ISSN 1995-0802, Lobachevskii Journal of Mathematics, 2022, Vol. 43, No. 9, pp. 2433–2443. © Pleiades Publishing, Ltd., 2022.

Improved Maximum Likelihood Estimation of ARMA Models

Leonardo Di Gangi[1*], Matteo Lapucci[1], Fabio Schoen[1***], and Alessio Sortino[1****]**

(Submitted by A. I. Volodin)

[1] *Global Optimization Laboratory, DINFO, Università degli Studi di Firenze, Florence, Italy*

Received November 12, 2021; revised November 20, 2021; accepted December 9, 2021

Abstract—In this paper we propose a new optimization model for maximum likelihood estimation of causal and invertible ARMA models. Through a set of numerical experiments we show how our proposed model outperforms, both in terms of quality of the fitted model as well as in the computational time, the classical estimation procedure based on Jones reparametrization. We also propose a regularization term in the model and we show how this addition improves the out of sample quality of the fitted model. This improvement is achieved thanks to an increased penalty on models close to the non causality or non invertibility boundary.

DOI: 10.1134/S1995080222120101

Keywords and phrases: *ARMA models, maximum likelihood estimation, bound-constrained optimization, Jones reparametrization, close-to-the-boundary solutions.*

1. INTRODUCTION

A zero mean ARMA process of order (p, q) is defined through the following stochastic difference equation:

$$Y_t - \phi_1 Y_{t-1} - \ldots - \phi_p Y_{t-p} = \theta_1 \epsilon_{t-1} + \ldots + \theta_q \epsilon_{t-q} + \epsilon_t, \quad \epsilon_t \sim \mathcal{WN}(0, \sigma^2), \tag{1}$$

or in compact form [5] as $\Phi(B)Y_t = \Theta(B)\epsilon_t$, where $\Phi(\cdot)$ and $\Theta(\cdot)$ are the pth and qth-degree polynomials

$$\Phi(z) = 1 - \phi_1 z - \ldots - \phi_p z^p, \tag{2}$$

$$\Theta(z) = 1 + \theta_1 z + \ldots + \theta_p z^q, \tag{3}$$

and B is the backward shift operator (see [5, 6, 13]). In Eq. (1), $\phi = (\phi_1, \ldots, \phi_p)$ are the parameters concerning the autoregressive part, while analogously $\theta = (\theta_1, \ldots, \theta_q)$ are the parameters of the moving average component. As it is typically assumed, the error terms in Eq. (1) are modeled as a zero mean Gaussian white noise process of variance σ^2.

The interest towards this class of statistical models is justified by their employment in a multitude of fields like business planning, finance, transportation systems, demography and medicine. With special reference to real-time forecasting systems, it is very important to develop computationally efficient estimation methods focused on improving the numerical stability of the related fitting procedure and the predictive ability of the ARMA models.

ARMA models estimation has a very long history [1, 2, 5, 9, 12, 14, 15, 26]. Maximum likelihood estimation is usually performed for its advantageous asymptotic properties. A closed form expression of the ARMA exact likelihood function was firstly given in [26]. Afterwards, the focus shifted to finding expressions of the exact likelihood being more suitable for its computation [2, 9]. Finally, in the late 70's, the computational advantages of computing the exact likelihood by means of Kalman Filter [20] have

*E-mail: leonardo.digangi@unifi.it
**E-mail: matteo.lapucci@unifi.it
***E-mail: fabio.schoen@unifi.it
****E-mail: alessio.sortino@unifi.it

been pointed out in [15]. To date, Kalman Filter algorithm, initialized according to the Gardner method [12], represents the state-of-the-art of the methods employed to compute the exact likelihood.

As it is usually required in forecasting applications, the estimation of (ϕ, θ) needs to take into account the causality and invertibility conditions [6] which act like constraints in the search space. These constraints are usually handled by means of the Jones reparametrization [19] which converts the original constrained ARMA estimation problem into an unconstrained one.

In this paper we propose to fit causal and invertible ARMA models by exact maximum likelihood estimation avoiding the employment of the Jones reparametrization [19]. This is achievable solving a bound constrained optimization problem. The benefits of our formulation are both lower computational fitting times and better numerical stability w.r.t. the classical unconstrained approach. Furthermore, we propose the addition of a quadratic regularization term to the ARMA exact likelihood function. This term improves the predictive ability of the fitted ARMA models.

The rest of the paper is organized as follows. Section 2 contains a review of the Jones reparametrization method. In Section 3 the notion of closeness of (ϕ, θ) to the feasibility boundary is defined. In Section 4 our bound constrained maximum likelihood estimation approach is provided. In Section 5, extensive computational experiments which assess the reliability of the proposed method are reported. Finally, the overall conclusions are remarked in Section 6.

