Penalised inference for autoregressive moving average models with time-dependent predictors

Hamed Haselimashhadi and Veronica Vinciotti
Department of Mathematics, Brunel University London, UK

Abstract: Linear models that contain a time-dependent response and explanatory variables have attracted much interest in recent years. The most general form of the existing approaches is of a linear regression model with autoregressive moving average residuals. The addition of the moving average component results in a complex model with a very challenging implementation. In this paper, we propose to account for the time dependency in the data by explicitly adding autoregressive terms of the response variable in the linear model. In addition, we consider an autoregressive process for the errors in order to capture complex dynamic relationships parsimoniously. To broaden the application of the model, we present an l_1 penalized likelihood approach for the estimation of the parameters and show how the adaptive lasso penalties lead to an estimator which enjoys the oracle property. Furthermore, we prove the consistency of the estimators with respect to the mean squared prediction error in high-dimensional settings, an aspect that has not been considered by the existing time-dependent regression models. A simulation study and real data analysis show the successful applications of the model on financial data on stock indexes.
Keywords: time series, high dimensional models, lasso

1 Introduction

This paper deals with fitting a general time series-regression model using l_1 regularized inference. In the context of linear models, l_1 penalized approaches have received great interest in recent years as they allow to perform variable selection and parameter estimation simultaneously for any data, including high-dimensional datasets, where classical approaches for parameter estimation break down, e.g. [14, 7, 10, 17, 12]. [17] have shown that a model where penalties are adapted to each individual regressor enjoys oracle properties. Most of the advances in regularized regression models have been for the case of independent and identically distributed data. A recent line of research has concentrated on regularized models in time dependent frameworks. Amongst

these, [15] showed the successful application of l_1 penalised inference in the context of autocorrelated residuals for a fixed order, by proposing the model

$$y_t = \sum_{i=1}^{\mathrm{r}} x'_{ti}\beta_i + \sum_{j=1}^{q} \theta_j \epsilon_{t-j} + e_t,$$

and studied the properties of this model in low-dimensional settings. [11] studied the theoretical properties of a regularized autoregressive process on Y_t for both low and dimensional cases, whereas [13] studied the l_1 estimation of vector autoregressive models. In both cases, no exogenous variables are included in the model. [9] studied the asymptotic properties of adaptive lasso in high dimensional time series models when the number of variables increases as a function of the number of observations. Their model covers a lagged regression in the presence of exogenous variables, but does not consider autocorrelated residuals. Recently, [16] proposed an extension of the model of [15] by adding a moving average term, that is they propose a model of the form

$$y_t = \sum_{i=1}^{\mathrm{r}} x'_{ti}\beta_i + \epsilon_t, \qquad \epsilon_t = \sum_{j=1}^{q} \theta_j \epsilon_{t-j} + e_t + \sum_{j=1}^{q} \phi_j e_{t-j}.$$

Similarly to [15], they proved the consistency of the model in low-dimensional cases. Despite the generality of this model, considering an ARMA process for the errors results in a complex model with a challenging implementation.

In this paper, we propose to account for the time dependency in the data by explicitly adding autoregressive terms of the response variable in the linear model, as in [11], as well as an autocorrelated process for residuals, as in [15], in order to capture complex dynamics parsimoniously. In particular, given fixed orders p and q, we propose the model

$$y_t = x'_t\beta + \sum_{j=1}^{p} \phi_j y_{t-j} + \sum_{i=1}^{q} \theta_i \epsilon_{t-i} + e_t. \tag{1}$$

We name the terms in the right hand side of (1) as **REG**ression term, **A**uto**R**egressive term and **M**oving **A**verage term respectively and call the resulting model REGARMA. We assume that all time dependent components in REGARMA are stationary and ergodic. Figure (1) illustrates a schematic view of the model.

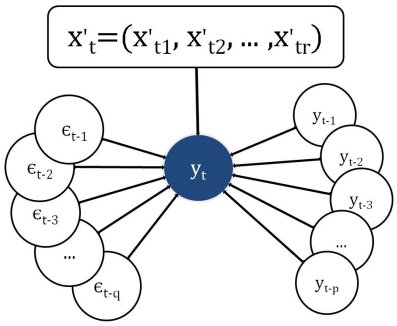

Figure 1: Schematic illustration of a REGARMA(p,q) model

In Section 2, we formulate the model and present an l_1 penalized likelihood approach for the estimation of the parameters. In Section 3, we prove the asymptotic properties of the model and show how the adaptive lasso penalty leads to estimators which enjoy the oracle property. Furthermore, we prove the consistency of the model with respect to the mean squared prediction error in high-dimensional settings, an aspect that has not been considered by the existing time-dependent regression models. In section 4, we discuss the implementation of REGARMA. A simulation study, given in section 5, will accompany the theoretical results. In section 6 we apply the model to two real datasets in finance and macroeconomic, respectively. Finally, we draw some conclusions in section 7.

2 L_1 penalised parameter estimation of REGARMA

The general form of REGARMA consists of a lagged response variable, covariates and autocorrelated residuals. Consider the following Gaussian REGARMA model of order p and q,

$$y_t = x_t'\beta + \sum_{j=1}^{p} \phi_j y_{t-j} + \sum_{i=1}^{q} \theta_i \epsilon_{t-i} + e_t, \qquad e_t \overset{iid}{\sim} N(0,\sigma^2), \qquad t = 1,2,3,\ldots,T$$

where x_t' is the t^{th} row of the matrix of r predictors $X'_{T\times r}$, $\{y_t\}$ and $\{\epsilon_t\}$ follow stationary time series processes, that is all roots of the polynomials

$1 - \sum_{i=0}^{p} \phi_i L^i$ and $1 - \sum_{i=0}^{q} \theta_i L^i$ are unequal and outside the unit circle, e_ts are independent and identical Gaussian noises with mean of zero and finite fourth moments, and p and q are both less than the number of observations T. Moreover, we assume that the errors and explanatory variables in X are independent of each other. To remove the constants from the model we follow the literature on regularized models, e.g. [14, 5], and standardize the covariates and response to zero means and unit variance.

Given the first $T_\circ = p + q$ observations, maximizing the l_1 penalized conditional likelihood of the model is equivalent to minimizing

$$Q_n(\Theta) = \sum_{t=T_\circ+1}^{T} \left((y_t - x_t'\beta) - \sum_{i=1}^{p} \phi_i y_{t-i} - \sum_{j=1}^{q} \theta_j \epsilon_{t-j} \right)^2 + \sum_{i=1}^{r} \lambda|\beta_i| + \sum_{j=1}^{p} \gamma|\phi_j| + \sum_{k=1}^{q} \tau|\theta_k| \tag{2}$$

where λ, γ, τ are tuning parameters and $\Theta = (\beta', \phi', \theta')$ is the vector of regression, autoregressive and moving average parameters. Following the literature, and given the superior properties of adaptive lasso models [17], we also propose an adaptive version of REGARMA penalised estimation as follows

$$Q_n^*(\Theta) = \sum_{t=T_\circ+1}^{T} \left((y_t - x_t'\beta) - \sum_{i=1}^{p} \phi_i y_{t-i} - \sum_{j=1}^{q} \theta_j \epsilon_{t-j} \right)^2 + \sum_{i=1}^{r} \lambda_i^*|\beta_i| + \sum_{j=1}^{p} \gamma_j^*|\phi_j| + \sum_{k=1}^{q} \tau_k^*|\theta_k|$$

where $\lambda_i^*, \gamma_j^*, \tau_k^*, i = 1, 2, \ldots, r; j = 1, 2, \ldots, p; k = 1, 2, \ldots, q$ are tuning parameters.

2.1 Matrix representation of the model

For convenience, we write the model in matrix representation. Let $H' = (H_{(p)}, H_{(q)}, X')$ be a $n \times (p + q + r)$ matrix including lags of autoregressive $(H_{(p)})$, moving average $(H_{(q)})$, and explanatory variables (X'). Let $\Theta = (\phi', \theta', \beta')$ denote the vector of corresponding parameters, $e' = (e_{T_0+1}, e_2, \ldots, e_T)$ be the vector of errors, $T_\circ = p + q$ and $n = T - T_\circ$, as previously defined. Then, in matrix form, the model can be written as

$$Y = H'\Theta + e$$

and the l_1 penalized conditional likelihood given the first T_0 observation is equivalent to

$$Q_n(\Theta) = L(\Theta) + \lambda'|\beta| + \gamma'|\phi| + \tau'|\theta|,$$

where $L(\Theta) = e'e, \lambda' = \{\lambda\}_{1\times r}, \gamma' = \{\gamma\}_{1\times p}, \tau' = \{\tau\}_{1\times q}$. Similarly, the adaptive form of the model is given by

$$Q_n^*(\Theta) = L(\Theta) + \lambda'^*|\beta| + \gamma'^*|\phi| + \tau'^*|\theta|, \tag{3}$$

where the parameters are given by

$\lambda^{*'} = (\lambda_1^*, \lambda_2^*, \ldots, \lambda_r^*), \gamma^{*'} = (\gamma_1^*, \gamma_2^*, \ldots, \gamma_p^*), \tau^{*'} = (\tau_1^*, \tau_2^*, \ldots, \tau_q^*), \Theta = (\beta', \phi', \theta')$.

