A note on forecasting demand using the multivariate exponential smoothing framework

Federico Poloni* and Giacomo Sbrana‡

Abstract

Simple exponential smoothing is widely used in forecasting economic time series. This is because it is quick to compute and it generally delivers accurate forecasts. On the other hand, its multivariate version has received little attention due to the complications arising with the estimation. Indeed, standard multivariate maximum likelihood methods are affected by numerical convergence issues and bad complexity, growing with the dimensionality of the model. In this paper, we introduce a new estimation strategy for multivariate exponential smoothing, based on aggregating its observations into scalar models and estimating them. The original high-dimensional maximum likelihood problem is broken down into several univariate ones, which are easier to solve. Contrary to the multivariate maximum likelihood approach, the suggested algorithm does not suffer heavily from the dimensionality of the model. The method can be used for time series forecasting. In addition, simulation results show that our approach performs at least as well as a maximum likelihood estimator on the underlying VMA(1) representation, at least in our test problems.

Keywords: Multivariate Exponential Smoothing, EWMA, Forecasting demand,

1 Introduction

Simple exponential smoothing represents an important benchmark model when forecasting the demand for goods and services. The most attractive feature of this model is its ease of computation. Unfortunately, the same feature does not hold for its multivariate version due to the complications arising with the estimation. Indeed, this represents an obstacle for practitioners, discouraging the employment of this model in empirical analysis. This paper addresses this relevant issue by providing a feasible and accurate estimation method for a multivariate exponential smoothing model.

*Dipartimento di Informatica, Università di Pisa. Largo Pontecorvo 3, 56127 Pisa, Italy. E-mail fpoloni@di.unipi.it

†Corresponding author Tel.: +33 232824673.

‡Neoma Business School. 1, rue du Maréchal Juin, 76130 Mont-Saint-Aignan, France. E-mail gsb@neoma-bs.fr

We focus on the following state-space representation of an unrestricted multivariate simple exponential smoothing model (See [Harvey, 1991, Chapter 8])

$$\begin{aligned} y_t &= \mu_t + \epsilon_t, \\ \mu_t &= \mu_{t-1} + \eta_t, \end{aligned} \tag{1}$$

with $y_t, \mu_t, \epsilon_t, \eta_t \in \mathbb{R}^N$. The noises η_t and ϵ_t characterizing the system are independent and identically distributed with expected value equal to zero and

$$\operatorname{cov} \begin{bmatrix} \epsilon_t \\ \eta_t \end{bmatrix} = \begin{bmatrix} \Sigma_\epsilon & 0 \\ 0 & \Sigma_\eta \end{bmatrix} \tag{2}$$

where $\Sigma_\epsilon > 0$, $\Sigma_\eta > 0$ and 0 are $N \times N$ matrices. The nonstationary system (1) is known as the *structural process*, and the covariances as in (2) are called *structural parameters*. The system can be reparametrized as a first order integrated vector moving average process (i.e. integrated VMA(1)), the so-called *reduced form*, using the Wold representation theorem

$$z_t := y_t - y_{t-1} = \eta_t + \epsilon_t - \epsilon_{t-1} = u_t - \Theta u_{t-1}, \quad \mathbb{E}\left[u_t u_t^T\right] = \Sigma_u, \tag{3}$$

for a suitable $\Theta, \Sigma > 0 \in \mathbb{R}^{N \times N}$, and an innovation process u_t which is uncorrelated, but not in general independent. The parameters can be chosen so that (3) is *invertible*, i.e., all the eigenvalues of Θ have modulus smaller than 1. This version can be recast in the more familiar exponentially weighted moving average form (EWMA)

$$\hat{y}_t = (I - \Theta) y_{t-1} + \Theta \hat{y}_{t-1}, \qquad \text{for } t = 1, 2, \ldots, T, \tag{4}$$

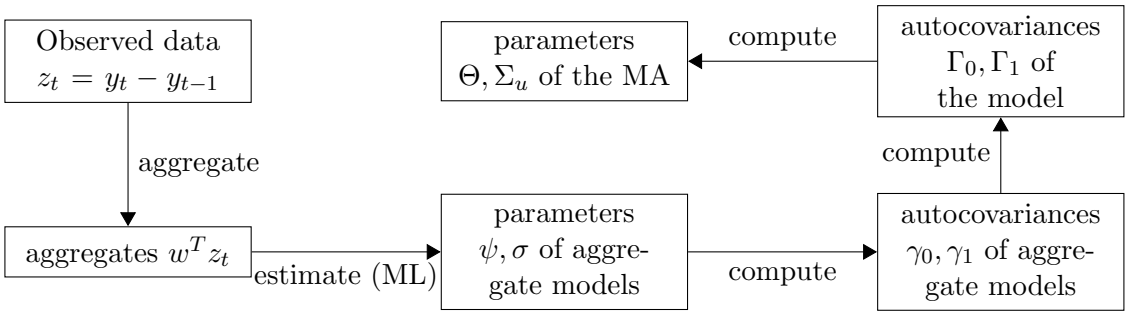
where $\hat{y}_t$ denotes the forecast of y_t.

Our method is based on the relation between the original model and several scalar aggregates of the form $x_t = w_t^T z_t$, for suitable $w \in \mathbb{R}^N$. It is a consequence of Wold's decomposition theorem [Lütkepohl, 2005, Section 2.1.3] that each of these models is a MA(1), i.e.,

$$x_t = v_t - \psi v_{t-1}, \quad \mathbb{E}\left[v_t^2\right] = \sigma.$$

(Notice that we use here σ to denote a variance, rather than a standard deviation, for notational consistency with Σ for multivariate processes.)

Visually, we can represent the estimation procedure using Figure 1:



Namely, we aggregate the model using several vectors w, estimate each of the univariate (scalar) models $x_t = w_t z_t$ as a MA(1), and then make some algebraic computations to derive the parameters Θ and Σ_u from them.

In order to make these closed-form computations possible, we need to derive explicit, closed-form relations that allow us to

- express the parameters Θ, Σ_u as a function of the autocovariances $\Gamma_0 := \mathbb{E}\left[z_t z_t^T\right]$ and $\Gamma_1 := \mathbb{E}\left[z_t z_{t-1}^T\right]$. This step is described in Section 3.
- express Γ_0 and Γ_1 as a function of the autocovariances of the aggregate models. This step, together with a description of the whole estimation procedure, is shown in Section 4.

Asymptotic consistency and normality of the resulting estimator are proved in Section 5.

In contrast, a maximum likelihood (ML) estimator of the VMA model (3) would follow directly the "missing arrow" on our diagram between the observed data z_t and the parameters Θ, Σ_u. What we do instead essentially trades off one N-dimensional maximum likelihood procedure for several univariate ones. Since the ML estimator is affected by numerical convergence issues and bad complexity, growing with the dimensionality of the model [Kascha, 2012], estimating many small models rather than a large one is computationally favorable. In Section 6 we compare the performance of ML with those of our suggested estimator, called *META (Moment Estimation Through Aggregation)*. Simulation results show that the suggested approach is not only very simple and fast but is also remarkably efficient having performance that is as good as that of the standard multivariate maximum likelihood approach.

2 Literature review

As noted by De Gooijer and Hyndman [2006], "*There has been remarkably little work in developing multivariate versions of the exponential smoothing methods for forecasting.*" We argue that this is probably due to the difficulties of estimating parameters in large-dimensional system. Our framework, also known as exponentially weighted moving average (EWMA), has a long tradition in forecasting time series (Muth [1960]). Moreover, the EWMA belongs to the more general exponential smoothing family (see Gardner Jr. [2006] and Holt [2004]). Despite its simplicity, this family represents a valid candidate in forecasting demand (see for example Dekker et al. [2004], Fliedner and Lawrence [1995], Fliedner [1999], Makridakis and Hibon [2000], Moon et al. [2012], Moon et al. [2013]).

More recently, the multivariate version of the EWMA model has been considered as production planning framework when forecasting aggregate demand. To overcome the estimation difficulties in the multivariate case, two strategies have been suggested, the so-called *top-down* and *bottom-up* approaches (see for example Lütkepohl [1987], Widiarta et al. [2009] and Sbrana and Silvestrini [2013]). In detail, let w^T be a fixed weight vector, and suppose that we are interested in forecasting the aggregated process $x_t = w^T z_t$. For

instance, $w^T = \begin{bmatrix} 1/N & 1/N & \dots & 1/N \end{bmatrix}$ means that we are interested in the arithmetic mean of the observed variables. Then,

- The *top-down* approach consists in constructing directly $x_t = w^T z_t$ and forecasting this aggregate series applying a (scalar) estimator to $\{x_t\}_{t=1,2,\dots,N}$. Note that this loses information about the original process, since we do not use the single components of z_t but only an aggregate function of them; hence less accuracy is to be expected.