2. JONES REPARAMETRIZATION

When causality and invertibility conditions [6] hold, the parameters $\phi = (\phi_1, \ldots, \phi_p)$ and $\theta = (\theta_1, \ldots, \theta_q)$ are constrained to belong to the set $S_p \times S_q$, corresponding to the polynomial operator root conditions

$$S_p = \{\phi \in \mathbb{R}^p \mid 1 - \phi_1 z - \ldots - \phi_p z^p \neq 0 \; \forall\, z \in \mathbb{C} \text{ s.t. } |z| \leq 1\} \tag{4}$$

$$S_q = \{\theta \in \mathbb{R}^q \mid 1 + \theta_1 z + \ldots + \theta_p z^q \neq 0 \; \forall\, z \in \mathbb{C} \text{ s.t. } |z| \leq 1\}. \tag{5}$$

These feasible sets are easily identified for $p \leq 2$ and $q \leq 2$, but for $k > 2$ the form of S_k becomes complicated and for $k > 4$ the polynomial Eqs. (4), (5) cannot be solved analytically [23]. The geometry of the feasible set $S_p \times S_q$ is described in detail in [7, 27, 29]. To circumvent the problem of dealing with constraints (4) and (5) Barndorff-Nielsen and Schou [4] reparametrize $\phi = (\phi_1, \ldots, \phi_p)$ in terms of the partial autocorrelations $\rho = (\rho_1, \ldots, \rho_p)$ by means of the one-to-one continously differentiable Levinson mapping $\Upsilon(\cdot)$:

$$\phi_k^{(k)} = \rho_k, \quad k = 1, \ldots, p, \quad \phi_i^{(k)} = \phi_i^{(k-1)} - \rho_k \phi_{k-i}^{(k-1)}, \quad i = 1, \ldots, k-1. \tag{6}$$

In (6), causality is simply obtained by $\rho_k \in (-1, 1) \;\; \forall k = 1, \ldots, p$. Jones [19] introduces an additional mapping $J : \mathbb{R}^p \to (-1, 1)^p$, which allows to formulate the original problem as an unconstrained optimization problem introducing variables $u_k, \; k = 1, \ldots, p$:

$$\rho_k = \frac{1 - \exp(-u_k)}{1 + \exp(-u_k)}, \quad k = 1, \ldots, p. \tag{7}$$

Similar transformations can also be employed for the moving average parameters $\theta = (\theta_1, \ldots, \theta_q)$ in order to guarantee the invertibility condition. By writing the moving average polynomial (3) for the negative vector of MA parameters, $-\theta$, we get

$$\Theta(z) = 1 - (-\theta_1) z - \ldots - (-\theta_q) z^q, \tag{8}$$

and the following can be deduced

$$\begin{gathered} \theta_k^{(k)} = b_k, \quad k = 1, \ldots, q, \\ \theta_i^{(k)} = \theta_i^{(k-1)} + b_k \theta_{k-i}^{(k-1)}, \quad i = 1, \ldots, k-1, \end{gathered} \tag{9}$$

where the variables $b_k \in (-1, 1) \; \forall k = 1, \ldots, q$. Jones reparametrization for the moving average part is equivalent to (7):

$$b_k = \frac{1 - \exp(-w_k)}{1 + \exp(-w_k)}, \quad k = 1, \ldots, q. \tag{10}$$

In [19], the variables b_k are called partial moving average coefficients. The optimization of the exact loglikelihood in the causal and invertible feasible space is now carried out with respect to the variables $u = (u_1, \dots, u_p) \in \mathbb{R}^p$ and $w = (w_1, \dots, w_q) \in \mathbb{R}^q$.

Note that $\phi = \Upsilon(\rho)$, while $\theta = -\Upsilon(b)$. In fact, for any u and w, the evaluation of the exact likelihood function in a causal and invertible feasible point can be computed by means of the transformations (6), (7), (9), (10), and the Kalman recursions. Inverse Jones transformations are easily found by solving (7), (10) respectively for $u_k,\ k = 1, \dots, p$ and $w_k,\ k = 1, \dots, q$. On the other hand, Monhan [24] derives the expression of the inverse transformation $\Upsilon^{-1}(\cdot)$ of (6) which equivalently can be extended for the moving average part (9).

3. CLOSENESS TO THE FEASIBLITY BOUNDARY

In this Section, the notion of closeness of a feasible point $(\phi, \theta) \in S_p \times S_q$ to the set $\partial S_p \times \partial S_q$, i.e. the boundary of the invertibility and causality regions, is formalized. This will be useful later in this work, when investigating the relation between the closeness to the boundary and the numerical stability during the optimization of the Gaussian ARMA exact log-likelihood function.

It is partially documented [1)] that log-likelihood evaluation by Kalman filter may fail when a point (ϕ, θ) is close to the causality boundary. Furthermore, it is well known that closeness to the non-invertible region is problematic due to the presence of the so-called pile-up effect [21, 22, 28]. Indeed, when the true parameter of an MA(1) process is close to unity, the model can be estimated to be non-invertible with a unit root even when the true process is invertible, with a considerably high probability in a finite sample. Ansley and Newbold [3] confirm the presence of such effect in ARMA models too.