3 Theoretical properties of REGARMA and adaptive-REGARMA

In order to study the theoretical properties of REGARMA and adaptive-REGARMA, we define the true coefficients by $\Theta^\circ = (\beta^{\circ'}, \phi^{\circ'}, \theta^{\circ'})$ and assume that some of these coefficients are zero. The indexes of non-zero coefficients in each group of coefficients, β, ϕ and θ, are denoted by s_1, s_2 and s_3 respectively, whereas s_1^c, s_2^c, s_3^c are the complementary sets and contain the indexes of zero coefficients. We also define $\beta_{s_1}^\circ, \phi_{s_2}^\circ, \theta_{s_3}^\circ$ and their corresponding (REGARMA) estimations by $\hat{\beta}_{s_1}, \hat{\phi}_{s_2}, \hat{\theta}_{s_3}$. Similarly, adaptive-REGARMA estimations are denoted by $\hat{\beta}_{s_1}^*, \hat{\phi}_{s_2}^*, \hat{\theta}_{s_3}^*$. Finally, different combinations of model parameters are going to be used, with obvious meaning, in particular $\Theta_1^\circ = \{\beta_{s_1}^{\circ\prime}, \phi_{s_2}^{\circ\prime}, \theta_{s_3}^{\circ\prime}\}$, $\Theta_2^\circ = \{\beta_{s_1^c}^{\circ\prime}, \phi_{s_2^c}^{\circ\prime}, \theta_{s_3^c}^{\circ\prime}\}$, $\hat{\Theta}_1 = \{\hat{\beta}'_{s_1}, \hat{\phi}'_{s_2}, \hat{\theta}'_{s_3}\}$, $\hat{\Theta}_2 = \{\hat{\beta}'_{s_1^c}, \hat{\phi}'_{s_2^c}, \hat{\theta}'_{s_3^c}\}$, $\hat{\Theta}_1^* = \{\hat{\beta}_{s_1}^{*\prime}, \hat{\phi}_{s_2}^{*\prime}, \hat{\theta}_{s_3}^{*\prime}\}$, $\hat{\Theta}_2^* = \{\hat{\beta}_{s_1^c}^{*\prime}, \hat{\phi}_{s_2^c}^{*\prime}, \hat{\theta}_{s_3^c}^{*\prime}\}$.

3.1 Assumptions

To prove the theoretical properties of the estimators, in line with the literature, we make use of the following assumptions:

(a). e_ts are i.i.d Gaussian random variables with finite fourth moments

(b). The covariates, $X_i, i = 1, 2, 3, \ldots, r$, and response variable, Y, are *stationary* and *ergodic* with finite second order moments. Also, we assume that none of the roots of $1 - \sum_{i=1}^p \phi_i L^i$ and/or $1 - \sum_{j=1}^q \theta_j L^i$ are equal and outside of the unit circle

(c). $X_i, i = 1, 2, 3, \ldots, r$ are independent of the errors

(d). $\frac{1}{n} X'X \to_{a.s} \mathbb{E}(X'X) < \infty$ and $\max_{1\leq i\leq r} X_i'X_i < \infty$.

Assumptions $(a)-(b)$ are standard assumptions for dealing with stationary time series. Assumption $(c)-(d)$ are used to guarantee that explanatory variables have finite expectations.

3.2 Theoretical properties of REGARMA when $r < n$

In the following theorems, we extend the theorems of [15] to cover a model with a lagged response.

Theorem 1. *Assume* $\lambda_n\sqrt{n} \to \lambda_\circ$, $\gamma_n\sqrt{n} \to \gamma_\circ$, $\tau_n\sqrt{n} \to \tau_\circ$ *and* $\lambda_\circ, \gamma_\circ, \tau_\circ \geq 0$. *Then under assumptions* $a-d$, *it follows that* $\sqrt{n}(\hat{\Theta}-\Theta^\circ) \xrightarrow{d} \arg\min(k(\delta))$ *where*

$$\begin{aligned} k(\delta) = -2\delta' W + \delta' U_B \delta &+ \lambda_\circ \sum_{i=1}^{r} \{u_i sign(\beta_i^\circ) I(\beta_i^\circ \neq 0) + |u_i| I(\beta_i^\circ = 0)\} \\ &+ \gamma_\circ \sum_{j=1}^{p} \{v_j sign(\phi_j^\circ) I(\phi_j^\circ \neq 0) + |v_i| I(\phi_j^\circ = 0)\} \\ &+ \tau_\circ \sum_{k=1}^{q} \{w_k sign(\theta_k^\circ) I(\theta_k^\circ \neq 0) + |w_k| I(\theta_k^\circ = 0)\} \end{aligned}$$

with $\delta = (u', v', w')$ *is a vector of parameters in* $\mathbb{R}^{(r+p+q)}$, $W \sim MVN(O, \sigma^2 U_B)$ *and* $U_B = \mathbb{E}(HH')$.

The proof is given in the Appendix. Theorem (1) shows that the REGARMA estimator has a Knight-Fu type asymptotic property [7] and it implies that the tuning parameters in $Q_n(\Theta)$ cannot shrink to zero at a speed faster than $n^{-1/2}$. Otherwise, $\{\lambda_\circ, \gamma_\circ, \tau_\circ\}$ are zero and $k(\delta)$ becomes a standard quadratic function,

$$k(\delta) = -2\delta W + \delta' U_B \delta$$

which does not produce a sparse solution. In addition, the proof of theorem (1) requires the errors to be independent and identically distributed but we do not make a strong assumption on the type of distribution for the errors, due to the use of the martingale central limit theorem for large n.

[7] proves that a lasso optimization returns estimates of non-zero parameters that suffer an asymptotic bias. This applies also to the REGARMA model, as we show with the following remark.

Remark 1. *Consider a special case of REGARMA when* $\beta_i^\circ > 0, \quad 1 \leq i \leq r$ *but* $\theta_{j_1}^\circ = 0$ *and* $\phi_{j_2}^\circ = 0$ *for* $1 \leq j_1 \leq q$, $1 \leq j_2 \leq p$, $j_1, j_2 \in \mathbb{N}$. *If minimizing* $k(\delta)$ *can correctly identify* Θ, *it means that* $u \neq 0$ *and* $v, w = 0$. *That is,* $k(\delta)$ *must satisfy*

$$\begin{aligned}
\frac{\partial k(\delta)}{\partial u} &= \frac{\partial k(u,0,0)}{\partial u} \\
&= \frac{\partial}{\partial u}\Big(-2(u',0,0)W + (u',0,0)'U_B(u',0,0) + (n\lambda_n'|\beta^\circ + \frac{u}{\sqrt{n}}| - n\lambda_n'|\beta^\circ|)\Big) \\
&= -2W_{1:r} + 2u'U_{B_{1:r}} + \lambda_\circ 1_{r\times 1} = 0 \\
&\implies u' = \frac{1}{2}(2W_{1:r} - \lambda_\circ 1_{r\times 1})U_{B_{1:r}}^{-1}.
\end{aligned}$$

Then using Theorem 1, $\sqrt{n}(\hat{\beta} - \beta^\circ) \xrightarrow{d} \arg\min(k(\delta = u')) = MVN\Big(\mathbb{E}(u') \neq 0,\, U_{B_{1:r}}^{-1}\Big)$, *where* $U_{B_{1:r}}^{-1}$ *is the matrix with the first* r *rows of* U_B *corresponding to the* r *covariates.*

If $\lambda_\circ, \gamma_\circ$ and $\tau_\circ$ are positive, remark (1) shows that $Q_n(\Theta)$ suffers an asymptotic bias and is different from the oracle estimator, $MVN\Big(O, U_{B_{1:r}}^{-1}\Big)$. In other words, REGARMA is not asymptotically consistent unless $\lambda_\circ, \gamma_\circ, \tau_\circ \underset{n\to\infty}{\to} 0$. The following remark can be extended to other groups of coefficients.

3.3 Theoretical properties of adaptive-REGARMA when $r < n$

Following the notation of section 2, we consider the adaptive version of the penalised likelihood and estimate the model parameters by minimizing

$$Q_n^*(\Theta) = L_n(\Theta) + n\lambda'^*|\beta| + n\gamma'^*|\phi| + n\tau'^*|\theta|$$

where

$$\begin{aligned}
L_n(\Theta) &= \Big(Y - X'\beta - H_{(p)}\phi + H_{(q)}\theta\Big)'\Big(Y - X'\beta - H_{(p)}\phi + H_{(q)}\theta\Big) \\
\lambda^{*'} &= \{\lambda^*\}_{r\times 1}', \; \gamma^{*'} = \{\gamma^*\}_{p\times 1}', \; \tau^{*'} = \{\tau^*\}_{q\times 1}', \; \Theta = (\beta', \phi', \theta').
\end{aligned}$$

Following [15] and [3], we define the maximum and minimum penalties for significant and insignificant coefficients by

$$a_n = \max(\lambda^*_{i_1}, \gamma^*_{i_2}, \tau^*_{i_3}; \quad i_1 \in s_1, i_2 \in s_2, i_3 \in s_3),$$
$$b_n = min(\lambda^*_{i^c_1}, \gamma^*_{i^c_2}, \tau^*_{i^c_3}; \quad i^c_1 \in s^c_1, i^c_2 \in s^c_2, i^c_3 \in s^c_3),$$

and prove a number of results on the theoretical properties of adaptive REGARMA.

Theorem 2. *Assume $a_n = o(1)$ as $n \to \infty$. Then under assumptions $a - d$, there is a local minimiser $\hat{\Theta}^*$ of $Q^*_n(\Theta)$ such that*

$$(\hat{\Theta}^* - \Theta^\circ) = O_p(n^{-1/2} + a_n).$$

The proof of the theorem is in the Appendix. Let $\alpha_n = a_n + n^{-1/2}$, then, theorem (2) proves that there exists a $\sqrt{n} - consistent$ local minimiser $Q^*_n(\Theta)$, when the tuning parameters (for significant variables) of REGARMA converge to zero at the speed faster than $n^{-1/2}$ (since $n\alpha^2_n \to o(1)$).

As the next step, we prove that if the tuning parameter associated with insignificant variables in REGARMA shrink to zero at a speed slower than $n^{-1/2}$, then their associated REGARMA coefficients will be estimated exactly equal to zero with probability tending to 1.

