We modeled long memory with just one lag!

Luc Bauwens
CORE, Université catholique de Louvain

Guillaume Chevillon
ESSEC Business School

Sébastien Laurent
Aix-Marseille University (Aix-Marseille School of Economics), CNRS
EHESS, Aix-Marseille Graduate School of Management – IAE

March 7, 2022

Abstract

A large dimensional network or system can generate long memory in its components, as shown by Chevillon, Hecq and Laurent (2018, CHL) and Schennach (2018). These authors derive conditions under which the variables generated by an infinite dimensional vector autoregressive model of order 1, a VAR(1), exhibit long memory. We go one step further and show how these theoretical results can be put to practice for modeling and forecasting series with long range dependence that belong to a large network or system. We estimate the VAR(1) model equation by equation, by shrinking the parameters to generic conditions matching those of CHL and Schennach, by ridge and Bayesian estimations. We consider two large-dimensional applications where long memory has long been an established observation. Our proposal significantly outperforms ARFIMA and HAR models when forecasting a non-parametric estimate of the log of the integrated variance of 250 assets, as well as seasonally adjusted historic monthly streamflow series recorded in 97 locations of the Columbia river basin.

Keyword: Bayesian estimation, Ridge regression, VAR, ARFIMA, HAR, Forecasting.
JEL: C10, C32, C58.

1 Introduction

Ever since Granger (1966) and Nelson and Plosser (1982), the question of the degree of persistence in macroeconomic and financial variables has exhibited regular puzzles. Long memory (i.e., a dependence between observations decaying hyperbolically with their distance in time, see Beran, 1992) is often encountered in economic and financial time series (at least since Smith, 1938, and Cox and Townsend, 1947; see, e.g., Baillie, 1996, for an introduction) and long memory models are found to provide a good empirical representation of persistence that is stronger than ARMA models but weaker than unit-root processes. The econometric literature has found that its origin can take several forms, such as aggregation (Granger, 1980, Abadir and Talmain, 2002), linear modeling of a nonlinear process (e.g., Robinson and Zaffaroni, 1998, Miller and Park, 2010, Chen, Hansen and Carrasco, 2010), structural changes (e.g., Diebold and Inoue, 2001, Gourieroux and Jasiak, 2001, Perron and Qu, 2010) as well as resulting from agents' self-referential learning behaviors and forward expectations (Chevillon and Mavroeidis, 2017, 2018). More recently, Chevillon, Hecq and Laurent (2018, CHL henceforth) have shown that long memory can result from the marginalization of a large dimensional system. More specifically, they provide a parametric framework under which the variables of an n-dimensional vector autoregressive model of order 1, i.e., a VAR(1), can be individually modelled as a fractional white noise (see Granger and Joyeux, 1980) as n tends to infinity. Long memory may therefore be a feature of univariate or low dimensional models that vanishes when considering larger systems. In the context of network dynamics, Schennach (2018, Schennach henceforth) has found a related result of hyperbolic response of outputs to distant input shocks. These sources of long memory differ from other sources mentioned in the literature, in particular the aggregation mechanism of Granger (1980).

In this paper, we address the question of whether and how the asymptotic theoretical results of CHL and Schennach can be put to use in empirical works. Given the large dimensional nature of their models, inference in empirical works is likely to be imprecise. Hence, rather than attempting to test the specification of a large scale model using a finite data set, we provide an assessment of the proximity of the models to the data generating process by means of forecasting exercises. We provide in particular a set of techniques using classical and Bayesian inference which allow an empirical modeler to benefit from the asymptotic theoretical results of CHL and Schennach.

It is well known that a VAR(1) model can be estimated equation by equation, each equation being an AR(1) model augmented by the first lag of all the other variables in the system. We refer to these univariate equations as AR(1)-X models. Our objective is to test whether such AR(1)-X models can constitute a viable alternative to pure long memory models like the ARFIMA model or models known to approximate well the long memory like the HAR model of Corsi (2009). By careful estimation of the AR(1)-X models, we evaluate whether the new source of long memory identified by CHL and Schennach is empirically relevant, is useful for forecasting variables depicting long memory and therefore is a good candidate for approximating the data generating process of series displaying long memory. Given their asymptotic nature (in the cross-sectional dimension n, not in the sample size T), the results of CHL and Schennach involve systems so large that inference may be infeasible or highly imprecise in finite samples.

We propose two methods to estimate the AR(1)-X model, which shrink the parameters towards a set of constraints provided by the theory developed by CHL and Schennach. The first shrinking strategy relies on an L2 penalization of the AR(1)-X model (i.e., ridge regression) and is denoted RAR(1)-X (for Ridge AR(1)-X). The second one relies on an informative prior density in a Bayesian approach, denoted BAR(1)-X (for Bayesian AR(1)-X).

If the penalty weight in the RAR(1)-X model is set to zero or the prior variances in the BAR(1)-X model are set to plus infinity, the theory inspired restrictions have no influence, and therefore the RAR(1)-X and the BAR(1)-X are equivalent to the unrestricted AR(1)-X estimated by OLS which is expected to deliver poor results when the number of parameters to estimate is large. Alternatively, if the penalty term in the RAR(1)-X is set to plus infinity or the prior variances of the BAR(1)-X are set to zero, the restrictions are fully imposed. Values of the penalty weight or of the prior variances between these two extreme cases are worth considering in an empirical application, and optimal values of these parameters can be chosen by cross-validation.

We compare the proposed shrinking estimation strategies of the AR(1)-X model to OLS estimation and to three univariate models for short and long range dependence: the AR(1) model, the autoregressive fractionally integrated moving average (ARFIMA) model, and the heterogeneous autoregressive (HAR) model of Corsi (2009). For this purpose, we perform two empirical applications to the prediction of (i) the logarithm of a robust to jumps estimate of the daily integrated variance (i.e., the MedRV of Andersen, Dobrev, and Schaumburg, 2012) computed on 5-minute returns for 250 US stocks, and (ii), the logarithm of monthly seasonally adjusted series of river streamflows recorded and computed at 97 locations in the Columbia river basin over 90 years. Given that we compare models based on different information sets, and that these models are of reduced form type and aimed at forecasting, it makes sense to use comparison criteria based on forecasts. Thus we compare forecasts produced by different models using the mean absolute deviation (MAD) and mean square forecast error (MSFE) loss functions, and rely on the Model Confidence Set test of Hansen, Lunde and Nason (2011) to discriminate between the models.

The rest of this paper is organized as follows. Section 2 provides the theoretical framework needed to understand how a VAR can generate long memory when the dimension of the system becomes large. The theory implies restrictions on the VAR parameters that can be useful for improving estimation and forecasting. Section 3 explains how the theoretical restrictions are used in the estimation of the VAR parameters, either by defining an informative prior density for conducting Bayesian estimation, or by ridge estimation. Section 4 contains the empirical results. Conclusions are offered in the last section. Proofs and technical details are contained in an appendix.

2 Long memory in a VAR(1) model

This section presents the conditions derived by CHL and Schennach needed to prove that long-memory observed in a univariate time-series can be the result of the marginalization of an infinitely large VAR(1) system. Let the observable vector $\boldsymbol{y}_{n,t}$ of dimension n satisfy, for $t \geq 1$,

$$(\boldsymbol{I}_n - \boldsymbol{A}_n L)(\boldsymbol{y}_{n,t} - \boldsymbol{\mu}) = \boldsymbol{\epsilon}_{n,t}, \tag{1}$$

where $\boldsymbol{\epsilon}_{n,t}$ is a short memory process with zero expectation and variance-covariance matrix $\boldsymbol{\Sigma}_n$. While the specific assumptions differ in CHL and Schennach, we let, for clarity of exposition,

$$\boldsymbol{A}_n = \boldsymbol{T}_n + \eta_n \boldsymbol{D}_n,$$

where η_n is a vanishing scalar sequence, and $(\boldsymbol{T}_n)$ and $(\boldsymbol{D}_n)$ denote sequences of Toeplitz matrices that are, respectively, symmetric and antisymmetric. The entries of $\boldsymbol{T}_n$ are labelled as

$$\boldsymbol{T}_n = \begin{bmatrix} t_0^{(n)} & t_1^{(n)} & \cdots & t_{n-1}^{(n)} \\ t_1^{(n)} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & t_1^{(n)} \\ t_{(n-1)}^{(n)} & \cdots & t_1^{(n)} & t_0^{(n)} \end{bmatrix}. \tag{2}$$

Following the exposition by Schennach, the process $\boldsymbol{y}_{n,t}$ can be seen as generated by a network that lies in a space of dimension one. She also considers higher dimensions, but for the purpose of the analysis using financial and hydrological data, we restrict ourselves to a one dimensional linear network so each node lies in $\mathbb{Z}$. In the spirit of Diebold and Yılmaz (2009, 2014), who estimate connectedness within a network using a VAR, this amounts to a system that consists of an infinite but countable number of variables indexed by $j \in \mathbb{Z}$. We denote the limiting, infinite dimensional, vectors by $(\boldsymbol{y}_t, \boldsymbol{\epsilon}_t) = \lim_{n\to\infty} (\boldsymbol{y}_{n,t}, \boldsymbol{\epsilon}_{n,t})$, and the ith elements of $\boldsymbol{y}_t, \boldsymbol{\epsilon}_t$ by $y_t^{(i)}, \epsilon_t^{(i)}$, for $i \in \mathbb{Z}$ or $\mathbb{N}$. We next describe the two models that have been shown to generate long memory within an infinite dimensional VAR(1) model such as (1).

CHL Model: These authors assume that the effect of $\boldsymbol{D}_n$ vanishes at rate $\eta_n = o\left(n^{-2}\right)$. They make a parametric assumption for the entries of $\boldsymbol{T}_n$, namely that (i) there exists a sequence $\delta_n \in \left(0, \frac{1}{2}\right)$ satisfying $n^2\left(\delta_n - \frac{1}{2}\right) = o(1)$, such that (ii) $t_k^{(n)} = \mathrm{Re}\left[\frac{1}{n}\sum_{j=0}^{n-1} g\left(\delta_n, e^{i\frac{2\pi j}{n}}\right) e^{-\frac{2i\pi jk}{n}}\right]$ (for $k = 0, 1, \ldots n-1$), where for $(\delta, \omega) \in (0,1) \times \mathbb{R}$, $g\left(\delta, e^{i\omega}\right) \equiv 1_{\{0 \le u < \pi\delta\}} + 1_{\left\{\pi\left(\frac{3}{2}-\delta\right) < u \le \frac{3\pi}{2}\right\}}$, with $\omega = u \bmod 2\pi$. They show that, as $n \to \infty$, with $(n-1)/4 \in \mathbb{N}$, $t_0^{(n)} \to \frac{1}{2}$ and $t_k^{(n)} \to 0$ for $k \neq 0$. Under the assumption $\boldsymbol{\epsilon}_{n,t} \sim \mathsf{IID}(\boldsymbol{0}, \boldsymbol{\Sigma}_n)$ and $\boldsymbol{\Sigma}_n$ diagonal, they prove (in their Theorem 1) that, as $n \to \infty$, all components of $\boldsymbol{y}_{n,t}$ tend to independent fractional white noises with identical degrees of integration (all equal to one-half):

$$\boldsymbol{y}_{n,t} \Rightarrow \boldsymbol{\mu} + \Delta^{-1/2}\boldsymbol{\epsilon}_t,$$

where $\Delta = 1 - L$ and $\Rightarrow$ denotes weak convergence of the associated probability measures. Since the entries of $\boldsymbol{A}_n - \frac{1}{2}\boldsymbol{I}_n$ tend to zero as $n \to \infty$, the cross-sectional dependence between the elements of $\boldsymbol{y}_{n,t}$ vanishes as $n \to \infty$. Yet, as in this setting $\sum_{k=0}^{n-1} t_k^{(n)} = 1$ remains nonzero, the dependence across individual series is sufficient to generate long memory in each of the components of the multivariate processes.

Schennach Model: She considers the limiting structure where $\boldsymbol{T} = \lim_{n\to\infty} \boldsymbol{A}_n = \lim_{n\to\infty} \boldsymbol{T}_n$, i.e., the case of an infinite dimensional network. She assumes that $\boldsymbol{\epsilon}_t$ constitutes a short memory MA(∞) process. The entries (t_k) of $\boldsymbol{T}$ are assumed to satisfy $t_0 > 0$, $\sum_{j=0}^{\infty} t_j = 1$, and $\mathsf{card}\{j \in \mathbb{Z},\ t_j > 0\} < \infty$. She then proves (in her Theorem 4) that, for all i, j, there exists a $c_{ij} > 0$ such that, as $k \to \infty$

$$\frac{\partial y_{t+k}^{(i)}}{\partial \epsilon_t^{(j)}} = c_{ij}k^{-1/2} + O\left(k^{-3/2}\right),$$

i.e., the impulse response function of $y_{t+k}^{(i)}$ to a shock $\epsilon_t^{(j)}$ is hyperbolic and its decay corresponds to that of a process that is integrated of order $\frac{1}{2}$.

Both Schennach and CHL find long memory of fractional degree one-half in infinite dimensional networks. They use different approaches and assumptions, but rely on the Toeplitz nature of dependence across the infinite dimensional – yet countable – number of variables in the system or nodes in the network. Both of them consider so-called bistochastic matrices whose rows and columns sum to unity. Schennach focuses on the interactions within the limiting system/network while CHL consider the evolution in dynamics as the finite system/network grows larger. Both find that long memory arises only in the infinitely dimensional environment.

Schennach is less restrictive in her assumptions on $\boldsymbol{\epsilon}_t$. She does not specify the values of the entries of $\boldsymbol{A}$ but assumes that only a finite number of t_k are nonzero, so that a rotation of $\boldsymbol{A}$ is *banded* (i.e., all subdiagonals are zero beyond a point). Hence, in the (one-dimensional) networks she considers, each variable/node is only directly connected to a finite number of variables/nodes.