Inspired by the method of McLeod and Zhang [32] for testing the presence of a parameter estimate on the boundary of an MA(q) model, we define the closeness of a point (ϕ, θ) to the boundary of the invertible and the causal-stationary regions exploiting the parametrization of an ARMA(p, q) in terms of ρ and b:

$$(\phi, \theta) = (\Upsilon(\rho), -\Upsilon(b)),$$
$$(\phi, \theta) \in S_p \times S_q \iff (\rho, b) \in (-1, 1)^p \times (-1, 1)^q.$$

$\Upsilon(\cdot)$ is not one-to-one on the hypercube boundary [4]. However, as elegantly shown in [32], $\Upsilon(\cdot)$ maps the boundary of $(-1, 1)^p$ onto ∂S_p. Since $\Upsilon(\cdot)$ is a continuously differentiable function in $[-1, 1]^p$, the closeness of an estimate $\phi \in S_p$ to the non causal-stationary boundary ∂S_p can be defined respectively in terms of the partial autocorrelations ρ. The same reasoning holds for the moving average part.

As reported in [32], $\phi \in \partial S_p$ if and only if $\|\rho\|_\infty = 1$ and similarly $\theta \in \partial S_q$ if and only if $\|b\|_\infty = 1$. Now, by fixing a threshold parameter $\tau > 0$, closeness of $(\phi, \theta) = (\Upsilon(\rho), -\Upsilon(b)) \in S_p \times S_q$ to the boundary $\partial S_p \times \partial S_q$ is defined by the following conditions:

(i) $(\phi, \theta) \in S_p \times S_q$ is close to ∂S_p if and only if $1 - \|\rho\|_\infty < \tau$;

(ii) $(\phi, \theta) \in S_p \times S_q$ is close to ∂S_q if and only if $1 - \|b\|_\infty < \tau$;

(iii) $(\phi, \theta) \in S_p \times S_q$ is close to both ∂S_p and ∂S_q if and only if $1 - \|\rho\|_\infty < \tau$ and $1 - \|b\|_\infty < \tau$.

A point $(\phi, \theta) \in S_p \times S_q$ which does not satisfy any of the above conditions (i), (ii), (iii) is defined as a strictly feasible point of $S_p \times S_q$.

[1)] See, e.g., https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/KalmanLike and https://bugs.r-project.org/bugzilla/show_bug.cgi?id=14682

4. THE PROPOSED APPROACH

We propose to fit causal and invertible ARMA(p, q) models by solving the following bound constrained optimization problem

$$\max_{\rho,b,\sigma^2} \ell\left(\Upsilon(\rho), -\Upsilon(b), \sigma^2\right)$$
$$\text{s.t. } \rho \in [-1+\varepsilon, 1-\varepsilon]^p, \quad b \in [-1+\varepsilon, 1-\varepsilon]^q, \quad \sigma \in \mathbb{R}_+. \tag{11}$$

Optimizing w.r.t. the partial autocorrelation and the partial moving average coefficients avoids the use of the Jones reparametrization (7), (10). Note that this formulation cuts off a small part of the feasible space $S_p \times S_q$. However, as highlighted by thorough numerical experiments that we will describe in the following Section, our formulation provides some nice advantages:

- it allows to save a significant amount of running time, as there is no more the need to compute Eqs. (7) and (10) p and q times respectively, each time the log-likelihood has to be computed during the optimization process (note that every gradient computation by finite differences requires $2(p+q)$ objective evaluations);
- it allows to avoid solutions too close to the feasibility boundary that typically lead to numerical errors.

We furthermore propose to include in the objective function of Problem (11) a Tikhonov regularization term:

$$\max_{\rho,b,\sigma^2} \ell\left(\Upsilon(\rho), -\Upsilon(b), \sigma^2\right) - \lambda(||\rho||_2^2 + ||b||_2^2)$$
$$\text{s.t. } \rho \in [-1+\varepsilon, 1-\varepsilon]^p, \quad b \in [-1+\varepsilon, 1-\varepsilon]^q, \quad \sigma \in \mathbb{R}_+. \tag{12}$$

We will experimentally show in the following that, in our context, this term not only discourages solutions close to the feasibility boundary, but it also improves the predictive ability of ARMA models.

5. COMPUTATIONAL EXPERIMENTS

In what follows the approximation parameter ε is set to 10^{-2}; we fixed the closeness parameter $\tau = 2\varepsilon$ in (i), (ii), (iii), so that it is still possible for models (11) and (12) to produce points that are close to the border of the original feasible set.