- The *bottom-up* approach consists in applying a scalar estimator to each of the N time series $\{(z_t)_1\}_{t=1,2,\dots,T}, \{(z_t)_2\}_{t=1,2,\dots,T}, \dots \{(z_t)_N\}_{t=1,2,\dots,T}$, and forecasting each of them individually to obtain a forecast of the aggregated process $x_t = w^T z_t$. Again, this method ignores the interdependence among the variables.

There is a vast literature comparing top-down and bottom-up approaches when forecasting aggregate demand. Without attempting to survey all the contributions on this topic, we refer to Fliedner and Lawrence [1995], Fliedner [1999], Weatherford et al. [2001], Dekker et al. [2004], Zotteri et al. [2005], Zotteri and Kalchschmidt [2007], Widiarta et al. [2009], Chen and Blue [2010]. More recently Moon et al. [2012], Moon et al. [2013] consider in details alternative forecasting methods, such as the simple exponential smoothing, for predicting the demand for spare parts in the South Korean Navy.

A necessary condition to compare top-down and bottom-up approaches is the knowledge of the parameters of the multivariate demand planning framework. Indeed, once data are available, practitioners are faced with the challenge to estimate the parameters of the system whose dimension might be large. Therefore, a relevant gap left by Sbrana and Silvestrini [2013] is that they do not provide any indication on how to derive the parameters of the framework using the available data (i.e. y_t). Indeed, quoting their conclusions: *"this paper contains useful results assuming full knowledge of the parameters of the multivariate exponential smoothing. We are aware that this represents an ideal situation since, in empirical analysis, practitioners do not have such information and misspecification issues do usually arise [...]"*. This note fills this important empirical gap by providing an efficient and fast estimation procedure for the exponentially weighted moving average model, based on the same aggregation techniques used in their paper.

3 Closed-form results

It is easy to see that the autocovariances of the process z_t, expressed with both parametrizations (1)–(2) and (3), are given by

$$\begin{aligned} \Gamma_0 &:= \mathbb{E}\left[z_t z_t^T\right] = \Sigma_u + \Theta\Sigma_u\Theta^T = \Sigma_\eta + 2\Sigma_\epsilon, \\ \Gamma_1 &:= \mathbb{E}\left[z_t z_{t-1}^T\right] = -\Theta\Sigma_u = -\Sigma_\epsilon. \end{aligned} \tag{5}$$

It is important in the following that Γ_1 is a symmetric matrix. Hence it is easy to derive the following relations that express the structural parameters as a function of the reduced

ones:

$$\Sigma_\epsilon = \Theta\Sigma_u, \quad \Sigma_\eta = \Sigma_u + \Theta\Sigma_u\Theta^T - 2\Sigma_\epsilon.$$

The inverse relationship, i.e., how to construct the reduced parameters Θ, Σ_u in terms of the structural ones $\Sigma_\eta, \Sigma_\epsilon$, is less obvious; we present it in the following result.

Proposition 1. *Consider the model* (1)–(2) *and its reparametrization* (3)*, and let* $Q := \Sigma_\eta\Sigma_\epsilon^{-1}$*. Then, the following relation holds*

$$\begin{aligned} \Theta &= \tfrac{1}{2}\left(Q + 2I - \left(Q^2 + 4Q\right)^{\frac{1}{2}}\right), \\ \Sigma_u &= \Theta^{-1}\Sigma_\epsilon. \end{aligned} \tag{6}$$

The matrix square root in this expression is well-defined since $Q^2 + 4Q$ *is diagonalizable with all positive eigenvalues.*

Proof. Note that Γ_0 can be expressed as

$$\Gamma_0 = \Sigma_u + \Gamma_1\Sigma_u^{-1}\Gamma_1, \tag{7}$$

since $\Gamma_1 = -\Theta\Sigma_u = -\Sigma_u\Theta^T$. Post-multiplying (7) by Σ_u^{-1}, we have

$$-\Gamma_0\Gamma_1^{-1}\Theta = \Gamma_0\Sigma_u^{-1} = I + \Gamma_1\Sigma_u^{-1}\Gamma_1\Sigma_u^{-1} = I + \Theta^2.$$

Therefore Θ satisfies the quadratic matrix equation

$$\Theta^2 + \Gamma_0\Gamma_1^{-1}\Theta + I = 0, \tag{8}$$

with $\Gamma_0\Gamma_1^{-1} = (\Sigma_\eta + 2\Sigma_\epsilon)(-\Sigma_\epsilon)^{-1} = -Q - 2I$. The matrix Q is always diagonalizable with positive eigenvalues, since it is the product of two positive-definite matrices [Horn and Johnson, 1990, Theorem 7.6.3]. Hence we can set $Q = PDP^{-1}$, $D = \mathrm{diag}(d_1, d_2, \ldots, d_N)$, with $d_i > 0$ for each $i = 1, 2, \ldots, N$. Pre- and post-multiplying (8) by P^{-1} and P, we get

$$\tilde{\Theta}^2 - (D + 2I)\tilde{\Theta} + I = 0, \quad \tilde{\Theta} := P^{-1}\Theta P.$$

The solutions of this matrix equation are given by diagonal matrices $\tilde{\Theta} = \mathrm{diag}(g_1, g_2, \ldots, g_N)$, where

$$g_i^2 - (d_i + 2)g_i + 1 = 0, \quad i = 1, 2, \ldots, N.$$

The usual formula for the quadratic solution gives

$$g_i = \frac{d_i + 2 \pm \sqrt{d_i^2 + 4d_i}}{2}. \tag{9}$$

Note that g_i are the eigenvalues of $\tilde{\Theta}$, and hence of Θ. Since $d_i < \sqrt{d_i^2 + 4d_i} < d_i + 2$ whenever $d_i > 0$, we have $0 < g_i < 1$ if we choose the minus sign and $g_i > 1$ if we choose the plus sign. Hence we choose the minus sign to obtain invertibility of the resulting system (3).

Putting back together the matrices, we get

$$\Theta = P\frac{1}{2}\left(D + 2I - (D^2 + 4D)^{1/2}\right)P^{-1} = \frac{1}{2}\left(Q + 2I - (Q^2 + 4Q)^{1/2}\right).$$

The matrix square root is well defined since $Q^2 + 4Q$ has positive eigenvalues $d_i^2 + 4d_i$, $i = 1, 2, \ldots, N$.

Finally, the second equation in (6) follows from the second one in (5), since we have already observed that $g_i < 0$ and thus Θ is nonsingular. □

Remark 2. *The results as in (6) are the multivariate extension of the univariate results (see for example [Muth, 1960] and [Harvey, 1991, p. 68]) with Q representing a "signal to noise" matrix ratio. In general, Q is not a symmetric matrix and therefore neither Θ is.*

The expression for Θ in (6) is useful for forecasting the system (4). Indeed, using the lag operator L (such that $Ly_t = y_{t-1}$) we can write the optimal linear forecasting for (4) as

$$y_{t+1}^F = (I - \Theta)(I - \Theta L)^{-1} y_t = (I - \Theta)\sum_{j=0}^{\infty} \Theta^j y_{t-j}. \tag{10}$$

Corollary 3. *The reduced form parameters can be expressed in terms of the autocorrelations of z_t as*

$$\begin{aligned} \Theta &= -\frac{1}{2}\left(\Gamma_0\Gamma_1^{-1} + \left(\Gamma_0\Gamma_1^{-1}\Gamma_0\Gamma_1^{-1} - 4I\right)^{\frac{1}{2}}\right), \\ \Sigma_u &= 2\left(\Gamma_0\Gamma_1^{-1} + \left(\Gamma_0\Gamma_1^{-1}\Gamma_0\Gamma_1^{-1} - 4I\right)^{\frac{1}{2}}\right)^{-1}\Gamma_1. \end{aligned} \tag{11}$$

Therefore, Γ_0, Γ_1 is the only information needed to obtain Θ and Σ_u. The reader might be tempted to use this result as an estimator, computing sample covariances $\hat{\Gamma}_0 = \frac{1}{T}\sum_{t=1}^{T} z_t z_t^T$, $\hat{\Gamma}_1 = \frac{1}{T}\sum_{t=1}^{T-1} z_t z_{t-1}^T$ and substituting them into (11). In empirical analysis, however, these sample covariances might not be accurate enough. To solve this issue, in the next section we provide a method to derive these moments more accurately.