Theorem 3. *Assume $b_n\sqrt{n} \to \infty$ and $||\hat{\Theta}^* - \Theta^\circ|| = O_p(n^{-1/2})$ then*

$$Pr(\hat{\beta}^*_{s^c_1} = 0) \to 1, \quad Pr(\hat{\phi}^*_{s^c_2} = 0) \quad \to 1, \quad Pr(\hat{\theta}^*_{s^c_3} = 0) \to 1.$$

The proof of the theorem is in the Appendix. Theorem (2) and (3) indicate that $\sqrt{n} - consistent$ estimator $\hat{\Theta}^*$ satisfies $Pr(\hat{\Theta}^*_2 = 0) \to 1$ under certain conditions on the tuning parameters, leading to the following result:

Theorem 4. *Assume $a_n\sqrt{n} \to 0$ and $b_n\sqrt{n} \to \infty$. Then, under assumptions $a - d$, the component $\hat{\Theta}^*_1$ of the local minimiser of $\hat{\Theta}^*$ in Theorem 3 satisfies*

$$\sqrt{n}(\hat{\Theta}^*_1 - \Theta^\circ_1) \xrightarrow{d} MVN(O, \sigma^2 U_0^{-1})$$

where U_0 is the sub-matrix U_B corresponding to Θ°_1.

The proof of the theorem is in the Appendix. Theorem (4) implies that if a_n tends to zero at the speed faster than $\sqrt{n}$ and simultaneously b_n increases at the speed slower than $\sqrt{n}$, then adaptive REGARMA is asymptotically an oracle estimator. In the next subsection, we consider the theoretical properties of adaptive REGARMA for high-dimensional problems.

3.4 Theoretical properties of adaptive REGARMA when $n \ll r$

In the proofs of the low-dimensional results (refer to proof of theorem 1), we rely on a unique path of reaching the maximum of the log likelihood. This is not true in high-dimensional cases, so different results are needed in this case. In this section we follow a similar strategy to [2] to prove theorems in the high dimensional case, an aspect which has not been considered by existing time-dependent regression models, such as those of [15] and [16].

In order to study the consistency of REGARMA in high-dimensional situations, we show that under assumptions $a-d$, REGARMA is consistent with respect to the mean squared prediction error.
Without loss of generality, we define the REGARMA model as a constrained optimization [14]. Thus, we have

$$\begin{aligned} &\min\{(y - X'\beta - H_{(p)}\phi - H_{(q)}\theta)'(y - X'\beta - H_{(p)}\phi - H_{(q)}\theta)\} \\ &\text{Subject to } \sum_{j=1}^{r}|\beta_j| \le K_\lambda, \quad \sum_{k=1}^{p}|\phi_k| \le K_\gamma, \quad \sum_{l=1}^{q}|\theta_l| \le K_\tau, \qquad (4) \\ &\textit{and} \qquad K_\lambda \ge 0, \quad K_\gamma \ge 0, \quad K_\tau \ge 0 \end{aligned}$$

where there is a one-to-one correspondence between λ, γ and τ in REGARMA, and K_λ, K_γ and K_τ in (4). Define the *Mean Squared Prediction Error*, $(MSPE)$, and its estimated value, $\widehat{MSPE}$, by

$$MSPE(\hat\beta, \hat\phi, \hat\theta) = \mathbb{E}(\|\hat Y - Y^\circ\|^2), \qquad \widehat{MSPE}(\hat\beta, \hat\phi, \hat\theta) = \frac{1}{n}\|\hat Y - Y^\circ\|^2$$

where Y° and $\hat Y$ are the REGARMA predictions of Y based on the true parameters $(\beta^\circ, \phi^\circ, \theta^\circ)$ and REGARMA estimates $(\hat\beta, \hat\phi, \hat\theta)$ from (4), respectively. Then the following theorem holds.

Theorem 5. *Under assumptions $a-d$ and $\|X\|_\infty \le M_1, \|H_{(p)}\|_\infty \le M_2$, $\|H_{(q)}\|_\infty \le M_3$ and $M_{max} = \sup\{M_1, M_2, M_3\}$, let $\hat\beta, \hat\phi$ and $\hat\theta$ be the REGARMA estimates, and $K_{max} = \sup\{K_\lambda, K_\gamma, K_\tau\}$ such that*

$$\sum_{j=1}^{r}|\beta_j| \le K_\lambda < \infty, \quad \sum_{k=1}^{p}|\phi_k| \le K_\gamma < \infty, \quad \sum_{l=1}^{q}|\theta_l| \le K_\tau < \infty.$$

Then

$$\widehat{MSPE}(\hat{\beta},\hat{\phi},\hat{\theta}) \leq \frac{2K_{max}M_{max}\sigma}{\sqrt{n}}\left(\sqrt{2\log(2r)}+\sqrt{2\log(2p)}+\sqrt{2\log(2q)}\right). \quad (5)$$

The proof of the theorem is in the Appendix. Note that in the situation where $p = 0$ and $q = 0$, equation (5) results in the standard lasso consistency formula in [2]. When K_{max} is correctly chosen, equation (5) also shows that REGARMA is prediction consistent when $\max\{\log(r), \log(p), \log(q)\} \ll n$. It is also possible to extend this result to $MSPE$.

Remark 2. *Under the same conditions as Theorem (5),*

$$MSPE(\hat{\beta},\hat{\phi},\hat{\theta}) \leq \frac{2K_{max}M_{max}\sigma}{\sqrt{n}}\sum_{i=1}^{3}\left(\sqrt{2\log(2a_i)}\right)+8K^*\sum_{i,j=1}^{3}\left(M_iM_j\sqrt{\frac{2log(2a_ia_j)}{n}}\right),$$

where $K_{\max}$ *and* M_{max} *are defined as before,* K^* *is defined in the Appendix and* $a_1 = r, a_2 = p$ *and* $a_3 = q$.

Remark (2) shows that if $\underset{i,j=1,2,3}{max}\log(a_ia_j) \ll n$ then REGARMA is consistent. Given that relatively small orders p and q are sufficient for most time series analyses, the consistency of the estimator is mainly dominated by the high-dimensional regression part. If $M_1 \geq M_2 \geq M_3$ then the above equation approximately reduces to a form similar to the standard lasso results in [2],

$$MSPE(\hat{\beta},\hat{\phi},\hat{\theta}) \leq \frac{2K_\lambda M_1\sigma}{\sqrt{n}}\sqrt{2\log(2r)}+8K^*M_1^2\sqrt{\frac{2log(2r^2)}{n}}.$$

4 Algorithm

Since $Q_n(\Theta) \subseteq Q_n^*(\theta)$, that is REGARMA is a subset of adaptive-REGARMA, and given the improved properties of adaptive REGARMA, we mainly focus on adaptive REGARMA in this section. Our formulation of the model lends itself naturally to its implementation, in contrast to the more complex implementation of the model of [16].

As the model contains regressions, moving averages and autoregressive coefficients, we use the two-step optimization procedure

$$\text{First step: } \hat{\epsilon} = Y - X'\hat{\beta} - H_{(p)}\hat{\phi}, \quad \text{Second step: } Y = X'\beta + H_{(p)}\phi + \hat{H}_{(q)}\theta.$$

Steps 1 and 2 provide a solution to REGARMA using the adaptive-Lasso algorithm of [17].

In terms of the selection of the penalties λ, γ and τ, these can be chosen using K-fold cross-validation or using an information criterion such as BIC or AIC, CP similarly to [15], [16] and [4]. The weights in adaptive-REGARMA are defined by using the (non-adaptive) REGARMA estimates. Some notes are needed about the selection of the orders p and q in the REGARMA model. We propose two general approaches to choose the optimal orders for the model: (a) setting an upper bound P and Q and choosing the model that minimizes BIC or AIC inside these bounds (b) setting an upper bound P and Q and letting the model choose the best orders by keeping or eliminating the time series coefficients under L_1 sparsity constraints. These two approaches are very similar but there is a slight difference between them: in the second approach, the fitting is based on $n = T - (P + Q)$ time points, whereas in the first approach, the number of time points depends on the orders p and q. Then a rule of thumb is to use the first approach when the number of observations is low and choose the second approach when there are enough observations.

We are in the process of implementing the methods into an R package. This is particularly needed in this area as, to the best of our knowledge, there is no implementation available. The current version of the package is available at http://people.brunel.ac.uk/~mastvvv/Software/.

5 Simulation study

We design a simulation study to compare the REGARMA model with existing methods. In the simulation, we:

1. Set the proportion of zero coefficients to 90%, 50% or 10%.
2. Assign unequal random numbers in $(-1, 1)$ to each non-zero coefficient.
3. Generate the design matrix, X, using stationary Gaussian processes, with $r = 25, 75, 200, 300, 400$ and $T = 50, 100, 150, 200, 250$.
4. Generate $e \sim \sigma \times N(0, 1)$ where $\sigma \in \{0.5, 1, 1.5\}$,

5. Set unequal AR and MA parameters, under the constraint that the roots of stationary polynomials site outside the unit circle and simulate data from a REGARMA model with $p \leq 3$ and $q \leq 3$.

6. Repeat each combination of models 10 times.

We compare the adaptive REGARMA model with adaptive lasso, as it is the closest model in the literature for which an implementation is available. Similar results were found in the comparison of the non-adaptive versions (results not reported). BIC was used to choose the optimal penalties, whereas the autoregressive and moving average orders were fixed as the true ones.

Figure (2) to (5) compare adaptive REGARMA and adaptive Lasso with respect to mean squared prediction error, BIC and mean squared error of $\hat{\beta}$ for $n = 50, 100, 150, 200, 250$, $\sigma = 0.5, 1, 1.5$ and $r = 25, 75, 200, 300, 400$. The figures show overall how REGARMA dominates lasso both for low and high-dimensional problems. Figure (2) shows that as the number of data points T increases, the relative outperformance of REGARMA versus lasso with respect to MSPE increases. Figures (3) shows how REGARMA achieves lower BIC values than lasso, particularly when $T < r$ but also for some high-dimensional cases. Figure (4) compares REGARMA and lasso with respect to the mean squared error of $\hat{\beta}$, averaged over the different regression coefficients. This plot shows the advantage of using REGARMA on time-dependent data in comparison with lasso. Finally, Figure (5) shows an outperformance of REGARMA over lasso, regardless of the level of noise σ.