All the experiments have been performed on a dataset of synthetically generated time series. We simulated a total of 2250 time series of different length $l \in \{100, 1000, 10000\}$ from ARMA (p, q) Gaussian processes up to a maximum order (p, q) of $(5, 5)$ and standard deviation $\sigma \in \{0.01, 0.1, 1\}$.

Specifically, for a given a combination of length, order and standard deviation, we generated 10 time series, each one representing a finite realization of a particular ARMA process with its structural autoregressive and moving average parameters (ϕ, θ). Each pair (ϕ, θ) is selected according to the methodology described in [18]. This methodology allows to choose (ϕ, θ) from a uniform distribution over the feasible set $S_p \times S_q$.

Firstly, we are interested in establishing the differences between solving problem (11) and the unconstrained one, based on Jones reparametrization, both from the standpoints of computational times and numerical stability. To this aim we carried out a multi-start strategy: for each time series, the fitting process is repeated 30 times from different randomly chosen starting points. These starting points are again obtained by uniform sampling over the feasible region. For a fair comparison, the two considered methods share the sets of starting points.

Secondly, we investigated the prediction performance of ARMA models close to the boundary. As usual, the performance is evaluated on a test set, after fitting on training data. Our test set for each time series is given by the last three observations (short term forecasting scenario). Similarly as above, the process of model estimation and computation of forecasts is repeated 30 times in a multi-start fashion. Note that, here, ARMA models have been fitted only by means of the classical Jones methodology. Indeed, our interest is to characterize both the forecasting performance of ARMA models close to the border and how frequently they are obtained in the standard setting.

Table 1. Two sided Wilcoxon signed-rank test. Null hypothesis: the median of the differences of the computational times $t_{\text{Jones}} - t_{\text{our}}$ is zero

Test statistic	P-Value
-34.3807	$< 1e{-}5$

Table 2. One sided Wilcoxon signed-rank test. Null hypothesis: the median of the differences of computational times $t_{\text{Jones}} - t_{\text{our}}$ is negative

Test statistic	P-Value
34.3807	$< 1e{-}5$

Our last experiment assesses the impact of the ℓ_2 regularization term in the short term forecasting. For each time series of our dataset, a single starting point to initialize the optimization is selected. The fitting procedure is then repeated for different values of the regularization hyperparameter λ in Eq. (12).

All the experiments were performed on a machine with Ubuntu Server 20.04 LTS OS, Intel Xeon E5-2430 v2 @ 2.50GHz CPU and 32GB RAM.

5.1. Fitting Procedure Runtimes

Our method provides a significant reduction of the computational time required to fit a time series with respect to the unconstrained fitting method of Jones. The time saving is estimated to be about 24% in relative terms.

This result is corroborated by the non parametric Wilcoxon signed-ranks test [8, 31]. We considered as fitting time for a time series the average runtime of successful runs (i.e., with no numerical error) of our multi-start procedure. Results of the Wilcoxon signed-ranks test are reported in Tables 1 and 2. These results point out that the median of the differences of fitting times between the two methods can be assumed to be positive, i.e., the constrained method has significantly lower fitting times.

5.2. Numerical Instability

Our fitting method prevents numerical issues during the optimization process of the ARMA exact likelihood function, thereby ensuring a higher level of computational stability.

The employment of the Jones reparametrization, where exponential operators are present, leads to a non-negligible probability of arithmetic issues, which almost always are divisions by zero and in rare cases overflows. Our method does not suffer at all from these issues.

The most critical errors, that completely undermine the fitting process, come from the Kalman Filter recursions. In general, it is well known that numerical instability often occurs in Kalman Filtering [30], especially related to the computation of the state covariance matrix.

Our experiments show that the closeness of a point (ϕ, θ) to the feasibility boundary is related to numerical instability within the Kalman Filter recursions. In particular, we observed a total of 19 `LinAlgError` errors (15 by the classical method, 4 by using our model (11)) because of the failed convergence of the SVD numerical computation.

In Tables 4 and 5 a detailed description of these errors is reported. The error may be due to the evaluation of the log-likelihood in that point or the computation in the same point of the gradient, since it is approximated by finite differences.

Two patterns are clear from Tables 4 and 5. Firstly, the classical method by Jones fails 4 times more frequently than ours. This means that our reformulation protects from the occurrence of most numerical errors. Secondly, regardless of the type of parametrization employed, it is evident that these numerical errors are related to points close the boundary $\partial S_p \times \partial S_q$ of the feasible set. Furthermore, by observing the first column of both tables, it seems that most errors inside the unconstrained framework happen even when fitting low order models.

