


Penalised inference for lagged dependent regression in the presence of autocorrelated residuals

Hamed Haselimashhadi¹ · Veronica Vinciotti¹ 

Received: 8 July 2016 / Accepted: 28 August 2017 / Published online: 9 September 2017
© The Author(s) 2017. This article is an open access publication

Abstract This paper deals with linear models for a time-dependent response and explanatory variables in a high-dimensional setting. We account for the time dependency in the data by explicitly adding autoregressive terms to the response variable in the model together with an autoregressive process for the residuals. We present a penalized likelihood approach for parameter estimation and discuss its theoretical properties. Finally, we show the successful application of the proposed methodology on simulated data and on two real applications, where we model air pollution and stock market indices, respectively. We provide an implementation of the method in the R package DREGAR, freely available on CRAN, <http://CRAN.R-project.org/package=DREGAR>.

Keywords Time series · High-dimensional models · Penalized likelihood

1 Introduction

This paper deals with fitting a 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 performing variable selection and parameter estimation simultaneously for any data, including high-dimensional datasets, where classical approaches for parameter estimation break down, e.g. [9, 13, 19, 21, 25]. In [25], it is 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, [22] showed the successful application of l_1 penalised inference in the context of autocorrelated residuals for a fixed order, by proposing the model

✉ Veronica Vinciotti
Veronica.Vinciotti@brunel.ac.uk

¹ Department of Mathematics, Brunel University, London, UK

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

and studied its asymptotic properties. We refer later to this model as REGAR(q). Nardi and Rinaldo [15] studied the theoretical properties of a regularized autoregressive process on Y_t for both low and high dimensional cases, whereas [1, 14, 20] studied the l_1 estimation of vector autoregressive models. Recently, [6] proposed an alternative to l_1 penalisation for vector autoregressive (VAR) models in the high-dimensional framework. In [1, 6, 15, 20], no exogenous variables are included in the model. In contrast to this, [12] studied the asymptotic properties of adaptive lasso in high dimensional time series models when the number of exogenous variables increases as a function of the number of observations and [17] considered the case of VAR models with exogenous variables. While both models cover a lagged regression in the presence of exogenous variables, they do not consider autocorrelated residuals. Recently, [23] proposed an extension of the model in [22] by adding a moving average term as follows

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

We refer later to this model as REGARMA(p,q). Similar to [22], 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 [15], as well as an autocorrelated process for residuals, as in [22], in order to capture complex dynamics parsimoniously. In particular, given fixed orders p and q , we propose the model

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

We name the resulting model DREGAR(p,q). The model is essentially a double autoregressive model with unbalanced weights for response and explanatory variables. To show this, we rewrite the model in (1), the REGARMA model of [23] and the REGAR model of [22] using the backward shift operator L :

$$\begin{aligned} \text{DREGAR: } L(\theta)L(\phi)y_t &= L(\theta)x'_t\beta + e_t, \\ \text{REGARMA: } L(\theta)y_t &= L(\theta)x'_t\beta + L(\phi)e_t, \\ \text{REGAR: } L(\theta)y_t &= L(\theta)x'_t\beta + e_t, \end{aligned}$$

where $L(\cdot)$ represents a stationary polynomial of L and $L(\theta)L(\phi)$ represents a special case of an $AR(p + q)$ process. From these equations, one can see how REGAR and REGARMA impose the same autoregressive structure on both response and covariates, whereas DREGAR assumes different structures on each of them. We found this aspect to be particularly advantageous on a number of analyses of real datasets, where DREGAR fits the data better than competitive models, with two of these examples reported in Sect. 6. In contrast to REGAR and DREGAR, REGARMA contains a moving average process on the errors. The MA component, however, induces a higher level of complexity in the parameter estimation and in the proofs of the theoretical results.

All three models belong to the general framework of ARMAX [11], which is common in the system identification and signal processing literature [16] where inference is typically

performed in the low-dimensional case. The focus of this paper is on statistical inference for the DREGAR model, in particular for the high-dimensional case where maximum likelihood estimation fails. In particular, we devise a penalised likelihood approach for parameter estimation, in the same spirit to the REGAR and REGARMA contributions. In Sect. 2, we formulate the model and present an l_1 penalized likelihood approach for the estimation of the parameters. In Sect. 3, we prove the asymptotic properties of the model DREGAR(p,0). In Sect. 4, we discuss the implementation of DREGAR. A simulation study, given in Sect. 5, will accompany the theoretical results. In Sect. 6 we apply the model to two real datasets, one on air pollution and another on the stock market, and compare the fit of the model with REGARMA and REGAR models. Finally, we draw some conclusions in Sect. 7.

2 L_1 penalised estimation

The general form of DREGAR consists of a lagged response variable, covariates and auto-correlated residuals. Consider the following Gaussian DREGAR model of order p and q ,

$$y_t = x'_t\beta + \sum_{j=1}^p \phi_j y_{t-j} + \epsilon_t, \quad \epsilon_t = \sum_{i=1}^q \theta_i \epsilon_{t-i} + e_t, \quad e_t \stackrel{iid}{\sim} N(0, \sigma^2), \quad t = 1, 2, \dots, T$$

where x'_t is the t th row of the design matrix containing 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; $e_t, t = 1, \dots, T$ are independent and identically normally distributed noises with mean zero and known finite fourth moments, and $p + q < T$. Moreover, we assume that the error and explanatory variables are mutually independent for all time points. To remove the constants from the model we follow the literature on regularized models, e.g. [7,21], and standardize the covariates to have zero mean and unit variance and the response to zero mean.

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

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

where λ, γ, τ are tuning parameters and $\Theta = (\beta, \phi, \theta)$ is the vector of unknown parameters. Following the literature, and given the superior properties of adaptive lasso [25], we also propose an adaptive version of the likelihood

$$Q_n^*(\Theta) = \sum_{t=T_o+1}^T \left((y_t - x'_t\beta) - \sum_{i=1}^p \phi_i y_{t-i} - \sum_{j=1}^q \theta_j \epsilon_{t-j} \right)^2 + \sum_{i=1}^r \lambda_i^* |\beta_i| + \sum_{j=1}^p \gamma_j^* |\phi_j| + \sum_{k=1}^q \tau_k^* |\theta_k| \tag{3}$$

where $\lambda_i^*, \gamma_j^*, \tau_k^*, i = 1, 2, \dots, r; j = 1, 2, \dots, p; k = 1, 2, \dots, q$ 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 response ($H_{(p)}$), residuals ($H_{(q)}$), and explanatory variables (X'). Let $\Theta = (\phi, \theta, \beta)'$ denote the vector of corresponding parameters, $e' = (e_{T_0+1}, e_{T_0+2}, \dots, e_T)$ be the vector of errors, $T_o = p + q$ and $n = T - T_o$, 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 observations 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

$$Q_n^*(\Theta) = L(\Theta) + \lambda^*|\beta| + \gamma^*|\phi| + \tau^*|\theta|, \tag{4}$$

where the parameters are given by $\lambda^* = (\lambda_1^*, \lambda_2^*, \dots, \lambda_r^*)$, $\gamma^* = (\gamma_1^*, \gamma_2^*, \dots, \gamma_p^*)$, $\tau^* = (\tau_1^*, \tau_2^*, \dots, \tau_q^*)$, $\Theta = (\beta, \phi, \theta)'$.

3 Theoretical properties of the model

In order to study the theoretical properties of DREGAR and adaptive-DREGAR, we define the true coefficients by $\Theta^\circ = (\beta^\circ, \phi^\circ, \theta^\circ)'$ and assume that some of the coefficients are zero. The indices 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 indices of zero coefficients. We also define $\beta_{s_1}^\circ, \phi_{s_2}^\circ, \theta_{s_3}^\circ$ and their corresponding (DREGAR) estimations by $\hat{\beta}_{s_1}, \hat{\phi}_{s_2}, \hat{\theta}_{s_3}$. Similarly, adaptive-DREGAR 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 $\Theta_1^\circ = (\beta_{s_1}^\circ, \phi_{s_2}^\circ, \theta_{s_3}^\circ)'$, $\Theta_2^\circ = (\beta_{s_1^c}^\circ, \phi_{s_2^c}^\circ, \theta_{s_3^c}^\circ)'$, $\hat{\Theta}_1 = (\hat{\beta}_{s_1}, \hat{\phi}_{s_2}, \hat{\theta}_{s_3})'$, $\hat{\Theta}_2 = (\hat{\beta}_{s_1^c}, \hat{\phi}_{s_2^c}, \hat{\theta}_{s_3^c})'$, $\hat{\Theta}_1^* = (\hat{\beta}_{s_1}^*, \hat{\phi}_{s_2}^*, \hat{\theta}_{s_3}^*)'$, $\hat{\Theta}_2^* = (\hat{\beta}_{s_1^c}^*, \hat{\phi}_{s_2^c}^*, \hat{\theta}_{s_3^c}^*)'$.