By contrast CHL rely on *i.i.d.* shocks and make parametric assumptions on $\boldsymbol{T}_n$. In their setting, variables/nodes are directly connected to *all* other variable/nodes, but with a connection that becomes weaker as the dimension of the system/network increases.

Then Schennach's result is that all response functions of all variables to all shocks exhibit hyperbolic decay, whereas CHL's applies only to the responses of variables to their idiosyncratic shocks in the VAR system.

3 How to use the theory for estimation?

The theory summarized in Section 2 provides parametric restrictions on the matrix $\boldsymbol{A}_n$ of a VAR(1) system of n variables, which imply that the variables have long memory properties when n tends to infinity. The stylized "long memory" restrictions (i.e., implying that the variables of the system have long memory properties) on the matrix $\boldsymbol{A}_n$ are that the diagonal elements are close to 0.5, the other elements are close to 0, and the sum of the elements of each row is equal to 1.

Estimating $\boldsymbol{A}_n$ is needed to obtain forecasts of $\boldsymbol{y}_n$. We present here some methods to shrink the estimates of $\boldsymbol{A}_n$ towards the theoretical results of CHL and Schennach reminded in Section 2. An obvious way to strictly impose the restrictions consists in paramaterizing explicitly the elements of $\boldsymbol{A}_n$ through the structure proposed by CHL or Schennach. This means in practice that all the elements of $\boldsymbol{A}_n$ depend on a scalar δ_n that can be estimated by minimum distance or by maximum likelihood (ML). This is certainly too restrictive and we want a certain degree of flexibility around these restrictions.

Between the least restrictive strategy of ignoring the restrictions and estimating the VAR by OLS, and the totally restrictive strategy mentioned above, there exists intermediate methods. One of them is penalized regression (e.g., ridge or lasso), where the least squares criterion is augmented with restrictions whose strength is modulated through penalty parameters. The resulting estimator is shrunk in the direction of the restrictions. Since the theoretical restrictions we consider here do not imply the exclusion of specific variables, the relevant penalization in our context is ridge, so we do not consider lasso.

Bayesian estimation provides another intermediate method where the restrictions are embedded in a prior density, in such a way that they hold a priori on average (through the prior expectation of the parameters), but with some degree of uncertainty (through prior positive variances on the

parameters or functions thereof). Depending on the degree of tightness of the prior, the prior information pulls data information more or less strongly in the direction of the restrictions.

Ridge regression and Bayesian estimation are exposed respectively in Subsections 3.2 and 3.3. The next subsection sets the details of the econometric model and its estimation, which in both methods is an "equation by equation" approach to the estimation of the VAR system, denoted, respectively, RAR(1)-X and BAR(1)-X. Appendix C explains the relation between the ridge and Bayesian estimators.

3.1 Econometric framework

We consider the estimation of a VAR(1) process, written for time t (dropping the subscript n on $\boldsymbol{A}_n$ and on the processes) as

$$\boldsymbol{y}_t = \boldsymbol{\tau} + \boldsymbol{A}\boldsymbol{y}_{t-1} + \boldsymbol{\epsilon_t}, \tag{3}$$

for the vector $\boldsymbol{y}_t$ consisting of n variables. The estimation of the parameters $\boldsymbol{\tau}$ and $\boldsymbol{A}$ is performed "equation by equation", instead of globally for the system. Assuming $\boldsymbol{\epsilon_t}$ is multivariate Gaussian with zero expectation and constant covariance matrix $\boldsymbol{\Sigma}$, the separate estimation by OLS of each equation is equivalent to the maximum likelihood estimation of the system, even if $\boldsymbol{\Sigma}$ is not diagonal. For Bayesian estimation, the equation by equation method is not equivalent to the joint estimation of all equations, but the latter method is much heavier in computing time for the dimensions we are interested in (e.g., 250 in the first empirical application).

A typical equation of the VAR(1) system is an AR(1)-X regression equation, that is written for date t as

$$y_t = \gamma_0 + \boldsymbol{\gamma}'\boldsymbol{x}_t + \epsilon_t, \tag{4}$$

where y_t is a variable of the system, γ_0 is the intercept parameter, $\boldsymbol{x}_t$ is the column vector containing the first lag of the n variables of the system (including the first lag of y_t), $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \dots \gamma_n)'$ is the vector of n slope coefficients of $\boldsymbol{x}_t$, and ϵ_t is an error term assumed to be Gaussian with zero expectation and constant variance σ^2. By convention, for any variable of the VAR, $\boldsymbol{x}_t$ is ordered in such a way that its first element is the lagged dependent variable (y_{t-1}), and $\boldsymbol{\gamma}$ is ordered accordingly: its first element (γ_1) is the autoregressive coefficient of the dependent variable, and the remaining elements are the coefficients of the other lagged variables. For example, if y_t is the first element of $\boldsymbol{y}_t$, $\boldsymbol{\gamma}'$ is the first row of the matrix $\boldsymbol{A}$ of the VAR(1) system, and γ_0 is the first element of $\boldsymbol{\tau}$.

For T observations, the AR(1)-X equation is written in the standard regression notation

$$\boldsymbol{Y} = \boldsymbol{Z}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{5}$$

where $\boldsymbol{Y} = (y_1, y_2, \dots, y_T)'$, $\boldsymbol{\epsilon} = (\epsilon_1, \dots, \epsilon_T)' \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I_T})$, $\boldsymbol{Z}$ is a $T \times k$ matrix, with $k = 1 + n$ and t-th row equal to $(1, \boldsymbol{x}_t')$, and $\boldsymbol{\beta} = (\gamma_0, \boldsymbol{\gamma}')'$.

The estimation of $\boldsymbol{\beta}$ by OLS is likely to be imprecise when n is large and to affect negatively the quality of forecasts. To take advantage of the parametric restrictions implied by the theories of CHL and Schennach, we opt for shrinking the elements of the vector $\boldsymbol{\beta} = (\gamma_0, \boldsymbol{\gamma}')'$ in (5) so that

C1: the autoregressive coefficient (γ_1) is close to 0.5,

C2: the other elements of $\boldsymbol{\gamma}$ are close to 0,

C3: the sum of the elements of $\boldsymbol{\gamma}$ is equal to 1.

In what follows, we explain how we implement this by ridge or Bayesian estimation.

3.2 Ridge estimation

To achieve **C1** and **C2**, we define a vector $\boldsymbol{\beta}_0$ as the shrinkage target of $\boldsymbol{\beta}$ such that

$$\boldsymbol{\beta}_0 = (0, d_0, a_0, \ldots, a_0)', \tag{6}$$

where $a_0 = (1 - d_0)/(n - 1)$ is repeated $n - 1$ times. The scalar d_0 in $(0, 1)$ is the target for the autoregressive coefficient and it determines the target a_0 of the other coefficients, which are shrunk to a value that is close to zero when n is large. We use two penalty parameters to control the shrinkage strength: λ_d^2 for the autoregressive parameter, and λ_a^2 for the other coefficients. The penalty term is defined as

$$\lambda_d^2(\gamma_1 - d_0)^2 + \lambda_a^2 \sum_{i=2}^{n} (\gamma_i - a_0)^2 = (\boldsymbol{\beta} - \boldsymbol{\beta}_0)' \boldsymbol{\Lambda}_k (\boldsymbol{\beta} - \boldsymbol{\beta}_0), \text{ where } \boldsymbol{\Lambda}_k = \text{diag}(0, \lambda_{\text{d}}^2, \lambda_{\text{a}}^2, \ldots, \lambda_{\text{a}}^2). \tag{7}$$

In this way, the last n elements of $\boldsymbol{\beta}$ are shrunk to the corresponding elements of $\boldsymbol{\beta}_0$, but the first element of $\boldsymbol{\beta}$ is not shrunk, the value (zero) of the first element of $\boldsymbol{\beta}_0$ being practically irrelevant.

The choice of $\boldsymbol{\beta}_0$ implies that the sum of the last n coefficients is equal to 1 in the target, but the penalty is distributed over the n coefficients. To better achieve **C3**, we add the penalty term $\lambda_s^2(\boldsymbol{\iota}'\boldsymbol{\beta} - 1)^2$, where λ_s^2 is a penalty parameter and $\boldsymbol{\iota} = (0, 1, 1, \ldots, 1)'$ is a vector of k elements. More generally, by writing the penalty as $\lambda_s^2(\boldsymbol{\iota}'\boldsymbol{\beta} - \boldsymbol{\iota}'\boldsymbol{\beta}_0)^2$, we cover the possibility that the target value be different from 1.

The extended ridge (ER) estimator is obtained by minimizing the objective function

$$(\boldsymbol{Y} - \boldsymbol{Z}\boldsymbol{\beta})'(\boldsymbol{Y} - \boldsymbol{Z}\boldsymbol{\beta}) + (\boldsymbol{\beta} - \boldsymbol{\beta}_0)' \boldsymbol{\Lambda}_k (\boldsymbol{\beta} - \boldsymbol{\beta}_0) + \lambda_s^2(\boldsymbol{\iota}'\boldsymbol{\beta} - \boldsymbol{\iota}'\boldsymbol{\beta}_0)^2, \tag{8}$$

and can be shown to be (see Appendix A)

$$\boldsymbol{\beta}_{ER} = \left(\boldsymbol{Z}'\boldsymbol{Z} + \boldsymbol{\Lambda}_k + \lambda_s^2 \boldsymbol{\iota}\boldsymbol{\iota}'\right)^{-1} (\boldsymbol{Z}'\boldsymbol{Y} + \boldsymbol{\Lambda}_k \boldsymbol{\beta}_0 + \lambda_s^2 \boldsymbol{\iota}\boldsymbol{\iota}'\boldsymbol{\beta}_0). \tag{9}$$

As usual, the ridge estimator simplifies to the OLS estimator if all the penalty parameters are set to zero.

The values of d_0, λ_d^2, λ_a^2, and λ_s^2 can be chosen by cross-validation on a training sample. A grid of values is set a priori for each of them. For each point of the grid, the estimator is computed using 80 percent of the training sample, forecasts are computed for the last 20 percent, and a forecast loss function is computed. The chosen triplet is the value minimizing the loss function over the grid. After this, the estimation is performed on a subsequent sample, and forecasts are computed and evaluated over a post-estimation sample. Details are provided in Section 4.

3.3 Bayesian estimation

Bayesian estimation is based on a prior density for $\boldsymbol{\beta}$ and σ^2, and the likelihood function, the latter resulting from the assumption of normality of the error terms. Since the theory does not provide information on σ^2, its prior "density" $p(\sigma^2)$ is chosen to be the usual "non-informative" prior:

$$p(\sigma^2) \propto 1/\sigma^2. \tag{10}$$

The prior density of $\boldsymbol{\beta}$ is designed to include the theory restrictions **C1**-**C3**. We opt for a Gaussian density for three reasons: it is convenient for computing the posterior density (see Appendix B);

the implementation of the restrictions is easy through four scalar parameters, as explained below; and the restrictions do not require an asymmetric density. The prior density is proportional to

$$\exp[-\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)'\boldsymbol{Q}_0(\boldsymbol{\beta}-\boldsymbol{\beta}_0)]\exp[-\frac{1}{2}h_0(\boldsymbol{\beta}'\boldsymbol{\iota}-\boldsymbol{\beta}_0'\boldsymbol{\iota})^2]. \tag{11}$$

The vector $\boldsymbol{\beta}_0$ is defined as (3.2); it depends on the scalar hyperparameter d_0 (see below). To explain the prior, let us first fix the scalar hyperparameter h_0 to zero, and discuss the first Gaussian kernel of (11), which corresponds to restrictions **C1** and **C2**. Then $\boldsymbol{\beta}_0$ is the prior expectation, and $\boldsymbol{Q}_0$ is the prior precision matrix. We specify this matrix to be diagonal:

$$\boldsymbol{Q}_0 = \mathrm{diag}(0, 1/\mathrm{s}_\mathrm{d}^2, 1/\mathrm{s}_\mathrm{a}^2, \ldots, 1/\mathrm{s}_\mathrm{a}^2), \tag{12}$$

so that s_d is the prior standard deviation of the autoregressive coefficient and s_a is the the prior standard deviation of the other coefficients. The strength with which the restrictions **C1** and **C2** are imposed depends on the values of s_d and s_a, respectively. Values close to zero correspond to a strong prior belief in favor of the restrictions. For the intercept term, the prior precision is set to zero, so that the data information dominates the prior information on this term.

Although the prior expectation $\boldsymbol{\beta}_0$ embeds the restriction **C3** that the sum of the last n elements of $\boldsymbol{\beta}$ is equal to 1, the prior variance of this sum is equal to $s_d^2 + (n-1)s_a^2$. Hence to obtain a small variance, s_a must be fixed to a very small value, which may be in contradiction with the restriction **C2**. Indeed, the latter requires not to shrink excessively to zero the coefficients of the last n elements of $\boldsymbol{\beta}$. The second Gaussian kernel of (11) is designed to avoid the conflict between the two restrictions, by adding a prior parameter that controls the strength imposed on the unit sum, independently of the strength imposed on the individual coefficients. Notice that in the second exponential function of (11), we have written $\boldsymbol{\beta}_0'\boldsymbol{\iota}$ after the minus sign, instead of 1, to cover the case where one wants this target to be different from 1, that is, the case where one defines $\boldsymbol{\beta}_0$ differently from (3.2).

If $\boldsymbol{Q}_0$ in the first kernel is a null matrix, the second kernel specifies that the prior mean of the sum of the last n elements of $\boldsymbol{\beta}$ is equal to $\boldsymbol{\beta}_0'\boldsymbol{\iota}$ (= 1 if $\boldsymbol{\beta}_0$ is given by (3.2)), and that its prior precision is equal to h_0. Hence a large value of h_0 corresponds to a strongly informative prior on the target for the sum of the coefficients.

