Statistica Sinica

Multiscale Autoregression on Adaptively Detected Timescales

Rafal Baranowski, Yining Chen and Piotr Fryzlewicz

London School of Economics and Political Science

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 is designed to offer simplicity and interpretability on the one hand, and modelling flexibility on the other. Our theory permits the longest timescale to increase with the sample size. A simulation study is presented to show the usefulness of our approach. Some possible extensions are also discussed, including the Adaptive Multiscale Vector AutoRegressive model (AMVAR) for multivariate time series, which demonstrates promising performance in the data example on UK and US unemployment rates. The R package **amar** (Baranowski et al., 2022) provides an efficient implementation of the AMAR framework.

Key words and phrases: multiscale modelling, regularised autoregression, piecewise-constant approximation, time series.

1. Introduction

1.1 Motivation and main idea

Autoregression in time series modelling is arguably the most frequently used device to characterise temporal dependence in data. The classical linear autoregressive model of order p, known as 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},...,X_{t-p}, plus white-noise-like innovation ε_t, that is $X_t = \beta_1 X_{t-1} + ... + \beta_p X_{t-p} + \varepsilon_t$ for $t=1,...,T$. However, in situations where the application of this model yields a large or even moderate 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 directly on the individual variables X_{t-1},...,X_{t-p}.

Motivated by this, we propose what we call a multiscale approach to time series autoregression, in which we include features of the path X_1,...,X_{t-1} that live on multiple timescales as linear regressors for X_t. To fix ideas, here we take these multiscale features to be the recent averages of X_t over multiple time spans (N.B. possible extensions will be discussed in Section 5), which are not necessarily known to the analyst a priori and need to be estimated from the data. This leads to the following **A**daptive **M**ultiscale **A**uto**R**egressive model of order q, abbreviated as

AMAR(q), for X_t:

$$X_t = \alpha_1 \frac{X_{t-1}+...+X_{t-\tau_1}}{\tau_1} + ... + \alpha_q \frac{X_{t-1}+...+X_{t-\tau_q}}{\tau_q} + \varepsilon_t, \quad t=1,...,T, \qquad (1.1)$$

where the number of timescales q, the timescales $1 \leq \tau_1 < \tau_2 < ... < \tau_q$ and the scale coefficients $\alpha_1,...,\alpha_q \in \mathbb{R}\backslash\{0\}$ are unknown (NB. zero is excluded for model identifiability), and where the innovations $\{\varepsilon_t\}$ follow a white noise process, which we take to mean a sequence of random variables that are uncorrelated, having zero-mean and a finite (but non-zero) variance. The number of scales q can possibly be much smaller than the largest timescale τ_q. 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.

In essence, the AMAR(q) model is a multiscale, sparsely parameterised, version of the AR(τ_q) process, which permits the longest timescale τ_q to be large in practice. These properties make the AMAR framework particularly suitable for the exploratory analysis of processes in which a seasonal component may be suspected, or for the modelling of time series which exhibit low-frequency trends (which may give them a non-stationary appearance) accompanied by higher-frequency oscillations. We shall illustrate these claims in Section S1 of the supplementary materials.

1.2 Literature review

We now provide an overview of other related literature. Reinsel (1983) considers 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 be 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, $t=1,...,mT$, the fine-level process, where the integer $m>1$ is known, and the coarse-level aggregate process $X_t = m^{-1}\sum_{j=1}^{m} Y_{tm-j} + \varepsilon_t$ for $t=1,...,T$, where the noise term $\varepsilon_t \sim \mathcal{N}(0,\sigma^2)$ is independently and identically distributed (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. 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 MIxed DAta Sampling (MIDAS) 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 $b_1(\cdot;\boldsymbol{\beta}),...,b_p(\cdot;\boldsymbol{\beta})$ are given functions of the lagged observations recorded at a higher frequency and of a

low-dimensional vector of unknown parameters $\boldsymbol{\beta}=(\beta_1,...,\beta_q)^T$, and where ε_t is random noise. For each recorded observation of X_t, m values of Y_i are sampled. 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 $1\leq\tau_1<...<\tau_q$ are known integers. One important difference between this and the AMAR framework is that $\tau_1,...,\tau_q$ in our model are unknown.

In Heterogeneous AutoRegressive (HAR) modelling (Corsi, 2009), the quantity of interest is regressed on its past realised averages over given known multiple timescales. The author shows 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 Corsi et al. (2012) for a review of HAR modelling of realised volatility.

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 instead of their spans as in AMAR.

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. The use of LSTM models in time series forecasting is less well explored and the theoretical understanding of their behaviour in the context of time series modelling is limited, but see Petnehazi (2019) for a recent review. The complexity of LSTM models means that large samples are typically required to train them.

In addition, one could view the AMAR model as 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 (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 approaches based on the total variation penalty (e.g. fused LASSO) is not optimal for this task.

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 $\tau_1,...,\tau_q$ 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 $\tau_1,...,\tau_q$ is achieved by a change-point detection algorithm, more precisely, a "narrowest-over-threshold"-type (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. Besides, the AMAR framework can be extended in different ways to handle more complicated data structure, including multivariate time series. The promising performance of this particular extension, named Adaptive Multiscale Vector Au-

toRegressive model (AMVAR), is also demonstrated in this paper. The R package **amar** (Baranowski et al., 2022) provides an efficient implementation of our proposal.

2. Methodology and theory

2.1 Model framework

Recall that AMAR(q) is an instance of a sparsely parametrised AR model. Therefore, for any $p \geq \tau_q$, (1.1) can be rewritten as

$$X_t = \beta_1 X_{t-1} + ... + \beta_p X_{t-p} + \varepsilon_t, \quad t = 1,...,T, \tag{2.2}$$

$$\beta_j = \sum_{k:\tau_k \geq j} \frac{\alpha_k}{\tau_k}, \quad j = 1,...,p, \tag{2.3}$$

where $\{\varepsilon_t\}$ is a white noise process. Here we refer to (2.2) and (2.3) as an AR(p) representation of the AMAR(q) process. Also note that $\beta_j = 0$ for $j = \tau_q + 1,...,p$.

Let $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, ..., \hat{\beta}_p)^T$ be the Ordinary Least Squares (OLS) estimator of $\boldsymbol{\beta} = (\beta_1,...,\beta_p)^T$. Then $\hat{\beta}_j$'s can be trivially decomposed as

$$\hat{\beta}_j = \beta_j + (\hat{\beta}_j - \beta_j), \quad j = 1,...,p. \tag{2.4}$$

The coefficients $\beta_1,...,\beta_p$ form a piecewise-constant vector with change-points at the timescales $\tau_1,...,\tau_q$, 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 the AMAR models. First, we choose an adequate 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 the series $\beta_1,...,\beta_p$, 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.

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 (i.e. the vector $(\hat{\beta}_1,...,\hat{\beta}_p)$ here) that contain at most a single change-point each. This is a key fact that makes our version of the NOT method easily amenable to a theoretical analysis in the AMAR framework.

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. For the development of our theory, we work in a framework where the number of timescales q, the timescales $\tau_1,...,\tau_q$ and the coefficients $\alpha_1,...,\alpha_q$ possibly depend on the sample size T under Gaussian innovations, and are fixed under heavy-tailed innovations. However, for the economy of notation, we shall suppress the dependence of these quantities on T in the remainder of the paper.

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 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 re-

gions of piecewise constancy, and thereby estimate the unknown timescales $\tau_1,...,\tau_q$.

2.2 Stationarity

Recall that the characteristic polynomial of any AR(p) is $b(z) := b(z;\beta_1,...,\beta_p) = 1-\sum_{j=1}^{p}\beta_j z^j$ for $z \in \mathbb{C}$, where $\mathbb{C}$ denotes the complex plane. Also, the unit circle is denoted by $\mathbb{T} = \{z \in \mathbb{C} : |z| = 1\}$.