6 Real data analysis

For the first application, we consider REGARMA in a low-dimensional problem. In particular, we consider financial data on daily returns of the *Istanbul Stock Exchange*(ISE) with seven other international indices, *SP, DAX, FTSE, NIKKEI, BOVESPA, MSCE EU, MSCI EM*, for a period of two years from 2009 to 2011. The data are publicly available at `http://archive.ics.uci.edu/ml` and are considered also by [1]. The goal of the analysis is to detect the most effective indices in relation to the ISE index.
We set a maximum order of 4 for both p and q and use BIC to select the optimal penalty parameters (i.e. method b on page 11). Table (1) shows a comparison of REGARMA with adaptive lasso. For REGARMA, we consider

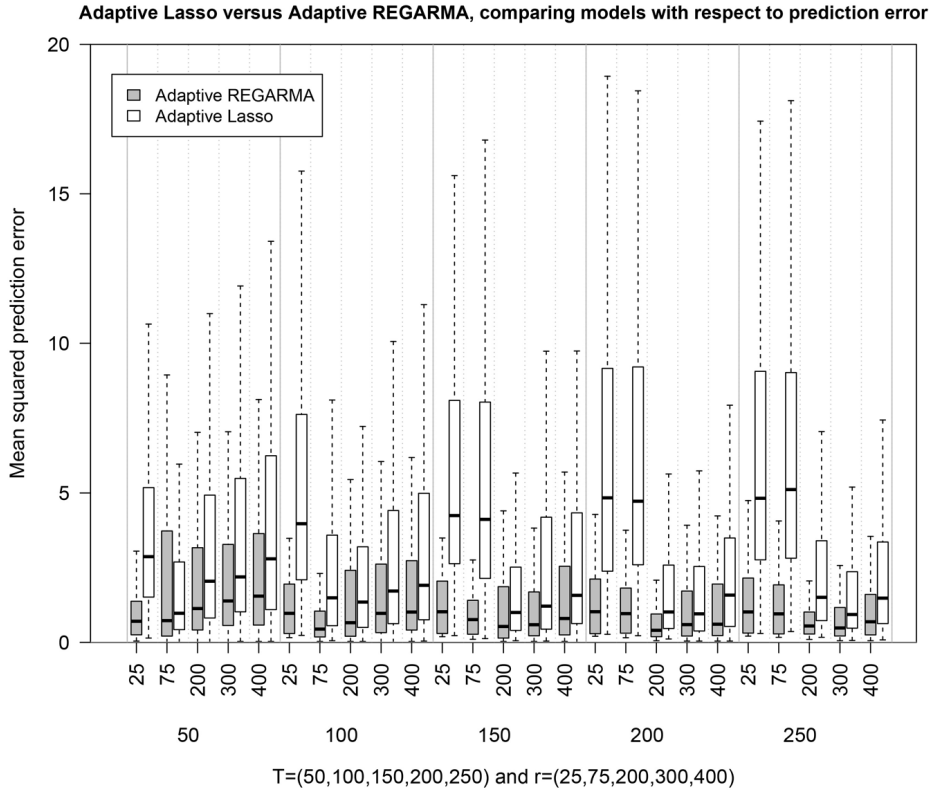

Figure 2: Comparison of adaptive lasso and adaptive REGARMA with respect to mean squared prediction error on simulated data with different values of r and T.

also the sub-model with only autoregressive terms, REGAR, and the one with only moving average terms, REGMA (which is essentially the model of [15]), as well as the full REGARMA model. All four models choose BOVESPA, EU and EM as the most effective indices for the Istanbul exchange market. These are within the 6 variables selected by [1]. From Table (1) and the residual analysis in Figure (6), we can conclude that the REGARMA family shows a better performance with respect to Mean Squared Error (MSE), Mean Absolute Error (MAE) and BIC compared to adaptive lasso.

For a hight dimensional example, we consider S&P500 indices. S&P500 is one of the leading stock market index for US equity: it is based on 500 leading companies and captures approximately 80% coverage of the available

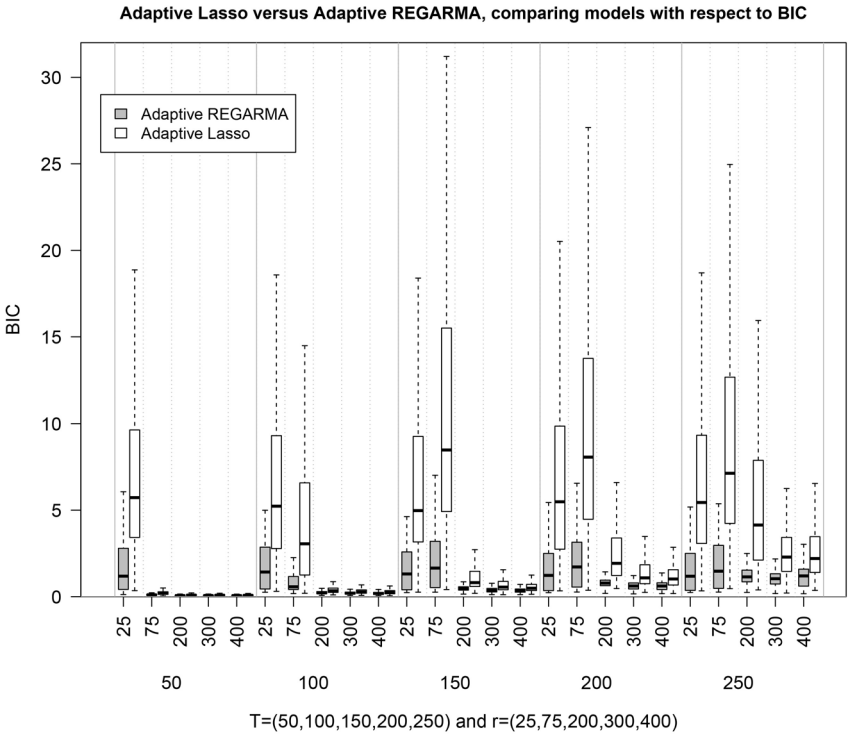

Figure 3: Comparison of adaptive lasso and adaptive REGARMA with respect to BIC on simulated data with different values of r and T.

market capitalization. The goal of the analysis is to find the S&P500 indices most related to the *AT&T Inc* index based on monthly data in a period of fourteen years from 2000 to 2014. These data are publicly available at http://thomsonreuters.com/. After removing variables with the majority of missing values, the dataset contains 416 variables and 170 datapoints. As before, we apply adaptive lasso and adaptive REGARMA to these data and choose the optimal penalties by BIC. Moreover, we set a maximum order of 4 for both p and q and let the model choose the optimal orders (using method *b* on page 11).

Table (2) summarises the results in terms of MSE, MAE, BIC and the number of non-zero coefficients. Moreover, Figure (7) illustrates the residual analysis of these four models. Both Table (2) and Figure (7) show an improved performance of REGARMA compared to the other methods.

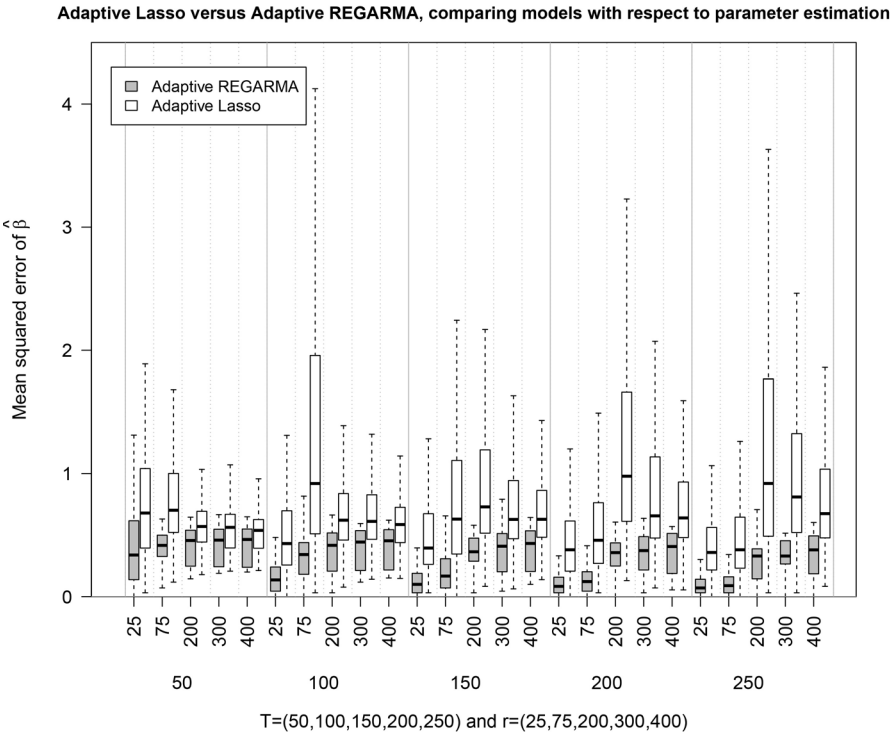

Figure 4: Comparison of adaptive Lasso and adaptive REGARMA with respect to mean squared error of $\hat{\beta}$ on simulated data for different values of r and T.

7 Conclusion

In this paper we extend the idea of regression-time series models proposed in [15] to a more general class of models, thus covering a wide spectrum of applications involving multivariate time-dependent data. In particular, we study an autoregressive moving average model with time-dependent explanatory variables and present l_1 penalised inference for the estimation of its parameters. Our model lends itself naturally to parameter estimation and implementation, contrary to the linear regression with ARMA errors of [16]. We prove asymptotic properties of the proposed model in low and high dimensional situations, with the latter not considered by the existing literature on time-series regression models. We test the performance of the model on a simulation study and show a successful application on financial data.