It is well-known that the product of two Gaussian kernels is a kernel of a Gaussian density. Hence, (11) is the kernel of the Gaussian density (see Appendix A)

$$\boldsymbol{\beta} \sim N_k(\boldsymbol{\beta}_0, \boldsymbol{V}_0), \tag{13}$$

where

$$\boldsymbol{V}_0 = (\boldsymbol{Q}_0 + h_0\boldsymbol{\iota}\boldsymbol{\iota}')^{-1}. \tag{14}$$

Notice that the expectation of $\boldsymbol{\beta}$ is $\boldsymbol{\beta}_0$, the same as in the first kernel. If $h_0 > 0$, the prior covariance matrix is not diagonal: actually, the covariances are negative, which is what is needed to reduce the prior standard deviation of $\boldsymbol{\beta}'\boldsymbol{\iota}$ with respect to its value when the prior covariance matrix is diagonal. For example, if $d_0 = 0.50$, $s_d = s_a = 0.02$, $h_0 = 5000$, $n = 250$, $\boldsymbol{\beta}_0 = (0, 0.5, 0.002008(249\text{ times}))$, $\boldsymbol{Q}_0^{-1} = \mathrm{diag}(100, 0.02^2(250\text{ times}))$, the diagonal of $\boldsymbol{V}_0$ is $(100, 0.01996^2(250\text{ times}))$, the off-diagonal elements are equal to 0 in the first line (and column), and the other covariances are equal to $-1.59681/10^6$ (the corresponding correlation coefficient being equal to -0.004008). The prior standard deviation of $\boldsymbol{\beta}'\boldsymbol{\iota}$ is equal to 0.014128, much smaller than its value of 0.317 when the prior is $N_k(\boldsymbol{\beta}_0, \boldsymbol{Q}_0^{-1})$, where $\boldsymbol{Q}_0^{-1}$ is defined as $\mathrm{diag}(0, s_d^2, s_a^2, \ldots, s_a^2)$.

In summary, the prior density (13), when $\boldsymbol{\beta}_0$ is defined by (3.2) and $\boldsymbol{Q}_0$ by (12), is fully determined by the four scalar hyperparameters d_0, s_d, s_a, and h_0, whatever the dimension n of the VAR. These hyperparameters can be fixed to some values, as in the example above, or they can be chosen for each equation of the VAR by a cross-validation procedure similar to the procedure defined in the last paragraph of the previous subsection.

The computation of the posterior mean of $\boldsymbol{\beta}$ for the prior (10)-(13) is performed by the simple Gibbs sampling algorithm defined in Appendix B. The prior is not conjugate since $\boldsymbol{V}_0$ is not proportional to σ^2. It becomes conjugate if (13) is replaced by

$$\boldsymbol{\beta}|\sigma^2 \sim N_k(\boldsymbol{\beta}_0, \sigma^2 \boldsymbol{V}_0). \tag{15}$$

The posterior mean corresponding to this conjugate prior is

$$\left(\boldsymbol{Z}'\boldsymbol{Z} + \boldsymbol{Q}_0 + h_0\boldsymbol{\iota}\boldsymbol{\iota}'\right)^{-1}\left(\boldsymbol{Z}'\boldsymbol{Y} + \boldsymbol{Q}_0\boldsymbol{\beta}_0 + h_0\boldsymbol{\iota}\boldsymbol{\iota}'\boldsymbol{\beta}_0\right), \tag{16}$$

where (14) has been used. If we set $\boldsymbol{Q}_0 = \boldsymbol{\Lambda}_k$ (by setting $\lambda_d^2 = 1/s_d^2$ and $\lambda_a^2 = 1/s_a^2$) and $h_0 = \lambda_s^2$, this posterior mean is exactly the ER estimator (9). With the non-conjugate prior, one can only derive the conditional (to σ^2) posterior mean of $\boldsymbol{\beta}$, which can be expressed (see Appendix B) as

$$\boldsymbol{\beta}_*(\sigma^2) = \left(\frac{\boldsymbol{Z}'\boldsymbol{Z}}{\sigma^2} + \boldsymbol{Q}_0 + h_0\boldsymbol{\iota}\boldsymbol{\iota}'\right)^{-1}\left(\frac{\boldsymbol{Z}'\boldsymbol{Y}}{\sigma^2} + \boldsymbol{Q}_0\boldsymbol{\beta}_0 + h_0\boldsymbol{\iota}\boldsymbol{\iota}'\boldsymbol{\beta}_0\right). \tag{17}$$

This differs from (16) only by the presence of σ^2. The Gibbs sampler defined in Appendix B is a way to marginalize $\boldsymbol{\beta}_*(\sigma^2)$ with respect to σ^2. The resulting unconditional posterior mean of $\boldsymbol{\beta}$ is then different from the corresponding posterior mean/ER estimator when the prior is conjugate.

3.4 Forecasting

After obtaining a point estimate of $\boldsymbol{\beta}$ for an equation of the VAR system, such as the OLS estimator, the extended ridge estimator, or the posterior mean, a one-step ahead point forecast of y_{t+1} is simply obtained using a point estimate of (4), and the regressor $\boldsymbol{x}_{t+1}$ observable at time t. This is equivalent to using the point estimates of all equations to form the estimated $\boldsymbol{\tau}$ and $\boldsymbol{A}$ of the VAR system (3), and then computing one-step ahead point forecasts as $\hat{\boldsymbol{y}}_{t+1} = \hat{\boldsymbol{\tau}} + \hat{\boldsymbol{A}}\boldsymbol{y}_t$.

To compute h-step ahead forecasts, with $h > 1$, one can use iterated multistep forecasting or direct multistep forecasting. An iterated h-step ahead forecast is based on the estimated VAR and computed recursively as $\hat{\boldsymbol{y}}_{t+h} = \hat{\boldsymbol{\tau}} + \hat{\boldsymbol{A}}\hat{\boldsymbol{y}}_{t+h-1}$. This approach amounts to compute $\hat{\boldsymbol{A}}^h$, i.e., to forecast all variables even if one is interested in only a subset of them (even just a single one). Hence the forecast of a variable of interest may be contaminated by erroneous and imprecise forecasts of the other variables (see, e.g., Schorfheide, 2005; Chevillon and Hendry, 2005).

If the objective is to forecast a subset of the series, or if one wishes to avoid the drawback inherent in the iterated multistep method highlighted above, the direct multistep forecasting method is preferable. The method consists in directly projecting $\boldsymbol{y}_t$ on its lag $\boldsymbol{y}_{t-h}$, as in

$$\boldsymbol{y}_t = \boldsymbol{\tau}_h + \boldsymbol{A}_h\boldsymbol{y}_{t-h} + \boldsymbol{u}_t. \tag{18}$$

Ignoring that $\boldsymbol{A}_h = \boldsymbol{A}^h$, a typical equation of (18) can be cast in the form of 4) and (5), adapting the definitions of $\boldsymbol{Y}$, $\boldsymbol{x}_t$ and $\boldsymbol{Z}$, and ignoring the dependence in $\boldsymbol{u}_t$ induced by recursive substitution. For $h > 1$, we denote the equation corresponding to (5) by

$$\boldsymbol{Y}_{(h)} = \boldsymbol{Z}_{(h)}\boldsymbol{\beta}_h + \boldsymbol{u}_{(h)}. \tag{19}$$

Hence, the system (18) can be estimated equation by equation, by OLS, ridge and Bayesian estimation, as for $h = 1$. By proceeding in this spirit, no direct use is made in estimation of the relation $\boldsymbol{A}_h = \boldsymbol{A}^h$, because this would imply that the regression coefficients of the different equations of (18) are nonlinear functions of the same parameters (those of $\boldsymbol{A}$), so that equation by equation estimation would be pointless. In brief, the parameter $\boldsymbol{\beta}_h$ is not treated as a function of the underlying parameters of $\boldsymbol{A}$.

Nevertheless, for ridge and Bayesian estimations, we let the target towards which $\boldsymbol{\beta}_h$ is shrunk depend on h and denote it by $\boldsymbol{\beta}_{h,0}$. This $\boldsymbol{\beta}_{h,0}$ relates to the first row of $\boldsymbol{A}_0^h$, like $\boldsymbol{\beta}_0$ is directly the first row of $\boldsymbol{A}_0 = d_0\boldsymbol{I}_n + a_0\left(\boldsymbol{J}_n - \boldsymbol{I}_n\right)$, where $\boldsymbol{J}_n$ is a matrix of ones, and $a_0 = (1-d_0)/(n-1)$. Practically, we set the last n elements of $\boldsymbol{\beta}_{h,0}$ to be close to the first row of $\boldsymbol{A}_0^h$ when n is large relative to h: this is achieved by setting (see Appendix C)

$$\boldsymbol{\beta}_{h,0} = \left(0, d_0^h, \frac{1-d_0^h}{n-1}, ..., \frac{1-d_0^h}{n-1}\right)'. \tag{20}$$

The extended ridge estimator for the corresponding $\boldsymbol{\beta}_h$ is defined as in (9), replacing $\boldsymbol{\beta}_0$ by $\boldsymbol{\beta}_{h,0}$, the penalty parameters and the value of d_0 being chosen by cross-validation for each horizon h. For Bayesian estimation, we use the same type of prior as when $h = 1$ (i.e., (10) and(13)), also replacing $\boldsymbol{\beta}_0$ by $\boldsymbol{\beta}_{h,0}$. Forecasts for specific elements of $\boldsymbol{y}_t$ can readily be formed by estimating only specific rows of (18), so that forecasts are obtained from the corresponding individual equations, as in the case $h = 1$.

4 Empirical illustrations

In this section, we provide two applications to data where long memory has been documented in the literature and for which multiple series supposedly belonging to the same system are available. In both cases, for a large number of series, we compare out-of-sample forecasts obtained from the AR(1)-X equation (4) by three estimation methods: OLS, ridge, and Bayesian estimation, as defined in Section 3. We also include in the comparison the forecasts of three benchmark models, which are purely univariate time series models in the sense that they specify y_t as a function of the (infinite) past of y_t only. The six models and their estimation method are listed below:

1. AR(1): $y_t = \gamma_0 + \gamma_1 y_{t-1} + \epsilon_t$, estimated by OLS.
2. ARFIMA(1,d,0): $(1-L)^d(y_t - \gamma_0 - \gamma_1 y_{t-1}) = \epsilon_t$, estimated by Gaussian maximum likelihood.
3. HAR (Corsi, 2009): $y_t = \gamma_0 + \gamma_1 y_{t-1} + \gamma_2 \frac{1}{5}\sum_{i=1}^{5} y_{t-i} + \gamma_3 \frac{1}{21}\sum_{i=1}^{21} y_{t-i} + \epsilon_t$, estimated by OLS.
4. AR(1)-X: $y_t = \gamma_0 + \gamma_1 y_{t-1} + \sum_{i=2}^{n} \gamma_i x_{i,t-1} + \epsilon_t$, estimated by OLS. This is the model defined in (4).
5. RAR(1)-X: This model is identical to the AR(1)-X. The estimator is the extended ridge estimator defined by (9), see Section 3.2. Recall that in this case we shrink γ_1 towards d_0 with penalty parameter λ_d, γ_i toward $(1-d_0)/n - 1$ $(\forall i > 1)$ with the penalty λ_a, and $\sum_{i=1}^{n}\gamma_i$ towards 1 with a penalty of λ_S. The penalty parameters (i.e., λ_d, λ_a and λ_S) and d_0 are chosen by cross-validation as explained at the end of Section 3.2; details are provided in Appendix D.

6. BAR(1)-X: This specification is also identical to the AR(1)-X but the estimation is performed by the Bayesian method presented in Section 3.3. The prior for the variance of ϵ_t is non-informative, see (10), and the prior for the regression coefficients $\boldsymbol{\beta} = (\gamma_0, \gamma_1, \gamma_2, \ldots, \gamma_n)'$ is the Gaussian density defined by (13) together with (3.2), (14) and (12). More specifically, the prior on γ_0 is quasi-noninformative (with a mean of 0 and a variance of 100), the prior mean of γ_1 is set equal to d_0, and the prior mean of γ_i $\forall i > 1$ is set to $(1 - d_0)/(n - 1)$. The prior precision of γ_1 is $1/s_d^2 + h_0$, the prior precision of γ_i $(\forall i > 1)$ is $1/s_a^2 + h_0$. The co-precisions (the off-diagonal elements of the inverse of V_0) are all set to h_0. The larger h_0, the smaller the prior variance that the sum of the last elements of $\boldsymbol{\beta}$ is close to the sum of their prior means (equal to 1 for (3.2)). The prior parameters d_0, s_d, s_a and h_0 are chosen by cross-validation (see Appendix D for details).

For forecast horizons $h > 1$, we use iterated multistep forecasts (i.e., recursive substitution) for the first three methods (AR(1), ARFIMA and HAR) and direct multistep forecasts for the last three AR(1)-X, RAR(1)-X and BAR(1)-X; as discussed in Subsection 3.4, this avoids contaminating forecasts across variables when additional (non autoregressive) regressors are present.

The out-of-sample forecasts (at several horizons) are compared to the observed values using both the mean squared error (MSE) and the mean absolute deviation (MAE) loss functions. These loss functions are defined for each model m as

$$MSE_h^{(m)} = \frac{1}{T_h}\sum_{t=1}^{T_h}(\hat{y}_{t,h}^{(m)} - y_t)^2, \quad MAE_h^{(m)} = \frac{1}{T_h}\sum_{t=1}^{T_h}|\hat{y}_{t,h}^{(m)} - y_t|, \tag{21}$$

where h is the forecast horizon, T_h is the number of forecasts, and $\hat{y}_{t,h}^{(m)}$ is the forecast of y_t at horizon h by model m. The comparison tool is the model confidence set (MCS) procedure of Hansen, Lunde, and Nason (2011) and we also report rolling windows of the average loss functions.

In the first application, y_t is the logarithm of a measure of daily realized volatility for a set of 250 U.S. company stocks. In the second application, it is the logarithm of the monthly seasonally adjusted river streamflows at 97 locations in the Columbia river basin.