We now discuss the stationarity of the AMAR models with q and $\alpha_1,...,\alpha_q$ being fixed. Since AMAR is a special form of AR, $\{X_t\}$ that follows the AMAR model has a stationary (and causal) solution if and only if the roots of $b(z;\beta_1,...,\beta_{\tau_q})$ are outside $\mathbb{T}$, where $\beta_1,...,\beta_{\tau_q}$ are defined in Equation (2.3). Here any $z \in \mathbb{C}$ is outside $\mathbb{T}$ if and only if $|z| > 1$. Furthermore, a simplified sufficient condition for stationarity is given in the following result.

Proposition 2.1 *Given $\{X_t\}$ follows the AMAR(q) model in Equation* (1.1) *with $\alpha_1,...,\alpha_q$. If $\sum_{j=1}^{q}|\alpha_j| < 1$, then $\{X_t\}$ has a causal stationary solution. Suppose all the α_j's are non-negative, then the converse is true.*

We remark that when all the α_j's are non-negative, previous observations have non-negative effects on the current one. In this case, $\sum_{j=1}^{q}\alpha_j < 1$ would be a sufficient (but not necessary) condition for stationarity. Furthermore, the above proposition holds even when $q = \infty$. When q (or τ_q) increases with T, the stationarity property of AMAR would need to be discussed in a setup that involves triangular arrays. These details are omitted for notational convenience.

Finally, we offer visual insights into the behaviour of AMAR with a single scale in Figure 1. Here realisations for different values of α_1 (from 0.5 to 0.95, the latter corresponds to series that are near unit-root) and τ_1 (from 1 to 10) with standard Gaussian noise are plotted. It appears that the longer the scale, the noisier the appearance; the overall shape (driven by the low frequencies) is preserved, but the details (driven by the high frequencies) are increasingly obscured by noise. In addition, even though all the series plotted in Figure 1 are weakly stationary, some of them exhibit behaviour that mimics non-stationarity, at least visually, when τ_1 if large, even for a moderate α_1. This hints at the usefulness of AMAR in the modelling of near unit-root or certain non-stationary series. Regarding this, see also additional numerical results in the supplements. Moreover, insights into other more complex special cases can be found in Section S1 of the supplementary materials.

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

As a prelude to the study of the behaviour of our proposed AMAR scale and coefficient estimation procedure, 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 (2.2), with all bounds explicitly depending on T, p and the other parameters of the AR(p) process. For any vector $\mathbf{v}=(v_1,...,v_k)^T\in\mathbb{R}^k$, the Euclidean norm is denoted by $\|\mathbf{v}\|=\sqrt{\sum_{j=1}^k v_k^2}$. The following theorem holds.

Theorem 2.1 *Suppose that* $\{X_t\}_{t=1}^T$ *follows the AR(p) model* (2.2) *with the*

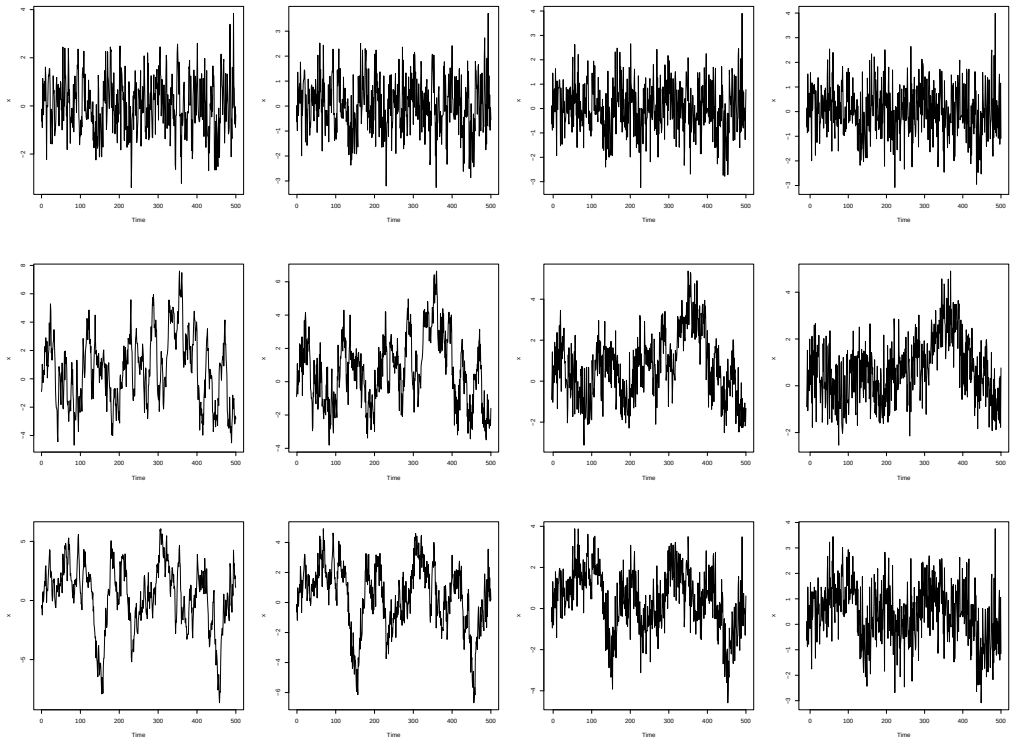

Figure 1: Simulated sample paths, of length $T=500$, from the single-scale AMAR model $X_t=\alpha_1\frac{X_{t-1}+...+X_{t-\tau_1}}{\tau_1}+\varepsilon_t$, for $t=1,...,T$, with $\alpha_1=0.5,0.9,0.95$ (respectively from top to bottom) and $\tau_1=1,2,5,10$ (respectively from left to right). The same random seed is used to generate path for each row.

innovations $\varepsilon_1,...,\varepsilon_T$ *being i.i.d.* $\mathcal{N}(0,\sigma^2)$ *with* $\sigma>0$. *Also assume that the process is stationary and causal. Let* $\hat{\boldsymbol{\beta}}=(\hat{\beta}_1,...,\hat{\beta}_p)^T$ *be the OLS estimator of the vector of the autoregressive coefficients* $\boldsymbol{\beta}=(\beta_1,...,\beta_p)^T$. *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*

$$\mathbb{P}\left(\left\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right\|\leq\kappa_1(\underline{b}/\overline{b})^2\|\boldsymbol{\beta}\|\frac{p\log(T)\sqrt{\log(T+p)}}{\sqrt{T}-\kappa_2 p\log(T)}\right)\geq 1-\frac{\kappa_3}{T}, \tag{2.5}$$

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 (2.4) converge to zero with $T\to\infty$, provided that $\frac{p\log(T)\sqrt{\log(T+p)}}{\sqrt{T}-\kappa_2 p\log(T)}\to 0$. Also note that

this result holds for any $\sigma > 0$, as the OLS estimate has the "self-normalising" property in the current setting, i.e. it remains unchanged when we scale the entire observed series $\{X_t\}$ (and thus σ) by a constant.

We remark that in a setting where both the order p and the autoregressive coefficients in model (2.2) 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 $\{\varepsilon_t\}$ is a martingale difference sequence with bounds on the conditional second moments. 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.