3.1 Assumptions

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

- (a) The response variable is assumed to be stationary and ergodic with finite second order moment. Further, we assume that the two polynomials $1 - \sum_{i=1}^p \phi_i L^i$ and $1 - \sum_{i=1}^q \theta_i L^i$ have all the roots unequal and outside the unit circle.
- (b) Covariates are assumed to be mutually independent of each other and of the error term. Additionally, we assume that $x_{.s}, s = 1, \dots, r$ are generated from stationary and ergodic processes.
- (c) e_t s are i.i.d Gaussian random variables with finite fourth moments.
- (d) $\frac{1}{n} X'X \xrightarrow{a.s} \mathbb{E}(X'X) < \infty$ and $\max_{1 \leq i \leq r} x_i x_i' < \infty$.

The first three assumptions guarantee that the mean and variance of the whole system remain unchanged over time. The last assumption guarantees the existence and convergence of the sample moments.

3.2 Theoretical properties of l_1 penalized DREGAR(p,0)

In this section we focus on the theoretical properties of l_1 penalized DREGAR estimators. In particular, we concentrate on the theoretical properties of DREGAR(p,0) as we can prove that there is asymptotically no bias in this model. This model differs from REGAR(p) [22] as it imposes an autoregressive process on the response whereas REGAR(p) considers the case of autocorrelated residuals (i.e. similar to a DREGAR(0,q) model). In the next two sections, we distinguish the cases of DREGAR(p,0) and adaptive-DREGAR(p,0), respectively.

3.3 Asymptotic properties of DREGAR(p,0)

Theorem 1 *Under assumptions [a–d] and assuming $\lambda_n\sqrt{n} \rightarrow \lambda_\circ$, $\gamma_n\sqrt{n} \rightarrow \gamma_\circ$, and $\lambda_\circ, \gamma_\circ \geq 0$, it follows that $\sqrt{n}(\hat{\Theta} - \Theta^\circ) \xrightarrow{d} \arg \min k(\delta)$ where*

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

and $\delta = (u', v')$ is a vector of parameters in $\mathbb{R}^{(r+p)}$, $W \sim \text{MVN}(O, \sigma^2U_B)$ and $U_B = \text{Cov}(X, H_{(p)})$.

Proof Let

$$k_n(\delta) = Q_n(\Theta^\circ + n^{-1/2}\delta) - Q_n(\Theta^\circ). \tag{5}$$

Note that k_n reaches the minimum at $\sqrt{n}(\hat{\Theta} - \Theta^\circ)$. From (2),

$$k_n(\delta) = \left(L_n \left(\Theta^\circ + \frac{\delta}{\sqrt{n}} \right) - L_n(\Theta^\circ) \right) \tag{6a}$$

$$+ \left(n\lambda'_n \left| \beta^\circ + \frac{u}{\sqrt{n}} \right| - n\lambda'_n |\beta^\circ| \right) \tag{6b}$$

$$+ \left(n\gamma'_n \left| \phi^\circ + \frac{v}{\sqrt{n}} \right| - n\gamma'_n |\phi^\circ| \right). \tag{6c}$$

The last two terms have limits:

$$(6b) = \left(\sqrt{nu}\lambda'_n \frac{|\beta^\circ + u/\sqrt{n}| - |\beta^\circ|}{u/\sqrt{n}} \right)$$

$$\xrightarrow{n \rightarrow \infty} \lambda_\circ \sum_{i=1}^r \{(u_i \text{sign}(\beta_i^\circ)I(\beta_i^\circ \neq 0)) + |u_i|I(\beta_i^\circ = 0)\}.$$

$$(6c) \xrightarrow{n \rightarrow \infty} \gamma_\circ \sum_{j=1}^p \{(v_j \text{sign}(\phi_j^\circ)I(\phi_j^\circ \neq 0)) + |v_j|I(\phi_j^\circ = 0)\}.$$

As for the first term:

$$(6a) = -e'e + \left((y - H_{(p)}\phi^\circ - X'\beta^\circ) - (X', H_{(p)})\frac{\delta}{\sqrt{n}} \right)' \times \left((y - H_{(p)}\phi^\circ - X'\beta^\circ) - (X', H_{(p)})\frac{\delta}{\sqrt{n}} \right).$$

Setting $A = (X', H_{(p)})$ and $e = y - H_{(p)}\phi^\circ - X'\beta^\circ$,

$$Q_n \left(\Theta^\circ + \frac{\delta}{\sqrt{n}} \right) - Q_n(\Theta^\circ) = \left(e' - \frac{\delta'}{\sqrt{n}}A' \right) \left(e - A\frac{\delta}{\sqrt{n}} \right) - e'e + (6b) + (6c),$$

which is equivalent to

$$\left(\frac{\delta' A'}{\sqrt{n}} \right) \left(\frac{A\delta}{\sqrt{n}} \right) - \left(\frac{\delta' A'}{\sqrt{n}} \right) e - e' \left(\frac{A\delta}{\sqrt{n}} \right) + (6b) + (6c). \tag{7}$$

From left to right, we prove that the first term in (7) is bounded and the next two terms follow (asymptotically) normal distributions:

$$\left(\frac{\delta' A'}{\sqrt{n}} \right) \left(\frac{A\delta}{\sqrt{n}} \right) \rightarrow O(1) \tag{8}$$

$$e' \left(\frac{A\delta}{\sqrt{n}} \right) = \left(\frac{\delta' A'}{\sqrt{n}} \right) e \rightarrow f_1. \tag{9}$$

Similar calculations to [5] show that (8) tends to $\delta'U_B\delta$ where U_B is the covariance matrix of $(X', H_{(p)})$, which is bounded ($O(1)$). Defining S_n as a function of n ,

$$S_n = \left(\frac{\delta' A'}{\sqrt{n}} \right) e = \frac{1}{\sqrt{n}}(u'X + v'H'_{(p)})e,$$

and using assumptions [a–d] and the central limit theorem for martingales result in

$$S_n \xrightarrow{d} \delta'W,$$

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

$$-(9) \xrightarrow{d} -2\delta'W.$$

Substituting all results in Eq. (5),

$$k_n(\delta) \xrightarrow{d} -2\delta'N(O, \sigma^2U_B) + \delta'U_B\delta + \lambda_\circ \sum_{i=1}^r \{(u_i \text{sign}(\beta_i^\circ)I(\beta_i^\circ \neq 0)) + |u_i|I(\beta_i^\circ = 0)\} + \gamma_\circ \sum_{j=1}^p \{(v_j \text{sign}(\phi_j^\circ)I(\phi_j^\circ \neq 0)) + |v_j|I(\phi_j^\circ = 0)\}.$$