4.1 Daily realized volatilities of U.S. stocks

The dataset consists of transaction prices at the 1-second sampling frequency for $n = 250$ large capitalization stocks from the NYSE, AMEX and NASDAQ, covering the period from 2005-01-03 to 2017-07-24 (3,276 trading days). The trading session runs from 9:30 EST until 16:00 EST and stocks are ordered by decreasing average daily transactions volume, i.e., with the most liquid coming first. We aggregated the data at the 5-minute frequency and computed the MedRV estimator of Andersen, Dobrev, and Schaumburg (2012), a non-parametric robust to jumps estimator of the integrated variance. If $r_{t,i}$ is the ith 5-minute return of a given stock on a day t containing M (e.g., 78) of such returns, $log(MedRV_t)$ (denoted by y_t hereafter) is computed as the logarithm of $MedRV_t = \frac{\pi}{6-4\sqrt{3}+\pi}\frac{M}{M-2}\sum_{i=3}^{M}\mathrm{med}(|\mathrm{r_{t,i}}|, |\mathrm{r_{t,i-1}}|, |\mathrm{r_{t,i-2}}|)^2$, where $\mathrm{med}(\cdot)$ denotes the median. Notice that VAR models for the logarithm of realized variances have been used for instance by Anderson and Vahid (2007).

The six competing models are estimated on rolling windows of $T = 1,000$ observations. They are estimated first on the sample spanning the period from 2005-01-03 to 2008-10-31, and h-step ahead forecasts of y_t are computed for ten horizons ($h = 1, 2, \ldots, 10$) leading to a total number

of 2,277-h forecasts. The parameters estimated on each window are kept constant to produce 25 consecutive forecasts and then re-estimated on the next window of T observations. To speed up the estimation, the four tuning parameters of the RAR(1)-X and BAR(1)-X models are only estimated once by cross-validation on the first window of T observation and then kept constant. The rolling is continued until the last possible window of the full sample. The models are estimated for each of the $n = 250$ available series.

The presence of long memory in the volatility of the log-returns of financial assets is a well recognized stylized fact (see Baillie, Bollerslevand Mikkelsen, 1996, Breidt, Crato and de Lima, 1998, Comte and Renault, 1998, among others). For the sake of illustration, the average value (over the 250 series) of the estimated d parameters of the ARFIMA$(1, d, 0)$ obtained on the full sample is about 0.48 (with a standard deviation of 0.02).

To make sure that our empirical results are not specific to the chosen forecasting period, we compare the forecasting power of the competing models on rolling windows. More specifically, Figures 1 and 2 show, respectively, the averages (over the 250 stocks) of the MSE and MAE loss functions for a sequences of rolling samples of 250 forecasts, and three forecast horizons (i.e., $h = 1, 5$ and 5). Figures 3 and 4 report, for the same three forecasting horizons, the time evolution of the frequencies at which each model belongs to the MCS at the confidence level of 75% (named MCS75 in the sequel), and again sequences of rolling samples of 250 forecasts. A frequency of 50 (percent) for model m at date t means that the model m is in the MCS75 for fifty percent of the 250 series; the MCS75 in question being obtained using the loss function computed from the 250 forecasts ending at date t. Notice that the MCS test is not applied to every consecutive window of 250 forecasts but every 25-th windows to facilitate the computations, so that 82 values are plotted for each competing model.

To help summarize the results, in addition to these 4 figures, Table 1 reports the average value (over the 82 windows) of the time evolution of the frequencies at which each model belongs to the MCS75. These values correspond to the average value of the six lines plotted in Figures 3 and 4 as well as the other forecasting horizons, i.e., for $h = 1, \ldots, 10$.

Some comments, which apply equally to both loss functions, follow.

- AR(1) and AR(1)-X are strongly outperformed by the other models over the forecast period. Their average losses are larger (often strongly) than those of the other models. The frequencies of inclusion of these models in the MCS75 are very often smaller than 10 percent, and almost never above 20. This is confirmed in Table 1, where these two models are by far the least present on average in MCS75, whatever the forecast horizon and the choice of the loss function.

- ARFIMA and HAR perform comparably, especially considering their average losses. Their frequencies of inclusion in the MCS75 are also similar, but sometimes more different than the losses. In broad outline, these frequencies fluctuate between 25 and 50 percent until mid-2012, and then between 50 and 70 percent. Table 1 shows that on average these two models belong to the MCS75 in about 50% of the cases, whatever the forecast horizon and the choice of the loss function.

- RAR(1)-X and BAR(1)-X perform comparably and better than ARFIMA and HAR, with smaller losses and higher frequencies. The latter are most of the time between 65 and 75%, though for horizons 5 and 10, the RAR frequencies are higher (by 10 to 20 points) than the BAR frequencies in 2012 and 2013, and again from March 2016. Interestingly, Table 1

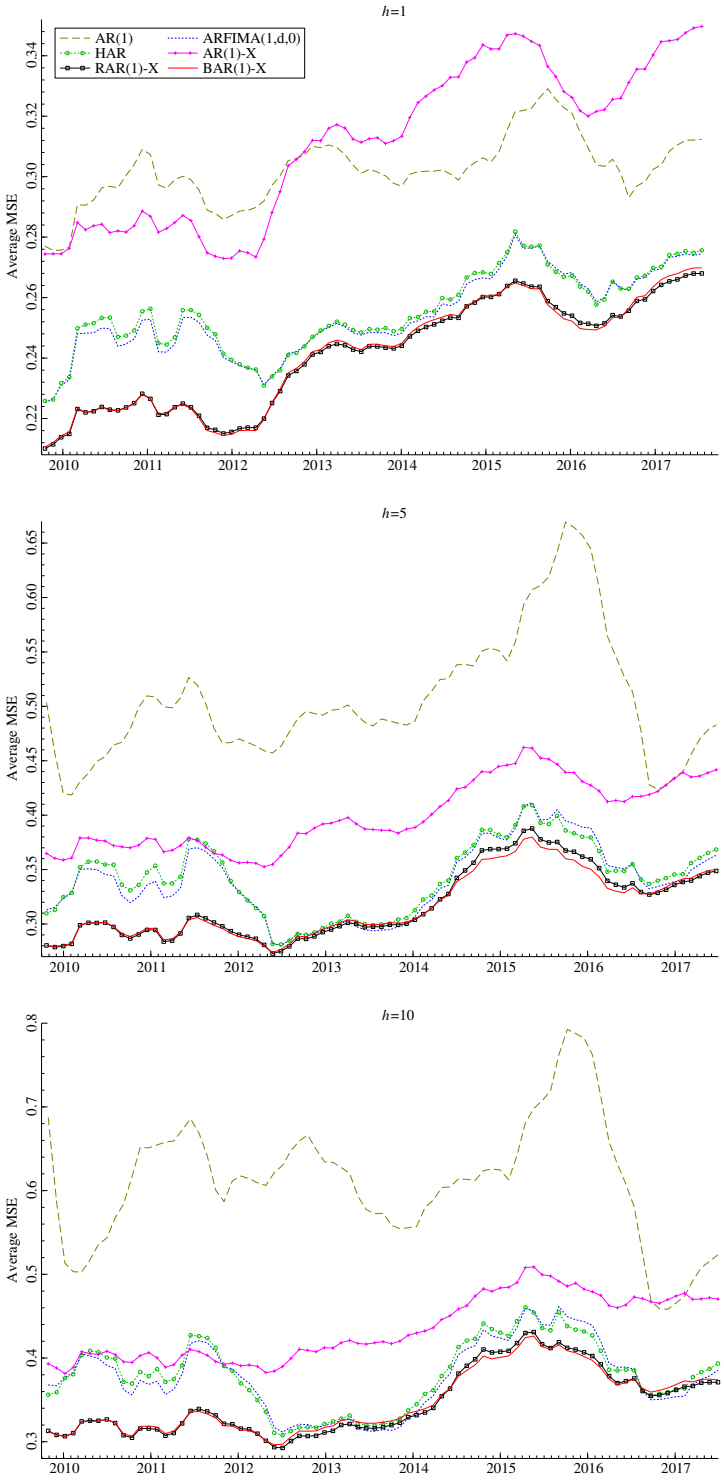

Figure 1: Average MSE (over the 250 series) computed on rolling windows of 250 observations. The three panels are respectively for $h = 1, 5$ and 10.

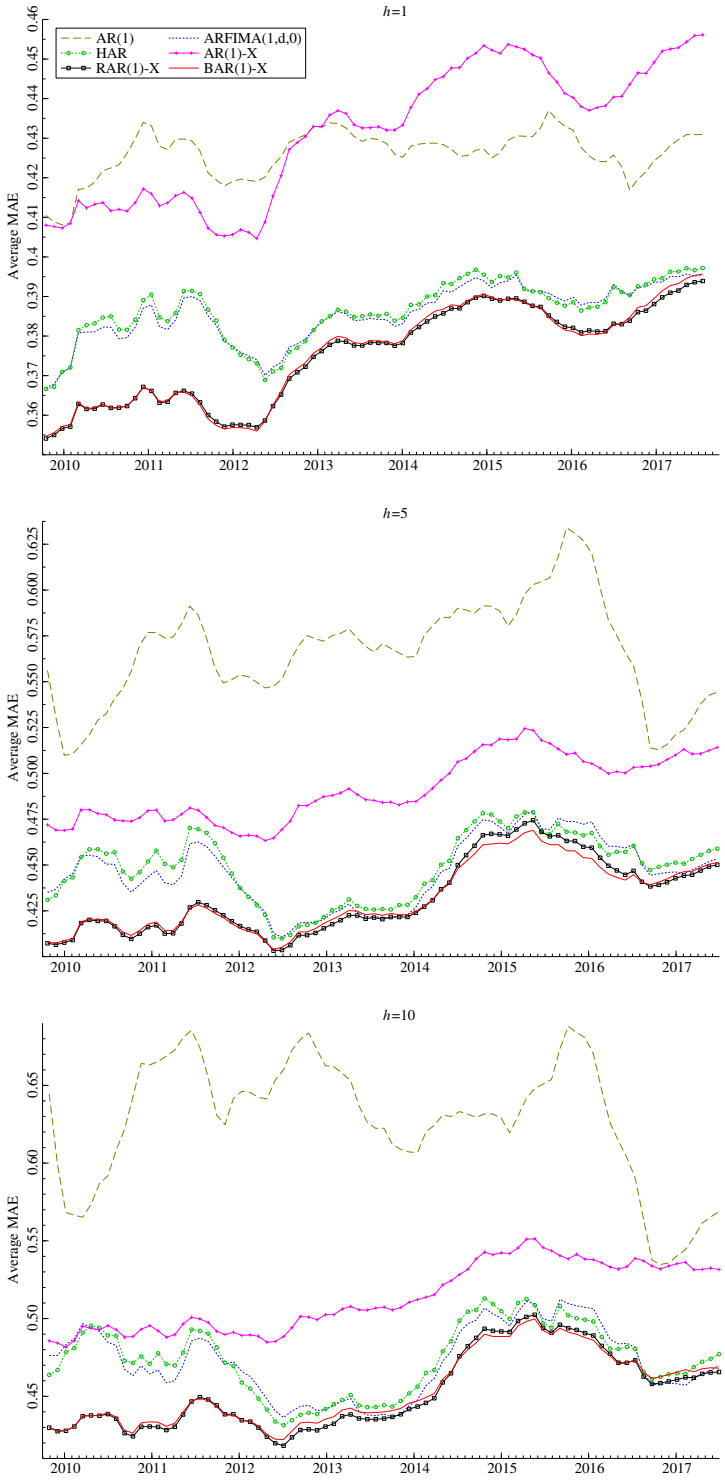

Figure 2: Average MAE (over the 250 series) computed on rolling windows of 250 observations. The three panels are respectively for $h = 1, 5$ and 10.

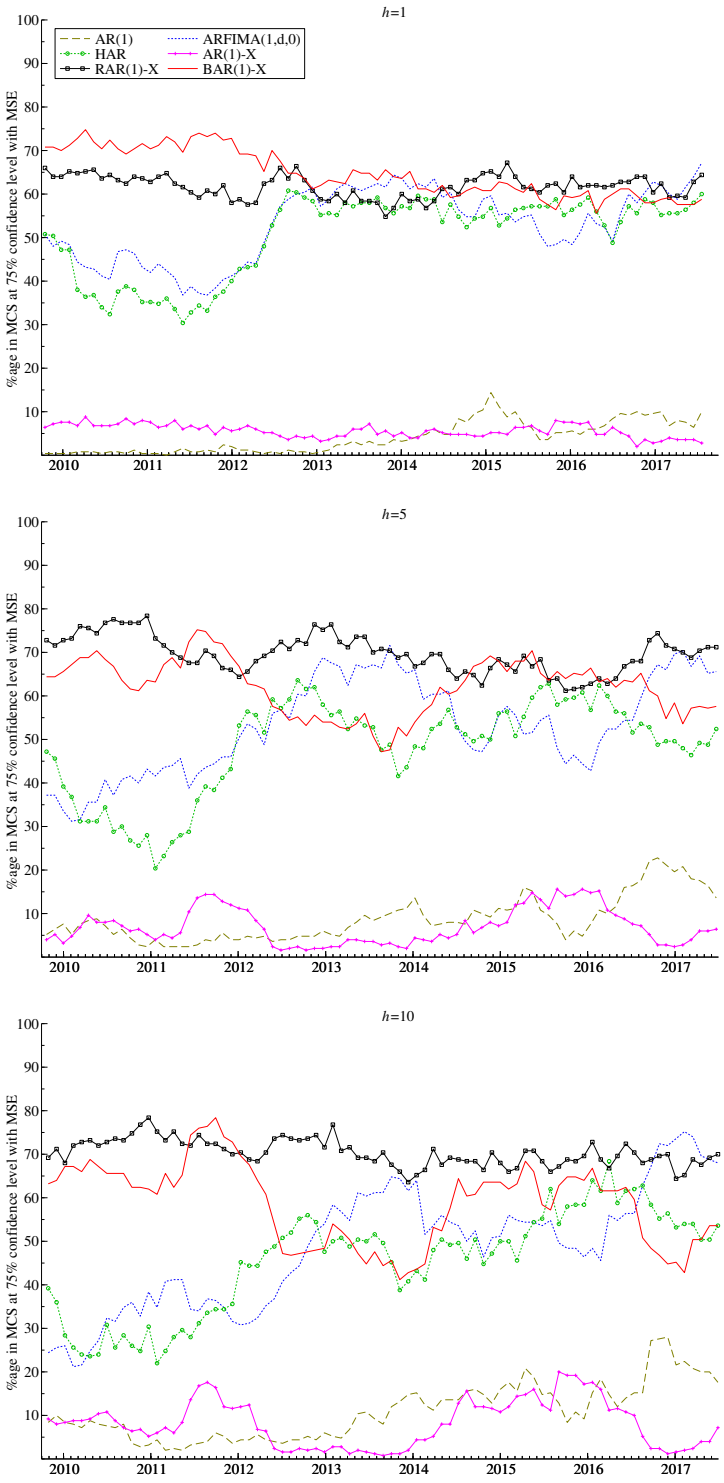

Figure 3: Frequencies (over the 250 series), at each date, at which each model belongs to the MCS (at 75% confidence level) for MSE loss function. The three panels are respectively for $h = 1, 5$ and 10.