2.4 AMAR estimation algorithm

2.4.1 Timescale estimation

To estimate the timescales $\tau_1,...,\tau_q$, at which the change-points in model (2.4) are located, we adapt the Narrowest-Over-Threshold (NOT) approach of Baranowski et al. (2019), with the cumulative sum (CUSUM) contrast function $\mathcal{C}^b_{s,e}(\cdot)$ suitable for the piecewise-constant model, defined by

$$\mathcal{C}^b_{s,e}(\mathbf{v}) = \left| \sqrt{\frac{e-b}{(e-s+1)(b-s+1)}} \sum_{t=s}^{b} v_t - \sqrt{\frac{b-s+1}{(e-s+1)(e-b)}} \sum_{t=b+1}^{e} v_t \right|. \qquad (2.6)$$

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 (2.4) due to the complex dependence structure in $\hat{\beta}_j - \beta_j$, we show in Section 2.4.2 that here NOT estimators 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 (with its choice to be discussed in Section 3.1). The NOT procedure for the estimation of the timescales in the AMAR(q) model is described in Algorithm 1, which serves as a key ingredient of the AMAR estimation algorithm, given in Section 2.4.2. Core to this approach is a particular blend of global and local treatment of the data $\hat{\boldsymbol{\beta}}$ in the search for the multiple scales that may be present in the true $\boldsymbol{\beta}_0$. At the global stage, we look at the behaviour of $\hat{\boldsymbol{\beta}}$ over a large number of subintervals (either drawn randomly or systematically), $(\hat{\beta}_s, \ldots, \hat{\beta}_e)$, where $1 \leq s < e \leq p$. On each subinterval, we assume, possibly erroneously, that only one feature (i.e. scale) is present and use a contrast function (in this setting, CUSUM-based) to find the most likely location of the scale. We retain those subsamples for which the contrast exceeds a certain specified threshold, and discard the others. Amongst the retained subsamples, we search for the one drawn on the narrowest interval, i.e. one for which $e - s$ is the smallest. The focus on the narrowest interval constitutes the local part of the method, which ensures

that with high probability, at most (and also at least, with appropriate choice of threshold) one scale is present in the selected interval. Having detected the first scale, our Algorithm 1 then proceeds recursively to the left and to the right of it, and stops, on any current subinterval, if no contrasts can be found that exceed the threshold. More details regarding the intuitions and the construction of the contrast function under different settings can be found in Baranowski et al. (2019).

Algorithm 1 NOT algorithm for estimation of timescales in AMAR models

Input: Estimates $\hat{\boldsymbol{\beta}}=(\hat{\beta}_1,...,\hat{\beta}_p)^T$; s and e are the start- and end-points of an interval of interest; F_T^M is a set of intervals within $[1,p]$; and a given threshold ζ_T and $\mathcal{S}=\emptyset$ (as an initiation).

Output: Set of estimated timescales $\mathcal{S}=\{\hat{\tau}_1,...,\hat{\tau}_{\hat{q}}\}\subset\{1,...,p\}$, where $\hat{\tau}_1,...,\hat{\tau}_{\hat{q}}$ are in increasing order.

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

Algorithm 2 AMAR algorithm

Input: Data X_1,...,X_T, p; threshold ζ_T, and M (needed only if $p>500$).

Output: Estimates of the relevant scales $\hat{\tau}_1$,...,$\hat{\tau}_{\hat{q}}$ and the corresponding AMAR coefficients $\hat{\alpha}_1$,...,$\hat{\alpha}_{\hat{q}}$.

procedure AMAR($\{X_1,...,X_T\}$, p, ζ_T)

Step 1 Find $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \ldots, \hat{\beta}_p)^T$, the OLS estimates of the autoregressive coefficients in the AR(p) representation of AMAR(q).

Step 2 Let F_T^M be a set of all $M = p(p-1)/2$ intervals within [1,p] (i.e. [1,2],...,[1,p],[2,3],...,[2,p],...,[$p-1$,p]). If p is large (e.g. >500), we take F_T^M to be a set of M intervals whose start- and end-points have been drawn independently and uniformly from $\{1,...,p\}$ with replacement.

Step 3 Call NOT($\hat{\boldsymbol{\beta}}$,1,p,F_T^M,ζ_T) from Algorithm 1 to find the estimates of the timescales; Sort them in increasing order to obtain $\hat{\tau}_1$,...,$\hat{\tau}_{\hat{q}}$.

Step 4 With the timescales in (1.1) set to $\{\hat{\tau}_1,...,\hat{\tau}_{\hat{q}}\}$, find $\hat{\alpha}_1$,...,$\hat{\alpha}_{\hat{q}}$, the OLS estimates of the scale coefficients α_1,...,α_q.

end procedure

2.4.2 Parameter estimation

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 et al., 2022). The choice of all the input parameters is discussed in Section 3.1. As a remark, we note that in Step 4, finding the AMAR coefficients via OLS amounts to the same procedure as refitting the OLS estimates of the AR coefficients (e.g. $\hat{\boldsymbol{\beta}}$) subject to equality constraints of having the coefficients to be the same from the (τ_k+1)-th to the τ_{k+1}-th time-lag for all k .

2.5 Theoretical properties

2.5.1 Gaussian innovations

The following two quantities will together measure the difficulty of our change-point problem detection problem (with the convention $\tau_0=0$ and $\tau_{q+1}=p$):

$$\delta_T := \min_{j=1,\ldots,q+1} |\tau_j - \tau_{j-1}|, \tag{2.7}$$

$$\underline{\alpha}_T := \min_{j=1,\ldots,q} |\beta_{\tau_j+1} - \beta_{\tau_j}| = \min_{j=1,\ldots,q} |\alpha_j| \tau_j^{-1}. \tag{2.8}$$

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

(A1) $\{X_t\}$ is stationary and follows the AMAR(q) model given in (1.1) with the innovations ε_t being i.i.d. $\mathcal{N}(0,\sigma^2)$ for some $\sigma>0$.

(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 $b(z)$ lie outside the unit circle $\mathbb{T}$. Furthermore, there exists constants $\underline{c}_2,\bar{c}_2>0$ such that $\underline{c}_2 \leq \min_{z\in\mathbb{T}}|b(z)| \leq \max_{z\in\mathbb{T}}|b(z)| \leq \bar{c}_2$ uniformly in T.

(A4) $\delta_T^{1/2}\underline{\alpha}_T \succ T^{\theta-\frac{1}{2}}(\log(T))^{3/2} =: \underline{\lambda}_T$, where θ is as in (A2), and where δ_T and $\underline{\alpha}_T$ are given by (2.7) and (2.8), respectively. Here $f(T)\succ g(T)$ means that $\liminf_{T\to\infty} f(T)/g(T)=\infty$.

Some comments regarding these assumptions are in order. First, 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 is shown later, Theorem 2.2 could possibly be extended to cover more general distributional scenarios for the noise ε_t.

Second, 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, $t=1,...,T$, 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, sufficiently far from being unit-root. Besides, the upper bound on $\max_{z\in\mathbb{T}}|b(z)|$ implies that $\|\boldsymbol{\beta}\|$ is uniformly bounded from below, in view of the Parseval's identity (see Lemma S4.1 in the supplementary materials).

Third, Assumption (A4) controls both the minimum spacing between the timescales and the size of the jumps in (2.3). 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.

Theorem 2.2 *Let assumptions (A1) – (A4) hold, and let $\hat{q}$ and $\hat{\tau}_1,...,\hat{\tau}_{\hat{q}}$ 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 $C_1\underline{\lambda}_T<\zeta_T<C_2\delta_T^{1/2}\underline{\alpha}_T$,*

and $M > 36T\delta_T^{-2}\log(T\delta_T^{-1})$ *(only if needed by Algorithm 2), then for all sufficiently large* T *we have*

$$\mathbb{P}\left(\hat{q}=q, \max_{j=1,\ldots,q}|\hat{\tau}_j-\tau_j|\leq\epsilon_T\right)\geq 1-C_4T^{-1}, \tag{2.9}$$

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

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, for example, the case when both the number of scales q and the scale coefficients $\alpha_1,\ldots,\alpha_q$ in (1.1) 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 (2.3), so the signal strength decreases to 0 when $T\to\infty$. Here $\delta_T^{1/2}\underline{\alpha}_T \sim T^{-\theta/2}$, thus (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, (2.9) simplifies to $\mathbb{P}(\hat{q}=q, \hat{\tau}_j=\tau_j \forall j=1,\ldots,q) \geq 1-C_4T^{-1}$, when T is sufficiently large. Furthermore, in the much simpler setting where all the locations of the timescales are fixed, Theorem 2.2 concludes that with high probability $\hat{q}=q$ and $\hat{\tau}_j=\tau_j$ for all $j=1,\ldots,q$. As a consequence, one could establish further that all the estimated autoregressive and scale coefficients (i.e. β_i's and α_i's) converging at the rate of $T^{-1/2}$. However, in gen-

eral, we would expect the convergence rate of $\hat{\alpha}_i$'s to α_i's to be slower than $O(T^{-1/2})$ when either q or τ_q (or both) increases with T. Due to its theoretical nature, we leave the complete characterisation of the asymptotic behaviours of $\hat{\alpha}_i$'s to future research.