Up to now, we have proved $k_n(\delta) \xrightarrow{d} k(\delta)$. 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 $\arg \min\{k_n(\delta)\} = O_p(1)$ [8,9]. This follows from

$$\begin{aligned} k_n(\delta) &= \left(\frac{\delta' A'}{\sqrt{n}}\right) \left(\frac{A\delta}{\sqrt{n}}\right) - \left(\frac{\delta' A'}{\sqrt{n}}\right) e - e' \left(\frac{A\delta}{\sqrt{n}}\right) \\ &\quad + \left(n\lambda'_n \left|\beta^\circ + \frac{u}{\sqrt{n}}\right| - n\lambda'_n |\beta^\circ|\right) + \left(n\gamma'_n \left|\phi^\circ + \frac{v}{\sqrt{n}}\right| - n\gamma'_n |\phi^\circ|\right) \\ &\geq \left(\frac{\delta' A'}{\sqrt{n}}\right) \left(\frac{A\delta}{\sqrt{n}}\right) - \left(\frac{\delta' A'}{\sqrt{n}}\right) e - e' \left(\frac{A\delta}{\sqrt{n}}\right) - (n\lambda'_n |un^{-1/2}| - (n\gamma'_n |vn^{-1/2}|)) \\ &\geq \left(\frac{\delta' A'}{\sqrt{n}}\right) \left(\frac{A\delta}{\sqrt{n}}\right) - \left(\frac{\delta' A'}{\sqrt{n}}\right) e - e' \left(\frac{A\delta}{\sqrt{n}}\right) - (\lambda'_\circ + \epsilon_\circ)|u| - (\gamma'_\circ + \epsilon_\circ)|v| \\ &= k_n^*(\delta), \end{aligned}$$

where $\epsilon_\circ > 0$ is a vector of positive constants. The fourth term in $k_n^*(\delta)$ for example, comes from the fact that $\forall \epsilon_\circ > 0, \exists N, \text{ if } n \geq N, |\lambda^\circ - \sqrt{n}\lambda_n| < \epsilon_\circ$. Then, $\sqrt{n}\lambda_n < \lambda^\circ + \epsilon_\circ$. 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. \square

This theorem shows that the DREGAR estimator has the Knight and Fu [9] asymptotic property and it implies that the tuning parameters in $Q_n(\Theta)$ do not shrink to zero at the speed faster than $n^{-1/2}$. In the proof of Theorem 1, the errors must be independent and identically distributed and we do not make a specific assumption about the type of distribution. In other words, the central limit theorem for martingale guarantees the convergence to the normal distribution.

As shown in [9], minimizing l_1 penalized likelihood in the linear model leads to unavoidable bias in the estimates of the non-zero parameters. In the following remark, we show this also in the context of the DREGAR model.

Theorem 2 Taking a special case where $\beta_i^\circ > 0, 1 \leq (\forall i \in \mathbb{N}) \leq r$ and $\phi_{i_2}^\circ = 0$ for $1 \leq j_1 \leq q, 1 \leq j_2 \leq p, j_1, j_2 \in \mathbb{N}$, assuming that there are enough observations and that the minimizer $k(\delta)$ correctly identifies the coefficients, that is, $u \neq 0$ and $v = 0$, then, $k(\delta)$ must satisfy

$$\begin{aligned} \frac{\partial k(\delta)}{\partial u} &= \frac{\partial k(u, 0)}{\partial u} \\ &= \frac{\partial}{\partial u} (-2(u', 0)W + (u', 0)'U_B(u', 0) + (6b) + (6c)) \\ &= -2W_{1:r} + 2u'U_{B_{1:r}} + \lambda_\circ 1_{r \times 1} = 0 \\ \rightarrow u' &= \frac{1}{2}(2W_{1:r} - \lambda_\circ 1_{r \times 1})U_{B_{1:r}}^{-1}. \end{aligned}$$

$$\begin{aligned} \text{Using Theorem (1)} : \sqrt{n}(\hat{\beta} - \beta^\circ) &\xrightarrow{d} \arg \min k(\delta = u') \\ &= \text{MVN} \left(\mathbb{E}(u') \neq 0, U_{B_{1:r}}^{-1} \right), \end{aligned}$$

where $U_{B_{1:r}}$ is the first r rows of U_B corresponding to the r covariates. From the final equation, DREGAR(p,0) suffers an asymptotic bias, provided the tuning parameter is positive. In other words, lasso regularization of DREGAR(p,0) is not asymptotically consistent. In the next section we discuss the adaptive-DREGAR(p,0) where a fixed level penalty term is replaced by a weighted (adaptive) one. We show that under certain conditions adaptive-DREGAR(p,0) is consistent and enjoys the oracle property.

3.3.1 Adaptive DREGAR(p,0) model

Recall from Eq. (3) that parameter estimation in adaptive-DREGAR(p,q) involves the minimization of

$$Q_n^*(\Theta) = \sum_{t=T_0+1}^T \left((y_t - x_t'\beta) - \sum_{i=1}^p \phi_i y_{t-i} - \sum_{j=1}^q \theta_j \epsilon_{t-j} \right)^2 + \sum_{i=1}^r \lambda_i^* |\beta_i| + \sum_{j=1}^p \gamma_j^* |\phi_j| + \sum_{k=1}^q \tau_k^* |\theta_k|$$

where $\lambda_i^*, \gamma_j^*, \tau_k^*$ are tuning parameters and $\Theta = (\beta, \phi, \theta)'$ is parameter space.

To prove the asymptotic property of adaptive-DREGAR(p,0) we follow [9,22] and define,

$$a_n = \max(\lambda_{i_1}^*, \gamma_{i_2}^*; \quad i_1 \in s_1, i_2 \in s_2)$$

$$b_n = \min(\lambda_{i_1^c}^*, \gamma_{i_2^c}^*; \quad i_1^c \in s_1^c, i_2^c \in s_2^c),$$

where a_n and b_n are maximum and minimum penalties for significant and insignificant coefficients respectively.

Theorem 3 *Let $a_n = o(1)$ as $n \rightarrow \infty$. Then under assumptions [a–d] there is a local minimiser $\hat{\Theta}^*$ of $Q_n^*(\Theta)$ such that*

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

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

$$R_n(\delta) = Q_n^*(\Theta^\circ + \alpha_n \delta) - Q^*(\Theta^\circ)$$

$$\geq L_n(\Theta^\circ + \alpha_n \delta) - L_n(\Theta^\circ) + K_1$$

$$\geq L_n(\Theta^\circ + \alpha_n \delta) - L_n(\Theta^\circ) + K_2$$

$$\geq L_n(\Theta^\circ + \alpha_n \delta) - L_n(\Theta^\circ) + K_3$$

where

$$K_1 = n \sum_{i \in s_1} \lambda_i^* (|\beta_i^\circ + \alpha_n u_i| - |\beta_i^\circ|) + n \sum_{j \in s_2} \gamma_j^* (|\phi_j^\circ + \alpha_n v_j| - |\phi_j^\circ|),$$

$$\text{(Using triangular inequality): } K_2 = -n\alpha_n \sum_{i \in s_1} \lambda_i^* |u_i| - n\alpha_n \sum_{j \in s_2} \gamma_j^* |v_j|,$$

$$\text{(Penalties } \leq \alpha_n \text{ by definition): } K_3 = -n\alpha_n^2(r_o + p_o)d. \tag{10}$$

The last equation holds because of the decreasing speed of α_n . On the other hand, similar calculations to Theorem 1 results in

$$L_n(\Theta^\circ + \alpha_n \delta) - L_n(\Theta^\circ) \rightarrow n\alpha_n^2 \{\delta' U_B \delta + o_p(1)\}. \tag{11}$$

Because (11) dominates (10), then for any given $\eta > 0$, there is a large enough constant d such that

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

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 $\|\hat{\Theta}^* - \Theta^\circ\| = O_p(\alpha_n)$ (see [22, Lemma 1], [5]).

The proof is completed. □

Theorem 3 implies that there exist a \sqrt{n} -consistent local minimiser $Q_n^*(\Theta)$, when tuning parameters (for significant variables) in DREGAR(p,0) converge to zero at the speed *faster* than $n^{-1/2}$.

In the next step we prove that under the case where the tuning parameter associated with insignificant variables in DREGAR(p,0) shrink to zero at a speed *slower* than $n^{-1/2}$, then their associate coefficients will be estimated exactly equal to zero with probability tending to 1. Further, in the next theorem we show that by increasing the penalties on the zero parameters at a certain speed, the probability of these coefficients to be estimated exactly zero tends to one.