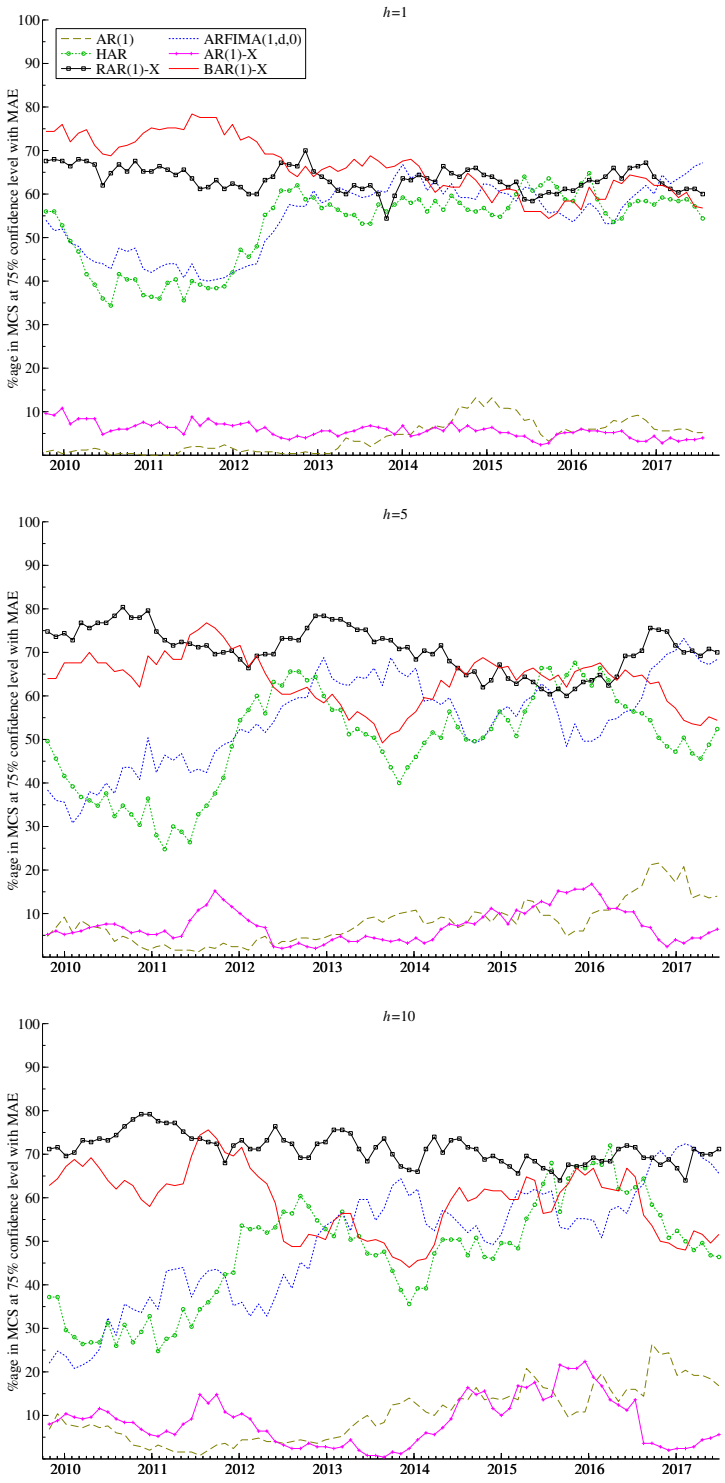

Figure 4: Frequencies (over the 250 series), at each date, at which each model belongs to the MCS (at 75% confidence level) for MAE loss function. The three panels are respectively for $h = 1, 5$ and 10.

suggests that for $h = 1$, BAR(1)-X is on average the most frequently in the MCS75 but for $h > 1$, RAR(1)-X is even better than BAR(1)-X, with an average frequency in the MCS75 around 70%, i.e., 20 points higher than ARFIMA and HAR.

In brief, the use of the theoretical constraints in the AR(1)-X model through the proposed Bayesian and ridge estimation methods strongly improves the model forecasting performance with respect to OLS. The bad performance of the latter is due to a lack of precision because 251 coefficients are estimated using 1,000 observations, whereas the shrinkage methods impose a relevant theoretical structure on the estimated coefficients. The performance of the shrinkage methods is also most of the time significantly superior to that of the ARFIMA and HAR models; this difference can be attributed to the use of a larger, but relevant, information set.

Table 1: Average frequencies (over the 250 series and the 82 rolling windows), at which each model belongs to the MCS (at 75% confidence level)

h	AR(1)	ARFIMA(1,d,0)	HAR	AR(1)-X	RAR(1)-X	BAR(1)-X
			MSE			
1	4.005	52.946	50.463	5.580	61.756	64.698
2	5.412	44.573	45.605	4.874	72.430	56.094
3	6.602	48.015	47.131	6.281	73.427	56.746
4	7.620	50.074	46.731	6.973	68.716	63.319
5	8.716	53.156	48.168	7.002	69.664	62.069
6	9.205	52.020	48.672	7.664	68.316	61.931
7	9.630	50.716	48.677	7.284	70.306	60.459
8	9.906	51.274	48.780	7.714	71.649	58.474
9	10.202	49.877	46.874	8.202	70.800	59.733
10	10.894	48.435	45.467	8.183	70.341	58.578
			MAE			
1	4.107	54.688	52.771	5.717	63.488	66.298
2	5.027	45.773	47.136	5.106	73.585	56.706
3	6.247	48.706	47.901	6.859	74.281	58.202
4	7.007	51.659	48.163	7.264	69.807	64.588
5	7.911	54.573	49.748	7.279	70.637	63.570
6	8.316	52.963	49.881	8.084	69.457	63.042
7	8.854	51.812	50.074	7.857	71.802	61.877
8	9.294	52.454	51.175	7.970	72.765	60.805
9	9.595	50.499	49.146	8.227	71.990	61.017
10	10.222	49.427	47.304	8.607	71.210	59.096

4.2 Monthly river streamflows in the Columbia river basin

Natural streamflows play a significant role in shaping biological communities and they regulate ecological processes in local ecosystems. In most industrialized economies streamflows are modified as a result of human activity (agricultural, industrial, ...), and regulated as such. Forecasting future flows is essential for planning dam discharges and adaptation.

In the hydrology community, many studies have been carried out on the test for long memory streamflow processes, following the seminal paper of Hurst (1951) on dimensioning dams for the Nile river, and which pioneered the literature on long memory. For instance, Montanari, Rosso and Taqqu (1997) applied ARFIMA modelling to the monthly and daily inflows of Lake Maggiore, Italy. Depending on the modelling strategy, their confidence interval for the degree of long memory varies with a $[.35, 45]$ range. This is a feature that has often been documented in the hydrology and streamflow forecasting literatures. Ooms and Franses (2001) documented that monthly river flow data displays long memory, in addition to pronounced seasonality based on simple time series plots and periodic sample autocorrelations. Wang et al. (2002) investigated the long memory property of two daily streamflows of the Yellow River in China and found that both daily streamflow processes exhibit strong long memory. It must be noted that long memory is not found in all hydrological datasets, depending on the data considered, the frequency and length of observation (see, for instance, Rao and Bhattacharya, 1999, and Montanari et al., 2000), but as mentioned in the doctoral thesis of Wen Wang (2006) at the Technological University of Delft, ARFIMA models remained at the time the main contenders for forecasting streamflows (though some neural network based techniques may help capturing some nonlinearities). Over the last 15 years, the literature has explored machine learning techniques (artificial neural networks, support vector machines, ...) for forecasting hydrological series and have found mixed evidence depending on the situations. To assess these results Papacharalampous et al. (2019) perform an extensive comparison of 20 prototypical multistep forecasting models (11 'stochastic', i.e., extensions of ARMA models, and 9 'machine learning' models) over hundreds of simulated and empirical datasets and using 18 accuracy metrics. Their findings are that (i) most empirical series exhibit a degree of long memory between 0 and 0.45, with a median close to 0.2 (see their Figure 1), (ii) the most accurate 'stochastic' and machine learning techniques perform similary, and (iii) ARFIMA models belong to the class of most accurate 'stochastic' techniques (see their Figure 18).

To illustrate our modelling approach, we assess its forecasting accuracy using the Modified Streamflow dataset of the Columbia river basin provided by the Bonneville Power Administration (BPA), the United States Army Corps of Engineers and the U.S. Bureau of Reclamations. To quote the BPA: "Since irrigation practices have changed since the historical streamflows were observed, the historical streamflows have been adjusted to account for current levels of irrigation depletions." Hence "Modified streamflows are historical streamflows that would have been observed if current irrigation depletions (as of year 2018) existed in the past and if the effects of river regulation were removed." These modified flows allow for intertemporal comparisons of the natural inflows since they are adjusted to a common level of irrigation development and evaporation in upstream reservoirs and lakes, and they reflect no regulation by dams. They are recorded and computed at 97 locations in the Columbia river basin over 90 years (October 1928-December 2018, i.e., $1,083$ monthly observations). We model and forecast the logarithm of the monthly series and we adjust them for seasonal variations using X12arima in Oxmetrics version 8.10.

To confirm the presence of long memory in the data, we estimated on the 97 series and on the full sample an ARFIMA$(0, d, 0)$ and an ARFIMA$(1, d, 0)$ model by maximum likelihood. The average $\widehat{d}$ is equal to 0.45 with a standard deviation of 0.06 for the former and 0.21 with a standard deviation of 0.19 for the latter.

We now report the results of a forecasting comparison of the six models listed at the beginning of this section. The estimation and forecasting are organized as described in the previous subsection, with rolling windows of 400 observations for estimation but because both the number of series and

the number of observations are smaller than in the previous application, all models are reestimated each time a new observation becomes available. The first window corresponds to the period October 1928-January 1962. We obtain a total of 683-h forecasts for the 97 series, where in this application, $h = 1, 2, \ldots, 6$.

Figures 5-8 report the forecasting results in the same way as in Figures 1-4, but for horizons $1, 3$ and 5. For this application, the MCS test is applied every 5-th window of 50 forecasts leading to a total of 125 tests. Theses results lead to the following observations:

- The AR(1)-X model (estimated by OLS) has the worst forecasting performance, whatever the forecasting horizon.
- The ARFIMA, HAR, RAR(1)-X and BAR(1)-X have average losses much lower than the AR(1)-X.
- The BAR(1)-X is the model that belongs the most often in the MCS75. For instance, for $h = 1, 2$ and 3, the BAR(1)-X belongs to the MCS75 on average in more than 90% of the cases (see Table 2) against 60 to 75% for the AR(1), ARFIMA(1,d,0) and HAR models and against 78% for RAR(1)-X.