4 Moment estimation through aggregation (*META*)

Consider a generic multivariate MA(1) process

$$z_t = u_t - \Theta u_{t-1}, \quad \mathbb{E}\left[u_t u_t^T\right] = \Sigma_u, \tag{12}$$

and define its autocovariance matrices $\Gamma_k := \mathbb{E}\left[z_t z_{t-k}^T\right]$; due to the structure of the process, $\Gamma_k = 0$ for $|k| > 1$, and $\Gamma_0 = \Gamma_0^T$.

We are interested in aggregate processes, that is, scalar processes of the form $x_t := w^T z_t$, for some vector $w \in \mathbb{R}^n$. This form includes in particular the components $(z_t)_1, (z_t)_2, \ldots, (z_t)_N$ of the vector process z_t, which are obtained by setting $w = e_i$, for $j = 1, 2, \ldots, N$, where e_i is the i-th vector of the canonical basis, that is, the i-th column of I_N.

It turns out that if $\Gamma_1 = \Gamma_1^T$ (as is the case in our EWMA setting, due to (5)), then we can recover these covariances by knowing those of some special aggregate processes.

Lemma 4. *Let z_t be a VMA(1) process* (12)*, and suppose that $\Gamma_1 = \Gamma_1^T$. Given a vector $w \in \mathbb{R}^N$, $w \neq 0$, define the aggregate $x_t^{(w)} := w^T z_t$, and let $\gamma_k^{(w)}$ be its covariances. Then, the entries of Γ_k are given by*

$$(\Gamma_k)_{i,j} = \begin{cases} \gamma_k^{(e_i)} & i = j, \\ \frac{1}{2}\left(\gamma_k^{(e_i+e_j)} - \gamma_k^{(e_i)} - \gamma_k^{(e_j)}\right) & i \neq j. \end{cases} \tag{13}$$

In particular, they are uniquely determined given the covariances of the $\frac{N(N+1)}{2}$ scalar processes constructed with vectors $w \in \mathcal{W}$,

$$\mathcal{W} := \{e_i \colon 1 \leq i \leq N\} \cup \{e_i + e_j \colon 1 \leq i < j \leq N\}.$$

Proof. Note that $\gamma_k^{(w)} = \mathbb{E}\left[w^T z_t z_{t-k}^T w\right] = w^T \Gamma_k w$. Hence, $\gamma_k^{(e_i)} = (\Gamma_k)_{ii}$, and $\gamma_k^{(e_i+e_j)} = (\Gamma_k)_{ii} + (\Gamma_k)_{ij} + (\Gamma_k)_{ji} + (\Gamma_k)_{jj} = (\Gamma_k)_{ii} + 2(\Gamma_k)_{ij} + (\Gamma_k)_{jj}$. □

Each aggregate process can be reparametrized as a scalar MA(1) itself (see [Lütkepohl, 1987]); hence, one can write

$$x_t^{(w)} = v_t^{(w)} - \psi^{(w)} v_{t-1}^{(w)}, \quad \mathbb{E}\left[(v_t^{(w)})^2\right] = \sigma^{(w)}, \tag{14}$$

for suitable white noise sequences $v_t^{(w)}$. Note that, although each $v_t^{(w)}$ is a white noise sequence on its own, two generic entries $v_{t_1}^{(w_1)}$ and $v_{t_2}^{(w_2)}$, for given t_1, t_2 and $w_1 \neq w_2$, might be correlated.

One can use this representation to express the autocovariances as a function of these parameters:

$$\begin{aligned} \gamma_0^{(w)} &= (1 + (\psi^{(w)})^2)\sigma^{(w)}, \\ \gamma_1^{(w)} &= -\psi^{(w)}\sigma^{(w)}. \end{aligned} \tag{15}$$

This approach suggests an estimation procedure as follows. Given T observations of the process (12):

1. For each of the $N(N+1)/2$ vectors $w \in \mathcal{W}$, construct the aggregate data $x_t^{(w)} = w^T z_t$, and estimate the MA(1) model (14), obtaining $\hat{\psi}^{(w)}$ and $\hat{\sigma}^(w)$.
2. For each w, construct $\hat{\gamma}_0^{(w)}$ and $\hat{\gamma}_1^{(w)}$ using the formulas (15).
3. Recover estimates $\hat{\Gamma}_0$ and $\hat{\Gamma}_1$ using (13).
4. Recover estimates $\hat{\Theta}$ and $\hat{\Sigma}_u$ using (11).

The advantage of steps 1–3 of this procedure with respect to an estimator based on the sample moments $\frac{1}{T}\sum_{t=1}^T z_t z_{t-k}^T$ is that a maximum likelihood estimator as above yields more accurate values for the asymptotic moments.

For the sake of simplicity, in order to provide intuition to the reader, we give an example using a bivariate model. Consider the following system with two variables

$$x_{1t} = u_{1t} + \phi_{11}u_{1t-1} + \phi_{12}u_{2t-1},$$
$$x_{2t} = u_{2t} + \phi_{21}u_{1t-1} + \phi_{22}u_{2t-1}.$$

The previous model can be reparametrized equation-by-equation as

$$x_{1t} = \upsilon_{1t} + \psi_1\upsilon_{1t-1},$$
$$x_{2t} = \upsilon_{2t} + \psi_2\upsilon_{2t-1}$$

with $\mathbb{E}\left[\upsilon_{it}^2\right] = \sigma_i$. Finally consider the MA(1) process derived from the simple aggregation of the two components

$$x_t = x_{1t} + x_{2t} = a_t + \alpha a_{1t-1}$$

with $\mathbb{E}\left[a_t^2\right] = \sigma_a$. Using the results above, we can now rewrite Γ_0 and Γ_1 as function of the parameters of the aggregate models as follows

$$\Gamma_0 = \begin{bmatrix} (1+\psi_1^2)\sigma_1 & \frac{1}{2}[(1+\alpha^2)\sigma_a - (1+\psi_1^2)\sigma_1 - (1+\psi_2^2)\sigma_2] \\ \frac{1}{2}[(1+\alpha^2)\sigma_a - (1+\psi_1^2)\sigma_1 - (1+\psi_2^2)\sigma_2] & (1+\psi_2^2)\sigma_2 \end{bmatrix},$$
$$\Gamma_1 = \begin{bmatrix} \psi_1\sigma_1 & \frac{1}{2}[\alpha\sigma_a - \psi_1\sigma_1 - \psi_2\sigma_2] \\ \frac{1}{2}[\alpha\sigma_a - \psi_1\sigma_1 - \psi_2\sigma_2] & \psi_2\sigma_2 \end{bmatrix}.$$

5 Asymptotic properties

As a first result, we prove that the aggregate MA processes that we estimate are well-behaved.

Lemma 5. *Suppose that the process* (12) *is invertible. Then, for each* $w \in \mathbb{R}^N$ *with* $w \neq 0$*, the process* (14) *is invertible, and* $\sigma^{(w)} > 0$*.*

Proof. Invertibility of (12) means that its autocovariance generating function [Brockwell and Davis, 2006, §3.5] $\Gamma(z)$ is nonsingular for each z on the unit circle, i.e.,

$$\Gamma(z) > 0 \quad \text{if } |z| = 1.$$

The autocovariance generating function of (14) is

$$(1 - z\psi^{(w)})\sigma^{(w)}(1 - z^{-1}\psi^{(w)}) = \gamma^{(w)}(z) = w^T\Gamma(z)w.$$

Since $\Gamma(z)$ is a positive-definite matrix for $|z| = 1$, we also have that $w^T\Gamma(z)w > 0$. Hence $(1 - z\psi^{(w)})\sigma^{(w)}(1 - z^{-1}\psi^{(w)}) > 0$ whenever $|z| = 1$, and this implies that $\sigma^{(w)} > 0$ and that there is an invertible representation with $|\psi^{(w)}| < 1$ for (14). □

Therefore in the following we assume without further mention that $|\psi^{(w)}| < 1$.