2.5.2 Heavy-tailed innovations

In the settings where the innovations follow more heavy-tailed distributions, consistency of our procedure could still be established. For simplicity, we shall assume that the number of scales q is fixed, so that the presented results would have much simpler dependence on the tail behaviour of the innovation distributions. The assumptions we impose under this setup are given below.

(B1) $\{X_t\}$ is stationary and follows the AMAR(q) model given in (1.1) with the innovations ε_t being i.i.d. following a symmetric distribution Z with regularly varying tail probabilities of index α, such that $\mathbb{P}(|Z|>z)=z^{-\alpha}L(z)$ for any $z>0$ with any positive $\alpha\neq 2$, and $L(\cdot)$ is a slowly varying function at ∞, i.e. $\lim_{z\to\infty}\frac{L(az)}{L(z)}=1$ for any $a>0$.

(B2) $\alpha_1,...,\alpha_q$, $\tau_1,...,\tau_q$ and p are fixed, with $p>\tau_q$.

(B3) The roots of the corresponding characteristic polynomial $b(z)$ lie outside the unit circle $\mathbb{T}$.

Regarding Assumption (B1), we note that it covers many scenarios of heavy-tailed distributions, including generalized Pareto and Cauchy. Here a smaller

α implies heavier tails. For instance, when $\alpha > 4$, the innovation distribution has finite fourth-moment, while a distribution with $\alpha \in (2,4)$ has finite variance. Here the case of $\alpha = 2$ (i.e. the boundary of an infinite variance) is not included to simplify our analysis further. In the setting of autoregressive models, much heavier tails (e.g. with $\alpha < 2$) actually tend to make the parameter estimation (fundamentally via autocorrelation) more accurate, which intuitively is due to the fact that observations would be more spread-out. See Yohai and Maronna (1977), Hannan and Kanter (1977), Davis and Resnick (1985, 1986). Consequently, our Algorithm 2 would still work as intended, as established in the following result.

Theorem 2.3 *Let assumptions (B1) – (B3) 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 (with F_T^M taken as the set of all $p(p-1)/2$ intervals within $[1,p]$). For any sufficiently small $\epsilon > 0$, there exist constants C_1, C_2 such that if $C_1 T^{-\max(1/2,1/\alpha)+\epsilon} < \zeta_T < C_2 \underline{\alpha}_T$, then as $T \to \infty$,*

$$\mathbb{P}\left(\hat{q} = q, \max_{j=1,\ldots,q} |\hat{\tau}_j - \tau_j| = 0 \right) \to 1.$$

3. Practicalities and simulated examples

3.1 Parameter choice

Threshold ζ_T. This threshold is one of the input parameters required in Algorithm 1 and Algorithm 2. The minimum rate of magnitude permitted by Theorem 2.2, that is $\zeta_T = CT^{-1/2}(\log(T))^{3/2}$ can be used (say, with $C = 0.5$), though a more

careful choice would be required for different setups. In practice, we advocate choosing the threshold using the Schwarz Information Criterion (SIC) as outlined below.

For any $\zeta_T > 0$ (and a fixed p), 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. Specifically, with the estimated timescales $\hat{\tau}_1(\zeta_T),...,\hat{\tau}_{\hat{q}(\zeta_T)}(\zeta_T)$ and corresponding scales coefficients $\hat{\alpha}_1(\zeta_T),...,\hat{\alpha}_{\hat{q}(\zeta_T)}(\zeta_T)$,

$$\hat{X}_t(\zeta_T) = \hat{\alpha}_1(\zeta_T)\frac{X_{t-1}+...+X_{t-\hat{\tau}_1(\zeta_T)}}{\hat{\tau}_1(\zeta_T)}+...+\hat{\alpha}_{\hat{q}(\zeta_T)}(\zeta_T)\frac{X_{t-1}+...+X_{t-\hat{\tau}_{\hat{q}(\zeta_T)}(\zeta_T)}}{\hat{\tau}_{\hat{q}(\zeta_T)}(\zeta_T)},$$

where we set the values of the unobserved $X_0, X_{-1},...$ to be the sample mean of the series.

We then select the threshold that minimises the SIC defined as follows:

$$\mathrm{SIC}(\zeta_T) = T\log\left(\sum_{t=1}^{T}(X_t - \hat{X}_t(\zeta_T))^2\right) + 2\hat{q}(\zeta_T)\log(T), \tag{3.10}$$

where (3.10) is minimised over ζ_T such that $\hat{q}(\zeta_T) \leq q_{\max} = 10$. Unless stated otherwise, we take this as our default approach in the remainder of this article and in Section S2 of the supplementary materials.

Number M of random intervals. As outlined in Algorithm 2, we normally use all the intervals unless p is extremely large. This would be computationally feasible for most applications. However, when p is large (say > 500), we would follow the recommendation in Baranowski et al. (2019) by setting $M = 10000$.

The autoregressive order p. We refrain from giving a universal recipe for the choice of p. In the real data example reported later, we choose the p that

corresponds to a large "natural" time span. If such choice is not obvious, then in principle, the SIC criterion (3.10) can be minimised with respect to both ζ_T and p. Here to reduce the computational burden, in practice, instead of going through all possible values of p (and finding the corresponding threshold that minimises that particular SIC), one possibility would be to search for p only on a grid with its elements increasing exponentially from 1 up to the order of $T^{1/2}$, e.g. $\{1,2,4,8,...\}$.

3.2 Computational complexity of the AMAR algorithm

The calculation of the OLS estimates in Steps 1 and 4 of Algorithm 2 takes $O(Tp^2)$ operations. The values of $\mathcal{C}_{s,e}^b(\cdot)$ can be computed for all b in $O(e-s)$ operations, hence the complexity of Step 3 is $O(Mp)$. This term is typically dominated by $O(Tp^2)$, and therefore the usual computational complexity of the AMAR algorithm is $O(Tp^2)$. We make use of an efficient implementation of OLS estimation available from the R package **RcppEigen** (Bates and Eddelbuettel, 2013).

3.3 Simulation study

We illustrate the finite sample behaviour and performance Algorithm 2 in a comprehensive simulation study. The data are simulated from (1.1) for the following four scenarios. In all these scenarios, the noise ε_t follows i.i.d. $\mathcal{N}(0,1)$.

(M1) Two timescales at $\tau_1 = 1$ and $\tau_2 = 3$, with the corresponding coefficients $\alpha_1 = 0.3$, $\alpha_2 = 0.6$ (i.e. $\boldsymbol{\beta} = (0.5,0.2,0.2)^T$).

(M2) Two timescales at $\tau_1 = 2$ and $\tau_2 = 5$, with the corresponding coefficients $\alpha_1 = 1.9$, $\alpha_2 = -1$ (i.e. $\boldsymbol{\beta} = (0.75, 0.75, -0.2, -0.2, -0.2)^T$).

(M3) Three timescales at $\tau_1 = 1$, $\tau_2 = 5$ and $\tau_3 = 14$, with the corresponding coefficients $\alpha_1 = 0.4$, $\alpha_2 = -1$, $\alpha_3 = 1.4$ (i.e. $\boldsymbol{\beta} = (0.4, -0.1, -0.1, -0.1, -0.1, 0.1, ..., 0.1)^T$).

(M4) Seasonal model with four timescales at $\tau_1 = 1$, $\tau_2 = 6$, $\tau_3 = 7$ and $\tau_4 = 8$, with the corresponding coefficients $\alpha_1 = 0.5$, $\alpha_2 = -4.8$, $\alpha_3 = 8.4$, $\alpha_4 = -3.2$ (i.e. $\boldsymbol{\beta} = (0.5, 0, ..., 0, 0.8, -0.4)^T$, so $\varepsilon_t = (1 - 0.8B^7)(1 - 0.5B)X_t$).

(M5) A single timescale at $\tau_1 = 10$ with $\alpha_1 = 0.9$ (i.e. $\boldsymbol{\beta} = (0.09, ..., 0.09)^T$), as illustrated in Figure 1.