Theorem 4 *Let $b_n \sqrt{n} \rightarrow \infty$ and $\|\hat{\Theta}^* - \Theta^\circ\| = O_p(n^{-1/2})$ then*

$$Pr(\hat{\beta}_{s_1^c}^* = 0) \rightarrow 1, \quad Pr(\hat{\phi}_{s_2^c}^* = 0) \rightarrow 1.$$

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

$$\begin{aligned} \left. \frac{\partial Q_n^*(\Theta)}{\partial \beta_i} \right|_{\hat{\Theta}^*} &= \frac{\partial L_n(\hat{\Theta}^*)}{\partial \beta_i} - n\lambda_i^* \text{sign}(\hat{\beta}_i^*) \\ &= \frac{\partial L_n(\Theta^\circ)}{\partial \beta_i} + nU_i(\hat{\Theta}^* - \Theta^\circ)\{1 + o_p(1)\} - n\lambda_i^* \text{sign}(\hat{\beta}_i^*) \end{aligned} \quad (12)$$

where U_i is the i th row of U_B and $i \in s_1^c$. The second term in (12) is a direct result of adding a $\pm X'\beta, \pm H_{(p)}\phi$ to $L_n(\hat{\Theta}^*)$. By using the central limit theorem, the first term in Eq. (12), $\sum_i e_i x'_{ii}$, is of order $O_p(n^{1/2})$ and the second term is $O_p(n^{1/2})$. Furthermore, both terms are dominated by $n\lambda_i^*$ since $b_n \sqrt{n} \rightarrow \infty$ (expansion of [9,22]). Then the sign of $\frac{\partial Q_n^*(\hat{\Theta}^*)}{\partial \beta_i}$ is dominated by the sign of $\hat{\beta}_i^*$, from which $\hat{\beta}_i^* = 0$ in probability. Analogously, we can show that $Pr(\hat{\phi}_{s_2^c}^*) \xrightarrow{P} 1$.

The proof is completed. □

Theorem 4 shows that adaptive-DREGAR(p,0) is capable of producing sparse solutions. Theorems 3 and 4 indicate that a \sqrt{n} -consistent estimator $\hat{\Theta}^*$ must satisfy $Pr(\hat{\Theta}_2^* = 0) \rightarrow 1$. Then, adaptive-DREGAR(p,0) is a sparse model.

Theorem 5 *Assume $a_n \sqrt{n} \rightarrow 0$ and $b_n \sqrt{n} \rightarrow \infty$. Then, under assumptions [a–d] we have*

$$\sqrt{n}(\hat{\Theta}_1^* - \Theta_1^\circ) \xrightarrow{d} \text{MVN}(O, \sigma^2 U_0^{-1}),$$

where U_0 is the sub-matrix U_B corresponding to Θ_1° , and $\hat{\Theta}_1^*$ corresponds to non-zero elements of $\hat{\Theta}^*$.

Proof From Theorems 3 and 4, one can conclude that $Pr(\hat{\Theta}_2^* = 0) \xrightarrow{P} 1$. Thus, the minimiser $Q_n^*(\Theta) \xrightarrow{pr \rightarrow 1} Q_n^*(\Theta_1)$. This implies that the lasso estimator $\hat{\Theta}_1^*$ satisfies the following equation

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

From Theorem 3, $\hat{\Theta}_1^*$ is a \sqrt{n} -consistent estimator. Thus a Taylor expansion of the above equation yields

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

where $p(\cdot)$ is the tuning function

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

For n sufficiently large $p(\hat{\Theta}_1^*) = p(\Theta_1^\circ)$. 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. □

Theorem 5 implies that, adaptive DREGAR(p,0) is asymptotically an oracle estimator provided a_n tends to zero at the speed faster than \sqrt{n} (or $a_n \sqrt{n} \rightarrow 0$) and simultaneously b_n increases at the speed slower than \sqrt{n} (or $b_n \sqrt{n} \rightarrow \infty$).

4 Implementation

The formulation of the model lends itself naturally to its implementation, in contrast to other time series models such as [23]. As the model contains residuals, which are unknown, we apply a two-step optimization procedure

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

Repeating steps 1 and 2 iteratively provides the solution to DREGAR.

The tuning parameters λ , γ and τ can be chosen by K-fold cross-validation or by an information criterion such as AIC, BIC or eBIC. For our model these are given by:

$$\begin{aligned} AIC &= -2L(\Theta) + 2\text{par} \\ BIC &= -2L(\Theta) + \log(T)\text{par} \\ eBIC &= -2L(\Theta) + \log(T)\text{par} + \log(\text{par}), \end{aligned}$$

where par is the number of non-zero estimated parameters. For the simulation and real data analyses, we use eBIC which was found to have a good performance by [4]. For adaptive-DREGAR, we use $\lambda^* = \omega/|\tilde{\beta}|$, $\gamma^* = \omega/|\tilde{\phi}|$ and $\tau^* = \omega/|\tilde{\theta}|$, with $\tilde{\beta}$, $\tilde{\phi}$ and $\tilde{\theta}$ the unpenalized or lasso estimations of the parameters. We assume the same ω for both terms, so that we can simplify the problem to the ordinary adaptive-lasso problem, and select this tuning parameter by one of the criteria mentioned above.

A final choice for model selection is setting the orders p and q . We propose two general approaches to choose the optimal orders: (a) setting an upper bound P and Q and choosing

the model that achieves the minimum eBIC inside the grid, (b) setting an upper bound P and Q and letting the model choose the optimal orders by keeping or eliminating the coefficients under l_1 sparsity constraints. 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.

The method is implemented in the R package *DREGAR*, available in CRAN <http://CRAN.R-project.org/package=DREGAR>.

5 Simulation study

We design a simulation study to compare the (adaptive) DREGAR with (adaptive) REGAR [22] and (adaptive) lasso [25]. To this end we propose the following configuration:

1. Generate the design matrix, X , using a stationary Gaussian process with $r = 100$ and $T = 50, 100, 1000$. That is, high-dimensionality is considered in terms of the number of exogenous variables.
2. Set 90% of β coefficients to zero. Assign unequal random numbers in $(-1, 1)$ to the non-zero coefficients.
3. Generate data from the DREGAR(2,2) model

$$\begin{aligned}
 y_t &= 0.5y_{t-1} + 0.2y_{t-2} + X\beta + \epsilon_t \\
 \epsilon_t &= 0.3\epsilon_{t-1} - 0.1\epsilon_{t-2} + e_t \\
 e_t &\sim N(0, \sigma^2),
 \end{aligned}$$

with $\sigma^2 = 0.5, 1$ and 1.5 .

4. Sample 1500 data points from the above model so that the first 50, 100 or 1000 observations are used for parameter estimation (training set) and the rest $n = 1500 - T$ points are left for evaluating the model performance (test set).
5. Select tuning parameters by minimizing *eBIC* and fix the maximum orders P and Q to 3 (i.e. allowing also for variable selection for ϕ and θ).

We repeat each combination of models 100 times and calculate mean squared error of $\hat{\beta}$ and the prediction mean squared error, defined by

$$PMSE = \frac{1}{n} \sum_1^n (y_{\text{test}} - \hat{y}_{\text{test}})^2,$$

where \hat{y}_{test} is calculated using the two steps discussed in the implementation. We compare DREGAR(3,3) with lasso and with a DREGAR(0,6) model, which has the same number of parameters as DREGAR(3,3) and is the closest in the DREGAR family to a REGAR model [22].

Figures 1 and 2 show overall how adaptive DREGAR dominates lasso and REGAR for low and high-dimensional problems in terms of both prediction error and MSE of $\hat{\beta}$. Table 1 shows the full set of results for PMSE with a better performance of DREGAR across the range of simulations.