The (quasi)-maximum likelihood estimator on the representation (14), using zero initial values for simplicity, is given by (dropping the $\cdot^{(w)}$ superscript for ease of notation)

$$(\hat{\psi}, \hat{\sigma}) := \arg\min_{(\tilde{\psi},\tilde{\sigma})} \sum_{t=1}^{T} \ell_t(\tilde{\psi}, \tilde{\sigma}),$$

with the unconditional negative log-likelihood function ℓ_t given for each pair of reals $\tilde{\psi}, \tilde{\sigma}$ by

$$\ell_t(\tilde{\psi}, \tilde{\sigma}) := \frac{1}{2}\log\tilde{\sigma} + \frac{\tilde{v}_t^2}{2\tilde{\sigma}}, \quad \tilde{v}_t := \sum_{k=0}^{t-1} \tilde{\psi}^k x_{t-k}. \tag{16}$$

The $\tilde{v}_t$ satisfy the linear recurrence $\tilde{v}_1 = x_1$, $\tilde{v}_t = x_t + \tilde{\psi}\tilde{v}_{t-1}$, and are a function of $\tilde{\psi}$ and of the observations. We set for brevity $\tilde{v}'_t := \frac{\partial}{\partial\tilde{\psi}}\tilde{v}_t$. Notice that $\tilde{v}'_t$ is a linear function of $\tilde{v}_1, \ldots, \tilde{v}_{t-1}$, as can be proved by induction using the relation

$$\tilde{v}'_t = \tilde{v}_{t-1} + \tilde{\psi}\tilde{v}'_{t-1}.$$

Moreover, when $\tilde{\psi} = \psi$ (the correct value), then $\tilde{v}_t = v_t$. We first evaluate the Hessian of the likelihood at the exact system parameters (ψ, σ): by ergodicity,

$$\begin{aligned}\frac{1}{T}\sum \nabla^2 \ell_t(\psi,\sigma) &\to \mathbb{E}\left[\begin{bmatrix} \frac{\partial^2}{\partial\tilde{\psi}^2}\ell_t(\psi,\sigma) & \frac{\partial^2}{\partial\tilde{\psi}\partial\tilde{\sigma}}\ell_t(\psi,\sigma) \\ \frac{\partial^2}{\partial\tilde{\psi}\partial\tilde{\sigma}}\ell_t(\psi,\sigma) & \frac{\partial^2}{\partial\tilde{\sigma}^2}\ell_t(\psi,\sigma)\end{bmatrix}\right] \\ &= \begin{bmatrix} \mathbb{E}\left[\frac{1}{\sigma}\left({v'_t}^2 + \frac{\partial^2 v_t}{\partial\psi^2}v_t\right)\right] & \mathbb{E}\left[-\frac{1}{\sigma^2}v'_t v_t\right] \\ \mathbb{E}\left[-\frac{1}{\sigma^2}v'_t v_t\right] & \mathbb{E}\left[\frac{1}{2\sigma^2}\left(\frac{2v_t^2}{\sigma} - 1\right)\right]\end{bmatrix} = \begin{bmatrix} \frac{1}{1-\psi^2} & 0 \\ 0 & \frac{1}{2\sigma^2}\end{bmatrix}. \end{aligned} \tag{17}$$

In evaluating these expected values, we have used the following facts:

- $\mathbb{E}\left[v_t^2\right] = \sigma$.
- v'_t and $\frac{\partial^2 v_t}{\partial\tilde{\psi}^2}$ are uncorrelated from v_t, since they are linear functions of $v_1, v_2, \ldots, v_{t-1}$.
- $\mathbb{E}\left[{v'_t}^2\right] = \frac{\sigma}{1-\psi^2}$. The simplest way to prove this relation is through

$$\begin{aligned}\mathbb{E}\left[{v'_t}^2\right] = \mathbb{E}\left[\left(v_{t-1} + \psi v'_{t-1}\right)^2\right] &= \mathbb{E}\left[v_{t-1}^2 + 2v_{t-1}\psi v'_{t-1} + \psi^2 {v'_{t-1}}^2\right] \\ &= \sigma + 0 + \psi^2\mathbb{E}\left[{v'_{t-1}}^2\right],\end{aligned}$$

  and by stationarity $\mathbb{E}\left[{v'_t}^2\right] = \mathbb{E}\left[{v'_{t-1}}^2\right]$.

We continue by proving that the estimates of the scalar parameters $\hat{\psi}^{(w)}, \hat{\sigma}^{(w)}$ are consistent and asymptotically normal. The consistency part is easier, since we can consider each aggregated process independently.

Lemma 6. *Let the model* (12) *be stationary and ergodic, with* $\Sigma_u > 0$. *Then, the maximum likelihood estimators* $\hat{\psi}^{(w)}, \hat{\sigma}^{(w)}$ *are asymptotically consistent.*

Proof. Let us consider the generic combination $x_t = w^T z_t$, $t = 1, 2, \dots, T$; as stated above, this is a MA(1) process with weak (uncorrelated but not independent) white noise v_t.

We make use the general results on ML consistency in [Ling and McAleer, 2010, Theorem 1(a)]. Since the Hessian (17) is asymptotically nonsingular, the maximizing point is isolated. We have

$$\tilde{v}_t = \sum_{i\geq 0} \tilde{\psi}^i x_{t-i} = \sum_{i\geq 0} \tilde{\psi}^i (v_{t-i} - \psi v_{t-i-1}) = v_t + \sum_{i\geq 0} \tilde{\psi}^i (\tilde{\psi} - \psi) v_{t-i},$$

hence

$$\mathbb{E}\left[\tilde{v}_t^2\right] = \sigma \left(1 + \sum_{i\geq 0} \tilde{\psi}^{2i} (\tilde{\psi} - \psi)^2\right).$$

If we restrict the parameter set to a compact set with $\tilde{\psi} < 1$, the sum converges and thus $\sup_{\tilde{\psi}} \mathbb{E}\left[\tilde{v}_t^2\right] < \infty$. Hence the hypotheses in Ling and McAleer [2010] hold and each aggregated process is asymptotically consistent. □

Establishing asymptotic normality is more involved: since the $v_t^{(w)}$ are neither independent nor uncorrelated from each other, we cannot rely on the classical central limit results. We use instead a central limit result for weakly dependent sequences from Peligrad and Utev [2006], which we summarize and report as follows.

Theorem 7. *For an i.i.d. sequence of random variables* $(Y_i)_{i\in\mathbb{Z}}$, *denote by* $\mathcal{F}_a^b$ *the* σ*-field generated by* Y_t *with* $a \leq t \leq b$ *and define* $\xi_t = f(Y_t, Y_{t-1}, \dots)$, $t \in \mathbb{Z}$. *Assume that* $\mathbb{E}[\xi_0] = 0$, $\mathbb{E}\left[\xi_0^2\right] < \infty$, *and*

$$\sum_{t=1}^{\infty} \frac{1}{\sqrt{t}} \|\xi_0 - \mathbb{E}\left[\xi_0 \mid \mathcal{F}_{-t}^0\right]\|_2 < \infty. \tag{18}$$

Then,

$$\frac{1}{\sqrt{T}} \sum_{t=1}^{T} \xi_t \to N(0, \mathbb{E}\left[\xi_0^2\right]). \tag{19}$$

Indeed, Peligrad and Utev [2006] contains a stronger result on triangular sequences (Corollary 5); our statement (19) is a special case that can be obtained by setting

$$a_i = \begin{cases} 1 & i = 0 \\ 0 & \text{otherwise} \end{cases}$$

in the thesis of their Theorem 1, so that $b_n = \sqrt{n}$.

In the process (3), the i.i.d. variables are $Y_t = \begin{bmatrix} \epsilon_t \\ \eta_t \end{bmatrix}$, and we aim to prove that each of the $\nabla \ell_t(\psi, \sigma)$ can be chosen as a ξ_t that satisfies the above condition (18). We start with a couple of lemmas.