(M6) Two timescales at $\tau_1 = 1$ and $\tau_2 = \lfloor T^{0.4} \rfloor$, (which increases with T), with the corresponding coefficients $\alpha_1 = \alpha_2 = 0.49$ (i.e. $\boldsymbol{\beta} = (0.49 + 0.49/\lfloor T^{0.4} \rfloor, 0.49/\lfloor T^{0.4} \rfloor ..., 0.49/\lfloor T^{0.4} \rfloor)^T$), as illustrated in Figure 5 in the supplementary materials.

These scenarios were designed to cover combinations of timescales of different lengths. Here $\boldsymbol{\beta}$ is selected as such that the series are stationary but also strongly autocorrelated with $\sum_{j=1}^{p} \beta_j \approx 0.9$ (or more in (M6)). We believe that this is the regime where AMAR models are most useful, and is in the lines with what one would get from fitting some of the real data in practice, as shown in Section 4.

We consider a few different aspects of the estimators obtained with Algorithm 2 with different numbers of observations $T = 400, 800, 1500, 3000$. We assess the

accuracy in terms of the number of the fitted timescales $\hat{q}$, the Hausdorff distance D_H between the fitted timescale locations $\{\hat{\tau}_1,...,\hat{\tau}_{\hat{q}}\}$ and the true ones $\{\tau_1,...,\tau_q\}$, as well as the Euclidean distance between the fitted parameter vector $\hat{\boldsymbol{\beta}}$ and the true one $\boldsymbol{\beta}$. We also compare the mean squared prediction errors (MSPE) of the fitted models with the oracles. For the sake of fair comparison, for each of the simulated series (with length T), after model fitting, we further draw $T^*=100$ observations at the end of the series and use these observations solely for the purpose of out-of-sample mean squared prediction error estimation, given as $\sum_{i=1}^{T^*}(\hat{X}_{T+i}-X_{T+i})^2/T^*$, where for every $i=1,...,T^*$, the predicted value of X_{T+i} is given as

$$\hat{X}_{T+i}=\hat{\alpha}_1\frac{X_{T+i-1}+...+X_{T+i-\hat{\tau}_1}}{\hat{\tau}_1}+...+\hat{\alpha}_q\frac{X_{T+i-1}+...+X_{T+i-\hat{\tau}_{\hat{q}}}}{\hat{\tau}_{\hat{q}}}.$$

We then report the ratio between the out-of-sample mean squared prediction error and $\sum_{i=1}^{T^*}\varepsilon_{T+i}^2/T^*$, which is the mean squared prediction error from the oracle model.

Here for our proposed AMAR approach, we select both the threshold and p via the Schwarz Information Criterion as mentioned previously with the maximum number of timescales $q_{\max}=10$. With regard to the competitors, we also report results obtained using the fused LASSO (N.B. details can be recalled from our literature review in Section 1.2), where $\boldsymbol{\beta}$ is estimated by minimising

$$\sum_{j=p+1}^{T}(X_j-\beta_1'X_{j-1}-\cdots-\beta_p'X_{j-p})^2+\lambda\Big(\sum_{j=1}^{p-1}|\beta_{j+1}'-\beta_j'|+|\beta_p'|\Big)$$

with respect to $\boldsymbol{\beta}'=(\beta_1',...,\beta_p')^T\in\mathbb{R}^p$, where λ is picked by cross-validation. Finally, we also consider the autoregressive model selected via AIC (i.e. among

AR(1), ..., AR(p)). Note that for the AIC, we do not enforce the parameters to be constant in between consecutive timescale locations; as such, only the corresponding $\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|$ and the mean squared prediction errors are computed. All the numerical experiments are repeated 1000 times and the results are summarised in Table 1 and Table 2.

Model (M1)										
	$E\lvert\hat{q}-q\rvert$		$E(D_H)$		$E\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|$			$\frac{\text{MSPE(fitted)}}{\text{MSPE(oracle)}}-1$		
Method	AMAR	Fused	AMAR	Fused	AMAR	Fused	AIC	AMAR	Fused	AIC
$T=400$	0.172	6.07	0.593	16.1	0.0159	0.0206	0.0156	0.0133	0.0226	0.0138
	(0.014)	(0.088)	(0.047)	(0.05)	(0.0008)	(0.00049)	(0.0006)	(0.00093)	(0.0012)	(0.00093)
$T=800$	0.051	8.65	0.181	24.1	0.0035	0.0114	0.00749	0.0046	0.0154	0.00802
	(0.0072)	(0.12)	(0.03)	(0.056)	(0.00026)	(0.00027)	(0.00028)	(0.00048)	(0.00089)	(0.00061)
$T=1500$	0.018	12.5	0.085	34.1	0.00116	0.00613	0.00445	0.00138	0.00764	0.00393
	(0.0042)	(0.16)	(0.03)	(0.051)	(0.000088)	(0.00014)	(0.0002)	(0.00024)	(0.00062)	(0.00041)
$T=3000$	0.012	20.2	0.072	50.2	0.000546	0.0029	0.00207	0.000662	0.00429	0.002
	(0.0034)	(0.21)	(0.035)	(0.052)	(0.000027)	(0.000063)	(0.000088)	(0.00017)	(0.00046)	(0.00028)
Model (M2)										
	$E\lvert\hat{q}-q\rvert$		$E(D_H)$		$E\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|$			$\frac{\text{MSPE(fitted)}}{\text{MSPE(oracle)}}-1$		
Method	AMAR	Fused	AMAR	Fused	AMAR	Fused	AIC	AMAR	Fused	AIC
$T=400$	0.303	7.32	1.33	14.1	0.02	0.0717	0.124	0.0281	0.0857	0.0763
	(0.018)	(0.064)	(0.072)	(0.062)	(0.0013)	(0.0019)	(0.0032)	(0.01)	(0.0043)	(0.0032)
$T=800$	0.194	9.39	0.764	21.8	0.00635	0.0595	0.061	0.00852	0.0615	0.0331
	(0.014)	(0.071)	(0.06)	(0.077)	(0.00071)	(0.001)	(0.0018)	(0.0013)	(0.0024)	(0.0016)
$T=1500$	0.108	10.9	0.921	31.6	0.00171	0.0535	0.0327	0.00666	0.0532	0.0165
	(0.01)	(0.069)	(0.11)	(0.092)	(0.00038)	(0.0011)	(0.0011)	(0.0038)	(0.0022)	(0.001)
$T=3000$	0.07	12.8	0.646	47.2	0.0000979	0.0504	0.0131	0.000793	0.0444	0.00571
	(0.0081)	(0.071)	(0.099)	(0.12)	(0.000021)	(0.0011)	(0.00048)	(0.0002)	(0.002)	(0.00056)
Model (M3)										
	$E\lvert\hat{q}-q\rvert$		$E(D_H)$		$E\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|$			$\frac{\text{MSPE(fitted)}}{\text{MSPE(oracle)}}-1$		
Method	AMAR	Fused	AMAR	Fused	AMAR	Fused	AIC	AMAR	Fused	AIC
$T=400$	0.711	5.76	1.37	5.12	0.0211	0.0204	0.0499	0.0296	0.0321	0.0567
	(0.035)	(0.077)	(0.046)	(0.038)	(0.00076)	(0.00041)	(0.00073)	(0.0016)	(0.0017)	(0.0022)
$T=800$	0.344	7.83	0.643	12.6	0.00699	0.012	0.0244	0.00922	0.0158	0.0215
	(0.026)	(0.11)	(0.034)	(0.068)	(0.00031)	(0.00024)	(0.00041)	(0.00075)	(0.00091)	(0.00099)
$T=1500$	0.083	10.1	0.31	22.4	0.00203	0.00704	0.013	0.0034	0.00984	0.0112
	(0.011)	(0.13)	(0.043)	(0.08)	(0.00011)	(0.00013)	(0.00022)	(0.0004)	(0.00066)	(0.00072)
$T=3000$	0.054	13.5	0.219	38.2	0.000673	0.00397	0.00648	0.0015	0.00628	0.00683
	(0.0082)	(0.16)	(0.045)	(0.084)	(0.000041)	(0.000074)	(0.00011)	(0.00023)	(0.00051)	(0.00048)

Table 1: Performance of different methods under (M1) – (M3), with estimated errors given in the brackets. Here $\hat{q}$ is the number of the fitted timescales, D_H is the Hausdorff distance between the fitted timescale locations $\{\hat{\tau}_1,...,\hat{\tau}_{\hat{q}}\}$ and the true ones $\{\tau_1,...,\tau_q\}$, $\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|$ is the Euclidean distance between the fitted parameter vector and the true one, and MPSE is the mean squared prediction errors of different models.