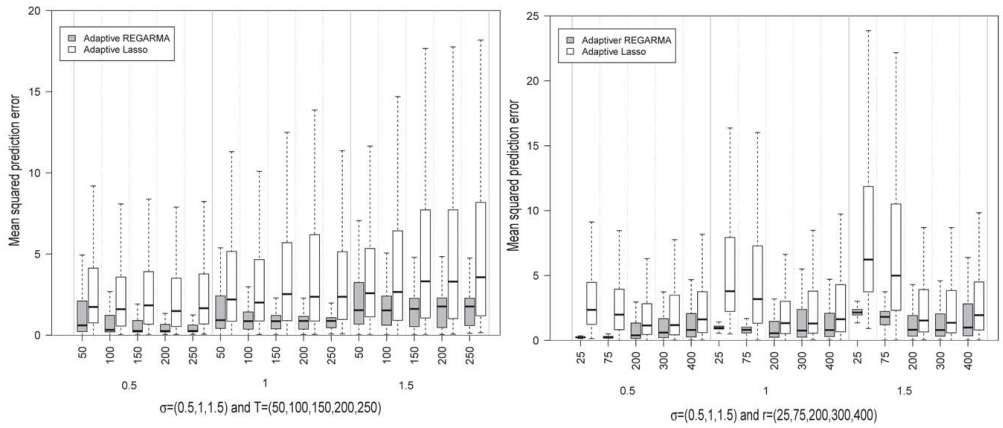

Figure 5: Comparison of adaptive Lasso and adaptive REGARMA with respect to mean squared prediction error on simulated data for different values of σ, r and T.

Table 1: Comparison of adaptive lasso and REGARMA models on Istanbul stock exchange (ISE) data.

MAX AR-MA orders: (4,4)	ADAPTIVE-LASSO	REGAR(2)	REGMA(1)	REGARMA(2,1)
MEAN SQUARED ERROR	0.4194	0.4191	0.4192	0.4079
MEAN ABSOLUTE ERROR	0.4932	0.4927	0.4927	0.4907
BIC	552.44	549.12	549.12	541.41

Table 2: Comparison of adaptive lasso and REGARMA models on *AT&T Inc* and S&P500 data.

MAX AR-MA orders: (4, 4)	ADAPTIVE-LASSO	REGAR(2)	REGMA(1)	REGARMA(2,3)
MEAN SQUARED ERROR	1.34	1.84	4.18	.083
MEAN ABSOLUTE ERROR	87.01	112.30	182.07	75.34
BIC	28.83	27.81	31.2	22.63
NON-ZERO COEFFICIENTS	272	266	282	263

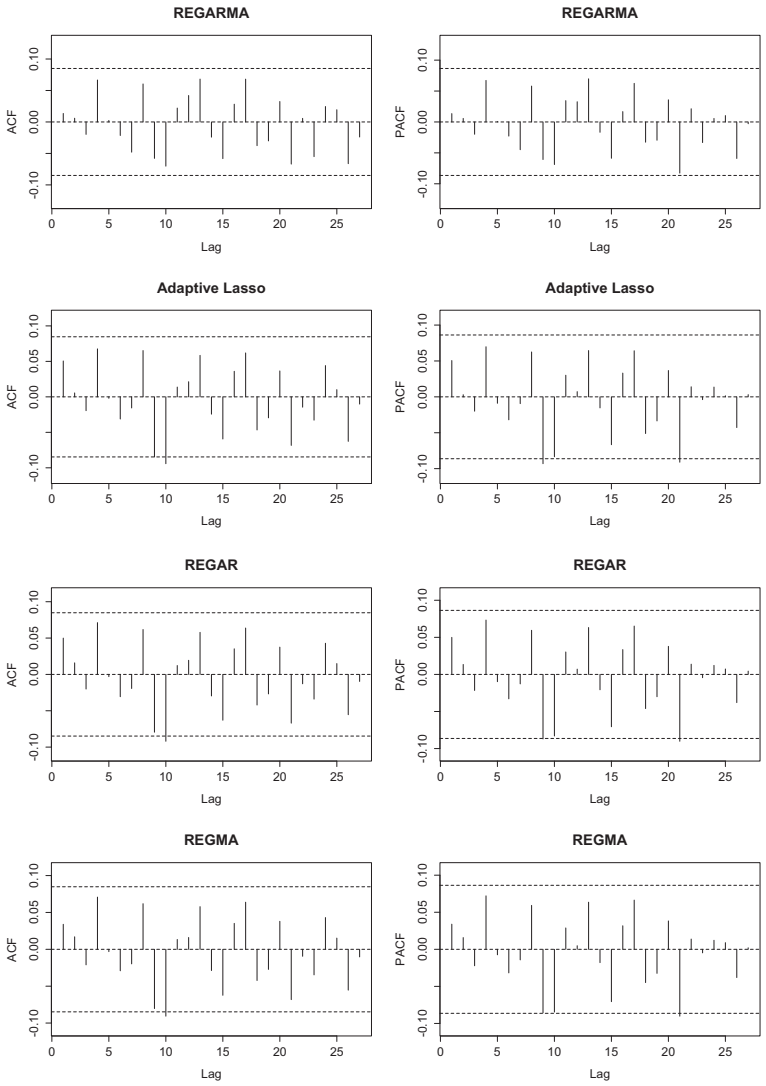

Figure 6: Residual analysis of Adaptive-Lasso, REGAR, REGMA and REGARMA on Istanbul stock exchange data.

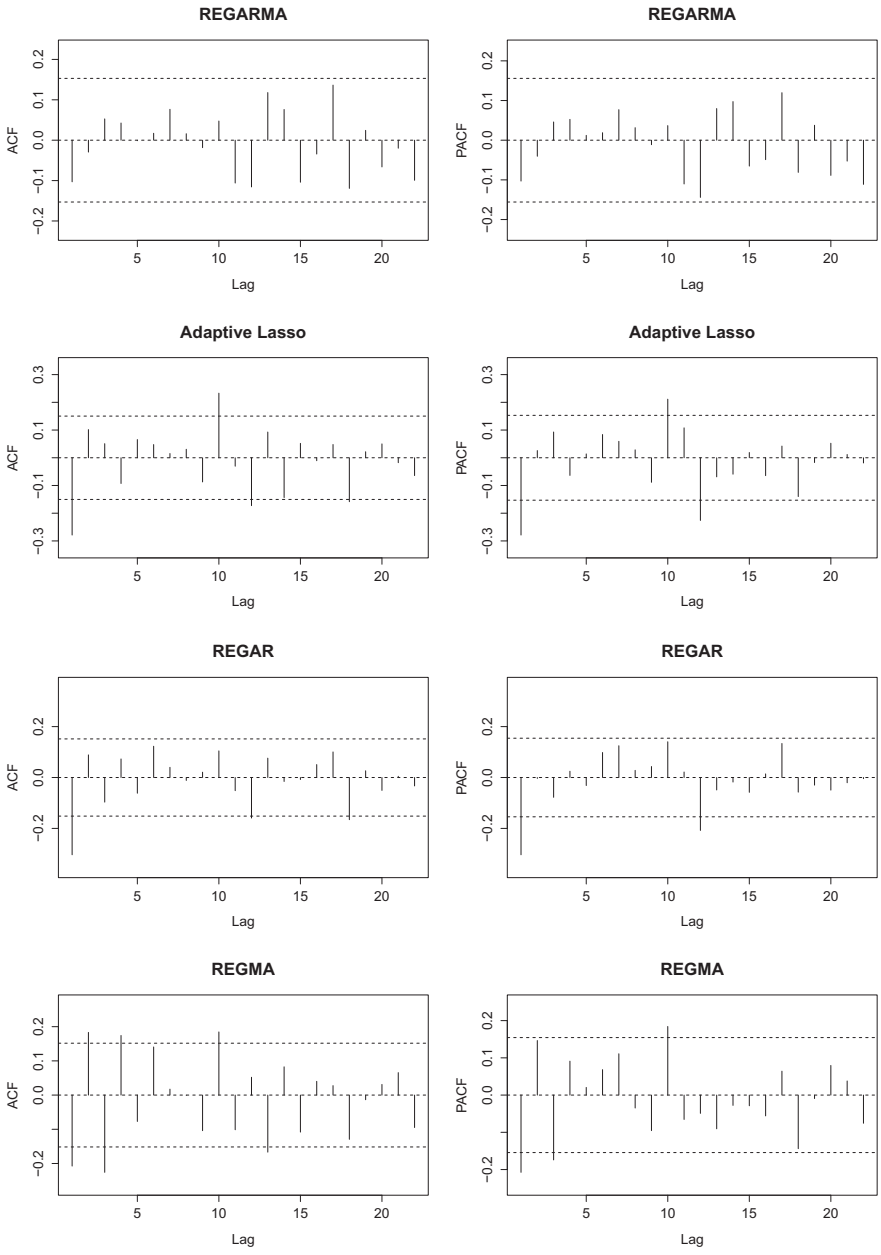

Figure 7: Residual analysis of Adaptive-Lasso, REGAR, REGMA and REGARMA on AT&T index.