Lemma 8. *Consider the process* (3)*, with i.i.d. variables* $Y_t = \begin{bmatrix}\epsilon_t \\ \eta_t\end{bmatrix}$*. For a fixed* $w \in \mathbb{R}^N$*,* $w \neq 0$*, define* $v_t^{(w)}$ *and* $\psi^{(w)}$ *as above. Then, there are vectors* $g_k, h_k \in \mathbb{R}^{2N}$ *such that*

$$v_t^{(w)} = \sum_{k=0}^{\infty} g_k Y_{t-k}, \tag{20}$$

$$v'^{(w)}_t = \sum_{k=0}^{\infty} h_k Y_{t-k}, \tag{21}$$

with $\|g_k\| = O((\psi^{(w)})^k)$ *and* $\|h_k\| = O(k(\psi^{(w)})^k)$ *for* $k \to \infty$*.*

Proof. Let us drop the superscript (w) for clarity. Using the lag operator L, one has

$$x_t = w^T z_t = w^T(\eta_t + (I - L)\epsilon_t) = \begin{bmatrix} w^T & w^T \end{bmatrix} Y_t - L \begin{bmatrix} 0 & w^T \end{bmatrix} Y_t.$$

Hence

$$\begin{aligned} v_t = (1 - \psi L)^{-1} x_t = \sum_{k\geq 0} \psi^k L^k x_t &= \sum_{k\geq 0} \psi^k L^k \begin{bmatrix} w^T & w^T \end{bmatrix} Y_t - \sum_{k\geq 0} \psi^k L^{k+1} \begin{bmatrix} 0 & w^T \end{bmatrix} Y_t \\ &= \begin{bmatrix} w^T & w^T \end{bmatrix} Y_0 + \sum_{k\geq 0} \psi^k L^{k+1} \begin{bmatrix} \psi w^T & (\psi - 1) w^T \end{bmatrix} Y^T. \end{aligned}$$

Similarly, starting from

$$v'_t = \sum_{k\geq 1} k\psi^{k-1} L^k x_t,$$

one gets the other result. □

Lemma 9. *Consider the process* (3)*, with i.i.d. variables* $Y_t = \begin{bmatrix}\epsilon_t \\ \eta_t\end{bmatrix}$ *with finite fourth moment. For a fixed* $w \in \mathbb{R}^N$*,* $w \neq 0$*, define* $v_t^{(w)}$ *and* $\psi^{(w)}$ *as above. Then, condition* (18) *holds for both* $\xi_t = v'^{(w)}_t v_t^{(w)}$ *and* $\xi_t = (v_t^{(w)})^2 - \sigma^{(w)}$*.*

Proof. We may write

$$v_0 = \underbrace{\sum_{k=0}^{t} g_k Y_{t-k}}_{:=p_t} + \underbrace{\sum_{k>t} g_k Y_{t-k}}_{:=q_t}, \tag{22}$$

where p_t is a function in the σ-field $\mathcal{F}^0_{-t}$ and q_t is independent from it, and similarly

$$v'_0 = \underbrace{\sum_{k=0}^{t} h_k Y_{t-k}}_{:=r_t} + \underbrace{\sum_{k>t} h_k Y_{t-k}}_{:=s_t}. \tag{23}$$

One has

$$\begin{aligned}\mathbb{E}\left[v_0'v_0 \mid \mathcal{F}_{-t}^0\right] &= \mathbb{E}\left[(p_t+q_t)(r_t+s_t) \mid \mathcal{F}_{-t}^0\right] \\ &= p_t r_t + \underbrace{\mathbb{E}\left[q_t \mid \mathcal{F}_{-t}^0\right]}_{=0} r_t + p_t \underbrace{\mathbb{E}\left[s_t \mid \mathcal{F}_{-t}^0\right]}_{=0} + \mathbb{E}\left[q_t s_t \mid \mathcal{F}_{-t}^0\right] = p_t r_t + \mathbb{E}\left[q_t s_t \mid \mathcal{F}_{-t}^0\right],\end{aligned}$$

thus

$$\begin{aligned}\left\|v_0'v_0 - \mathbb{E}\left[v_0'v_0 \mid \mathcal{F}_{-t}^0\right]\right\|_2 &= \left\|q_t r_t + p_t s_t + q_t s_t + \mathbb{E}\left[q_t s_t \mid \mathcal{F}_{-t}^0\right]\right\|_2 \\ &\leq \|q_t\|_2 \|r_t\|_2 + \|p_t\|_2 \|s_t\|_2 + 2\|q_t\|_2 \|s_t\|_2.\end{aligned}$$

Since the decompositions (20), (21), (22), (23) are into independent (orthogonal) components, one can estimate

$$\begin{aligned}\|p_t\|_2 &\leq \|v_0\|_2, & \|q_t\|_2 &= O\left(\sum_{k>t}\|g_k\|\right) = O(\psi^t), \\ \|r_t\|_2 &\leq \left\|\left(\frac{\partial}{\partial\psi}v_0\right)\right\|, & \|s_t\|_2 &= O\left(\sum_{k>t}\|h_k\|\right) = O(t\psi^t).\end{aligned}$$

When estimating the two sums, we used the fact that $\sum_{t\geq 0}\psi^t = (1-\psi)^{-1} < \infty$ and $\sum_{t\geq 0} t\psi^t = \psi(1-\psi)^{-2} < \infty$. Putting everything together, we have proved that

$$\left\|v_0'v_0 - \mathbb{E}\left[v_0'v_0 \mid \mathcal{F}_{-t}^0\right]\right\|_2 = O(t\psi^t) \quad \text{for } t \to \infty.$$

Hence the sum in (18) converges (indeed, even without the $\frac{1}{\sqrt{t}}$ term), and the condition is verified.

A similar reasoning works for $v_t^2 - \sigma$: we have

$$\left\|v_t^2 - \sigma - \mathbb{E}\left[v_t^2 - \sigma \mid \mathcal{F}_{-t}^0\right]\right\|_2 = \left\|2p_t q_t + \mathbb{E}\left[q_t^2 \mid \mathcal{F}_{-t}^0\right]\right\| = O(\psi^t).$$

The fourth moment finiteness is needed in order to have $\mathbb{E}\left[\xi_0^2\right] < \infty$. □

We have now all the tools to prove the asymptotic normality of the aggregated system parameters.

Theorem 10. *Consider the process* (3)*, with i.i.d variables* $Y_t = \begin{bmatrix}\epsilon_t \\ \eta_t\end{bmatrix}$ *with finite fourth moment. Suppose that the process is stationary and ergodic, and that* $\Sigma_\epsilon > 0$, $\Sigma_\eta > 0$. *Then, the maximum likelihood estimators* $\hat{\psi}^{(w)}$, $\hat{\sigma}^{(w)}$, *for all* $w \in \mathcal{W}$, *are jointly asymptotically normal.*

Proof. The first-order optimality conditions for the ML estimates state that $0 = \frac{1}{T}\sum \nabla \ell_t(\hat{\psi}, \hat{\sigma})$, where ∇ denotes taking a gradient with respect to the pair of parameters $(\tilde{\psi}, \tilde{\sigma})$. Using a multivariate Taylor expansion around (ψ, σ), we get

$$0 = \frac{1}{T}\sum \nabla \ell_t(\psi, \sigma) + \left(\frac{1}{T}\sum \nabla^2 \ell_t(\tilde{\psi}, \tilde{\sigma})\right)\left(\begin{bmatrix}\hat{\psi} \\ \hat{\sigma}\end{bmatrix} - \begin{bmatrix}\psi \\ \sigma\end{bmatrix}\right), \tag{24}$$

for a suitable pair $(\tilde{\psi}, \tilde{\sigma})$ lying in the segment that joins $(\hat{\psi}, \hat{\sigma})$ and (ψ, σ). If $(\hat{\psi}, \hat{\sigma})$ are close enough to the exact values, then by continuity the Hessian matrix is invertible and bounded, thus we can rewrite (24) as

$$\begin{bmatrix} \hat{\psi} \\ \hat{\sigma} \end{bmatrix} - \begin{bmatrix} \psi \\ \sigma \end{bmatrix} = -\left(\frac{1}{T} \sum \nabla^2 \ell_t(\tilde{\psi}, \tilde{\sigma}) \right)^{-1} \left(\frac{1}{T} \sum \nabla \ell_t(\psi, \sigma) \right). \tag{25}$$