We see that AMAR approach performs consistently better than the fused

Model (M4)										
	$E\lvert\hat{q}-q\rvert$		$E(D_H)$		$E\lVert\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\rVert$			$\frac{\text{MSPE(fitted)}}{\text{MSPE(oracle)}}-1$		
Method	AMAR	Fused	AMAR	Fused	AMAR	Fused	AIC	AMAR	Fused	AIC
$T=400$	0.098	11.1	0.199	11.8	0.00892	0.0483	0.0246	0.0145	0.0492	0.0358
	(0.012)	(0.073)	(0.027)	(0.017)	(0.00065)	(0.00086)	(0.00083)	(0.0011)	(0.0019)	(0.0017)
$T=800$	0.044	16.3	0.092	19.7	0.00397	0.0274	0.0107	0.00657	0.0274	0.0142
	(0.0085)	(0.1)	(0.019)	(0.019)	(0.0003)	(0.00045)	(0.00038)	(0.0006)	(0.0012)	(0.00088)
$T=1500$	0.035	24.2	0.291	29.8	0.00179	0.0182	0.00685	0.00333	0.0182	0.00812
	(0.006)	(0.14)	(0.059)	(0.017)	(0.00011)	(0.00027)	(0.00026)	(0.0004)	(0.00091)	(0.00066)
$T=3000$	0.023	36.3	0.129	45.8	0.000756	0.0106	0.00301	0.0017	0.0116	0.0043
	(0.0051)	(0.19)	(0.033)	(0.016)	(0.000023)	(0.00015)	(0.00012)	(0.00024)	(0.0007)	(0.0004)
Model (M5)										
	$E\lvert\hat{q}-q\rvert$		$E(D_H)$		$E\lVert\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\rVert$			$\frac{\text{MSPE(fitted)}}{\text{MSPE(oracle)}}-1$		
Method	AMAR	Fused	AMAR	Fused	AMAR	Fused	AIC	AMAR	Fused	AIC
$T=400$	0.217	3.11	1.64	6.95	0.0109	0.0106	0.0341	0.0164	0.0151	0.0398
	(0.017)	(0.085)	(0.073)	(0.1)	(0.00045)	(0.00038)	(0.00053)	(0.0028)	(0.00099)	(0.0016)
$T=800$	0.133	4.06	0.858	12.9	0.00414	0.00562	0.0166	0.00517	0.00833	0.0176
	(0.013)	(0.098)	(0.056)	(0.17)	(0.00022)	(0.00019)	(0.0003)	(0.00055)	(0.00069)	(0.001)
$T=1500$	0.099	5.06	0.704	22.1	0.00167	0.00331	0.00877	0.00237	0.00454	0.00908
	(0.012)	(0.11)	(0.076)	(0.25)	(0.00012)	(0.000097)	(0.00017)	(0.00033)	(0.00046)	(0.00065)
$T=3000$	0.052	7.07	0.331	38.9	0.000339	0.00171	0.00427	0.000788	0.00278	0.00452
	(0.0086)	(0.16)	(0.054)	(0.27)	(0.000043)	(0.000055)	(0.000087)	(0.00017)	(0.00035)	(0.00044)
Model (M6)										
	$E\lvert\hat{q}-q\rvert$		$E(D_H)$		$E\lVert\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\rVert$			$\frac{\text{MSPE(fitted)}}{\text{MSPE(oracle)}}-1$		
Method	AMAR	Fused	AMAR	Fused	AMAR	Fused	AIC	AMAR	Fused	AIC
$T=400$	0.407	6.87	2.3	9.26	0.0133	0.0336	0.0372	0.023	0.0435	0.06
	(0.024)	(0.097)	(0.054)	(0.05)	(0.00046)	(0.0018)	(0.00062)	(0.0016)	(0.0027)	(0.0027)
$T=800$	0.886	9.83	3.29	13.1	0.00902	0.0234	0.0252	0.015	0.0311	0.0296
	(0.035)	(0.13)	(0.071)	(0.054)	(0.00028)	(0.0017)	(0.00036)	(0.00098)	(0.0025)	(0.0013)
$T=1500$	0.455	13.8	3.08	19	0.00336	0.0174	0.0174	0.00668	0.0241	0.0193
	(0.028)	(0.17)	(0.1)	(0.057)	(0.00013)	(0.0016)	(0.00023)	(0.00055)	(0.0023)	(0.00098)
$T=3000$	0.642	20.9	3.52	28.8	0.00177	0.0134	0.011	0.00395	0.0187	0.0111
	(0.037)	(0.21)	(0.11)	(0.063)	(0.000064)	(0.0015)	(0.00012)	(0.00038)	(0.0021)	(0.00068)

Table 2: Performance of different methods under (M4) – (M6), with estimated errors given in the brackets. Here $\hat{q}$ is the number of the fitted timescales, D_H is the Hausdorff distance between the fitted timescale locations $\{\hat{\tau}_1,...,\hat{\tau}_{\hat{q}}\}$ and the true ones $\{\tau_1,...,\tau_q\}$, $\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|$ is the Euclidean distance between the fitted parameter vector and the true one, and MPSE is the mean squared prediction errors of different models.

LASSO for all aspects in all model settings and with all the sample sizes we consider. In fact, estimates from the fused LASSO do not seem to be consistent in terms of estimating the number and locations of the scales, indicating that the fused LASSO approach (with L_1 penalisation) is not appropriate for identifying jumps within the parameter vector. Rather interestingly, the AMAR approach also

seems to perform better than the approach based on the AIC in terms of the mean squared prediction errors, illustrating the usefulness and importance of taking into account additional structures in the parameters when they are available.

In Section S2 of the supplementary materials, we also report results from our sensitivity analysis where we look into the performance of our proposed approach with (i) different choices of $q_{\max}$ and (ii) a fixed p. In addition, We run experiments with series simulated from non-stationary AR models with unit roots. To summarise the findings here, AMAR is generally not sensitive to the choice of $q_{\max}$ (as long as the truth is no greater). Besides, a fixed p might lead to some very moderate improvement over our current approach of selection via SIC when T is small, but could be problematic when the chosen p is smaller than or close to τ_q. Finally, even in the setting of non-stationary observations, AMAR still performs much better than its competitors in most settings, though the reported results from all methods are associated with larger standard errors.

4. Real data examples

4.1 Stock returns

In this example, we demonstrate the strength of AMAR models for predicting the DAX stock index daily return over the traditional AR. We look at ten years of data from 1 January 2011 to 31 December 2020, with the first seven years of data

(70%) used for training, and the last three years of data (30%) used for testing. Here we shall work directly on the series of log-return, which we denote as $\{X_t\}$. The visual appearance of the series is illustrated in Figure 2.

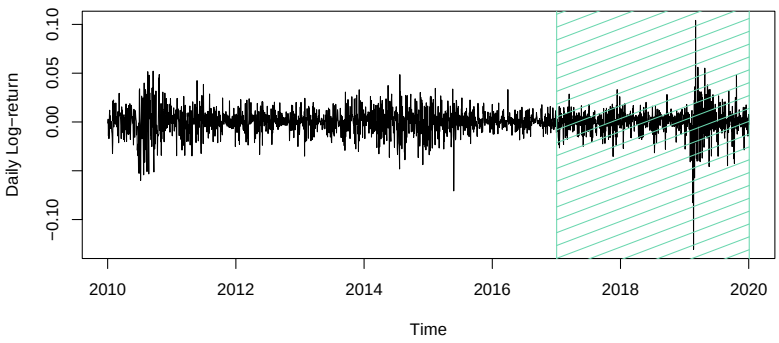

Figure 2: DAX daily log-return from January 2011 to December 2020. The series is divided into two parts for training and testing, with the part for testing highlighted in shade.

