face aging with hierarchical autoregressive memory. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41:654–668, 2019.

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

M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *J. R. Statist. Soc. B*, 68:49–67, 2006.

Y. Zeng, J. Yang, C. Peng, and Y. Yin. Evolving Gaussian process autoregression based learning of human motion intent using improved energy kernel method of EMG. *IEEE Transactions on Biomedical Engineering,* to appear, 2019.