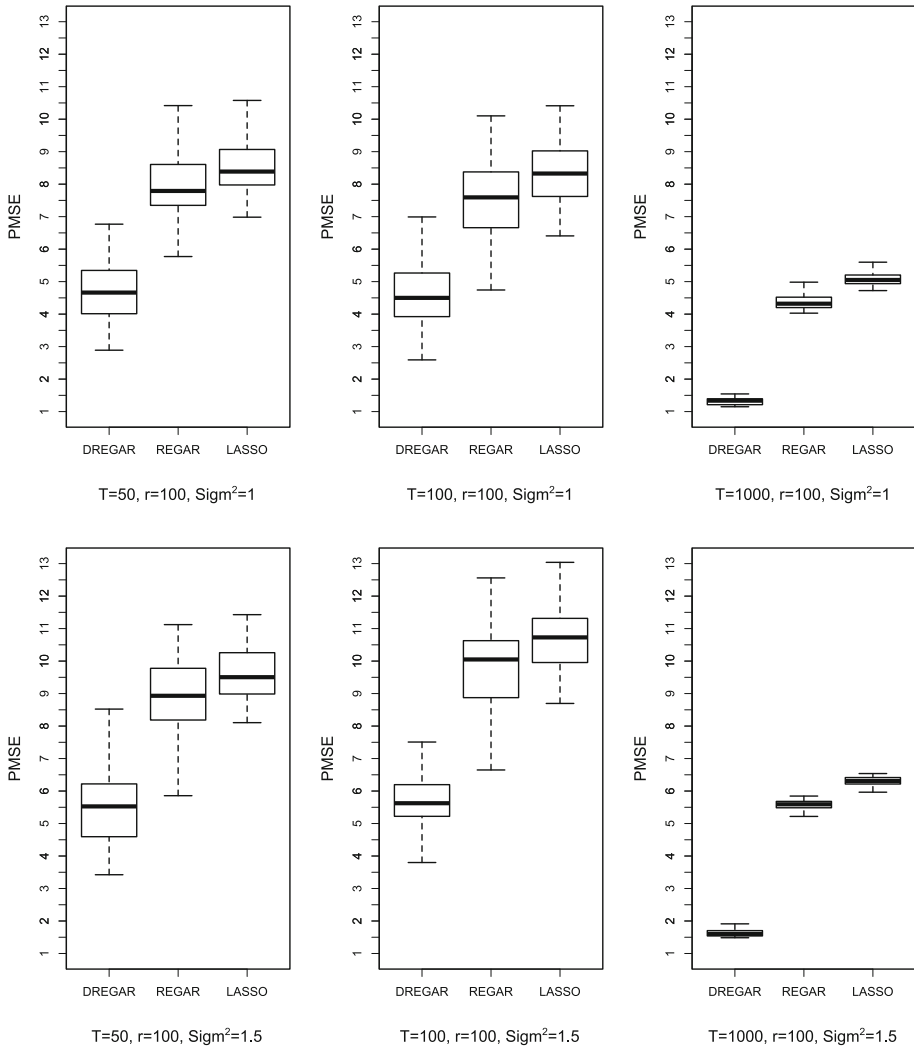


Fig. 1 Comparison of adaptive DREGAR, adaptive REGAR and adaptive lasso on the simulated data with respect to PMSE for different values of $T = 50, 100, 1000$ and $r = 100$ for 90% sparsity in the regression coefficients. The *top figures* refer to $\sigma^2 = 1$ and the *bottom figures* to $\sigma^2 = 1.5$

6 Real data illustration

6.1 Analysis of air pollution data

In this section, we show the performance of the model on the National Mortality, Morbidity and Air Pollution Study (NMMAPS) dataset. This dataset is publicly available from <http://www.ihapss.jhsph.edu/data/NMMAPS/> and contains daily mortality, air pollution, and weather data for 108 cities in the US from January 1, 1987 to December 31, 2000. The variables include six indicators for mortality (total non-accidental, cardiovascular disease, respiratory, pneumonia, chronic obstructive pulmonary disease, accidental),

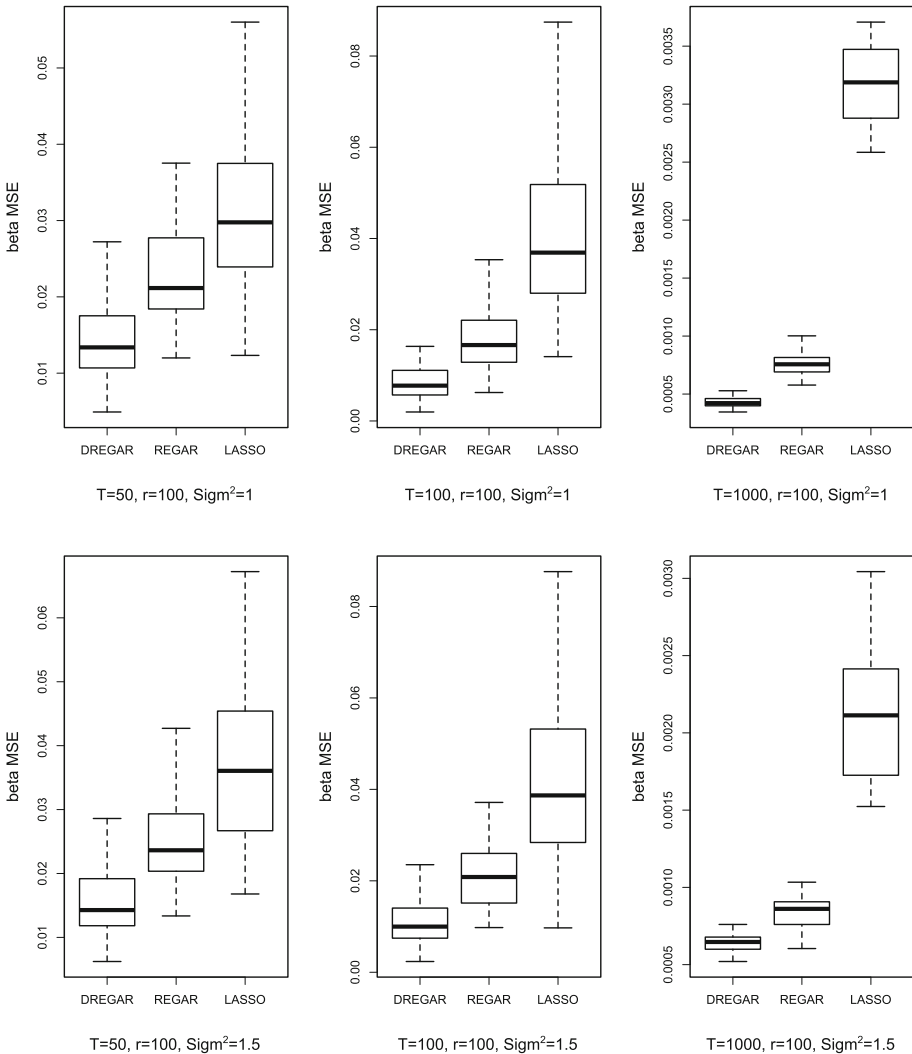


Fig. 2 Comparison of adaptive DREGAR, adaptive REGAR and adaptive lasso on the simulated data with respect to MSE of estimations for different values of $T = 50, 100, 1000$ and $r = 100$. The *top figures* refer to $\sigma^2 = 1$ and the *bottom figures* to $\sigma^2 = 1.5$

six indicators of air pollution (repairable particulates (PM10)/(PM25), carbon monoxide (CO), ozone (O₃), sulphur dioxide (SO₂), nitrogen dioxide (NO₂)) as well as three indicators of weather (temperature (T), dew point temperature (D) and relative humidity (H)). Similar to [23] we study the relationship between ground level of ozone and indicators of air pollution and weather conditions in Chicago in 1995. Differently to [23], we take the effect of carbon monoxide (CO) into account. The covariates in the model consist of NO₂, SO₂, CO, PM10, temperature and relative humidity as well as all two-ways interactions. We show the interactions by initials, for instance NS represents the interaction between NO₂ and SO₂. A total number of 365 observations and 21 covariates are included

Table 1 Comparing adaptive-DREGAR, adaptive-REGAR and adaptive-lasso with respect to PMSE for $T = 50, 100, 1000$ and $\sigma^2 = 0.5, 1, 1.5$

σ	T	DREGAR	REGAR	LASSO
0.5	50	3.61 (0.74)	6.04 (0.92)	6.7 (0.68)
	100	4.18 (3.77)	6.16 (2.73)	6.83 (3.32)
	1000	0.56 (0.13)	3.1 (0.31)	3.76 (0.37)
1	50	4.74 (0.91)	7.96 (1.06)	8.54 (0.76)
	100	4.61 (1.36)	7.59 (1.38)	8.38 (1.11)
	1000	1.32 (0.15)	4.37 (0.25)	5.1 (0.24)
1.5	50	5.46 (1.13)	8.89 (1.13)	9.62 (0.81)
	100	6.26 (3.49)	10.1 (2.8)	11.13 (2.97)
	1000	1.57 (0.09)	5.57 (0.14)	6.31 (0.13)