Table 3. Occurrence of numerical instability issues per 1000 runs

Method	Arithmetic issues	Kalman filter errors
Our	0	0.06
Jones reparametrization	2.65	0.22

Table 4. Numerical errors in Kalman filtering when using Jones reparametrization. The first three columns contain information about the ARMA process that generated the tested series and the series itself (orders p and q, series length, standard deviation of the white noise generator process). The fourth and fifth columns provide details about the optimization run: the starting point and the point where the error has been generated are characterized in terms of closeness to the feasibility boundary, according to the metrics introduced in Section 3. The sixth column provides the same information associated with the parameters of the model employed to generate the series

Model	Length	σ	Starting point	Error point	Ground truth point
ARMA(2, 1)	100	0.01	Strictly feasible	(iii)	Strictly feasible
ARMA(2, 1)	10 000	0.01	Strictly feasible	(iii)	(i)
ARMA(2, 1)	10 000	0.01	(i)	(ii)	Strictly feasible
ARMA(2, 1)	100	0.1	Strictly feasible	(iii)	Strictly feasible
ARMA(2, 1)	100	0.1	(ii)	(i)	Strictly feasible
ARMA(2, 1)	100	0.1	Strictly feasible	(i)	Strictly feasible
ARMA(2, 1)	1000	0.1	Strictly feasible	(iii)	Strictly feasible
ARMA(2, 1)	10 000	0.1	Strictly feasible	(iii)	Strictly feasible
ARMA(2, 1)	10 000	0.1	Strictly feasible	(iii)	Strictly feasible
ARMA(2, 1)	100	1	(i)	(iii)	Strictly feasible
ARMA(2, 1)	1000	1	Strictly feasible	(iii)	Strictly feasible
ARMA(2, 1)	10 000	1	Strictly feasible	(iii)	Strictly feasible
ARMA(2, 3)	10 000	1	(ii)	(iii)	Strictly feasible
ARMA(3, 2)	100	0.01	Strictly feasible	(iii)	Strictly feasible
ARMA(5, 1)	10 000	1	Strictly feasible	(i)	Strictly feasible

Table 5. Numerical errors in Kalman filtering when using model (11). The first three columns contain information about the ARMA process that generated the tested series and the series itself (orders p and q, series length, standard deviation of the white noise generator process). The fourth and fifth columns provide details about the optimization run: the starting point and the point where the error has been generated are characterized in terms of closeness to the feasibility boundary, according to the metrics introduced in Section 3. The last column provides the same information associated with the parameters of the model employed to generate the series

Model	Length	σ	Start point	Error point	Ground truth point
ARMA(4, 2)	10 000	1	Strictly feasible	(iii)	Strictly feasible
ARMA(4, 4)	1000	0.1	Strictly feasible	(iii)	(ii)
ARMA(5, 5)	100	0.1	Strictly feasible	(ii)	Strictly feasible
ARMA(5, 5)	1000	0.1	Strictly feasible	Strictly feasible	Strictly feasible

Table 6. Results from the two-sided Wilcoxon test at different horizons. Null hypothesis: the median of the differences of the MASE errors, $\text{MASE}_{\text{border}} - \text{MASE}_{\text{strictly feasible}}$, is zero

Error	Test statistic	P-Value
MASE(3)	−4.23197	2.31e−5
Scaled error(1)	−1.49874	0.13394
Scaled error(2)	−1.67521	0.09389
Scaled error(3)	−4.35523	1.33e−5

Table 7. Results from the one-sided Wilcoxon test at different horizons. Null hypothesis: the median of the differences of the MASE errors, $\text{MASE}_{\text{border}} - \text{MASE}_{\text{strictly feasible}}$, is negative

Error	Test statistic	P-Value
MASE(3)	4.23197	1.16e−5
Scaled error(1)	1.49874	0.06697
Scaled error(2)	1.67521	0.04695
Scaled error(3)	4.35523	$<$ 1e−5

Table 8. Average of ranks between different ARMA models performance w.r.t. different error metrics

Error	Jones	$\lambda = 0$	$\lambda = 1$	$\lambda = 2$	$\lambda = 4$	$\lambda = 8$	$\lambda = 16$
MASE(3)	4.228	4.201	4.056	3.947	3.882	**3.825**	3.862
Scaled error(1)	4.022	3.996	4.018	3.999	3.972	**3.968**	4.025
Scaled error(2)	4.082	4.095	4.01	3.972	3.958	**3.935**	3.948
Scaled error(3)	4.220	4.226	4.081	3.980	3.885	**3.798**	3.809

5.3. Forecasting with Almost-Border Models

As reported above, we employed again a multi-start approach to assess the predictive performance of close to the border ARMA models. For our analysis, we picked time series having at least one strictly feasible solution and at least a solution that meets one of the conditions (i), (ii), (iii). In doing so, we got a total of 614 time series with such features.

When multiple strictly feasible solutions are available, we considered the best one according to the exact log-likelihood value. The same is done when multiple solutions close to the border are obtained for a single time series. We then computed multi-step ahead predictions with the two selected models for each time-series.

