Automatic covariates selection in dynamic regression models with application to COVID-19 evolution

Ana Ezquerro [1] Germán Aneiros [2] Manuel Oviedo [3]

[1]University of A Coruña, ana.ezquerro@udc.es

[2]CITIC, Grupo MODES, Departamento de Matemáticas, University of A Coruña, german.aneiros@udc.es

[3]CITIC, Grupo MODES, Departamento de Matemáticas, University of A Coruña, manuel.oviedo@udc.es

XoveTIC 2022 Conference

Introduction

The **linear dynamic regression model** defines the linear dependence between a stochastic process Y_t and a set of processes $\mathcal{X} = \{X_t^{(1)}, ..., X_t^{(m)}\}$:

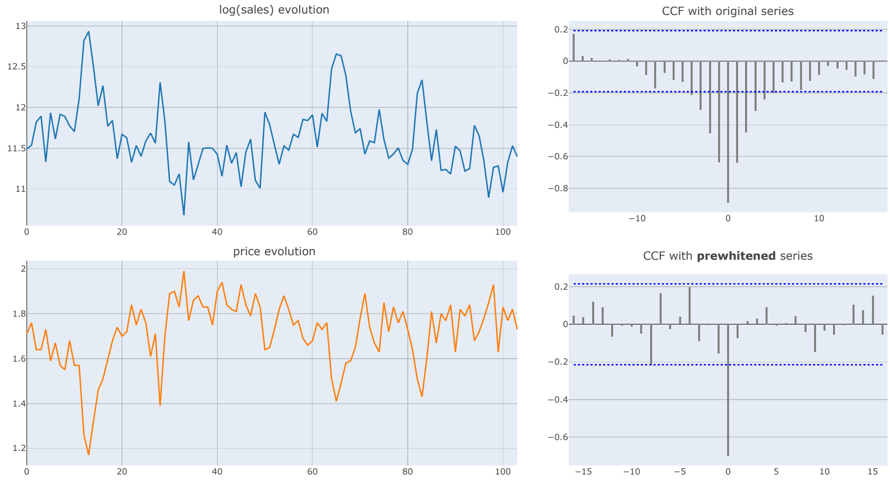
$$Y_t = \beta_0 + \beta_1 X_{t-r_1}^{(1)} + \cdots \beta_m X_{t-r_m}^{(m)} + \eta_t$$

constrained to $r_i \geq 0$ for $i = 1, ..., m$ and $\eta_t \sim \text{ARMA}(p, q)$[1].

- [Cryer and Chan, 2008] proposed the *prewhitening* as a technique for removing spurious correlation between processes in order to detect linear correlation.
- We propose a forward-selection method that iteratively adds regressor variables (from a set of candidates) Y_t is *significantly* dependent with.

[1]here we denote the AutoRegressive Moving Average model as ARMA.

Example of spurious correlation and prewhitening



Methodology

Let Y_t be the dependent variable and $\mathcal{X}$ the set of covariates candidates. Thus, selection proceeds as follows:

1. Initialization. Consider X_t^{best} as the covariate (lagged r moments) that minimizes the IC of the constructed model with Y_t:

$$X_t^{\text{best}} = \arg\min_{X_t \in \mathcal{X}} \left\{ \text{IC}\left(Y_t = \beta_0 + \beta_1 X_{t-r}^{\text{best}} + \eta_t \right) \right\}$$

2. Iteration. Use the regression errors (η_t) of the last model created to check if some correlation exists between the rest of the covariates not yet added to the model. Find again the "best" variable and add it to the model to obtain a new IC value. If this value improves the last one achieved, repeat this step. If it does not, stop the iteration.
3. Finalization. The errors of the last fitted model must satisfy the stationary property. In other case, consider the regular differentiation of all data and start again the procedure.

Simulation procedure

1. Seven time series (modelable by an ARIMA) were generated: six act as covariate candidates $\mathcal{X} = \{X_t^{(1)}, ..., X_t^{(6)}\}$ with random lags $r \in [0, 6]$, and the remaining as the errors of the model.
2. Formally, each simulation follows this scheme:

$$Y_t = \beta_0 + \beta_1 X_{t-r_1}^{(1)} + \beta_2 X_{t-r_2}^{(2)} + \beta X_{t-r_3}^{(3)} + \eta_t$$

   where $\eta_t \sim \text{ARMA}(p, q)$, $\beta_0, ..., \beta_3$ are randomly generated and $r_i \in [0, 6]$ for $i = 1, .., 3$.
3. Selection method was tested with different configurations:
   - Changing the IC with three different options: AIC, BIC or AICc.
   - Changing the method to check stationary: via the Dickey-Fuller test or via adjusting an ARIMA and checking the differentiation order.