This expansion (25) is valid for every $w \in \mathcal{W}$ that we use as the aggregation weights. Let β be the vector obtained by stacking the vectors $\begin{bmatrix} \psi^{(w_i)} \\ \sigma^{(w_i)} \end{bmatrix}$ for each $w_i \in \mathcal{W}$, one above the other, and $\hat{\beta}$ be similarly defined with their ML estimators. Stacking the Taylor expansions (25) one above the other and multiplying by $\sqrt{T}$ we get

$$\sqrt{T}(\hat{\beta} - \beta) = -M(\tilde{\beta})^{-1} \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \begin{bmatrix} \frac{\partial}{\partial \psi^{(w_1)}} \ell_t^{(w_1)} \\ \frac{\partial}{\partial \sigma^{(w_1)}} \ell_t^{(w_1)} \\ \frac{\partial}{\partial \psi^{(w_2)}} \ell_t^{(w_2)} \\ \frac{\partial}{\partial \sigma^{(w_2)}} \ell_t^{(w_2)} \\ \vdots \\ \frac{\partial}{\partial \sigma^{(w_{N(N+1)/2})}} \ell_t^{(w_{N(N+1)/2})} \end{bmatrix}, \tag{26}$$

with $M(\tilde{\beta})$ the block diagonal matrix containing $\frac{1}{T} \sum \nabla^2 \ell_t^{(w_i)}(\psi^{(\tilde{w}_i)}, \sigma^{(\tilde{w}_i)})$ in its diagonal blocks. We know from the proof of Lemma 6 and the consistency that $M(\tilde{\beta})$ converges almost surely to a diagonal matrix; if we prove that a central limit result holds for the vector sum in (26), then the thesis follows by Slutsky's theorem. Theorem 7 and Lemma 9 together give only a weaker result, i.e., that each component of the vector sum is asymptotically normal when considered alone. To prove *joint* normality, we need a modification of the above proof. We shall prove that each linear combination of its entries is asymptotically normal, and use the Cramer-Wold device [Brockwell and Davis, 2006, Proposition 6.3.1]. Let us take a generic linear combination

$$C_t := \sum_{i=1}^{N(N+1)/2} a_i \frac{\partial}{\partial \psi^{(w_i)}} \ell_t^{(w_i)} + \sum_{i=1}^{N(N+1)/2} b_i \frac{\partial}{\partial \sigma^{(w_i)}} \ell_t^{(w_i)}.$$

By the linearity of expectation and subadditivity of norms,

$$\begin{aligned} \left\| C_0 - \mathbb{E}\left[C_0 \mid \mathcal{F}_{-t}^0 \right] \right\|_2 \leq & \sum_{i=1}^{N(N+1)/2} |a_i| \left\| \frac{\partial}{\partial \psi^{(w_i)}} \ell_t^{(w_i)} - \mathbb{E}\left[\frac{\partial}{\partial \psi^{(w_i)}} \ell_t^{(w_i)} \mid \mathcal{F}_{-t}^0 \right] \right\|_2 \\ & + \sum_{i=1}^{N(N+1)/2} |b_i| \left\| \frac{\partial}{\partial \sigma^{(w_i)}} \ell_t^{(w_i)} - \mathbb{E}\left[\frac{\partial}{\partial \sigma^{(w_i)}} \ell_t^{(w_i)} \mid \mathcal{F}_{-t}^0 \right] \right\|_2, \end{aligned}$$

and we know from the proof of Lemma 9 that each term in the right-hand side is $O(t(\psi^{(w_i)})^t)$. Setting $\psi_{\max} := \max_{w\in\mathcal{W}}|\psi^{(w)}|$, we have therefore

$$\left\|C_0 - \mathbb{E}\left[C_0 \mid \mathcal{F}_{-t}^0\right]\right\|_2 = O(t\psi_{\max}^t),$$

hence our generic linear combination C_t satisfies the bound of Theorem 19 and thus is asymptotically normal. So, putting all together, in (26) $M(\tilde{\beta})^{-1}$ converges a.s. to a constant matrix and the scaled sum converges in probability to a normal vector, thus by Slutsky's theorem $\sqrt{T}(\hat{\beta} - \beta)$ is asymptotically normal. □

Hence we have the following asymptotic result for $\hat{\Theta}$ and $\hat{\Sigma}_u$.

Theorem 11. *Consider the EWMA process* (3)*; suppose that the noises η_t and ϵ_t are i.i.d. processes with variances $0 < \Sigma_\eta, \Sigma_\epsilon < \infty$, and that the resulting EWMA process is stationary and ergodic. Then, the META estimator $\hat{\Theta}, \hat{\Sigma}_u$ described in Section 4 is asymptotically consistent. If, in addition, the fourth moments of ϵ_t and η_t are finite, then the estimator is asymptotically normal.*

Proof. The estimated values $\hat{\Theta}$ and $\hat{\Sigma}_u$ are an a function of $\psi^{(w)}$ and $\sigma^{(w)}$ for $w \in \mathcal{W}$ as introduced above; the specific function, obtained by composing (13) and (**??**), is continuous and differentiable. Since it is continuous and these values are consistent by Lemma 6, the estimator is consistent. Under the additional hypothesis on the fourth moment, these values are also asymptotically normal by Theorem 10, thus by the delta method [Lütkepohl, 2005, Appendix C.5] asymptotic normality holds. □

6 META vs. Maximum Likelihood: some numerical experiments

In this section we provide some numerical experiments to compare the estimates obtained with the META method vis-a-vis a maximum likelihood estimator on the VMA(1) representation (3). We generated simulated data for a model of the form (1), using Gaussian noise with four different sets of covariance matrices, two bivariate and two trivariate ones, resulting in signal-to-noise ratios of different magnitude:

Model 1 $\Sigma_\eta = \begin{bmatrix} 1 & -0.5 \\ -0.5 & 1.5 \end{bmatrix}$, $\Sigma_\epsilon = \begin{bmatrix} 1.5 & -0.15 \\ -0.15 & 1 \end{bmatrix}$,

Model 2 $\Sigma_\eta = \begin{bmatrix} 1 & -0.5 \\ -0.5 & 1.5 \end{bmatrix}$, $\Sigma_\epsilon = \begin{bmatrix} 30 & -3 \\ -3 & 20 \end{bmatrix}$,

Model 3 $\Sigma_\eta = \begin{bmatrix} 1 & -0.5 & 0.3 \\ -0.5 & 1.5 & -0.2 \\ 0.3 & -0.2 & 1 \end{bmatrix}$, $\Sigma_\epsilon = \begin{bmatrix} 1.5 & -0.15 & -0.1 \\ -0.15 & 1 & 0.3 \\ -0.1 & 0.3 & 1.5 \end{bmatrix}$,

Model 4 $\Sigma_\eta = \begin{bmatrix} 1 & -0.5 & 0.3 \\ -0.5 & 1.5 & -0.2 \\ 0.3 & -0.2 & 1 \end{bmatrix}$, $\Sigma_\epsilon = \begin{bmatrix} 30 & -3 & -2 \\ -3 & 20 & 6 \\ -2 & 6 & 30 \end{bmatrix}$.

The aim here is generating two different types of models; Model 1 and Model 3 have roots of the matrix Θ being about half of those of Model 2 and Model 4 respectively. Moreover, in Model 1 and Model 3 Σ_u is much smaller than that of Model 2 and Model 4 respectively.

For each model, we generated time series of three different sample sizes $T = 200$, 400 and 1000, and estimated them using both the ML and META methods. Each experiment has been repeated 500 times, with different data, produced each time using new computer-generated random numbers. All simulations were carried out using MATHEMATICA 8 by Wolfram and its TIMESERIES 1.4.1 package[1]. The source files for the simulation are available upon request.

Since the parameter matrices for these simulated models are explicitly available (see Proposition 1), we can check how close the estimated Θ and Σ_u are to the real ones. As error measure, we used the relative error in the Frobenius norm (root mean squared error of the matrix entries)

$$RMSE = \frac{\left\|\hat{X} - X\right\|_F}{\|X\|_F}, \tag{27}$$

where X is the matrix of true parameters and $\hat{X}$ is the estimated one, and $\|\cdot\|_F$ is the Frobenius norm: for a $m \times n$ matrix X, $\|X\|_F := \left(\sum_{i=1}^{m}\sum_{j=1}^{n} X_{ij}^2\right)^{1/2}$. The average errors on the set of 500 repeated experiments under this error metric are reported in Table 1. The central columns contain the RMSE (27) multiplied by 1000. The last two columns report the average time (in seconds) taken by each estimation procedure.