First, we fit an AMAR model on $\{X_t\}$ using AMAR with the thresholds and p selected automatically via the approach outlined in Section 3.1. A three-scale AMAR model is selected, with $\tau_1 = 1, \tau_2 = 5$ and $\tau_3 = 27$.

However, for the purpose of interpretability, we note that a two-scale AMAR model might be preferred, with the short scale fixed at $\tau_1 = 1$ for this particular application. As such, we also fit different AMAR models on only the training data with $\tau_1 = 1$ and $\tau_2 = \{2,...,251\}$, and select the corresponding τ_2 that minimises the residual sum of squares. This results in a two-scale AMAR model with $\tau_2 = 5$ (and $\tau_1 = 1$).

Since our focus is on prediction, we also fit the traditional AR model with the order selected by the AIC. This results in an AR(6) model.

We then examine the performance of these fitted models on the testing data and measure their performance by both the rooted mean squared prediction error (RMSPE) and the hit rate. Here the hit rate is defined as the proportion of time the model predicts the sign of the daily log-return correctly, which is an important performance indicator for financial time series modelling. The results are reported in Table 3. In terms of both criteria, AMAR with two scales performs the best. Here AMAR with three scales appears slightly worse, while AR with its order selected by the AIC performs the worst. In addition, we remark that AMAR with $\tau_1 = 1$ and $\tau_2 = 5$ can be easily interpreted as having the daily log-return depending on the returns of both the previous trading day and the previous week, a fact that would potentially be appreciated by the practitioners. In summary, we believe that AMAR would be a promising alternative to the traditional AR models in modelling real data of this type.

Methods	AMAR (auto-selection for scales)	AMAR (with two scales)	AR (order via AIC)
RMSPE	0.014564	**0.014521**	0.014580
Hit Rate	0.5013	**0.5186**	0.4775

Table 3: Performance of different methods in terms of their rooted mean squared prediction error (RMSPE) and hit rate. Results from the better method are highlighted in bold.

Finally, we note that in this example, more complicated dependence structure, such as heteroscedasticity, has not been taken into account. In principle, the AMAR approach could be extended to the multiscale modelling of both the AR component and the ARCH-type errors.

4.2 UK and US unemployment data

In this example, we first consider the time series of seasonally adjusted UK monthly unemployment rate from 1960 to 2020. The data can be found in OECD (2022). The series is shown in the top plot of Figure 3. As before, here our aim is not to find models that best fit the data, but to compare AMAR with the AR alternatives normally used in practice, and to demonstrate the potential superiority and practicality of AMAR over other AR approaches. For this analysis, we report our findings on both the original series and the differenced series using AMAR and AR with its order selected via AIC. We also report prediction errors of the different models, where we use the last 5, 10, 20 or 30 years of data for testing and the remaining for training (without specifying the scales or orders a priori). We set the maximum AR order to be 48 (i.e. four years) for both methods. For the original series, AMAR fits a model with scales at 1,2 and 3, while AIC selects an AR(13). For the differenced series, AMAR fits a model with scales at 1 and 10, while AIC selects an AR(8). A closer look reveals that the sum of the fitted AR coefficients on the original series is close to one (>0.99) for both approaches, reflecting the possibility that the series might

not be stationary, while the sum on the differenced series is much smaller (at around 0.7). Looking at the quality of prediction in terms of rooted mean squared errors (at the original series level), we see that AMAR performs better than AR with order selected by AIC in both the original and differenced series in the testing periods of all lengths, though admittedly the difference between these two methods becomes much smaller when considering the differenced series. In fact, for the purpose of prediction, results from Table 4, suggest that it is more appropriate to model the differenced rather than the original series. However, we would like to point out that no matter which series (i.e. original or differenced) one prefers to work with in this particular example, AMAR always offers better predictive performance than the AR with the order selected using AIC, and possibly also comes with improved interpretability.

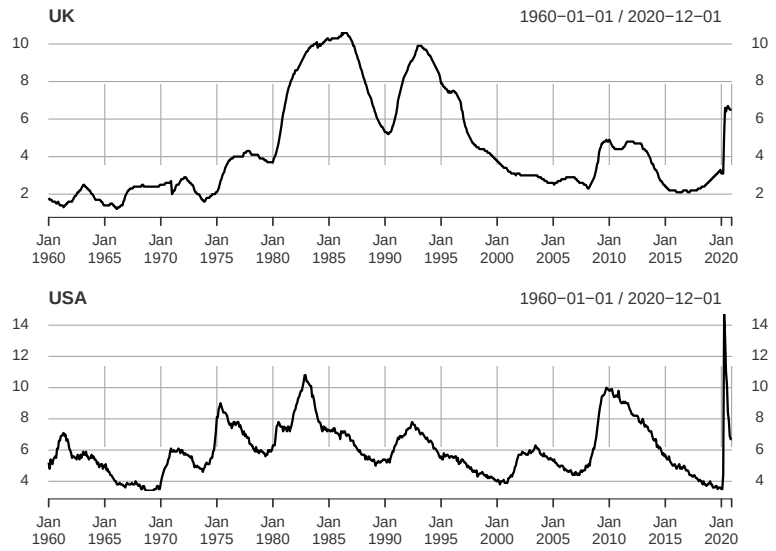

Figure 3: Seasonally adjusted monthly unemployment rates of UK and USA from January 1960 to December 2020.

Series	Methods	5 years	10 years	20 years	30 years
original	AMAR	**1.0267**	**0.7590**	**0.5849**	**0.5683**
	AR-AIC	1.2436	0.9452	0.7409	0.7636
differenced	AMAR	**0.3359**	**0.2469**	**0.1861**	**0.15295**
	AR-AIC	0.3753	0.2554	0.1874	0.1557

Table 4: Performance of different methods for different testing periods in terms of rooted mean squared prediction errors at the original series level. Here AMAR is the adaptive multiscale autoregression, while AR-AIC is the autoregressive model with order chosen by AIC. Results from the better method are highlighted in bold.

Next, to demonstrate the potential use of AMAR on multivariate time series, we additionally include the seasonally adjusted US monthly unemployment rate during the same period. The data can be found in U.S. Bureau of Labor Statistics (2022), with the series also shown in Figure 3. In view of our previous analysis, we only consider the differenced time series for the purpose of prediction. Let $X_{t,1}$ and $X_{t,2}$ represent the respective differenced UK and US unemployment rates at time t. Then the corresponding Adaptive Multiscale Vector AutoRegressive model (AMVAR) for bivariate observations can be written as

$$\begin{pmatrix} X_{t,1} \\ X_{t,2} \end{pmatrix} = \boldsymbol{\alpha}_1 \begin{pmatrix} \frac{X_{t-1,1}+...+X_{t-\tau_1,1}}{\tau_1} \\ \frac{X_{t-1,2}+...+X_{t-\tau_1,2}}{\tau_1} \end{pmatrix} + ... + \boldsymbol{\alpha}_q \begin{pmatrix} \frac{X_{t-1,1}+...+X_{t-\tau_q,1}}{\tau_q} \\ \frac{X_{t-1,2}+...+X_{t-\tau_q,2}}{\tau_q} \end{pmatrix} + \begin{pmatrix} \varepsilon_{t,1} \\ \varepsilon_{t,2} \end{pmatrix} \quad (4.11)$$

where $\tau_1,...,\tau_q$ are the scales, $\boldsymbol{\alpha}_1,...,\boldsymbol{\alpha}_q$ are 2×2 matrices, and where $\boldsymbol{\varepsilon}_t = (\varepsilon_{t,1},\varepsilon_{t,2})^T$ are noise vectors. The optimal selection of the number of scales and their locations for AMVAR is beyond the scope this paper. One simple approach would be to

perform scale selection for each univariate series, and combine all of them to be then used as the scales for AMVAR. Though there might be limitations in this approach, we believe it serves as a good starting point for further exploration, as demonstrated below. Recall that AMAR selects scales of 1 and 10 for the differenced UK unemployment series. In addition, AMAR selects a single scale of 11 for the differenced US unemployment series. Consequently, we use scales at 1, 10 and 11 for AMVAR. To facilitate comparison with our previous analysis, we focus on the differenced UK unemployment series, which we now explicit model as

$$X_{t,1} = \alpha_{1,(1,1)} X_{t-1,1} + \alpha_{1,(1,2)} X_{t-1,2} + \alpha_{2,(1,1)} \frac{X_{t-1,1} + ... + X_{t-10,1}}{10} + \alpha_{2,(1,2)} \frac{X_{t-1,2} + ... + X_{t-10,2}}{10} + \alpha_{3,(1,1)} \frac{X_{t-1,1} + ... + X_{t-11,1}}{11} + \alpha_{3,(1,2)} \frac{X_{t-1,2} + ... + X_{t-11,2}}{11} + \varepsilon_{t,1}.$$

