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

Ana Xiangning Pereira Ezquerro
ana.ezquerro@udc.es

Germán Aneiros Pérez
german.aneiros@udc.es

Manuel Oviedo de la Fuente
manuel.oviedo@udc.es

Departamento de Matemáticas, Curso 2021-2022

In time-series analysis, the well-known dynamic regression models allow formally modelling the dependence between a set of covariates and a dependent variable considering the intrinsic temporal component of all participant variables. Based on a previous study of [Cryer and Chan, 2008] and [Hyndman and Athanasopoulos, 2018], a forward-selection method is proposed for adding new significant covariates from a given set to a regression model with their respective optimal lags.

Formally, dynamic linear regression models define the lineal dependence between a stochastic proccess Y_t and a set of proccesses $\mathcal{X} = \{X_t^{(1)}, X_t^{(2)}, ..., X_t^{(m)}\}$ in times non-greater than t:

$$Y_t = \beta_0 + \beta_1 X_{t-r_1}^{(1)} + \beta_2 X_{t-r_2}^{(2)} + \cdots + \beta_m X_{t-r_m}^{(m)} + \eta_t$$

where $r_i \geq 0$, for $i = 1, ..., m$, and $\eta_t \sim$ ARMA(p,q).

[Cryer and Chan, 2008] proposed a method named *prewhitening* for removing spurious correlation between two processes X_t and Y_t and, thus, cleanly detecting the existence of lineal dependency between them, as well as the optimal lag r dependency occurs in. Following this methodology, our approach adds iteratively dependent processes to a model by checking if a significant correlation (following [Cryer and Chan, 2008] methodology) exists between a new proccess and the residuals η_t of the model.

Given a stochastic proccess Y_t and a set of proccesses $\mathcal{X} = \{X_t^{(1)}, X_t^{(2)}, ..., X_t^{(m)}\}$ which act as regressor variables in the model, and an information criterion for model evaluation, the method proceeds as follows:

1. Search in $\mathcal{X}$ the proccess with its optimal lag that best simple regression model produces,

based on the selected information criterion.

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

2. If X_t^{best} exists, construct the model $\mathcal{M} : Y_t = \beta_0 + \beta_1 X_{t-r}^{\text{best}} + \eta_t$, remove X^{best} from $\mathcal{X}$ and proceed iteratively:
   2.1. Search in $\mathcal{X}$ the proccess X_t^{best} such that:

   $$X_t^{\text{best}} = \arg\min_{X \in \mathcal{X}} \Big\{ \text{criteria}\big(\tilde{Y}_t = \beta_0 + \beta_1 X_{t-r} + \eta_t\big) \Big\}$$

   restricted to $\text{criteria}\big(\tilde{Y}_t = \beta_0 + \beta_1 X_{t-r} + \eta_t\big) < \text{criteria}(\mathcal{M})$.

   2.2. In case of finding such model, consider a new $\mathcal{M} : \tilde{Y}_t = \beta_0 + \beta_1 X_t^{\text{best}} + \eta_t$ and $\tilde{Y}_t = \eta_t$ and return to (2.1). Otherwise, stop the iteration.
3. The dynamic regression model is formed by the set of proccesses selected as X_t^{best} in each iteration of the algorithm.

This new proposal has been implemented and optimized in the statistical language R as a package, and it has been applied to multiple simulations to validate its performance. Finally, the obtained results from the IRAS database of Catalonia are presented to analyze the COVID-19 evolution.

References

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

[Hyndman and Athanasopoulos, 2018] Hyndman, R. J. and Athanasopoulos, G. (2018). *Forecasting: principles and practice*, chapter 9. OTexts.