Differences in predictive performance of these two distinct ARMA models are again investigated by means of the Wilcoxon signed-ranks test [8, 31]. We employed the mean absolute scaled error (MASE) [17] to measure the accuracy of forecasts. Indeed, the MASE can be used to compare forecast methods on a single series and, being scale-free, to compare forecast accuracy across series [16].

In our experiments, MASE at a given forecast horizon h is computed as

$$\text{MASE}(h) = \frac{1}{h} \frac{\sum_{t=n+1}^{h} |y_t - \hat{y}_t|}{\frac{1}{n-1}\sum_{t=2}^{n} |y_t - y_{t-1}|}. \tag{13}$$

We also reported the single absolute scaled errors for each different forecast horizon h:

$$\text{Scaled error}(h) = \frac{|y_{n+h} - \hat{y}_{n+h}|}{\frac{1}{n-1}\sum_{t=2}^{n} |y_t - y_{t-1}|}. \tag{14}$$

Table 9. Results of Friedman test for the difference in forecasting performance of various ARMA models w.r.t. different error metrics

Error	Test statistic	P-Value
MASE(3)	78.06724	$< 1e{-5}$
Scaled error(1)	1.57091	0.95465
Scaled error(2)	12.13886	0.05894
Scaled error(3)	94.93939	$< 1e{-5}$

Table 10. Posthoc analysis of the performance forecasting: pairwise comparison of the MASE(3) error

	Jones	$\lambda = 0$	$\lambda = 1$	$\lambda = 2$	$\lambda = 4$	$\lambda = 8$	$\lambda = 16$
Jones	1.00000	0.90000	0.10395	0.00100	0.00100	0.00100	0.00100
$\lambda = 0$	0.90000	1.00000	0.26546	0.00154	0.00100	0.00100	0.00100
$\lambda = 1$	0.10395	0.26546	1.00000	0.60537	0.10031	0.00630	0.04196
$\lambda = 2$	0.00100	0.00154	0.60537	1.00000	0.90000	0.48698	0.82448
$\lambda = 4$	0.00100	0.00100	0.10031	0.90000	1.00000	0.90000	0.90000
$\lambda = 8$	0.00100	0.00100	0.00630	0.48698	0.90000	1.00000	0.90000
$\lambda = 16$	0.00100	0.00100	0.04196	0.82448	0.90000	0.90000	1.00000

Results are reported in Tables 6 and 7. The observed P-value in the last row of Table 6 evidences that significant differences exist in forecast accuracy between strictly feasible ARMA (p, q) models and close-to-the-border ARMA (p, q) models. The significant differences involve only the MASE (3) error and the absolute scaled error at horizon $h = 3$: in both cases the associated P-values are strictly lower than the default significance level $\alpha = 0.05$. Furthermore, for these two metrics the one-sided test confirms that ARMA models close to the feasibility boundary perform poorer in terms of the predictive ability than the strictly feasible ARMA models.

Considering instead the remaining error metrics, results in Table 6 indicate that at forecast horizon $h = 1$ non substantial difference exists in forecast accuracy between the two types of ARMA models. Differences in predictive ability become more evident as the forecast horizon grows. From Table 6 we observe that at horizon 2, only assuming a significance level $\alpha = 0.1$, it is possible to deduce a statistically significant difference between the two ARMA models in forecasting performances.

The main conclusion of this experiment is that ARMA models close to the feasibility boundary perform poorer in terms of the predictive ability than the strictly feasible ARMA models. The practical meaning of this result is that caution is needed with close to the border ARMA models when forecasting is required. This is one of the motivations to modify our fitting model (11) by adding to the objective an ℓ_2 penalty term as in (12). We will discuss in depth the effects of this modification in the next section.

5.4. Forecasting with Regularized ARMA Models

The next and final experiment investigates the effect of the addition of an ℓ_2-regularization term from a forecasting accuracy perspective. Different values of the regularization hyperparameter λ in Eq. (12) give rise to different ARMA(p, q) models with diverse forecasting performances.

ARMA models are, in practice, fitted by iterative optimization algorithms that start at preliminary estimates obtained, for example, with the well-known Hannan and Rissanen (HR) method [14]. We consider this setting to carry out the experiment, in order to assess the impact of the regularization term in the common use cases.

Table 11. Posthoc analysis of the performance forecasting: pairwise comparison of the absolute scaled error at horizon $h = 2$

	Jones	$\lambda = 0$	$\lambda = 1$	$\lambda = 2$	$\lambda = 4$	$\lambda = 8$	$\lambda = 16$
Jones	1.00000	0.90000	0.90000	0.60131	0.46951	0.25145	0.37172
$\lambda = 0$	0.90000	1.00000	0.82448	0.48264	0.34176	0.16502	0.25839
$\lambda = 1$	0.90000	0.82448	1.00000	0.90000	0.90000	0.90000	0.90000
$\lambda = 2$	0.60131	0.48264	0.90000	1.00000	0.90000	0.90000	0.90000
$\lambda = 4$	0.46951	0.34176	0.90000	0.90000	1.00000	0.90000	0.90000
$\lambda = 8$	0.25145	0.16502	0.90000	0.90000	0.90000	1.00000	0.90000
$\lambda = 16$	0.37172	0.25839	0.90000	0.90000	0.90000	0.90000	1.00000