8 Appendix (Proof of theorems)

Proof 1 (Theorem 1)**.** *Assume* $\lambda_n\sqrt{n} \to \lambda_\circ$, $\gamma_n\sqrt{n} \to \gamma_\circ$, $\tau_n\sqrt{n} \to \tau_\circ$, *and* $\delta = (u', v', w')$. *Define*

$$k_n(\delta) = Q_n(\Theta^\circ + n^{-(1/2)}\delta) - Q_n(\Theta^\circ). \tag{6}$$

Note that k_n *achieves a minimum at* $\delta = \sqrt{n}(\hat{\Theta} - \Theta^\circ)$. *Using (2), it implies that*

$$\begin{aligned} k_n(\delta) =& \left(L_n(\Theta^\circ + \frac{\delta}{\sqrt{n}}) - L_n(\Theta^\circ) \right) && \text{(7a)} \\ &+ (n\lambda_n'|\beta^\circ + \frac{u}{\sqrt{n}}| - n\lambda_n'|\beta^\circ|) && \text{(7b)} \\ &+ (n\gamma_n'|\phi^\circ + \frac{v}{\sqrt{n}}| - n\gamma_n'|\phi^\circ|) && \text{(7c)} \\ &+ (n\tau_n'|\theta^\circ + \frac{w}{\sqrt{n}}| - n\tau_n'|\theta^\circ|). && \text{(7d)} \end{aligned}$$

The last three terms can be simplified as

$$\begin{aligned} (n\lambda_n'|\beta^\circ + \frac{u}{\sqrt{n}}| - n\lambda_n'|\beta^\circ|) &= \left(\sqrt{n}u\lambda_n'(\frac{|\beta^\circ + u/\sqrt{n}| - |\beta^\circ|}{u/\sqrt{n}}) \right) \\ &\underset{n\to\infty}{\to} \lambda_\circ \sum_{i=1}^{r} \{(u_i sign(\beta_i^\circ) I(\beta_i^\circ \neq 0)) + |u_i| I(\beta_i^\circ = 0)\}. \end{aligned}$$

Similarly, for the other two terms, we have

$$\begin{aligned} (7c) &\underset{n\to\infty}{\to} \gamma_\circ \sum_{j=1}^{p} \{(v_j sign(\phi_j^\circ) I(\phi_j^\circ \neq 0)) + |v_j| I(\phi_j^\circ = 0)\} \\ (7d) &\underset{n\to\infty}{\to} \tau_\circ \sum_{k=1}^{q} \{(w_k sign(\theta_k^\circ) I(\theta_k^\circ \neq 0)) + |w_k| I(\theta_k^\circ = 0)\}. \end{aligned}$$

For (7a), we have

$$\begin{aligned} (7a) =& - e'e + \left((Y - H_{(q)}\theta^\circ - H_{(p)}\phi^\circ - X'\beta^\circ) - (X', H_{(p)}, H_{(q)})\frac{\delta}{\sqrt{n}}) \right)' \\ &\times \left((Y - H_{(q)}\theta^\circ - H_{(p)}\phi^\circ - X'\beta^\circ) - (X', H_{(p)}, H_{(q)})\frac{\delta}{\sqrt{n}}) \right). \end{aligned}$$

Set $A = (X', H_{(p)}, H_{(q)})$ *and recall that* $e = Y - H_{(q)}\theta^\circ - H_{(p)}\phi^\circ - X'\beta^\circ$. *Then, we have*

$$Q_n(\Theta^\circ + \frac{\delta}{\sqrt{n}}) - Q_n(\Theta^\circ) = (e' - \frac{\delta'}{\sqrt{n}}A')(e - A\frac{\delta}{\sqrt{n}}) - e'e + (7b) + (7c) + (7d).$$

The right-hand side of the last equation is equivalent to

$$(\frac{\delta' A'}{\sqrt{n}})(\frac{A\delta}{\sqrt{n}}) - (\frac{\delta' A'}{\sqrt{n}})e - e'(\frac{A\delta}{\sqrt{n}}) + (7b) + (7c) + (7d). \tag{8}$$

From left to right, we now prove that the first term in (8) is bounded and the two other terms follow (asymptotically) normal distributions, i.e.

$$(\frac{\delta' A'}{\sqrt{n}})(\frac{A\delta}{\sqrt{n}}) \to O(1) \tag{9}$$

$$(\frac{\delta' A'}{\sqrt{n}})e \to f_1 \tag{10}$$

$$e'(\frac{A\delta}{\sqrt{n}}) \to f_1' = f_1. \tag{11}$$

Let $H_2' = \frac{A'}{\sqrt{n}}$. *Then,*

$$\begin{aligned} H_2'e &= \frac{1}{\sqrt{n}}(X', H_{(p)}, H_{(q)})'e \\ \sqrt{n}H_2'e &= (X', H_{(p)}, H_{(q)})'e \\ H_t^\circ &= \sqrt{n}H_{2_t}'e_t = (X_t', H_{(p)_t}, H_{(q)_t})'e_t. \end{aligned}$$

H_t° *is a martingale difference sequence(in short mds) because*

$$\begin{aligned} \mathbb{E}(H_t^\circ | t = t-1, t-2, \ldots, t-(p+q)) &= \mathbb{E}((X_t', H_{(p)_t}, H_{(q)_t})'e_t| < t) \\ &= (X_t', H_{(p)_t}, H_{(q)_t})'\mathbb{E}(e_t) = 0. \end{aligned}$$

In order to establish that the conditions of the mds central limit theorem are satisfied (refer to **(author?)** *[8, p 51] for the mds central limit theorem and conditions), define*

$$\bar{\mu} = \frac{1}{n}\sum_{t=T_0+1}^{T} H_t^\circ, \quad \bar{\sigma}^2 = \frac{1}{n}\sum_{t=T_0+1}^{T} Var(H_t^\circ) = \sigma^4 U_B.$$

To show the boundedness condition in the martingales central limit theorem, choose $\delta = 2$*, so that*

$$\mathbb{E}(|H_t^\circ|^4) = \mathbb{E}(e_t^4) E\left(X_t', H_{(p)_t}, H_{(q)_t}\right)^4.$$

Under assumption a, $\mathbb{E}(e_t^4) < \infty$ *and it can be shown that* $E\left(X_t', H_{(p)_t}, H_{(q)_t}\right)^4 < \infty$*, provided* y_t *and* x_t *are stationary and ergodic. Moreover*

$$\frac{1}{n}\sum_{t=T_0+1}^{T} e_t^2\left(X_t, H_{(p)_t}', H_{(q)_t}'\right)^2 =$$

$$\frac{1}{n}\sum_{t=T_0+1}^{T}(e_t^2-\sigma^2)\left(X_t', H_{(p)_t}, H_{(q)_t}\right)^2 + \sigma^2\frac{1}{n}\sum_{t=T_0+1}^{T}\left(X_t', H_{(p)_t}, H_{(q)_t}\right)^2. \quad (12)$$

The first term in (12) is a mds, which has mean zero. So using the weak law of large numbers, we have that $\frac{1}{n}\sum_{t=T_0+1}^{T}(e_t^2-\sigma^2)\left(X_t', H_{(p)_t}, H_{(q)_t}\right)^2 \xrightarrow{p} 0$*. The second term in the right hand side of (12) also tends to* $\sigma^2 U_B$*. As a result*

$$\frac{1}{n}\sum_{t=T_0+1}^{T} e_t^2\left(X_t, H_{(p)_t}', H_{(q)_t}'\right)^2 \xrightarrow{p} \sigma^2 U_B.$$

Therefore, by the central limit theorem for martingales, it follows that $\frac{1}{\sqrt{n}}H_2' e \xrightarrow{d} N(0, \sigma^2 U_B)$ *and*

$$(\frac{\delta' A'}{\sqrt{n}})e \xrightarrow{d} \delta' W,$$

where $\delta = (u', v', w')$ *and* $W \sim MVN(O, \sigma^2 U_B)$*. Then*

$$-((10)+(11)) \xrightarrow{d} -2\delta' W.$$

If $X_i, i = 1, 2, 3, \ldots, r$ *and* y_t *are stationary and ergodic, it is possible to show that (9) tends to* $\delta' U_B \delta$ *where* U_B *is the covariance matrix of* $(X', H_{(p)}, H_{(q)})$*, i.e. (9)*$\to O(1)$*. Finally,* $k_n(\delta)$ *in equation (6) converges to*

$$\begin{aligned} k_n(\delta) \xrightarrow{d} -2\delta' N(O, \sigma^2 U_B) + \delta' U_B \delta &+ \lambda_\circ \sum_{i=1}^{r}\{(u_i sgn(\beta_i^\circ) I(\beta_i^\circ \neq 0)) + |u_i| I(\beta_i^\circ = 0)\} \\ &+ \gamma_\circ \sum_{j=1}^{p}\{(v_j sgn(\phi_j^\circ) I(\phi_j^\circ \neq 0)) + |v_i| I(\phi_j^\circ = 0)\} \\ &+ \tau_\circ \sum_{k=1}^{q}\{(w_k sgn(\theta_k^\circ) I(\theta_k^\circ \neq 0)) + |w_k| I(\theta_k^\circ = 0)\}. \end{aligned}$$

Up to here, we have proved that $k_n(\delta) \xrightarrow{d} k(\delta)$. *To show that* $\arg\min(k_n(\delta)) = \sqrt{n}(\hat{\Theta} - \Theta^\circ) \xrightarrow{d} \arg\min(k(\delta))$ *it is enough to prove that* $\arg\min\{k_n(\delta)\} = O_p(1)$ *[6, 7]. To show this, note that*

$$\begin{aligned}
k_n(\delta) &= (\frac{\delta' A'}{\sqrt{n}})(\frac{A\delta}{\sqrt{n}}) - (\frac{\delta' A'}{\sqrt{n}})e - e'(\frac{A\delta}{\sqrt{n}}) + \\
&\quad (n\lambda_n'|\beta^\circ + \frac{u}{\sqrt{n}}| - n\lambda_n'|\beta^\circ|) + (n\gamma_n'|\phi^\circ + \frac{v}{\sqrt{n}}| - n\gamma_n'|\phi^\circ|) + (n\tau_n'|\theta^\circ + \frac{w}{\sqrt{n}}| - n\tau_n'|\theta^\circ|) \\
&\geq (\frac{\delta' A'}{\sqrt{n}})(\frac{A\delta}{\sqrt{n}}) - (\frac{\delta' A'}{\sqrt{n}})e - e'(\frac{A\delta}{\sqrt{n}}) - (n\lambda_n'|un^{-1/2}| - (n\gamma_n'|vn^{-1/2}|) - (n\tau_n'|wn^{-1/2}|) \\
&\geq (\frac{\delta' A'}{\sqrt{n}})(\frac{A\delta}{\sqrt{n}}) - (\frac{\delta' A'}{\sqrt{n}})e - e'(\frac{A\delta}{\sqrt{n}}) - (\lambda_\circ' + \epsilon_0)|u| - (\gamma_\circ' + \epsilon_0)|v| - (\tau_\circ' + \epsilon_0)|w| + f_n(\delta) \\
&= k_n^*(\delta)
\end{aligned}$$

where $\epsilon_0 > 0$ *is a vector of positive constants. The fourth term in* $k_n^*(\delta)$ *for example, comes from the fact that* $\forall \epsilon_0 > 0$, *there exists an* N *such that if* $n \geq N$, *then* $|\lambda^\circ - \sqrt{n}\lambda_n| < \epsilon_0$. *Then,* $\sqrt{n}\lambda_n < \lambda^\circ + \epsilon_0$. *In addition,* $k_n(0) = k_n^*(0)$ *and* $f_n(\delta) = o_p(1)$. *As a result* $\arg\min\{k_n^*(\delta)\} = O_p(1)$ *and* $\arg\min\{k_n(\delta)\} = O_p(1)$. *The proof of the theorem is completed.*