Table 2: Average frequencies (over the 97 series and the 125 rolling windows), at which each model belongs to the MCS (at 75% confidence level)

h	AR(1)	ARFIMA(1,d,0)	HAR	AR(1)-X	RAR(1)-X	BAR(1)-X
			MSE			
1	61.344	68.973	69.905	47.282	77.171	95.505
2	66.400	70.416	74.425	37.386	78.746	94.334
3	70.606	70.623	74.507	46.829	77.023	92.066
4	73.856	72.998	72.412	43.909	82.351	89.608
5	79.002	76.132	74.449	37.287	83.357	81.386
6	77.823	73.237	72.503	33.402	82.334	81.056
			MAE			
1	64.495	70.680	73.963	46.400	77.616	92.627
2	67.810	70.771	75.340	39.167	78.969	93.278
3	70.548	71.984	75.588	50.111	77.155	90.977
4	72.841	72.478	75.076	46.639	82.219	88.998
5	79.068	76.759	76.577	40.462	84.198	83.571
6	78.078	72.652	73.229	36.412	84.091	82.804

5 Conclusions

This paper considers a novel approach in empirical work for modeling a variable exhibiting long range dependence using a large cross-section of related variables, instead of using its own and long history. This approach is based on two recent theoretical contributions that show that long memory can be caused by dependences within a large network or system. We provide two estimation techniques that harness the informativeness of the theoretical models and use them to drive the

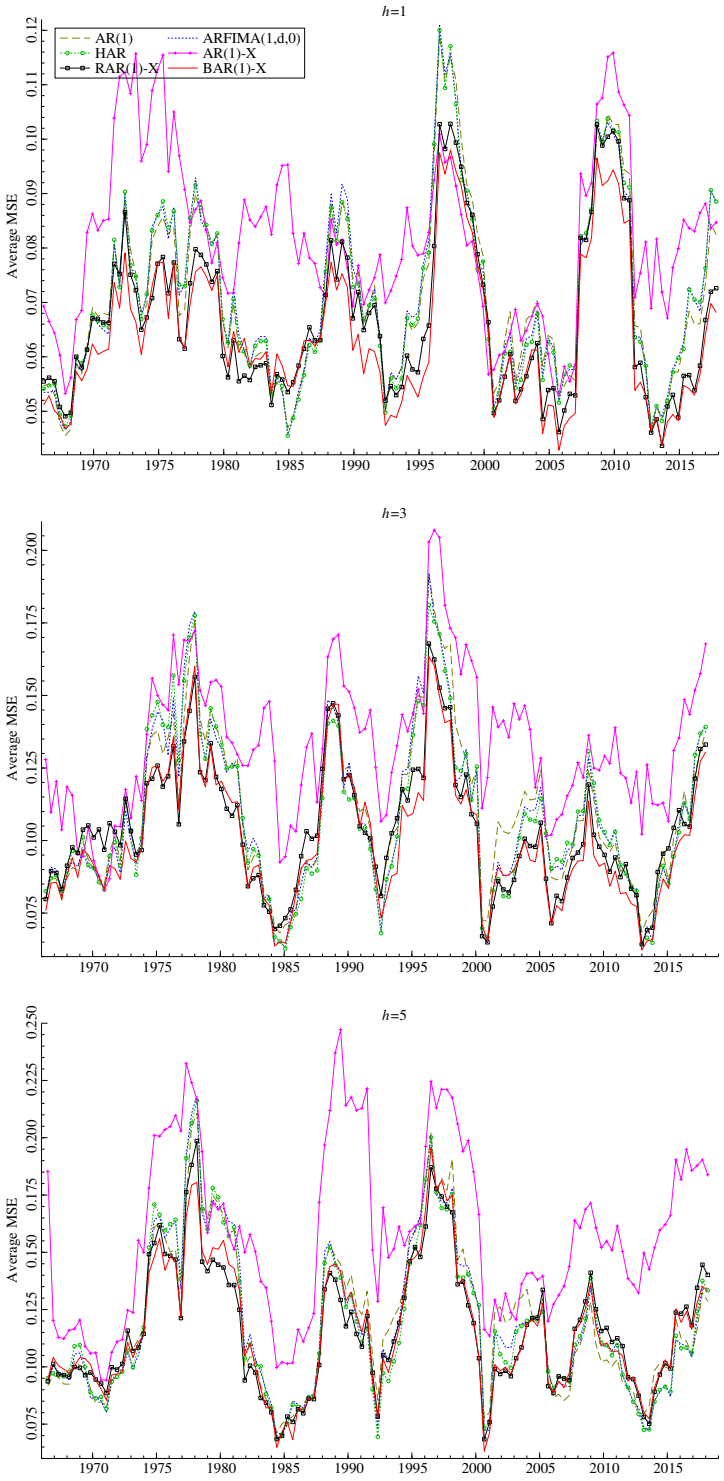

Figure 5: Average MSE (over the 97 series) computed on rolling windows of 50 observations. The three panels are respectively for $h = 1, 3$ and 5.

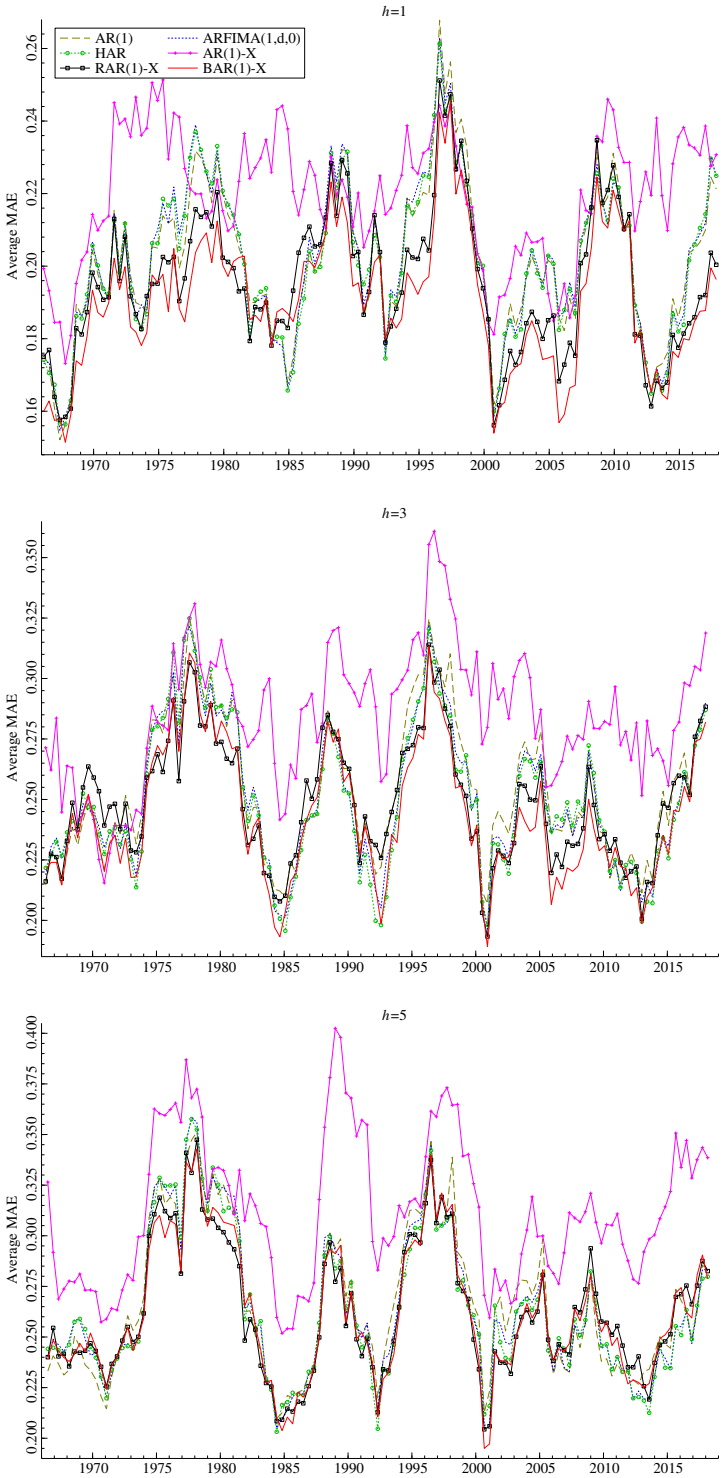

Figure 6: Average MAE (over the 97 series) computed on rolling windows of 50 observations. The three panels are respectively for $h = 1, 3$ and 5.

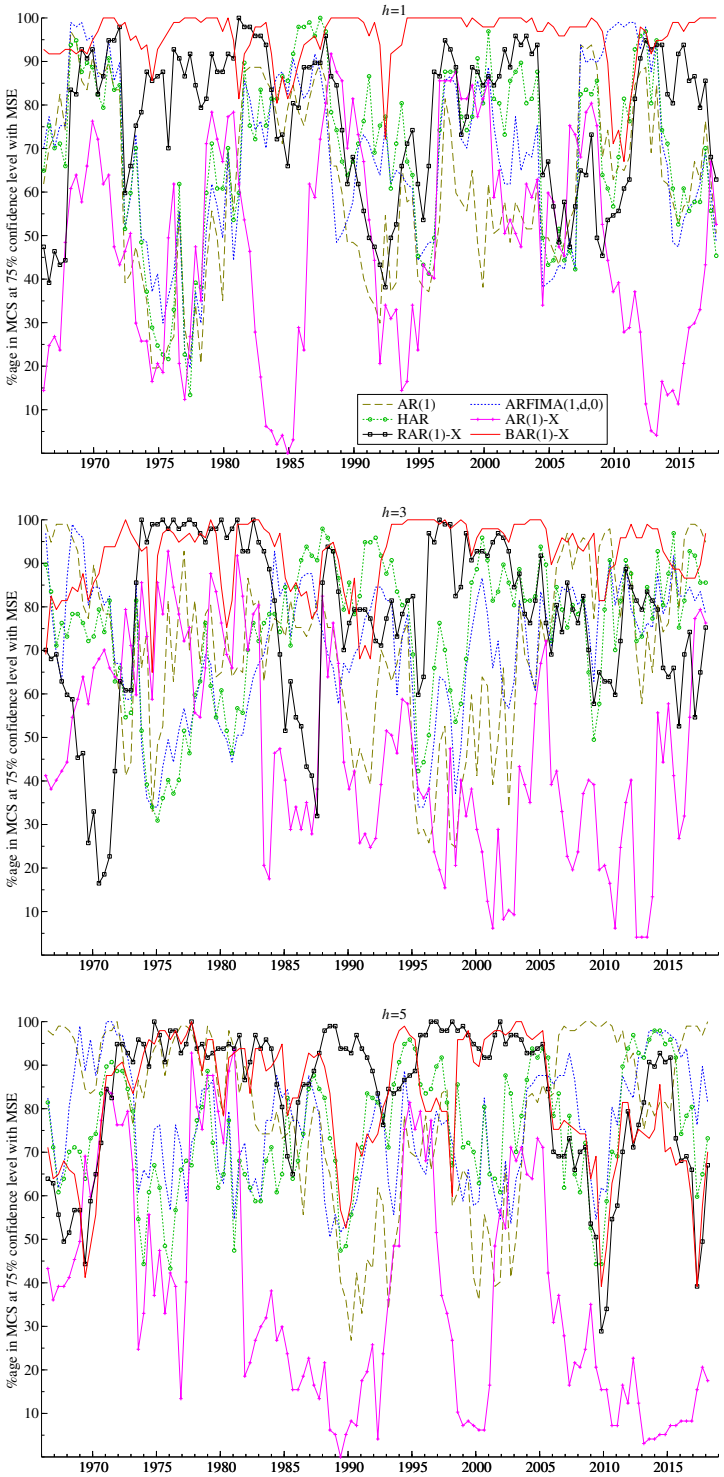

Figure 7: Frequencies (over the 97 series), at each date, at which each model belongs to the MCS (at 75% confidence level) for the MSE loss function. The three panels are respectively for $h = 1, 3$ and 5.

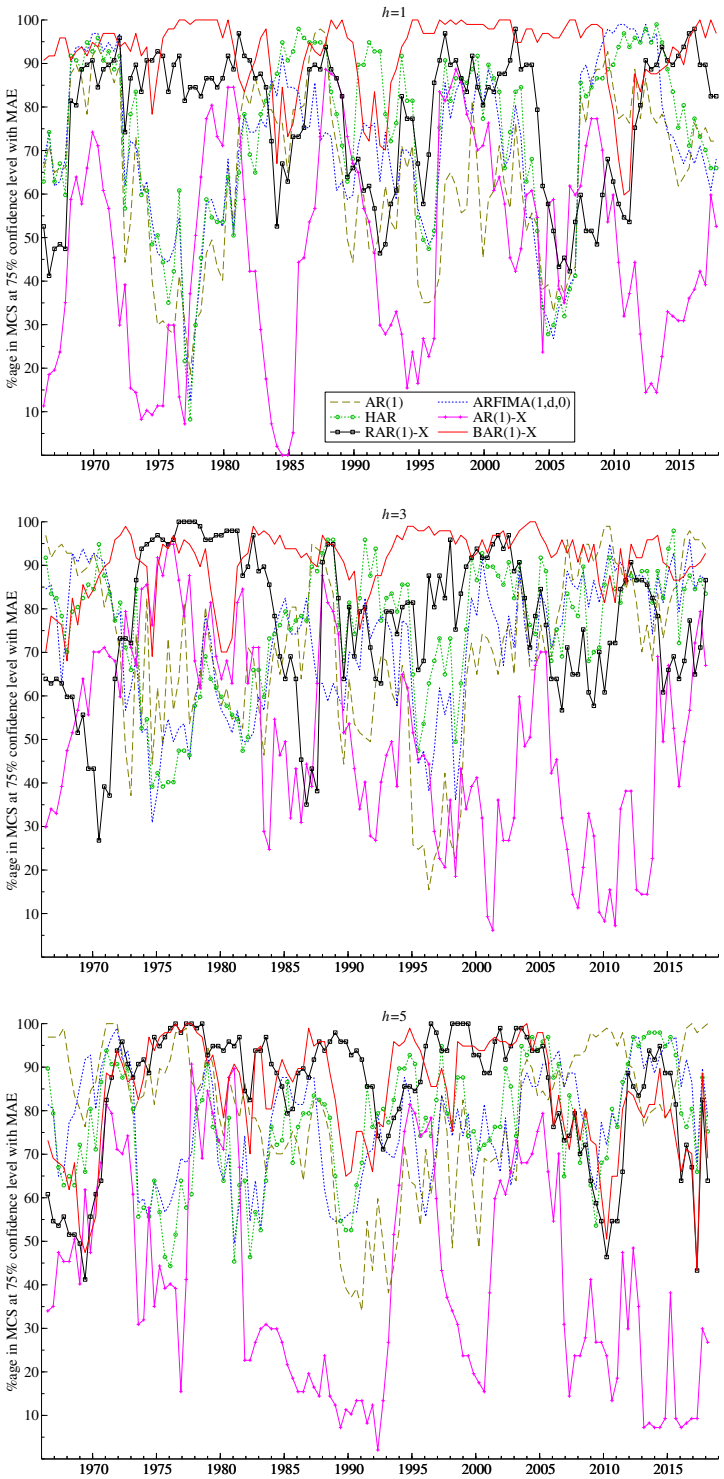

Figure 8: Frequencies (over the 97 series), at each date, at which each model belongs to the MCS (at 75% confidence level) for the MAE loss function. The three panels are respectively for $h = 1, 3$ and 5.

estimation, either via an extended ridge regression that shrinks the estimates toward a structure derived from the theory, or by using the latter to design an informative prior in a Bayesian setup.