Table 12. Posthoc analysis of the performance forecasting: pairwise comparison of the absolute scaled error at horizon $h = 3$

	Jones	$\lambda = 0$	$\lambda = 1$	$\lambda = 2$	$\lambda = 4$	$\lambda = 8$	$\lambda = 16$
Jones	1.00000	0.90000	0.31753	0.00357	0.00100	0.00100	0.00100
$\lambda = 0$	0.90000	1.00000	0.27263	0.00259	0.00100	0.00100	0.00100
$\lambda = 1$	0.31753	0.27263	1.00000	0.67435	0.03709	0.00100	0.00100
$\lambda = 2$	0.00357	0.00259	0.67435	1.00000	0.73116	0.07136	0.11154
$\lambda = 4$	0.00100	0.00100	0.03709	0.73116	1.00000	0.80825	0.90000
$\lambda = 8$	0.00100	0.00100	0.00100	0.07136	0.80825	1.00000	0.90000
$\lambda = 16$	0.00100	0.00100	0.00100	0.11154	0.90000	0.90000	1.00000

The classical Jones fitting method is compared with models (11) and (12), varying the values of the regularization parameter λ. For each time series, all optimization algorithms are started at the same initial point, identified using HR procedure.

We employed the Friedman test [8, 10, 11] to catch the differences between the methods. The test ranks the fitting methods for each time series separately, the best performing method (lowest error) getting the rank of 1, the second best rank 2 and so on. The null-hypothesis, states that all the fitting methods are equivalent and so their ranks should be equal. Table 8 reports the average of ranks over all the time series in our dataset, w.r.t. the metrics of interest (13) and (14).

We observe from Table 8 that for the MASE(3) and the absolute scaled error at horizon $h = 3$ the averages of ranks go down until a value of the hyperparameter $\lambda = 8$. For the other two errors the trend of the averages of the ranks seems quite stationary: this pattern finds confirmation from the results of Friedman test as it is shown in Table 9.

Friedman test, whose results are reported in Table 9, suggests that the forecasting performance of the considered fitting models statistically differ (assuming a significance level of $\alpha = 0.1$) for all the errors except for the absolute scaled forecasting error at horizon $h = 1$.

Therefore, based on these results we considered necessary to conduct post hoc-analysis w.r.t. the MASE(3), the absolute scaled forecasting error at horizon $h = 3$ and 2 (although the P-value in the latter case is not negligible).

Post-hoc analysis is performed by means of the Nemenyi test [8, 25]. Critical differences between two generic methods are assessed in terms of the differences between the averages of the ranks. Results of the Nemenyi test are reported in Tables 10, 11 and 12.

Regarding the absolute scaled error at horizon $h = 2$, results from the Nemenyi test indicate no significant differences between the fitting methods in terms of the forecasting performances. All the P-values reported in Table 11 are greater than 0.1.

On the other end, results about absolute scaled error at horizon $h = 3$ and the MASE(3) are equivalent. By observing both Tables 10 and 12, no significant difference is found between the two non regularized methods. Furthermore, no significant differences in forecasting performance have been identified between both the non regularized methods and the regularized one with $\lambda = 1$.

Instead, stronger regularization leads to significantly better forecasts w.r.t. the non regularized methods. Forecasting performance, as mentioned above, starts to deteriorate as the regularization hyperparameter grows to $\lambda = 16$. In summary, the constrained fitting method with regularization leads to causal and invertible ARMA models with better short term predictive ability than the non regularized ones.

6. CONCLUSIONS

Fitting causal and invertible ARMA models by constrained optimization in the partial autocorrelation and partial moving-average coefficients space has several advantages w.r.t. the classical unconstrained approach based on the Jones reparametrization. First of all, we observed that our approach leads to a significant reduction of the fitting times. Moreover, almost-border solutions are often avoided. Such solutions, as further experiments highlight, are bad both because they lead to numerical errors during the optimization of the ARMA exact log-likelihood and because they do not perform well at forecasting.

Based on these results we proposed ℓ_2-regularization to discourage almost-border solutions. As non parametric statistical tests assess, ℓ_2-regularization also improves the short term forecasting performances of causal and invertible ARMA models.