Proof 2 (Theorem 2)**.** *Let* $\alpha_n = n^{-1/2} + a_n$, *and* $\{\Theta^\circ + \alpha_n\delta : ||\delta|| \leq d, \delta = (u, v, w)'\}$ *be a ball around* Θ°. *Then for* $||\delta|| = d$ *we have*

$$\begin{aligned}
R_n(\delta) &= Q_n^*(\Theta^\circ + \alpha_n\delta) - Q^*(\Theta^\circ) \\
&\geq L_n(\Theta^\circ + \alpha_n\delta) - L_n(\Theta^\circ) + K \\
&\geq L_n(\Theta^\circ + \alpha_n\delta) - L_n(\Theta^\circ) + K' \\
&\geq L_n(\Theta^\circ + \alpha_n\delta) - L_n(\Theta^\circ) + K''
\end{aligned}$$

where

$$K = n\sum_{i\in s_1}\lambda_i^*(|\beta_i^\circ + \alpha_n u_i| - |\beta_i^\circ|) + n\sum_{j\in s_2}\gamma_j^*(|\phi_j^\circ + \alpha_n v_j| - |\phi_j^\circ|) + n\sum_{k\in s_3}\tau_k^*(|\theta_k^\circ + \alpha_n w_k| - |\theta_k^\circ|)$$

$$\textit{(Using triangular inequality)}: K' = -n\alpha_n\sum_{i\in s_1}\lambda_i^*|u_i| - n\alpha_n\sum_{j\in s_2}\gamma_j^*|v_j| - n\alpha_n\sum_{k\in s_3}\tau_k^*|w_k|$$

$$\textit{(Penalties} \leq \alpha_n \textit{ by definition)}: K'' = -n\alpha_n^2(r_\circ + p_\circ + q_\circ)d. \tag{13}$$

The last equation holds because of the decreasing speed of α_n. *Similar calculations to those in theorem (1) result in*

$$L_n(\Theta^\circ + \alpha_n\delta) - L_n(\Theta^\circ) \to n\alpha_n^2\{\delta' U_B\delta + o_p(1)\}. \tag{14}$$

Because (14) dominates (13), then for any given $\eta > 0$, there is a large constant d such that

$$Pr[\inf_{||\delta||=d} \{Q_n^*(\Theta^\circ + \alpha_n\delta)\} > Q_n^*(\Theta^\circ)] \geq 1 - \eta.$$

This result shows that with probability at least $1-\eta$, there is a local minimiser in the ball $\{\Theta^\circ + \alpha_n\delta : ||\delta|| \leq d\}$ and as a result a minimiser, $Q_n^*(\Theta)$, such that $||\hat{\Theta}^* - \Theta^\circ|| = O_p(\alpha_n)$. The proof is completed.

Proof 3 (Theorem 3)**.** *This proof follows from the fact that $Q_n^*(\hat{\Theta}^*)$ must satisfy*

$$\begin{aligned}\frac{\partial Q_n^*(\Theta)}{\partial \beta_i}\Big|_{\hat{\Theta}^*} &= \frac{\partial L_n(\hat{\Theta}^*)}{\partial \beta_i} - n\lambda_i^* sign(\hat{\beta}_i^*) \\ &= \frac{\partial L_n(\Theta^\circ)}{\partial \beta_i} + nU_i(\hat{\Theta}^* - \Theta^\circ)\{1 + o_p(1)\} - n\lambda_i^* sign(\hat{\beta}_i^*) \qquad (15)\end{aligned}$$

where U_i is th i^{th} row of U_B and $i \in s_1^c$. The second term in (15) is a direct result of (3) by adding a $\pm X'\beta^\circ, \pm H_{(p)}\phi^\circ$ and $\pm H_{(q)}\theta^\circ$ to $L_n(\hat{\Theta}^*)$. By using the central limit theorem, the first term of equation (15), $\sum_t e_t x'_{ti}$, will be of order $O_p(n^{1/2})$ and the second term of order $O_p(n^{1/2})$. Furthermore both are dominated by $n\lambda_i^*$ since $b_n\sqrt{n} \to \infty$. Then the sign of $\frac{\partial Q_n^*(\hat{\Theta}^*)}{\partial \beta_i}$ is dominated by the sign of $\hat{\beta}_i^*$ and $\hat{\beta}_i^* = 0$ in probability. Analogously, we can show that $Pr(\hat{\phi}_{s_2^c}^*) \xrightarrow{p} 1$ and $Pr(\hat{\theta}_{s_3^c}^*) \xrightarrow{p} 1$.
The proof is completed.

Proof 4 (Theorem 4)**.** *From Theorem (2) and (3) one can conclude that $Pr(\hat{\Theta}_2^* = 0) \xrightarrow{p} 1$. Thus, $Q_n^*(\Theta) \xrightarrow{with \quad pr\to 1} Q_n^*(\Theta_1)$. So it implies that the Lasso estimator $\hat{\Theta}_1^*$ satisfies the equation*

$$\frac{\partial Q_n^*(\Theta_1)}{\partial \Theta_1}|_{\Theta_1 = \hat{\Theta}_1^*} = 0.$$

From Theorem (2), $\hat{\Theta}_1^*$ is a $\sqrt{n}-$consistent estimator, thus a Taylor expansion of the above equation yields

$$\begin{aligned}0 &= \frac{1}{\sqrt{n}}\frac{\partial L_n(\hat{\Theta}_1^*)}{\partial \Theta_1} + F(\hat{\Theta}_1^*)\sqrt{n} \\ &= \frac{1}{\sqrt{n}}\frac{\partial L_n(\hat{\Theta}_1^\circ)}{\partial \Theta_1} + F(\hat{\Theta}_1^\circ)\sqrt{n} + U_0\sqrt{n}(\hat{\Theta}_1^* - \Theta_1^\circ) + o_p(1),\end{aligned}$$

where F is the first-order derivation of the tuning function

$$\sum_{i\in s_1}\lambda_i|\beta_i| + \sum_{j\in s_2}\gamma_j|\phi_j| + \sum_{k\in s_3}\tau_k|\theta_k|.$$

For n sufficiently large, $F(\hat{\Theta}_1^*) = F(\Theta_1^\circ)$, thus

$$\begin{aligned}(\Theta_1^\circ - \hat{\Theta}_1^*)\sqrt{n} &= \frac{U_0^{-1}}{\sqrt{n}}\frac{\partial L_n(\Theta_1^\circ)}{\partial\Theta_1} + o_p(1)\\ &\xrightarrow{d} N(0, \sigma^2 U_0^{-1}).\end{aligned}$$

The proof is completed.

Proof 5 (Theorem 5)**.** *Let $Y = (y_1, y_2, \ldots, y_n)'$, $\hat{Y} = (\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n)'$ and $Y^\circ = (y_1^\circ, y_2^\circ, \ldots, y_n^\circ)'$ be observations, REGARMA predictions and predictions from the true model, respectively, that is*

$$\begin{aligned}\hat{Y} &= X'\hat{\beta} + H_{(p)}\hat{\phi} + H_{(q)}\hat{\theta}\\ Y^\circ &= X'\beta^\circ + H_{(p)}\phi^\circ + H_{(q)}\theta^\circ.\end{aligned}$$

Define a set $C = \{X'\beta + H_{(p)}\phi + H_{(q)}\theta;\ \sum_{j=1}^r|\beta_j| \le K_\lambda, \sum_{k=1}^p|\phi_k| \le K_\gamma, \sum_{l=1}^q|\theta_l| \le K_\tau\}$. Note that C is a compact and convex subset of $\mathbb{R}^n$ and that $\hat{Y}$ is the projection of Y on C. As a result of the convexity of C, for any vector v in C, the vector $v - \hat{Y}$ must be at an obtuse angle to the vector $Y - \hat{Y}$. This means that

$$\left\langle (v - \hat{Y}), (Y - \hat{Y}) \right\rangle \le 0.$$

Since $Y^\circ \in C$, then the inner product of $(Y^\circ - \hat{Y})$ and $(Y - \hat{Y})$ is non-positive

$$\left\langle (Y^\circ - \hat{Y}), (Y - \hat{Y}) \right\rangle \le 0. \tag{16}$$

Using (16) and some simple algebra we have

$$\begin{aligned}\widehat{MSPE} &= \frac{1}{n}\|Y^\circ - \hat{Y}\|^2\\ &\le \frac{1}{n}\left\langle (Y - Y^\circ), (\hat{Y} - Y^\circ) \right\rangle\\ &\le \frac{1}{n}e'\left(X'(\hat{\beta} - \beta^\circ) + H_{(p)}(\hat{\phi} - \phi^\circ) + H_{(q)}(\hat{\theta} - \theta^\circ)\right)\\ &= \frac{1}{n}\left(e'X'(\hat{\beta} - \beta^\circ) + e'H_{(p)}(\hat{\phi} - \phi^\circ) + e'H_{(q)}(\hat{\theta} - \theta^\circ)\right)\\ &\le \frac{1}{n}\left(2K_\lambda \max_{1\le i\le r}|e'X_i| + 2K_\gamma \max_{1\le j\le p}|e'H_{(p)_j}| + 2K_\tau \max_{1\le k\le q}|e'H_{(q)_k}|\right).\end{aligned} \tag{17}$$

On the other hand, conditioning on X and history of y results in

$$
\begin{aligned}
e'X' &\sim N(O, \sigma^2(XX')) \\
e'H_{(p)} &\sim N(O, \sigma^2 H'_{(p)} H_{(p)}) \\
e'H_{(q)} &\sim N(O, \sigma^2 (H'_{(q)} H_{(q)})).
\end{aligned}
$$