Averages across 100 iterations are reported with standard deviations in brackets

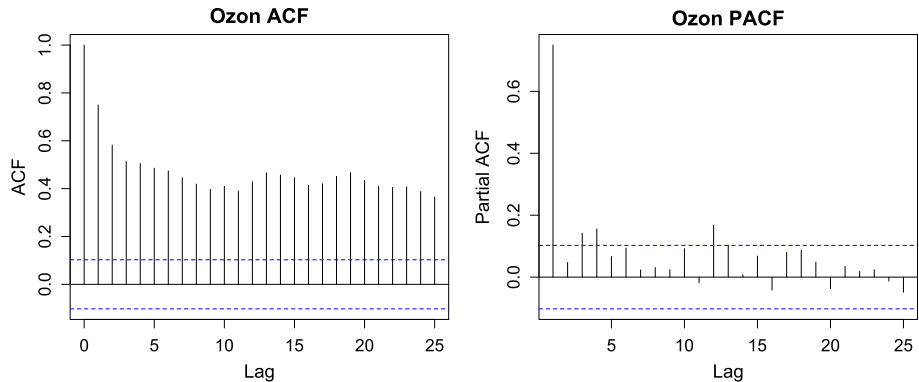


Fig. 3 The autocorrelation function (ACF) and partial autocorrelation function (PACF) of the air pollution data for the first 25 lags

in the analysis. All covariates and response are normalized to zero mean and unit variance.

Figure 3 shows a difficult choice for the maximum orders P and Q . We therefore follow the second approach in Sect. 4 and propose $P = 5$ and $Q = 5$ and let the model’s parameter inference guide the best orders. The parameters are estimated using the adaptive algorithm in Sect. 4, setting a maximum of 50 iterations and selecting the tuning parameters by minimizing eBIC.

We compare the optimal DREGAR(p,q) model with alternative models of similar complexity or natural subclasses. In particular, we consider DREGAR(p+q,0), DREGAR(p,0), DREGAR(0,q+p) and DREGAR(0,q). Note that the last two are the closest models to [22] in the DREGAR family. In addition, we consider standard non-dynamic models, namely adaptive-lasso and elastic-net. For the latter, we choose the optimal proportion of norms α over a range of 100 values. We compare the models on the basis of a number of commonly used criteria: eBIC, AIC, Quasi-likelihood Information Criteria (QIC) [18] and Consistent AIC (CAIC) [3].

Table 2 provides a detailed illustration of the parameter estimates as well as information on the comparison of the models. Time series coefficients in the middle-bottom of the table propose an order of four and three for DREGAR as well as DREGAR(1,0) and DREGAR(0,3) for the other models, suggesting that the maximum order of 5 for p and q is sufficient.

Table 2 (Top) Comparing adaptive-lasso, elastic net and adaptive-DREGAR models on the air pollution data with respect to eBIC, AIC, CAIC and QIC where the asterisk denotes the minimum value

Model comparison							
Model	eBIC	AIC	CAIC	QIC			
Lasso	2457.37	2408.09	2466.89	6.82			
enet	2493.52	2420.43	2508.63	6.70			
DREGAR(5,5)	2349.27*	2280.28*	2363.44*	6.40*			
DREGAR(0,10)	2355.88	2282.94	2370.99	6.49			
DREGAR(10,0)	2350.12	2296.90	2368.21	6.42			
DREGAR(0,5)	2399.92	2295.98	2422.66	6.80			
DREGAR(5,0)	2358.64	2297.48	2370.93	6.45			
Parameter estimation							
Variables	lasso	enet	p = 5 q = 5	p = 10 q = 0	p = 0 q = 10	p = 0 q = 5	p = 5 q = 0
Temp (T)	5.29	5.24	5.27	5.26	5.26	3.51	3.48
PM10(P)	0.00	-0.03	-0.06	0.00	0.00	0.00	0.00
SO2(S)	-10.53	-5.66	-6.09	-12.04	-11.78	-7.94	-8.43
NO2(N)	-2.87	-1.75	-1.76	-1.95	-2.16	-1.28	-1.64
Hum(H)	-1.10	-1.62	-1.61	-1.25	-1.26	-1.12	-1.11
CO(C)	-0.18	-1.95	-2.02	-0.46	-0.28	-1.83	-0.67
NS	0.00	-0.04	0.00	0.20	0.19	0.41	0.51
NP	-0.41	-1.47	-1.48	-0.49	-0.50	-0.85	-1.00
NT	0.00	-0.71	-0.78	0.00	0.00	0.00	0.00
NH	1.08	0.82	0.91	0.19	0.36	0.00	0.41
SP	0.00	0.71	0.69	0.00	0.00	0.40	0.49
ST	6.29	4.60	4.79	7.22	7.10	4.28	4.33
SH	5.34	1.59	1.85	6.06	5.87	4.48	4.88
PT	3.60	3.73	3.74	3.25	3.26	2.68	2.65
PH	0.00	0.00	0.00	-0.07	0.00	0.00	0.00
TH	0.00	0.00	0.00	0.09	0.05	0.00	0.00
CN	0.00	0.12	0.13	0.19	0.22	0.00	0.00
CS	0.00	0.84	0.83	0.00	0.00	0.00	0.00
CP	0.00	0.00	0.00	0.00	0.00	0.00	0.00
CT	-0.47	-0.45	-0.39	-1.25	-1.18	-0.11	-0.28
CH	-1.61	-0.01	0.00	-0.92	-1.18	0.00	-0.99
Time series coefficients							
	lasso	enet	p = 5 q = 5	p = 10 q = 0	p = 0 q = 10	p = 0 q = 5	p = 5 q = 0
-	-	-	$\phi_1 = 0.23$	$\phi_1 = 0.46$	$\theta = 0.36$	$\theta_1 = 0.35$	$\phi_1 = 0.46$
-	-	-	$\phi_2 = 0$	$\phi_2 = 0$	$\theta_2 = 0$	$\theta_2 = 0$	$\phi_2 = 0$
-	-	-	$\phi_3 = 0$	$\phi_3 = 0$	$\theta_3 = 0.08$	$\theta_3 = 0.09$	$\phi_3 = 0$
-	-	-	$\phi_4 = 0.09$	$\phi_4 = 0$	$\theta_4 = 0$	$\theta_4 = 0$	$\phi_4 = 0$

Table 2 continued

Time series coefficients							
lasso	enet	p = 5	p = 10	p = 0	p = 0	p = 5	
		q = 5	q = 0	q = 10	q = 5	q = 0	
–	–	$\phi_5 = 0$	$\phi_5 = 0$	$\theta_5 = 0$	$\theta_5 = 0$	$\phi_5 = 0$	
–	–	$\theta_1 = 0.26$	$\phi_6 = 0$	$\theta_6 = 0$	–	–	
–	–	$\theta_2 = 0.32$	$\phi_7 = 0$	$\theta_7 = 0$	–	–	
–	–	$\theta_3 = 0.16$	$\phi_8 = 0$	$\theta_8 = 0$	–	–	
–	–	$\theta_4 = 0$	$\phi_9 = 0$	$\theta_9 = 0$	–	–	
–	–	$\theta_5 = 0$	$\phi_{10} = 0$	$\theta_{10} = 0$	–	–	

Ljung–Box statistic							
	lasso	enet	p = 5	p = 10	p = 0	p = 0	p = 5
			q = 5	q = 0	q = 10	q = 5	q = 0
P-value	0	0	0.61	0.08	0.54	0.52	0.09

(Middle-top) Estimation of the regression coefficients. (Middle-bottom) Estimation of the time-dependent coefficients. (Bottom) Ljung–Box p-value for the null hypothesis of residuals following white noise

DREGAR(4,3) shows better results than DREGAR(p,0), and DREGAR(0,q) models with respect to model performance as shown in the top panel of the table. In line with [23], our results show several significant interactions, especially those between sulphur dioxide-temperature (ST) and humidity (SH), as well as between particulates and temperature (PT). However, we should stress that the two analyses are not directly comparable, since we consider an additional variable, CO, which shows a significant effect on the ozone ground level and a non-zero effect for the interaction with weather indicators, namely CT and CS. We further report the Ljung–Box test [2] statistics in the bottom of the Table 2. With the exception of lasso, elastic-net, DREGAR(5,0) and DREGAR(10,0), all other models show good fitting, i.e. no evidence against the white noise assumption. Figure 4 displays the scatterplot of fitted versus observed response for lasso and DREGAR(4,3), the residuals from the DREGAR(4,3) mode and the corresponding sample ACF and PACF. The small curvature in the scatter plot, mentioned also by [23], can be an indication of a particular weather condition that results in an interaction between primary pollutants. The residuals’ ACF and PACF suggest that the residuals are white noise as confirmed also by the p-value of the Ljung–Box test (0.61).