In two empirical applications in the context of (i) realized volatilities of stocks that are interdependent within financial markets; and (ii) streamflow series of the Columbia river basin, we show that our proposed modeling and estimation strategy improves upon standard techniques.

Importantly, our results show that it is possible to model variables that exhibit strong dependence over long horizons even in a dataset with a short time span, provided that the cross-sectional dimension is large. Indeed, information related to the distant past can be estimated using a large collection of variables within a system or a network.

Appendices

Appendix A: Proof of (9) and of (13)

Proof of (9): notice that $(\boldsymbol{\beta}'\boldsymbol{\iota} - \boldsymbol{\beta}_0'\boldsymbol{\iota})^2 = (\boldsymbol{\beta}'\boldsymbol{\iota} - \boldsymbol{\beta}_0'\boldsymbol{\iota})(\boldsymbol{\beta}'\boldsymbol{\iota} - \boldsymbol{\beta}_0'\boldsymbol{\iota})' = \boldsymbol{\beta}'\boldsymbol{\iota}\boldsymbol{\iota}'\boldsymbol{\beta} - 2\boldsymbol{\beta}'\boldsymbol{\iota}\boldsymbol{\iota}'\boldsymbol{\beta}_0 + \boldsymbol{\beta}_0'\boldsymbol{\iota}\boldsymbol{\iota}'\boldsymbol{\beta}_0$. By developing the quadratic forms, the ER objective function (8) is equal to $\boldsymbol{\beta}'\boldsymbol{Z}'\boldsymbol{Z}\boldsymbol{\beta} - 2\boldsymbol{\beta}'\boldsymbol{Z}'\boldsymbol{Y} + \boldsymbol{\beta}'\boldsymbol{\Lambda}_k\boldsymbol{\beta} - 2\boldsymbol{\beta}'\boldsymbol{\Lambda}_k\boldsymbol{\beta}_0 + \lambda_s^2\boldsymbol{\beta}'\boldsymbol{\iota}\boldsymbol{\iota}'\boldsymbol{\beta} - 2\lambda_s^2\boldsymbol{\beta}'\boldsymbol{\iota}\boldsymbol{\iota}'\boldsymbol{\beta}_0 + \boldsymbol{Y}'\boldsymbol{Y} + \lambda\boldsymbol{\beta}_0'\boldsymbol{\Lambda}_k\boldsymbol{\beta}_0 + \lambda_s^2\boldsymbol{\beta}_0'\boldsymbol{\iota}\boldsymbol{\iota}'\boldsymbol{\beta}_0$. Solving the first-order condition yields the solution (9).

Proof of (13): to show that that the kernel (11) corresponds to (13), we can write that (11) is equal to

$$\exp\{-\frac{1}{2}[(\boldsymbol{\beta}-\boldsymbol{\beta}_0)'\boldsymbol{Q}_0(\boldsymbol{\beta}-\boldsymbol{\beta}_0) + h_0(\boldsymbol{\beta}'\boldsymbol{\iota} - \boldsymbol{\beta}_0'\boldsymbol{\iota})(\boldsymbol{\beta}'\boldsymbol{\iota} - \boldsymbol{\beta}_0'\boldsymbol{\iota})']\} = K_0\exp[-\frac{1}{2}f(\boldsymbol{\beta})],$$

where K_0 does not depend on $\boldsymbol{\beta}$ and

$$f(\boldsymbol{\beta}) = \boldsymbol{\beta}'(\boldsymbol{Q}_0 + h_0\boldsymbol{\iota}\boldsymbol{\iota}')\boldsymbol{\beta} - 2\boldsymbol{\beta}'(\boldsymbol{Q}_0\boldsymbol{\beta}_0 + h_0\boldsymbol{\iota}\boldsymbol{\beta}_0'\boldsymbol{\iota}) = (\boldsymbol{\beta}-\bar{\boldsymbol{\beta}}_0)'\boldsymbol{V}_0{}^{-1}(\boldsymbol{\beta}-\bar{\boldsymbol{\beta}}_0) + C_0,$$

where $\boldsymbol{V}_0{}^{-1} = \boldsymbol{Q}_0 + h_0\boldsymbol{\iota}\boldsymbol{\iota}'$, $\bar{\boldsymbol{\beta}}_0 = \boldsymbol{V}_0(\boldsymbol{Q}_0\boldsymbol{\beta}_0 + h_0\boldsymbol{\iota}\boldsymbol{\beta}_0'\boldsymbol{\iota})$, and $C_0 = \bar{\boldsymbol{\beta}}_0'\boldsymbol{V}_0^{-1}\bar{\boldsymbol{\beta}}_0$ does not depend on $\boldsymbol{\beta}$. Hence, the prior density depends on $\boldsymbol{\beta}$ only through $\exp[-\frac{1}{2}(\boldsymbol{\beta}-\bar{\boldsymbol{\beta}}_0)'\boldsymbol{V}_0{}^{-1}(\boldsymbol{\beta}-\bar{\boldsymbol{\beta}}_0)]$, which is the kernel of the Gaussian density $N_k(\bar{\boldsymbol{\beta}}_0, \boldsymbol{V}_0)$. To show that this Gaussian density is the same as (13), we show that $\bar{\boldsymbol{\beta}}_0 = \boldsymbol{\beta}_0$:

$$\begin{aligned}
\bar{\boldsymbol{\beta}}_0 &= (\boldsymbol{Q}_0 + h_0\boldsymbol{\iota}\boldsymbol{\iota}')^{-1}(\boldsymbol{Q}_0\boldsymbol{\beta}_0 + h_0\boldsymbol{\iota}\boldsymbol{\beta}_0'\boldsymbol{\iota}) = (\boldsymbol{Q}_0^{-1} - \frac{h_0\boldsymbol{Q}_0^{-1}\boldsymbol{\iota}\boldsymbol{\iota}'\boldsymbol{Q}_0^{-1}}{1+h_0\boldsymbol{\iota}'\boldsymbol{Q}_0^{-1}\boldsymbol{\iota}})(\boldsymbol{Q}_0\boldsymbol{\beta}_0 + h_0\boldsymbol{\iota}\boldsymbol{\beta}_0'\boldsymbol{\iota}) \\
&= \boldsymbol{\beta}_0 + h_0\boldsymbol{Q}_0^{-1}\boldsymbol{\iota}\boldsymbol{\beta}_0'\boldsymbol{\iota} - \frac{1}{1+h_0\boldsymbol{\iota}'\boldsymbol{Q}_0^{-1}\boldsymbol{\iota}}(h_0\boldsymbol{Q}_0^{-1}\boldsymbol{\iota}\underbrace{\boldsymbol{\iota}'\boldsymbol{Q}_0^{-1}\boldsymbol{Q}_0\boldsymbol{\beta}_0}_{=\boldsymbol{\beta}_0'\boldsymbol{\iota}} + h_0\boldsymbol{Q}_0^{-1}\boldsymbol{\iota}\boldsymbol{\iota}'\boldsymbol{Q}_0^{-1}h_0\boldsymbol{\iota}\boldsymbol{\beta}_0'\boldsymbol{\iota}) \\
&= \boldsymbol{\beta}_0 + h_0\boldsymbol{Q}_0^{-1}\boldsymbol{\iota}\boldsymbol{\beta}_0'\boldsymbol{\iota}\Big(1 - \frac{1}{1+h_0\boldsymbol{\iota}'\boldsymbol{Q}_0^{-1}\boldsymbol{\iota}} - \frac{h_0\boldsymbol{\iota}'\boldsymbol{Q}_0^{-1}\boldsymbol{\iota}}{1+h_0\boldsymbol{\iota}'\boldsymbol{Q}_0^{-1}\boldsymbol{\iota}}\Big) = \boldsymbol{\beta}_0.
\end{aligned}$$

In the first line, the explicit form of the inverse of $\boldsymbol{Q}_0 + h_0\boldsymbol{\iota}\boldsymbol{\iota}'$ is obtained by applying the Sherman-Morrison formula.

Appendix B: Bayesian estimation of the AR-X(1) model

The results exposed in this appendix are included for ease of reference. They are well known, see e.g. Bauwens, Lubrano, and Richard (1999) for details.

For the regression equation (5), with the assumption of normality of the error term, the prior (10) and (13), the posterior density of $\boldsymbol{\beta}$ and σ^2 is proportional to

$$(\sigma^2)^{-(T+2)/2}\exp\{-\frac{\hat{s}}{2\sigma^2}\}\exp\{-\frac{1}{2}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})'\frac{\boldsymbol{Z}'\boldsymbol{Z}}{\sigma^2}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})\}\exp\{-\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)'\boldsymbol{V}_0^{-1}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\}, \quad (22)$$

where $\hat{\boldsymbol{\beta}}$ is the OLS estimator $(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{Y}$, and $\hat{s}$ is the sum of squared OLS residuals.

Because the prior density is not conjugate, the posterior marginal density of $\boldsymbol{\beta}$ is not available analytically. However, the posterior density of $(\boldsymbol{\beta}, \sigma^2)$ can be simulated by applying a Gibbs sampler iterating between $\boldsymbol{\beta}$ and $\boldsymbol{\sigma}^2$. Indeed, the posterior density of $\boldsymbol{\beta}$ conditional on σ^2 is Gaussian:

$$\boldsymbol{\beta}|\sigma^2, \boldsymbol{Y}, \boldsymbol{Z} \sim N_k(\boldsymbol{\beta}_*, \boldsymbol{V}_*), \quad (23)$$

where

$$\boldsymbol{V}_* = \left(\frac{\boldsymbol{Z}'\boldsymbol{Z}}{\sigma^2} + \boldsymbol{V}_0^{-1}\right)^{-1}, \quad (24)$$

$$\boldsymbol{\beta}_* = \boldsymbol{V}_*\left(\frac{\boldsymbol{Z}'\boldsymbol{Y}}{\sigma^2} + \boldsymbol{V}_0^{-1}\boldsymbol{\beta}_0\right) := \boldsymbol{\beta}_*(\sigma^2). \quad (25)$$

and the complementary conditional density of σ^2 is inverted-gamma:

$$\sigma^2|\boldsymbol{\beta} \sim IG(T, (\boldsymbol{Y}-\boldsymbol{Z}\boldsymbol{\beta})'(\boldsymbol{Y}-\boldsymbol{Z}\boldsymbol{\beta})). \quad (26)$$

The Gibbs sampling algorithm to generate S draws $(\boldsymbol{\beta}^{(s)}, (\sigma^2)^{(s)})$, for $s = 1, 2, \ldots, S$, from the posterior of the parameters (after S_0 warming-up draws) is organized as follows:

1. *Choose an initial value* $(\sigma^2)^{(0)}$ *(e.g.* $\hat{s}/(T-k-2)$*).*
2. *Set* $s = 1$.
3. *Draw successively* $\boldsymbol{\beta}^{(s)}$ *from* the Normal density (23) where β_* and Q_* are computed with $\sigma^2 = (\sigma^2)^{(s-1)}$, and $(\sigma^2)^{(s)}$ *from* $IG(T, \boldsymbol{Y}-\boldsymbol{Z}\boldsymbol{\beta}^{(s)})'(\boldsymbol{Y}-\boldsymbol{Z}\boldsymbol{\beta}^{(s)})$.
4. *Set* $s = s+1$ *and go to step 3 unless* $s > S_0 + S$.
5. *Discard the first* S_0 *values of* $\boldsymbol{\beta}^{(s)}$ *and* $(\sigma^2)^{(s)}$.

The posterior expectation of $\boldsymbol{\beta}$ is approximated by the mean of the S draws $\boldsymbol{\beta}^{(s)}$, or by the mean of the S conditional expectations $\boldsymbol{\beta}_*[(\sigma^2)^{(s)}]$.

Appendix C: Explanation of (20)

Using $a_0 = (1-d_0)/(n-1)$, $\boldsymbol{A}_0 = d_0\boldsymbol{I}_n + a_0(\boldsymbol{J}_n - \boldsymbol{I}_n) = \frac{nd_0-1}{n-1}\boldsymbol{I}_n + \frac{1-d_0}{n-1}\boldsymbol{J}_n$. Using $\boldsymbol{J}_n^h = n^{h-1}\boldsymbol{J}_n$,

$$\begin{aligned}\boldsymbol{A}_0^h &= \sum_{j=0}^{h}\frac{h!}{j!\,(h-j)!}\left[\left(\frac{nd_0-1}{n-1}\right)^{h-j}\left(\frac{1-d_0}{n-1}\right)^{j}\right]\boldsymbol{J}_n^j \\ &= \left(d_0 + \frac{d_0-1}{n-1}\right)^h\boldsymbol{I}_n + \frac{1}{n}\left[1-\left(d_0+\frac{d_0-1}{n-1}\right)^h\right]\boldsymbol{J}_n\end{aligned}$$

and hence $\boldsymbol{A}_0^h = \left(d_0^h + o\left(n^{-1}\right)\right)\boldsymbol{I}_n + \left(\frac{1-d_0^h}{n} + o\left(n^{-1}\right)\right)\boldsymbol{J}_n$, for $n >> h$, so that the firs row is then close to $\left(d_0^h, \frac{1-d_0^h}{n}, ..., \frac{1-d_0^h}{n}\right)'$. The target $\boldsymbol{\beta}_{h,0}$ in (20) is obtained by putting 0 as first element and dividing the last $n-1$ elements by $n-1$ (instead of n) to ensure that the sum of the target is exactly equal to 1.

Note that this is restricted to large n relative to h, as $\boldsymbol{A}^h = \frac{1+o(1)}{n}\boldsymbol{J}_n$ when $h >> n$.

Appendix D: Technical details

Model confidence set

The procedure of Hansen et al. (2011) is applied using the MAE and MSE loss functions defined in (21) to perform the hypothesis tests of equal predictive accuracy needed to obtain each model confidence set. These tests are performed at the 25% significance level, so that the resulting MCS is at the confidence level of 75%. The test statistic is the range statistic that requires a bootstrap procedure.