Using Lemma 3 in [2],

$$\mathbb{E}(\max_{i=1,2,3,\ldots,r} |e'X_i|) \leq M_1 \sigma \sqrt{2n \log(2r)} \tag{18}$$

$$\mathbb{E}(\max_{j=1,2,3,\ldots,p} |e'H_{(p)_j}|) \leq M_2 \sigma \sqrt{2n \log(2p)} \tag{19}$$

$$\mathbb{E}(\max_{k=1,2,3,\ldots,q} |e'H_{(q)_k}|) \leq M_3 \sigma \sqrt{2n \log(2q)}. \tag{20}$$

Substituting (18,19,20) in (17) result in

$$\frac{1}{n}\|Y^\circ - \hat{Y}\|^2 \leq \frac{1}{n}\left(2K_\lambda M_1 \sigma \sqrt{2n \log(2r)} + 2K_\gamma M_2 \sigma \sqrt{2n \log(2p)} + 2K_\tau M_3 \sigma \sqrt{2n \log(2q)}\right).$$

But $M_{max} = \sup\{M_1, M_2, M_3\}$ and $K_{max} = \sup\{K_\lambda, K_\gamma, K_\tau\}$, therefore

$$\frac{1}{n}\|Y^\circ - \hat{Y}\|^2 \leq \frac{2K_{max} M_{max} \sigma}{n}\left(\sqrt{2n \log(2r)} + \sqrt{2n \log(2p)} + \sqrt{2n \log(2q)}\right)$$

and

$$\widehat{MSPE}(\hat{\beta}, \hat{\phi}, \hat{\theta}) \leq \frac{2K_{max} M \sigma}{\sqrt{n}}\left(\sqrt{2 \log(2r)} + \sqrt{2 \log(2p)} + \sqrt{2 \log(2q)}\right). \tag{21}$$

The proof of the theorem is completed.

Lemma 1. *If $X_1, X_2, X_3, \ldots, X_m$ are m dependent mean zero random variables where $|X_i| \leq L, \forall i \in \{1, 2, 3, \ldots, m\}$. Then, $\forall \beta \in \mathbb{R}$,*

$$\mathbb{E}(e^{\beta \sum_{i=1}^m x_i}) \leq e^{(mL\beta)^2}.$$

Proof. This result extends the result of [2] from independent variables to dependent variables.

It is obvious that $\sum_{i=1}^m x_i \leq mL$ then,

$$\mathbb{E}(e^{\beta \sum_{i=1}^m x_i}) = \int_{-mL}^{mL} e^{\beta \sum_{i=1}^m x_i} d\mu(\sum_{i=1}^m x_i),$$

where μ is a probability distribution. On the other hand, $x \mapsto e^{x\beta}$ is a convex map. Define $t = \frac{\sum_{i=1}^m x_i}{2mL} + \frac{1}{2}$, then

$$e^{\beta \sum_{i=1}^m x_i} = e^{\beta\left(t(mL)-(1-t)(mL)\right)} \leq te^{\beta mL} + (1-t)e^{-\beta mL}.$$

Using the fact that $\mathbb{E}(\sum_{i=1}^m x_i) = \sum_{i=1}^m \mathbb{E}(x_i) = 0$,

$$\begin{aligned}\int_{-mL}^{mL} e^{\beta\left(t(mL)-(1-t)(mL)\right)} d\mu(\sum_{i=1}^m x_i) &\leq \int_{-mL}^{mL} te^{\beta mL} + (1-t)e^{-\beta mL} d\mu(\sum_{i=1}^m x_i) \\ &= \frac{e^{\beta mL} + e^{-\beta mL}}{2} = \cosh(\beta mL) \leq e^{(mL\beta)^2/2}.\end{aligned}$$

The proof is completed. □

Proof 6 (Remark (2))**.** *Consider*

$$\begin{aligned}\mathbb{E}(Y^\circ - \hat{Y})^2 =& E\Big((X'(\beta^\circ - \hat{\beta}) + H_{(p)}(\phi^\circ - \hat{\phi}) + H_{(q)}(\theta^\circ - \hat{\theta}))' \\ & (X'(\beta^\circ - \hat{\beta}) + H_{(p)}(\phi^\circ - \hat{\phi}) + H_{(q)}(\theta^\circ - \hat{\theta}))\Big). \quad (22)\end{aligned}$$

On the other hand

$$\begin{aligned}\frac{1}{n}\|Y^\circ - \hat{Y}\|^2 = \frac{1}{n}&(X'(\beta^\circ - \hat{\beta}) + H_{(p)}(\phi^\circ - \hat{\phi}) + H_{(q)}(\theta^\circ - \hat{\theta}))' \\ & (X'(\beta^\circ - \hat{\beta}) + H_{(p)}(\phi^\circ - \hat{\phi}) + H_{(q)}(\theta^\circ - \hat{\theta})). \quad (23)\end{aligned}$$

Combining (22) and (23) results in,

$$\begin{aligned}\mathbb{E}(Y^\circ - \hat{Y})^2 - \frac{1}{n}\|Y^\circ - \hat{Y}\|^2 \leq\ & 4K_\lambda^2 \max_{(1\leq j\leq r, 1\leq k\leq r)} |\mathbb{E}(X_j X_k') - \frac{1}{n} X_j X_k'| \\ & + 4K_\gamma^2 \max_{(1\leq j\leq p, 1\leq k\leq p)} |\mathbb{E}(H'_{(p)_j} H_{(p)_k}) - \frac{1}{n} H'_{(p)_j} H_{(p)_k}| \\ & + 4K_\tau^2 \max_{(1\leq j\leq q, 1\leq k\leq q)} |\mathbb{E}(H'_{(q)_j} H_{(q)_k}) - \frac{1}{n} H'_{(q)_j} H_{(q)_k}| \\ & + 8K_\lambda K_\gamma \max_{(1\leq j\leq r, 1\leq k\leq p)} |\mathbb{E}(X_j H_{(p)_k}) - \frac{1}{n} X_j H_{(p)_k}| \\ & + 8K_\gamma K_\tau \max_{(1\leq j\leq p, 1\leq k\leq q)} |\mathbb{E}(H'_{(q)_j} H_{(p)_k}) - \frac{1}{n} H'_{(q)_j} H_{(p)_k}| \\ & + 8K_\lambda K_\tau \max_{(1\leq j\leq r, 1\leq k\leq q)} |\mathbb{E}(X_j H_{(q)_k}) - \frac{1}{n} X_j H_{(q)_k}|. \quad (24)\end{aligned}$$

Take $K^* = max\{K_\lambda^2, K_\gamma^2, K_\tau^2, 2K_\lambda K_\gamma, 2K_\lambda K_\tau, 2K_\gamma K_\tau\}$ *and notes that each term in the max expressions is bounded (e.g.* $|\mathbb{E}(X_j X_k') - X_{ij} X_{ki}'| \leq 2M_1^2$*). Using Lemma (1) and lemma 3 of [2], each term in (24) is bounded by* $2M_i M_j \sqrt{\frac{2log(2a_i a_j)}{n}}$, $i \in \{1, 2, 3, \ldots, r\}$, $j \in \{1, 2, 3, \ldots, p\}$, $k \in \{1, 2, 3, \ldots, q\}$, $a_1 = r, a_2 = p$ *and* $a_3 = q$*. Then,*

$$MSPE(\hat{\beta}, \hat{\phi}, \hat{\theta}) \leq \frac{2K_{max} M_{max} \sigma}{\sqrt{n}} \sum_{i=1}^{3} \left(\sqrt{2\log(2a_i)} \right) + 8K^* \sum_{i,j=1}^{3} \left(M_i M_j \sqrt{\frac{2log(2a_i a_j)}{n}} \right).$$

The proof is completed.

References

[1] Oguz Akbilgic, Hamparsum Bozdogan, and M.Erdal Balaban. A novel hybrid rbf neural networks model as a forecaster. *Statistics and Computing*, 24(3):365–375, 2014.

[2] S. Chatterjee. Assumptionless consistency of the lasso. *arXiv*, 5817v3, 2013.

[3] Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001.

[4] Kei Hirose, Shohei Tateishi, and Sadanori Konishi. Efficient algorithm to select tuning parameters in sparse regression modeling with regularization. *arXiv:1109.2411*, 2011.

[5] Jian Huang, Shuangge Ma, and Cun-Hui Zhang. Adaptive lasso for sparse high-dimensional regression models. *Statistica Sinica*, 18(4):1603, 2008.

[6] Jeankyung Kim and David Pollard. Cube root asymptotics. *The Annals of Statistics*, 18(1):191–219, 03 1990.

[7] Keith Knight and Wenjiang Fu. Asymptotics for lasso-type estimators. *Annals of Statistics*, 28(5):1356–1378, 2000.

[8] V. Martin, S. Hurn, and D. Harris. *Econometric Modelling with Time Series: Specification, Estimation and Testing.* Themes in Modern Econometrics. Cambridge University Press, 2012.

[9] Marcelo C Medeiros and Eduardo F Mendes. Estimating high-dimensional time series models. Technical report, 2012.

[10] Nicolai Meinshausen and Peter Bühlmann. High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 34(3):1436–1462, 2006.

[11] Yuval Nardi and Alessandro Rinaldo. Autoregressive process modeling via the lasso procedure. *Journal of Multivariate Analysis*, 102(3):528 – 549, 2011.

[12] Trevor Park and George Casella. The Bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–686, 2008.

[13] S. Song and P. Bickel. Large vector auto regressions. *arXiv:1106.3915*, June 2011.

[14] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B*, 58(1):267–288, 1996.

[15] H. Wang, G. Li, and C. Tsai. Regression coefficient and autoregressive order shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B*, 69(1):63–78, 2007.

[16] Rongning Wu and Qin Wang. Shrinkage estimation for linear regression with arma errors. *Journal of Statistical Planning and Inference*, 142(7):2136–2148, 2012.

[17] Hui Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429, 2006.