Finally, we have also compared the fit of the best DREGAR model, DREGAR(4,3), with a DREGAR(0,7) model (the closest to a REGAR(7) model), in order to assess the benefit in having different autoregressive structures for the response and the predictors, a unique feature of the model that we propose in this paper. Without penalising the coefficients, the maximum likelihood for DREGAR(4,3) is -1106.884 and that of DREGAR(0,7) is -1110.832 , suggesting an improved fit for the DREGAR(4,3) model.

6.2 Analysis of stock market data

For the second real application we take an example from the stock market. Data are collected from yahoo finance (<https://finance.yahoo.com>) and contain 251 closing prices for 30 indices in the DowJones market in 2015. We take the IBM index as the response and the remaining

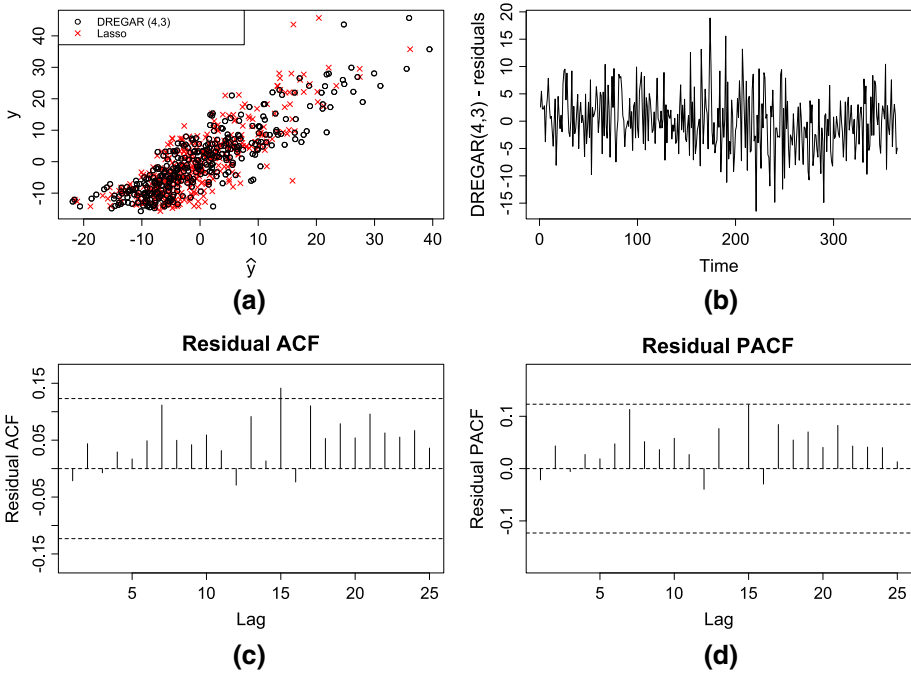


Fig. 4 Diagnostic plots on air pollution data. **a** Scatter plot of fitted versus observed response for adaptive DREGAR(4,3) and adaptive lasso, **b** DREGAR(4,3) residuals, **c** sample ACF and PACF for the DREGAR(4,3) residuals, **d** sample PACF of the DREGAR(4,3) residuals

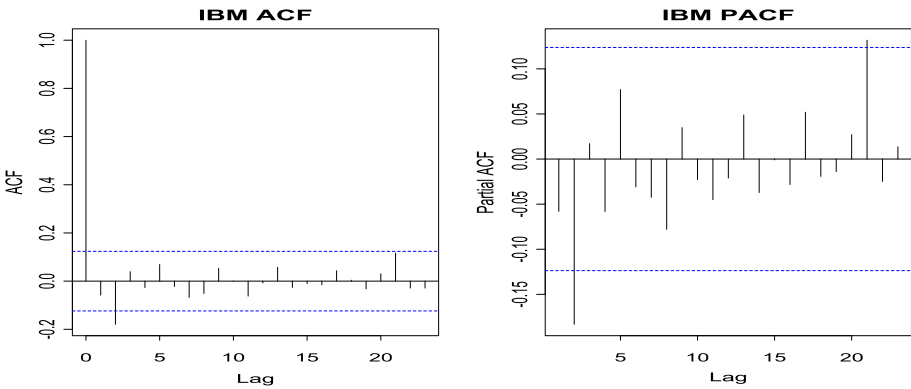


Fig. 5 The autocorrelation function (ACF) and partial autocorrelation function (PACF) of the IBM data for the first 25 lags

29 indices as the covariates and study their correlations via the DREGAR family of models. The variables are listed as follows: 3M (MMM), American Express (AXP), Alcoa (AA), AT&T (T), Bank of America (BAC), Boeing (BA), Caterpillar (CAT), Chevron (CVX), Cisco Systems (C), Coca-Cola (KO), DuPont (DD), ExxonMobil (XOM), General Electric (GE), Hewlett-Packard (HPQ), The Home Depot (HD), Intel (INTC), IBM (IBM), Johnson & Johnson (JNJ), JPMorgan Chase (JPM), Kraft (KRFT), McDonald’s (MCD), Merck (MRK),

Table 3 Comparison of adaptive-lasso, elastic-net, GARCH and adaptive-DREGAR for the DowJones30 dataset on the basis of BIC, AIC, CAIC, QIC, sparsity and Ljung–Box statistic

Model	eBIC	AIC	CAIC	QIC	Ljung–Box p-value	#Non-zero
GARCH(1,1)	889.50	726.70	912.03	2.7	0.01	29
Lasso	644.30	549.89	581.57	2.2	0.06	7
elastic-net	624.05	560.59	624.05	2.5	0.05	18
DREGAR(5,5)	561.27*	528.91*	598.43*	2.3	0.12	7
DREGAR(10,0)	569.94	542.61	610.10	2.4	0.11	7
DREGAR(0,10)	581.50	536.82	604.70	2.3	0.11	9
DREGAR(5,0)	586.31	537.58	605.47	2.4	0.12	8
DREGAR(0,5)	590.74	543.19	611.10	2.3	0.15	7

For the information criteria, the asterisk denotes the minimum

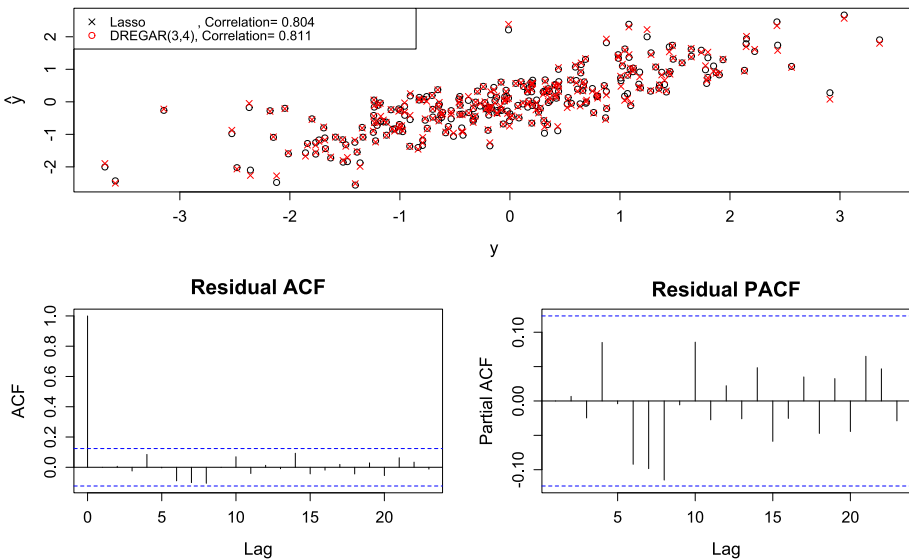


Fig. 6 Diagnostic plots for the DowJones30 analysis. (Top) Scatter plot of fitted versus observed y for adaptive DREGAR(3,4) and adaptive lasso. (Bottom) Sample ACF and PACF of the DREGAR(3,4) residuals

Microsoft (MSFT), Pfizer (PFE), Procter & Gamble (PG), General Motors (GM), United Technologies (UTX), Verizon (VZ), Wal-Mart (WMT), Walt Disney (DIS).