Notice that the error is considerably lower in Models 2 and 4, which have a lower signal-to-noise ratio Q and thus Θ with eigenvalues closer to the unit circle. This behaviour is expected in view of the properties of the maximum likelihood estimator for ARMA processes (cfr. [Lütkepohl, 2005, Section 7.2.3], and [Brockwell and Davis, 2006, §8.5] for a discussion of asymptotic efficiency in the scalar case). These empirical results show that META, whose derivation contains elements from both moment estimators and ML estimators, does not degrade in quality like the former when the eigenvalues are closer to the unit circle, but seems to maintain the higher asymptotic efficiency of the latter.

Overall, the results are clearly in favor of the META estimator. Indeed, not only the META estimator is extremely faster than the multivariate ML estimator, but it seems to outperform the rival estimator in terms of accuracy nearly all the times. There are only two cases where ML slightly outperforms META with respect to Σ_u (i.e., Model 1 and Model 3 with $T = 400$). On the contrary, regarding Θ, META is always more accurate than the ML approach. The last column of Table 1 shows the computational difficulties of the standard ML approach. On a normal computer, it took us about one week of computation time to obtain the ML results for Models 3 and 4 with $T = 1000$. In contrast, the same experiments with the META estimator took one hour and a half. These results clearly suggest the adoption of the suggested estimator, being not only extremely faster to compute, but also having very good small sample properties.

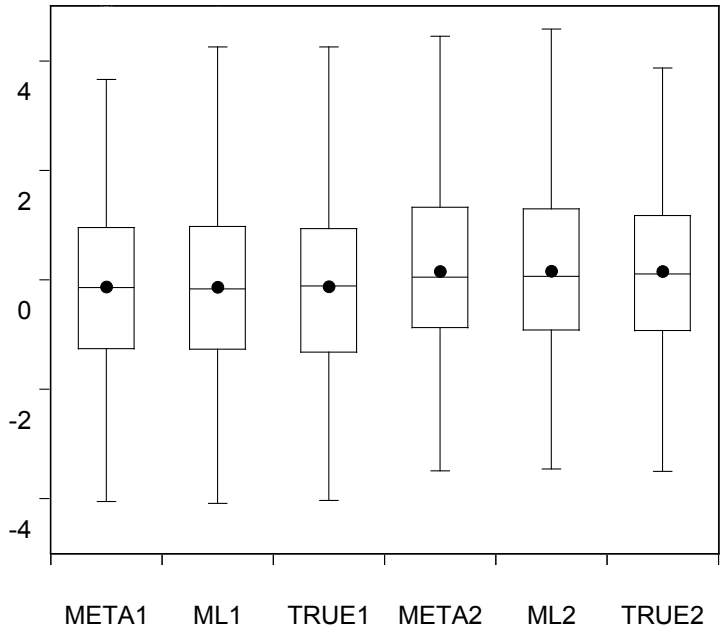
[1]Further information can be found in the Wolfram website: see http://media.wolfram.com/documents/TimeSeriesDocumentation.pdf

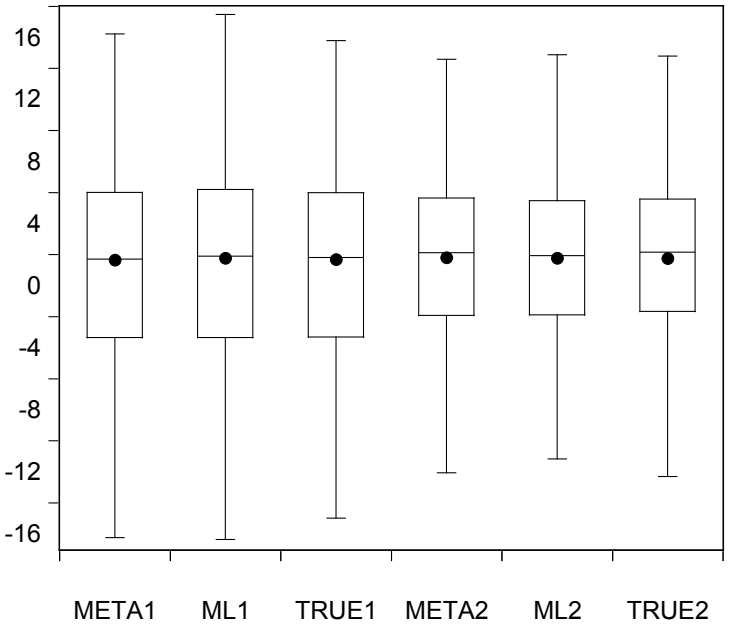
Table 1: Mean relative (normalized) RMSE of the estimates of the reduced parameters

	Sample size T	Θ META	Θ ML	Σ_u META	Σ_u ML	Time META	Time ML
Model 1	200	202.52	236.77	108.28	109.11	1.55	59.99
	400	121.41	138.13	83.31	82.93	3.13	107.28
	1000	80.83	101.53	48.65	48.96	8.34	226.64
Model 2	200	69.51	78.26	97.50	98.31	1.36	46.53
	400	48.26	56.48	80.91	81.52	3.74	113.36
	1000	28.01	34.07	47.60	48.02	7.63	210.06
Model 3	200	205.07	254.49	135.26	136.41	2.65	201.41
	400	162.95	187.40	93.48	93.21	5.77	316.74
	1000	93.85	108.92	60.08	61.13	12.62	523.67
Model 4	200	86.66	107.22	123.86	124.19	2.81	202.95
	400	57.03	67.25	95.13	96.75	5.66	295.93
	1000	29.91	37.04	61.78	62.13	13.12	541.36

The cells relative to Θ and Σ_u reports the average relative root mean squared error (27) of the estimated parameters, multiplied by 1000. The last two columns report the average number of seconds required for a single run of the estimation procedure. Lower is better.

As a second performance test, for Models 1 and 2 and $T = 200$, we compared the forecasting accuracy of the estimated parameters. In detail, we generated $T = 200$ observations of each model, estimated the parameters using the first 199 observations only, and used them to generate a prediction y^F_{200} of the last value y_{200} of the time series according to (10). Each of these experiments has been repeated $R = 200$ times, each time with a different randomly-generated time series of length 200. We report in Figures 2 and 3 the boxplot of the forecast error $y_{200} - y^F_{200}$ for Model 1 and 2 respectively. Each figure contains box plots for the two components $(y^F_{200} - y_{200})_1$ and $(y^F_{200} - y_{200})_2$; for example, *META1* and *META2* refer respectively to the forecast error of the first and second component of the multivariate time series when META is employed. As a term of comparison, we report the prediction error obtained with the *true* system matrix Θ, which is available explicitly in our simulation.





As expected, the forecast errors relative to Model 2 are much more dispersed than those for Model 1. Moreover, the prediction accuracy is almost identical in all three cases for each equation relative to both Model 1 and 2, showing that the META forecasts are essentially as good as the ML ones (and almost as good as the real system matrices).

7 Conclusions

Simple exponential smoothing has been shown to be a valid candidate in forecasting demand (see Dekker et al. [2004] and Moon et al. [2012]). This paper provides exact results linking reduced form parameters and autocovariances for the simple exponential smoothing in the multivariate framework. The results are used to provide a fast and efficient estimator, which seems to outperform the multivariate maximum likelihood on the underlying VMA representation in both time and accuracy. The technique used in the estimator allows one to reduce the problem from one N-dimensional maximum likelihood estimation to $N(N+1)/2$ scalar ML problems. This is especially convenient, since ML estimators for high-dimensional problems are slower to converge and more prone to numerical failures. The ML estimator has an expected complexity of $O(N^3T)$ per step, hence, under the reasonable assumption that the number of steps stays the same or decreases for the aggregate problems, passing to univariate problems rates to be even more effective when N is larger. An additional benefit is that the scalar estimation problems are separate and can be solved in parallel.

A key feature of the VMA models resulting from our exponential smoothing model is that the autocovariance Γ_1 is a symmetric matrix. This property is used in both Proposition 1 and Lemma 4. Our estimation procedure requires only this hypothesis, so it works without changes for any MA(1) model with $\Gamma_1 = \Gamma_1^T$. We are currently working on removing the assumption $\Gamma_1 = \Gamma_1^T$. This more general model arises, for instance, when the noises ϵ_t and η_t characterizing the system (1) are correlated; we leave this for future research. Another open problem is deriving the exact asymptotic covariance of the estimator, with the aim of comparing it to maximum likelihood in terms of asymptotic efficiency (cfr. [Brockwell and Davis, 2006, §8.5]). This task looks challenging, even in the case of Gaussian noise, since the noises $v^{(w)}$ of the aggregate processes are correlated and the computation would have to keep track of all these correlations.