REFERENCES

1. D. J. Aigner, "A compendium on estimation of the autoregressive moving average model from the series data," Int. Econ. Rev. **12**, 348–371 (1971).
2. C. F. Ansley, "An algorithm for the exact likelihood of a mixed autoregressive-moving average process," Biometrika **66**, 59–65 (1979).
3. C. F. Ansley and P. Newbold, "Finite sample properties of estimators for autoregressive moving average models," J. Econometr. **13**, 159–183 (1980).
4. O. Barndorff-Nielsen and G. Schou, "On the parametrization of autoregressive models by partial autocorrelations," J. Multivar. Anal. **3**, 408–419 (1973).
5. G. E. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, *Time Series Analysis: Forecasting and Control* (Wiley, Chichester, 2015).
6. P. J. Brockwell, R. A. Davis, and S. E. Fienberg, *Time Series: Theory and Methods: Theory and Methods* (Springer Science, New York, 1991).
7. P. L. Combettes and H. J. Trussell, "Best stable and invertible approximations for ARMA systems," IEEE Trans. Signal Process. **40**, 3066–3069 (1992).
8. J. Demšar, "Statistical comparisons of classifiers over multiple data sets," J. Machine Learn. Res. **7**, 1–30 (2006).
9. W. Dent, "Computation of the exact likelihood function of an arima process," J. Stat. Comput. Simul. **5**, 193–206 (1977).
10. M. Friedman, "The use of ranks to avoid the assumption of normality implicit in the analysis of variance," J. Am. Stat. Assoc. **32**, 675–701 (1937).
11. M. Friedman, "A comparison of alternative tests of significance for the problem of M ranking," Ann. Math. Stat. **11**, 86–92 (1940).
12. G. Gardner, A. C. Harvey, and G. D. Phillips, "Algorithm as 154: An algorithm for exact maximum likelihood estimation of autoregressive-moving average models by means of Kalman filtering," J. R. Stat. Soc., Ser. C **29**, 311–322 (1980).
13. J. D. Hamilton, *Time Series Analysis* (Princeton, New Jersey, 1994), Vol. 2.
14. E. J. Hannan and J. Rissanen, "Recursive estimation of mixed autoregressive-moving average order," Biometrika **69**, 81–94 (1982).
15. A. C. Harvey and G. D. Phillips, "Maximum likelihood estimation of regression models with autoregressive-moving average disturbances," Biometrika **66**, 49–58 (1979).
16. R. J. Hyndman et al., "Another look at forecast-accuracy metrics for intermittent demand," Int. J. Appl. Forecast. **4** (4), 43–46 (2006).
17. R. J. Hyndman and A. B. Koehler, "Another look at measures of forecast accuracy," Int. J. Forecast. **22**, 679–688 (2006).

18. M. Jones, "Randomly choosing parameters from the stationarity and invertibility region of autoregressive-moving average models," J. R. Stat. Soc., Ser. C **36**, 134–138 (1987).
19. R. H. Jones, "Maximum likelihood fitting of arma models to time series with missing observations," Technometrics **22**, 389–395 (1980).
20. R. E. Kalman, "A new approach to linear filtering and prediction problems," J. Basic Eng. **82**, 35–45 (1960).
21. K. M. Kang, "A comparison of estimators for moving average processes," Unpublished paper (1975).
22. C.-J. Kim and J. Kim, "Thepile-up problem'in trend-cycle decomposition of real GDP: Classical and bayesian perspectives," MPRA Paper No. 51118 (Univ. Library of Munich, 2013).
23. J. Marriott "Bayesian analysis of arma processes: Complete sampling-based inferences under full likelihood," in *Bayesian Statistics and Econometrics: Essays in Honor of Arnold Zellner,* Ed. by D. A. Berry, K. M. Chaloner, and J. K. Geweke, *Wiley Series in Probability and Statistics* (Wiley, Chichester, 1995).
24. J. F. Monahan, "A note on enforcing stationarity in autoregressive-moving average models," Biometrika **71**, 403–404 (1984).
25. P. Nemenyi, "Distribution-free multiple comparisons," Biometrics **18**, 263 (1962).
26. P. Newbold, "The exact likelihood function for a mixed autoregressive-moving average process," Biometrika **61**, 423–426 (1974).
27. B. Picinbono and M. Benidir, "Some properties of lattice autoregressive filters," IEEE Trans. Acoust. Speech, Signal Process. **34**, 342–349 (1986).
28. J. D. Sargan and A. Bhargava, "Maximum likelihood estimation of regression models with first order moving average errors when the root lies on the unit circle," Econometrica **51**, 799–820 (1983).
29. S. Shlien, "A geometric description of stable linear predictive coding digital filters (corresp.)," IEEE Trans. Inform. Theory **31**, 545–548 (1985).
30. F. Tusell et al., "Kalman filtering in R," J. Stat. Software **39** (2), 1–27 (2011).
31. F. Wilcoxon, "Individual comparisons by ranking methods," in *Breakthroughs in Statistics* (Springer, Berlin, 1992), pp. 196–202.
32. Y. Zhang and A. I. McLeod, "Fitting ma (q) models in the closed invertible region," Stat. Probab. Lett. **76**, 1331–1334 (2006).