We apply first differences of the log-prices to get stationary returns [10]. Figure 5 shows low orders of auto-correlations in the residuals as typical of financial data.

Adaptive DREGAR(5,5), DREGAR(10,0), DREGAR(0,10), DREGAR(5,0) and DREGAR(0,5) are applied to the data and the tuning parameters are selected using eBIC. In addition, we consider adaptive-lasso and elastic net as well as GARCH, which is typically used for financial data. For GARCH, we use the R package `rugarch` and choose the optimal model by searching among all models with maximum orders (2,2). The models are compared on the basis of eBIC, AIC, CAIC, QIC, Ljung–Box statistic and sparsity.

Table 3 shows that DREGAR(5,5) performs very well compared to other methods with respect to eBIC, AIC and CAIC as well as sparsity. Fitting adaptive DREGAR(5,5) to data

results in an order of 3 for the dynamic term and an order of 4 for the residuals. So the final selected model is DREGAR(3,4), suggesting that a maximum order of 5 for p and q is adequate also for this dataset. Amongst the top selected predictors, there are: MSFT (coefficient 0.3), HPQ (0.23), VZ (0.20), MMM (0.14), MRK (0.13) and CVX (0.10). Figure 6(top) shows observed versus fitted response for lasso and DREGAR(3,4). From this figure, DREGAR(3,4) has a better fit compared to lasso in terms of the correlation between the observed and fitted values ($\rho_{DREGAR(3,4)} = 0.811$, $\rho_{lasso} = 0.804$). Finally, the sample ACF and PACF at the bottom of Fig. 6 confirm the results from the Ljung–Box statistic, showing that the residuals from DREGAR(3,4) behave like white noise.

Similarly to the previous example, we compare the fit of the best DREGAR(3,4) with a DREGAR(0,7) model. Without penalising the coefficients, the maximum likelihood for DREGAR(3,4) is -243.98 and that of DREGAR(0,7) is -251.41 , suggesting an improved fit for the DREGAR(4,3) model.

7 Conclusion

This paper addressed the problem of dynamic regression in the presence of autocorrelated residuals by proposing an extension of the regression model of [22] with the inclusion of lags of the response. We showed that adding this dynamic term results in a structure more similar to a general ARMAX model than REGAR [22] and REGARMA [23] and with fewer difficulties in parameter estimations than REGARMA. Further, we proposed an l_1 penalized likelihood approach for variable selection for both regression and time-dependent coefficients and studied its theoretical properties. We proposed two iterative algorithms for parameter estimation and provided an R package that contains the implementations and simulation from the model. Finally, we show the applicability of the model and comparison with existing approaches in the simulation study as well as two real data applications.

Future work could extend the methods presented in this paper by estimating DREGAR coefficients using penalties that strike a trade-off between l_1 and l_2 norms, such as elastic net. We expect these methods to work well, as the l_2 penalty imposes less weight on small coefficients compared to the l_1 penalty. Such an extension is also expected to work well in the presence of correlation among the predictors. Moreover, it would be interesting to add GARCH-type errors to the model, similar to a recent contribution to the literature for the REGARMA model [24]. Finally, it would be of interest to extend the methodology to non-linear and non-stationary cases.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

References

1. Basu, S., Michailidis, G.: Regularized estimation in sparse high-dimensional time series models. *Ann. Stat.* **43**(4), 1535–1567 (2015)
2. Box, G.E., Pierce, D.A.: Distribution of residual autocorrelations in autoregressive-integrated moving average time series models. *J. Am. Stat. Assoc.* **65**(332), 1509–1526 (1970)
3. Bozdogan, H.: Model selection and Akaike's information criterion (AIC): the general theory and its analytical extensions. *Psychometrika* **52**(3), 345–370 (1987)

4. Chen, J., Chen, Z.: Extended Bayesian information criteria for model selection with large model spaces. *Biometrika* **95**(3), 759–771 (2008). doi:[10.1093/biomet/asn034](https://doi.org/10.1093/biomet/asn034)
5. Fan, J., Li, R.: Variable selection via penalized likelihood and its oracle properties. *J. Am. Stat. Assoc.* **96**(456), 1348–1360 (2001)
6. Han, F., Lu, h., Liu, H.: A direct estimation of high dimensional stationary vector autoregressions. *J. Mach. Learn. Res.* **16**, 3115–3150 (2015). <http://jmlr.org/papers/v16/han15a.html>
7. Huang, J., Ma, S., Zhang, C.H.: Adaptive lasso for sparse high-dimensional regression models. *Stat. Sin.* **18**(4), 1603 (2008)
8. Kim, J., Pollard, D.: Cube root asymptotics. *Ann. Stat.* **18**(1), 191–219 (1990)
9. Knight, K., Fu, W.: Asymptotics for lasso-type estimators. *Ann. Stat.* **28**(5), 1356–1378 (2000)
10. Kwiatkowski, D., Phillips, P.C., Schmidt, P., Shin, Y.: Testing the null hypothesis of stationarity against the alternative of a unit root. *J. Econom.* **54**(1), 159–178 (1992)
11. Ljung, L.: *System Identification: Theory for the User*. Pearson Education, Englewood Cliffs (1998)
12. Medeiros, M.C., Mendes, E.F.: Estimating high-dimensional time series models. Technical report (2012)
13. Meinshausen, N., Bühlmann, P.: High-dimensional graphs and variable selection with the lasso. *Ann. Stat.* **34**(3), 1436–1462 (2006)
14. Melnyk, I., Banerjee, A.: Estimating structured vector autoregressive models. In: Balcan, M.F., Weinberger, K.Q. (eds.) *Proceedings of the 33rd International Conference on Machine Learning: Proceedings of Machine Learning Research*, vol. 48. PMLR, New York, pp. 830–839 (2016)
15. Nardi, Y., Rinaldo, A.: Autoregressive process modeling via the lasso procedure. *J. Multivar. Anal.* **102**(3), 528–549 (2011)
16. Nelles, O.: *Nonlinear System Identification: From Classical Approaches to Neural Networks and Fuzzy Models*. Springer, Berlin (2013)
17. Nicholson, W.B., Matteson, D.S., Bien, J.: Varx-l: Structured regularization for large vector autoregressions with exogenous variables. *Int. J. Forecast.* **33**(3), 627–651 (2017)
18. Pan, W.: Akaike’s information criterion in generalized estimating equations. *Biometrics* **57**(1), 120–125 (2001)
19. Park, T., Casella, G.: The Bayesian lasso. *J. Am. Stat. Assoc.* **103**(482), 681–686 (2008)
20. Song, S., Bickel, P.: Large vector auto regressions (2011). [arXiv:1106.3915](https://arxiv.org/abs/1106.3915)
21. Tibshirani, R.: Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B* **58**(1), 267–288 (1996)
22. Wang, H., Li, G., Tsai, C.: Regression coefficient and autoregressive order shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B* **69**(1), 63–78 (2007)
23. Wu, R., Wang, Q.: Shrinkage estimation for linear regression with ARMA errors. *J. Stat. Plan. Inference* **142**(7), 2136–2148 (2012)
24. Yoon, Y.J., Lee, S., Lee, T.: Adaptive lasso for linear regression models with ARMA-GARCH errors. *Commun. Stat. Simul. Comput.* **46**(5), 3479–3490 (2017)
25. Zou, H.: The adaptive lasso and its oracle properties. *J. Am. Stat. Assoc.* **101**(476), 1418–1429 (2006)