Example of one simulation result

Figure 1: Code output when launching the function `drm.select()` that implements our approach

```
beta0 <- -0.6; beta1 <- 1.7; beta2 <- -2.2; beta3 <- 1.3; r1 <- 2; r3 <- 3
Y <- beta0 + beta1*lag(X1,-r1) + beta2*X2 + beta3*lag(X3,-r3) + residuals
xregs <- cbind(X1, X2, X3, X4, X5, X6)
ajuste <- drm.select(Y, xregs, ic='aicc', st_method='adf.test', show_info=F)
```

```
print(ajuste$history, row.names=F)
```

```
 var lag                ic
  X2   0 -1156.68486061937
  X1  -2 -2171.66958134745
  X3  -3 -3108.15443209894
```

```
print(ajuste, row.names=F)
```

```
Series: serie
Regression with ARIMA(0,0,4) errors

Coefficients:
         ma1     ma2  ma3     ma4  intercept      X2      X1      X3
      0.2498  0.3360    0  0.1589    -0.5947 -2.1868  1.6949  1.3083
s.e.  0.0304  0.0302    0  0.0300     0.0033  0.0105  0.0089  0.0320

sigma^2 = 0.002377:  log likelihood = 1562.15
AIC=-3108.3   AICc=-3108.15   BIC=-3069.26
```

Results of multiple simulations where $\eta_t \sim$ ARMA(p,q)

Table 1: Percentage data results when residuals are stationary

	AIC	**BIC**	**AICc**	**AIC**	**BIC**	**AICc**
adf.test	97.66%	97.66%	97.66%	3.66%	1.33%	3.66%
auto.arima	98.33%	98.33%	98.33%	3.66%	1.33%	3.66%
	correctly added (TP)			incorrectly added (FP)		

	AIC	**BIC**	**AICc**	**AIC**	**BIC**	**AICc**
adf.test	96.33%	98.66%	96.33%	2.33%	2.33%	2.33%
auto.arima	96.33%	98.66%	96.33%	1.66%	1.66%	1.66%
	correctly **not** added (TN)			incorrectly **not** added (FN)		

Results of multiple simulations where $\eta_t \sim$ ARIMA(p,d,q)

Table 2: Percentage data results when residuals are non-stationary

	AIC	**BIC**	**AICc**	**AIC**	**BIC**	**AICc**
adf.test	93.33%	93.33%	93.33%	4.33%	0.30%	4.33%
auto.arima	94.33%	94.66%	95.33%	5.00%	1.33%	5.00%
	correctly added (TP)			incorrectly added (FP)		

	AIC	**BIC**	**AICc**	**AIC**	**BIC**	**AICc**
adf.test	95.00%	98.66%	95.00%	6.66%	6.66%	6.66%
auto.arima	94.66%	99.66%	95.66%	4.66%	5.33%	4.66%
	correctly **not** added (TN)			incorrectly **not** added (FN)		

Application to COVID19 evolution

Table 3: Information about the dynamic regression model constructed via selection of multiple vaccination variables to model COVID19 evolution

Covariate	Lag	Coefficient est. (s.e)
`vac4565`	-3	-0.0410 (0.0057)
`vac6580`	-2	-0.0468 (0.0120)
`vac1845`	-6	-0.0901 (0.0047)
`vac1218`	Not included in the model	
`vac80`	Not included in the model	
residuals	ARIMA(4, 0, 0)	$\phi_1 = 2.0816(0.0810)$ $\phi_2 = -1.2837(0.1152)$ $\phi_4 = 0.1919(0.0432)$

- Lagged negative correlation between the vaccination data and COVID-19 evolution.
- Vaccination of young population (from 18 up to 45 years old) has a greater impact in the COVID19 evolution.
- Vaccination of the youngest and oldest range of ages has no significative impact in the COVID19 confirmed cases.

Conclusions and future work

- R implementation is openly available in:

  https://github.com/anaezquerro/dynamic-arimax

- This code has been optimized and some steps are parallelized.
- Widespread application in financial, economic and biomedical fields.
- Other covariates might be considered, such as discrete or functional variables.

Thanks to

- Department Collaboration Scholarship financed by Banco Santander.
- Mathematics Department of the University of A Coruña.

Cryer, J. D. and Chan, K.-S. (2008).
Time series analysis: with applications in R, chapter 11.
2.