For the application to daily realized volatilities, 10,000 bootstrap samples are used, with a block length of 5 observations to account for potential serial correlation and conditional heteroscedasticity in the losses. For the application to monthly river streamflows, the number of bootstrap samples is 1,000 and the block length is 3.

Data source

The data for the modified river streamflows of the Columbia river basin are available at https://www.bpa.gov/p/Power-Products/Historical- Streamflow-Data/Pages/Historical-Streamflow-Data.aspx

Cross validation

Table 3 reports the grids of the cross-validations performed to choose the values of the tuning parameters that determine the shrinkage for the RAR(1)-X and BAR(1)-X models. The grids are the same for both applications. The cross-validations are performed only on the first estimation window of the sample. It might be more at the advantage of both methods to renew the cross-validation for each new window of estimation, but this would increase the computation time considerably.

Table 3: Grids for the cross-validations

RAR(1)-X	d_0	0.2 to 0.55 by steps of 0.025
	λ_d^{-1}	0.01 to 0.05 by steps of 0.01
	λ_a^{-1}	0.01 to 0.05 by steps of 0.01
	λ_S^2	0 to 5,000 by steps of 1,000
BAR(1)-X	d_0	0.2 to 0.55 by by steps of 0.05
	s_d	0.01 to 0.05 by steps of 0.01
	s_a	0.01 to 0.05 by steps of 0.01
	h_0	0 to 5,000 by steps of 1,000

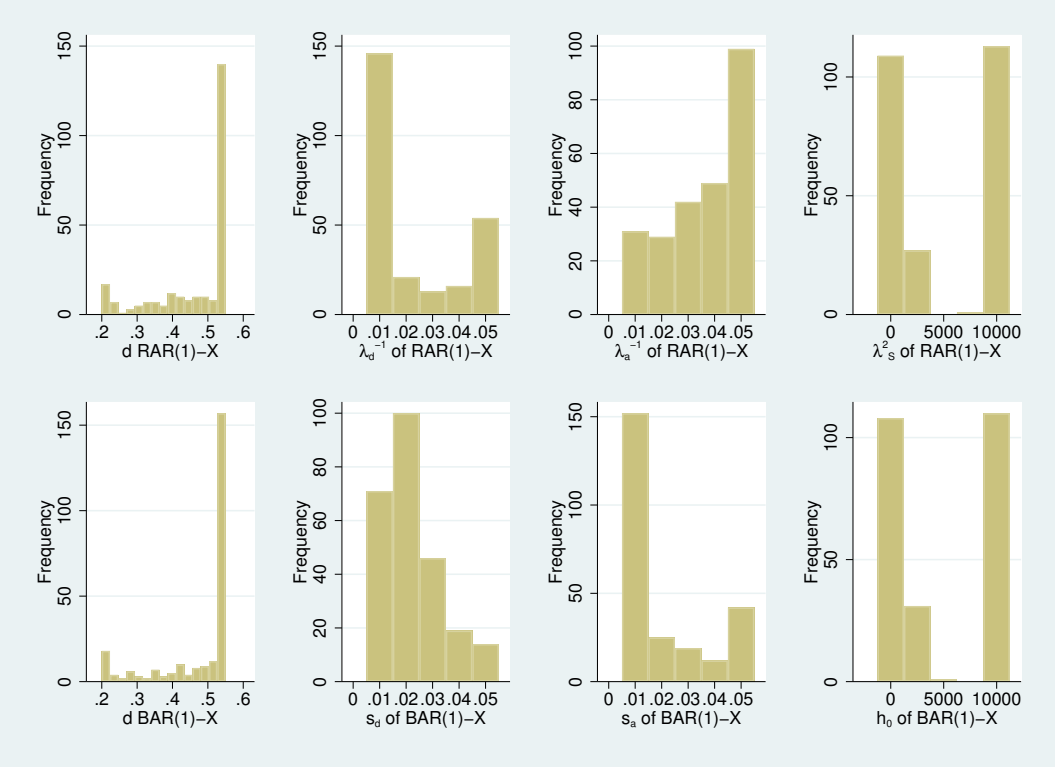

Figure 9: Histogram of the four tuning parameters estimated by cross-validation for the first application (i.e., log(MedRV)) on the first sample of 1,000 observations.

Figures 9 and 10 provide the histograms of the values obtained by the cross-validations, for RAR(1)-X and BAR(1)-X. The ordinates show the number of series, for example d_0 is equal to 0.55 for a bit less than 150 series (out of 250) for RAR and a bit more than 150 for BAR in the first application. In the second application, the cross-validation procedure chooses $d_0 = 0.55$ for about half of the series, and the value 0.2 for about 25 percent in the case of RAR (40 in the case of BAR).

The parameters $1/\lambda_d$ of RAR and s_d in BAR are selected at the lowest values of the grid (0.01 or 0.02) for about two-thirds of the series in the first application. In the second application, $1/\lambda_d$ is selected in equal proportions at the boundaries of the grid range (0.01 and 0.05), whereas s_d is selected mainly at the end of the range. The parameters $1/\lambda_a$ of RAR and s_a in BAR are selected differently between RAR and BAR and between applications 1 and 2.

The additional shrinkage of the sum of the coefficients toward 1 by the parameter λ_S^2 (RAR) or the equivalent parameter h_0 (BAR) is effective for around 120 series (about 48 percent) in the first application, but for very few series in the second application.

These results illustrate the flexibility of the cross-validation procedure.

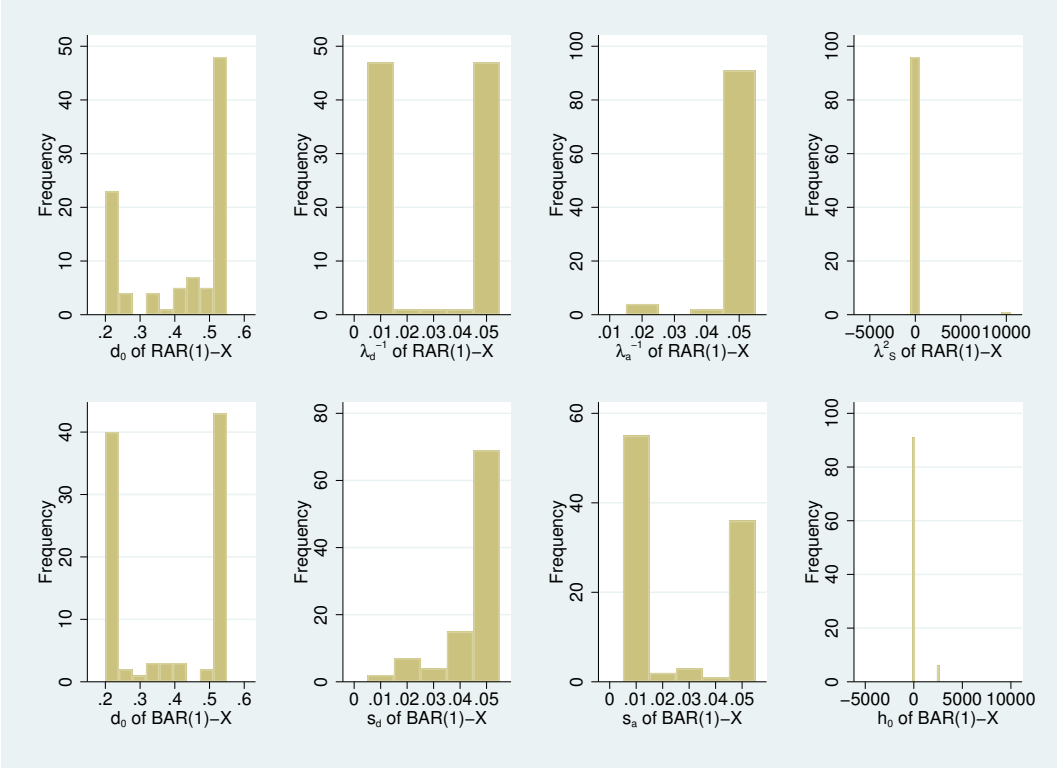

Figure 10: Histograms of the four tuning parameters estimated by cross-validation for the second application (i.e., river streamflows) on the first sample of 400 observations.

References

Abadir, K. M. and G. Talmain (2002). Aggregation, persistence and volatility in a macro model. *Review of Economic Studies 69*(4), 749–79.

Andersen, T., D. Dobrev, and E. Schaumburg (2012). Jump robust volatility estimation using nearest neighbor truncation. *Journal of Econometrics 169*(1), 75–93.

Anderson, H. M. and F. Vahid (2007). Forecasting the volatility of Australian stock returns: Do common factors help? *Journal of Business & Economic Statistics 25*(1), 76–90.

Baillie, R., T. Bollerslev, and H. Mikkelsen (1996). Fractionally integrated generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics 74*, 3–30.

Baillie, R. T. (1996). Long memory processes and fractional integration in econometrics. *Journal of Econometrics 73*, 5–59.

Bauwens, L., M. Lubrano, and J.-F. Richard (1999). *Bayesian Inference in Dynamic Econometric Models*. Oxford University Press.

Beran, J. (1992). Statistical methods for data with long-range dependence. *Statistical Science 7*(4), 404–416.

Breidt, F. J., N. Crato, and P. de Lima (1998). The detection and estimation of long memory in stochastic volatility. *Journal of Econometrics 83*(1-2), 325–348.

Chen, X., L. P. Hansen, and M. Carrasco (2010). Nonlinearity and temporal dependence. *Journal of Econometrics 155*(2), 155–169.

Chevillon, G., A. Hecq, and S. Laurent (2018). Generating univariate fractional integration within a large var (1). *Journal of Econometrics 204*(1), 54–65.

Chevillon, G. and D. F. Hendry (2005). Non-parametric direct multi-step estimation for forecasting economic processes. *International Journal of Forecasting 21*(2), 201–218.

Chevillon, G. and S. Mavroeidis (2017). Learning can generate long memory. *Journal of Econometrics 198*, 1–9.

Chevillon, G. and S. Mavroeidis (2018). Perpetual learning and apparent long memory. *Journal of Economic Dynamics and Control 90*, 343–365.

Comte, F. and E. Renault (1998). Long memory in continuous-time stochastic volatility models. *Mathematical Finance 8*, 291–323.

Corsi, F. (2009). A simple approximate long-memory model of realized volatility. *Journal of Financial Econometrics 7*(2), 174–196.

Cox, D. R. and M. W. H. Townsend (1947). The use of the correlogram in measuring yarn irregularities. *Proceedings of the Royal Society of Edinburgh, Section A 63*, 290–311.

Diebold, F. X. and A. Inoue (2001). Long memory and regime switching. *Journal of Econometrics 105*(1), 131–159.

Diebold, F. X. and K. Yılmaz (2009). Measuring financial asset return and volatility spillovers, with application to global equity markets. *The Economic Journal 119*, 158–171.

Diebold, F. X. and K. Yılmaz (2014). On the network topology of variance decompositions: Measuring the connectedness of financial firms. *Journal of Econometrics 182*, 119–134.

Gourieroux, C. and J. Jasiak (2001). Memory and infrequent breaks. *Economics Letters 70*, 29–41.

Granger, C. W. J. (1966). The typical spectral shape of an economic variable. *Econometrica 34*, 150–161.

Granger, C. W. J. (1980). Long memory relationships and the aggregation of dynamic models. *Journal of Econometrics 14*(2), 227–238.

Granger, C. W. J. and R. Joyeux (1980). An introduction to long-memory time series models and fractional differencing. *Journal of time series analysis 1*(1), 15–29.

Hansen, P., A. Lunde, and J. Nason (2011). The model confidence set. *Econometrica 79*, 453–497.

Hurst, H. E. (1951). Long-term storage capacity of reservoirs. *Transactions of the American society of civil engineers 116*(1), 770–799.

Miller, J. I. and J. Y. Park (2010). Nonlinearity, nonstationarity, and thick tails: How they interact to generate persistence in memory. *Journal of Econometrics 155*(1), 83 – 89.

Montanari, A., R. Rosso, and M. S. Taqqu (1997). Fractionally differenced arima models applied to hydrologic time series: Identification, estimation, and simulation. *Water resources research 33*(5), 1035–1044.

Nelson, C. R. and C. R. Plosser (1982). Trends and random walks in macroeconmic time series: some evidence and implications. *Journal of Monetary Economics 10*(2), 139–162.

Ooms, M. and P. H. Franses (2001). A seasonal periodic long memory model for monthly river flows. *Environmental Modelling & Software 16*(6), 559–569. Economics and Environmental Modelling.

Papacharalampous, G., H. Tyralis, and D. Koutsoyiannis (2019). Comparison of stochastic and machine learning methods for multi-step ahead forecasting of hydrological processes. *Stochastic Environmental Research and Risk Assessment 33*, 481–514.

Perron, P. and Z. Qu (2010). Long-memory and level shifts in the volatility of stock market return indices. *Journal of Business & Economic Statistics 28*(2), 275–290.

Rao, A. R. and D. Bhattacharya (1999). Hypothesis testing for long-term memory in hydrologic series. *Journal of Hydrology 216*(3-4), 183–196.

Robinson, P. M. and P. Zaffaroni (1998). Nonlinear time series with long memory: a model for stochastic volatility. *Journal of Statistical Planning and Inference 68*(2), 359–371.

Schennach, S. M. (2018). Long memory via networking. *Econometrica 86*(6), 2221–2248.

Schorfheide, F. (2005). VAR forecasting under misspecification. *Journal of Econometrics 128*(1), 99–136.

Smith, H. F. (1938). An empirical law describing heterogeneity in the yields of agricultural crops. *The Journal of Agricultural Science 28*(01), 1–23.

Wang, W., P. van Gelder, and J. Vrijling (2005). Long-memory in streamflow processes of the Yellow river. In *International conference on water economics, statistics and finance, Rethymno, Crete*, pp. 481–490. University of Crete.