On the other hand, fitting the data by the Vector Autoregressive (VAR) models and selecting order via AIC leads to a VAR(4). The rooted squared errors for prediction of different models are reported in Table 5, where we use the last five, ten, twenty or thirty years of data for testing and the remaining for training. Our results suggest that for the purpose of prediction, AMVAR performs better than VAR with order selected using AIC for various testing periods in this example under the multivariate setting. Importantly, this appears to be the case even without fine tuning the scale selection procedure. In addition, comparing results in Table 4 and Table 5, we note that AMVAR performs better than the univariate AMAR for

Series	Methods	5 years	10 years	20 years	30 years
differenced	AMVAR	**0.3023**	**0.2176**	**0.1682**	**0.1438**
	VAR-AIC	0.4383	0.3136	0.2271	0.1912
	AMAR	0.3359	0.2469	0.1861	0.15295

Table 5: Performance of different methods for different testing periods in terms of rooted mean squared prediction errors. Here AMVAR is the adaptive multiscale vector autoregression, while VAR-AIC is the vector autoregressive model with order chosen by AIC, both under the bivariate setting. We also include (univariate) AMAR for comparison. Results from the better method are highlighted in bold.

UK unemployment, indicating that including an extra regressor from a different time series (i.e. US unemployment rate) does indeed improve the predictive power.

5. Extensions and further discussions

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). As an example, we could consider features that are linearly-weighted averages, instead of the simple averages in (1.1) for AMAR(q). To give more details, for some $1 \leq \tau_1 < ... < \tau_q$, the model is given as

$$X_t = \alpha_1 \frac{\tau_1 X_{t-1} + (\tau_1 - 1) X_{t-2} + ... + X_{t-\tau_1}}{\tau_1(1+\tau_1)/2} + \cdots + \alpha_q \frac{\tau_q X_{t-1} + (\tau_q - 1) X_{t-2} + ... + X_{t-\tau_q}}{\tau_q(1+\tau_q)/2} + \varepsilon_t$$

Here the influence of the past observations on any given feature decays as the time gap between them and the present widens. The linear decaying form is just one

possible way of modelling that results in a more parsimonious parameter structure of the AR. The other appealing reason for linear decaying is that τ_1,...,τ_q can also be estimated using the previous framework, say, by simply changing the contrast function in the NOT algorithm from piecewise-constant contrast to piecewise-linear and continuous contrast.

As briefly illustrated in our second real data example, another interesting venue for applying the AMAR framework is time series data with multivariate, or even high-dimensional observations. In particular, in the high-dimensional setting, instead of applying the same averages (or features) across different components of $(X_{t,1},...,X_{t,p})^T$, some group structures (or even factors) within these components can be introduced to further enhance its interpretability.

Finally, it would also be of interest to investigate how estimation uncertainty could be quantified in AMAR (e.g. via block bootstrap), and whether the AMAR philosophy could be extended to multiscale features of some latent or hidden observations. One such example would be to consider multiscale autoregressive and moving average (ARMA) models, in which there are different scales on both the observed time series and the unobserved innovations, i.e. we observe $\{X_t\}$ following

$$\begin{aligned} X_t = &\alpha_1 \frac{X_{t-1}+...+X_{t-\tau_1}}{\tau_1} + ... + \alpha_{q_{\mathrm{AR}}} \frac{X_{t-1}+...+X_{t-\tau_{q_{\mathrm{AR}}}}}{\tau_{q_{\mathrm{AR}}}} \\ &+ \varepsilon_t + \beta_1 \frac{\varepsilon_{t-1}+...+\varepsilon_{t-\rho_1}}{\rho_1} + ... + \beta_{q_{\mathrm{MA}}} \frac{\varepsilon_{t-1}+...+\varepsilon_{t-\rho_{q_{\mathrm{MA}}}}}{\rho_{q_{\mathrm{MA}}}} \end{aligned}$$

where q_{AR} and q_{MA} are, respectively, the numbers of timescales for AR and MA, with the AR timescales $1 \leq \tau_1 < ... < \tau_{q_{\mathrm{AR}}}$ and the MA timescales $1 \leq \rho_1 < ... < \rho_{q_{\mathrm{MA}}}$, and with α_1,...,$\alpha_{q_{\mathrm{AR}}}$, β_1,...,$\beta_{q_{\mathrm{MA}}}$ being the scale coefficients.

Supplementary materials

Contains discussions and illustrations on some special cases of AMAR, further simulations, an additional real data example, and the proofs of the theoretical results.

Acknowledgements

We are extremely grateful to the associate editor and three anonymous reviewers for their valuable comments and suggestions, which have helped improve the manuscript.

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

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.

R. Baranowski, Y. Chen, and P. Fryzlewicz. **amar**: Adaptive Multiscale Autoregressive time series models, 2022. URL `https://github.com/ychen101/amar`. R package version 1.2.

M. Basseville, A. Benveniste, and A. Willsky. Multiscale autoregressive processes, part I: Schur-Levinson parametrizations. *IEEE Transactions on Signal Processing*, 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.

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.

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.

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

R. Davis and S. Resnick. Limit theory for moving averages of random variables with regularly varying tail probabilities. *The Annals of Probability*, 13:179 – 195, 1985.

R. Davis and S. Resnick. Limit theory for the sample covariance and correlation functions of moving averages. *The Annals of Statistics*, 14:533 – 558, 1986.

M. Ferreira, M. West, H. Lee, and D. Higdon. Multi-scale and hidden resolution time series models.

Bayesian Analysis, 1:947–967, 2006.

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

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. J. Hannan and M. Kanter. Autoregressive processes with infinite variance. *Journal of Applied Probability*, 14:411–415, 1977.

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

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

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.

H. Maeng and P. Fryzlewicz. Regularized forecasting via smooth-rough partitioning of the regression coefficients. *Electronic Journal of Statistics*, 13:2093–2120, 2019.

OECD. Registered unemployment rate for the United Kingdom. Retrieved from FRED, Federal Reserve Bank of St. Louis, `https://fred.stlouisfed.org/series/LMUNRRTTGBM156S`, 2022. Accessed: 2022-03-30.

G. Petnehazi. Recurrent neural networks for time series forecasting. *arXiv:1901.00069*, 2019.

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

R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society Series B (Statistical Methodology)*, 67:91–108, 2005.

U.S. Bureau of Labor Statistics. Unemployment rate. Retrieved from FRED, Federal Reserve Bank of St. Louis, `https://fred.stlouisfed.org/series/UNRATE`, 2022. Accessed: 2022-03-30.

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

V. J. Yohai and R. A. Maronna. Asymptotic behavior of least-squares estimates for autoregressive processes with infinite variances. *The Annals of Statistics*, 5:554–560, 1977.

M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society Series B (Statistical Methodology)*, 68:49–67, 2006.

Rafal Baranowski, London School of Economics and Political Science

E-mail: r.baranowski@lse.ac.uk

Yining Chen, London School of Economics and Political Science

E-mail: y.chen101@lse.ac.uk

Piotr Fryzlewicz, London School of Economics and Political Science

E-mail: p.fryzlewicz@lse.ac.uk