A limitation of this work is that the suggested estimator is valid for the single exponential smoothing, but not for the whole family of exponential smoothing models. For example, our estimator cannot be used for models that take into account for the presence of a stochastic trend, such as the local linear trend model or the cubic smoothing spline models (see Harvey [1991] and Hyndman et al. [2005]). This is because the lack of closed-form results for more complex models.

This manuscript has practical implications for practitioners involved in forecasting a multivariate production planning framework. Consider, for example, the case of a retail company providing a broad range of products to its customers. In order to reduce costs and to manage efficiently the production planning process, the company has to rely on accurate forecasts for the demand of each good/service as well as for the aggregate demand. Aggregated models, such as the top-down and bottom-up approaches, are often used because it is difficult and computationally intensive to handle a multivariate approach with full dependence between the variables. Instead, thanks to the algorithm suggested in this paper, estimation and forecasting can now be implemented with a multivariate exponential smoothing model without facing heavy computational issues.

References

P. J. Brockwell and R. A. Davis. *Time series: theory and methods*. Springer Series in Statistics. Springer, New York, 2006. ISBN 978-1-4419-0319-8; 1-4419-0319-8. Reprint of the second (1991) edition.

A. Chen and J. Blue. Performance analysis of demand planning approaches for aggregating, forecasting and disaggregating interrelated demands. *International Journal of Production Economics*, 128(2):586–602, Dec. 2010. ISSN 0925-5273. doi: 10.1016/j.ijpe.2010.07.006. URL `http://www.sciencedirect.com/science/article/pii/S0925527310002318`.

J. G. De Gooijer and R. J. Hyndman. 25 years of time series forecasting. *International Journal of Forecasting*, 22(3):443 – 473, 2006. ISSN 0169-2070. doi: http://dx.doi.org/10.1016/j.ijforecast.2006.01.001. URL `http://www.sciencedirect.com/science/article/pii/S0169207006000021`. Twenty five years of forecasting.

M. Dekker, K. van Donselaar, and P. Ouwehand. How to use aggregation and combined forecasting to improve seasonal demand forecasts. *International Journal of Production Economics*, 90(2):151–167, July 2004. ISSN 0925-5273. doi: 10.1016/j.ijpe.2004.02.004. URL `http://www.sciencedirect.com/science/article/pii/S0925527304000398`.

E. B. Fliedner and B. Lawrence. Forecasting system parent group formation: An empirical application of cluster analysis. *Journal of Operations Management*, 12(2):119–130, 1995.

G. Fliedner. An investigation of aggregate variable time series forecast strategies with specific subaggregate time series statistical correlation. *Computers & Operations Research*, 26(10):1133–1149, 1999.

E. Gardner Jr. Exponential smoothing: The state of the art – part ii. *International Journal of Forecasting*, 22(4):637–666, 2006. doi: 10.1016/j.ijforecast.2006.03.005.

A. C. Harvey. *Forecasting Structural Time Series Models and the Kalman Filter*. Cambridge University Press, 1991.

C. C. Holt. Author's retrospective on "forecasting seasonals and trends by exponentially weighted moving averages". *International Journal of Forecasting*, 20(1):11–13, Jan. 2004. ISSN 0169-2070. doi: 10.1016/j.ijforecast.2003.09.017. URL `http://www.sciencedirect.com/science/article/pii/S0169207003001158`.

R. A. Horn and C. R. Johnson. *Matrix analysis*. Cambridge University Press, Cambridge, 1990. ISBN 0-521-38632-2. Corrected reprint of the 1985 original.

R. J. Hyndman, M. L. King, I. Pitrun, and B. Billah. Local linear forecasts using cubic smoothing splines. *Australian & New Zealand Journal of Statistics*, 47(1):87–99, Mar. 2005. ISSN 1467-842X. doi: 10.1111/j.1467-842X.2005.00374.x. URL `http://onlinelibrary.wiley.com/doi/10.1111/j.1467-842X.2005.00374.x/abstract`.

C. J. Kascha. A comparison of estimation methods for vector autoregressive moving-average models. *Econometric Reviews*, 31(3):297–324, 2012. ISSN 0747-4938.

S. Ling and M. McAleer. A general asymptotic theory for time-series models. *Stat. Neerl.*, 64(1):97–111, 2010. ISSN 0039-0402. doi: 10.1111/j.1467-9574.2009.00447.x. URL http://dx.doi.org/10.1111/j.1467-9574.2009.00447.x.

H. Lütkepohl. *Forecasting aggregated vector ARMA processes*, volume 284 of *Lecture Notes in Economics and Mathematical Systems*. Springer-Verlag, Berlin, 1987. ISBN 3-540-17208-4. doi: 10.1007/978-3-642-61584-9. URL http://dx.doi.org/10.1007/978-3-642-61584-9.

H. Lütkepohl. *New introduction to multiple time series analysis*. Springer-Verlag, Berlin, 2005. ISBN 3-540-40172-5.

S. Makridakis and M. Hibon. The M3-competition: results, conclusions and implications. *International Journal of Forecasting*, 16(4):451–476, Oct. 2000. ISSN 0169-2070. doi: 10.1016/S0169-2070(00)00057-1. URL http://www.sciencedirect.com/science/article/pii/S0169207000000571.

S. Moon, C. Hicks, and A. Simpson. The development of a hierarchical forecasting method for predicting spare parts demand in the South Korean NavyA case study. *International Journal of Production Economics*, 140(2):794–802, Dec. 2012. ISSN 0925-5273. doi: 10.1016/j.ijpe.2012.02.012. URL http://www.sciencedirect.com/science/article/pii/S0925527312000709.

S. Moon, A. Simpson, and C. Hicks. The development of a classification model for predicting the performance of forecasting methods for naval spare parts demand. *International Journal of Production Economics*, 143(2):449–454, June 2013. ISSN 0925-5273. doi: 10.1016/j.ijpe.2012.02.016. URL http://www.sciencedirect.com/science/article/pii/S0925527312000746.

J. F. Muth. Optimal Properties of Exponentially Weighted Forecasts. *Journal of the American Statistical Association*, 55(290):299–306, June 1960. ISSN 01621459. doi: 10.2307/2281742. URL http://dx.doi.org/10.2307/2281742.

M. Peligrad and S. Utev. Central limit theorem for stationary linear processes. *Ann. Probab.*, 34(4):1608–1622, 2006. ISSN 0091-1798. doi: 10.1214/009117906000000179. URL http://dx.doi.org/10.1214/009117906000000179.

G. Sbrana and A. Silvestrini. Forecasting aggregate demand: Analytical comparison of top-down and bottom-up approaches in a multivariate exponential smoothing framework. *International Journal of Production Economics*, 146(1):185–198, Nov. 2013. ISSN 0925-5273. doi: 10.1016/j.ijpe.2013.06.022. URL http://www.sciencedirect.com/science/article/pii/S0925527313002922.

L. R. Weatherford, S. E. Kimes, and D. A. Scott. Forecasting for hotel revenue management: Testing aggregation against disaggregation. *The Cornell Hotel and Restaurant Administration Quarterly*, 42(4):53–64, Aug. 2001. ISSN 0010-8804. doi: 10.1016/S0010-8804(01)80045-8. URL http://www.sciencedirect.com/science/article/pii/S0010880401800458.

H. Widiarta, S. Viswanathan, and R. Piplani. Forecasting aggregate demand: An analytical evaluation of top-down versus bottom-up forecasting in a production planning framework. *International Journal of Production Economics*, 118(1):87–94, March 2009. URL http://ideas.repec.org/a/eee/proeco/v118y2009i1p87-94.html.

G. Zotteri and M. Kalchschmidt. Forecasting practices: Empirical evidence and a framework for research. *International Journal of Production Economics*, 108(12):84–99, July 2007. ISSN 0925-5273. doi: 10.1016/j.ijpe.2006.12.004. URL http://www.sciencedirect.com/science/article/pii/S0925527306003069.

G. Zotteri, M. Kalchschmidt, and F. Caniato. The impact of aggregation level on forecasting performance. *International Journal of Production Economics*, 9394:479–491, Jan. 2005. ISSN 0925-5273. doi: 10.1016/j.ijpe.2004.06.044. URL http://www.sciencedirect.com/science/article/pii/S092552730400266X.