Optimal lag selection for covariates in INGARCH models:

an application to the analysis of air quality effect on daily respiratory hospital admissions

Ana Martins[1], Manuel Scotto[2], Sónia Gouveia[1]

[1] IEETA, [2] CEMAT-IST

26th Portuguese Conference on Pattern Recognition, RecPad 2020, Évora, October 2020

Introduction

Model construction with covariates demands optimal criteria for covariate selection. In time series context, the association between a response and a predictor are usually lagged. Hence, lag selection is important for model construction. We consider two approaches: Fixed Lag (**FL**) approach, where optimal response/predictor lag is evaluated from the absolute cross-correlation function (CCF), previously to model construction; and, Changeable Lag (**CL**) approach, where optimal lag is selected based on AIC criterion among several candidate lags. Both approaches were implemented using the Block-Forward (**BF**) method. Briefly, covariates expected to induce the same effect on the response are included in one block and, at most, only one is included in the model.

Goal: Comparison of two strategies for lag selection using BF method: i) fixed lag (**FL**) and, ii) changeable lag (**CL**).

Data (2005-2017)

- Hourly air quality time series of $PM_{2.5}$, PM_{10}, NO_x, NO_2, CO, O_3 and SO_2), at 58 locations were downloaded from QualAr (`www.qualar.apambiente.pt`).
- Hourly temperature data at 23 spatial locations were made available by Instituto Português do Mar e da Atmosfera (`https://www.ipma.pt/`).
- Respiratory hospital admissions episodes registered in Portugal were provided by Administração Central do Sistema de Saúde (`http://www.acss.min-saude.pt`)
- Count time series were built with ICD-9:460–519 and ICD-10:J00-J99 codes from patients with address within the 20km influence circumference area of the air pollutant monitoring station. Temperature time series were paired based on euclidean distance.

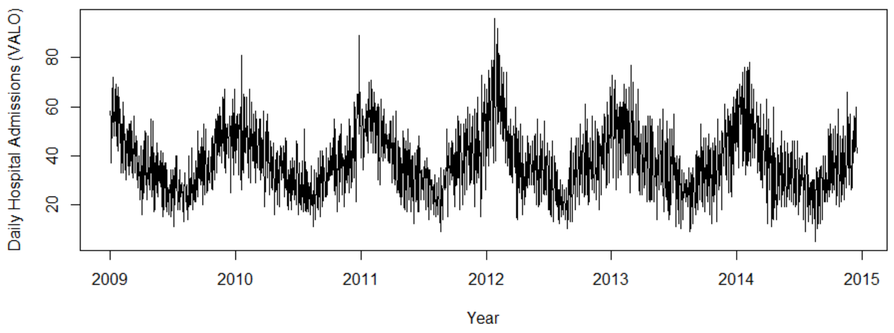

Fig. 1: Hospital Admission time series at Valongo, Portugal.

INGARCH Models

Let $Y_t|\mathcal{F}_{t-1} : \text{NB}(\lambda_t, \phi)$, where $\lambda_t := E(Y_t|\mathcal{F}_{t-1})$ and $\phi \in (0, \infty)$ is the dispersion parameter, $\mathcal{F}_{t-1} := \sigma(\, Y_s,\ \boldsymbol{X}_{s+1},\ s \leq t-1\,)$ is the joint history of the process (up to time $t-1$) and covariates (up to and including time t). The INGARCH model is

$$g(\lambda_t) = \beta_0 + \sum_{k=1}^{p} \beta_k\, \tilde{g}(Y_{t-k}) + \sum_{\ell=1}^{q} \alpha_\ell\, g(\lambda_{t-\ell}) + \boldsymbol{\eta}^T\, \boldsymbol{X}_t, \qquad (1)$$

where p and q are the INGARCH model orders, $\beta_0 > 0, \beta_k \geq 0,\ \alpha_\ell \geq 0,\ \forall_{k,\ell}$ and $\sum_{k=1}^{p} \beta_k + \sum_{\ell=1}^{q} \alpha_\ell < 1$. Also, $\boldsymbol{X}_t = (X_{t,1}, \ldots, X_{t,r})^T$ is a time-varying r-dimensional covariate vector for each time t and $\boldsymbol{\eta} := (\eta_1, ..., \eta_r)^T$ is the parameter vector.

- (p, q) pairs were selected by **minimising AIC** (p, q varied from 0 to 7).
- $\boldsymbol{X}_t = (X_{t,1}, \ldots, X_{t,r})^T$ were selected with **Block-Forward (BF) method**.

Block-Forward Method & Lag Selection

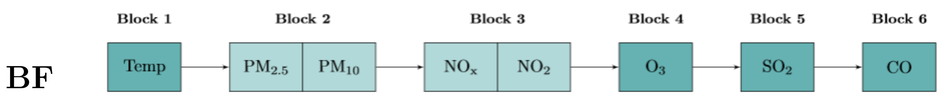

Fig. 2: Covariates' blocks in the BF method. Blocks order reflect the current knowledge of the effect of temperature and air pollutants on hospital admissions.

- Covariates are organised in blocks, where each block includes the covariates that are expected to induce a similar effect on Y_t.
- The significant covariate (per block) leading to the lowest AIC model enters the model, as long as the other covariates remain significant (at 5% significance level).

Acknowledgements: Ana Martins acknowledges Fundação para a Ciência e a Tecnologia (FCT) for financial support SFRH/BD/143973/2019.

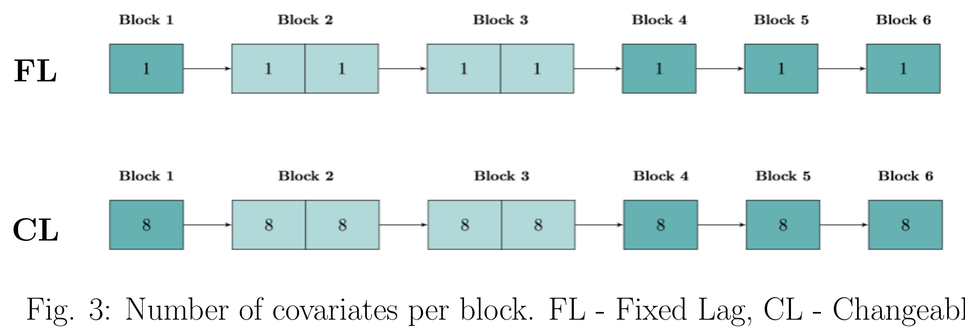

Fig. 3: Number of covariates per block. FL - Fixed Lag, CL - Changeable Lag.

- In **FL** approach the lag is selected prior to BF method by maximising the absolute values of the sample cross-correlation between the covariate and Y_t
- In **CL** approach lags from 0 to 7 seven are evaluated during BF method

Results

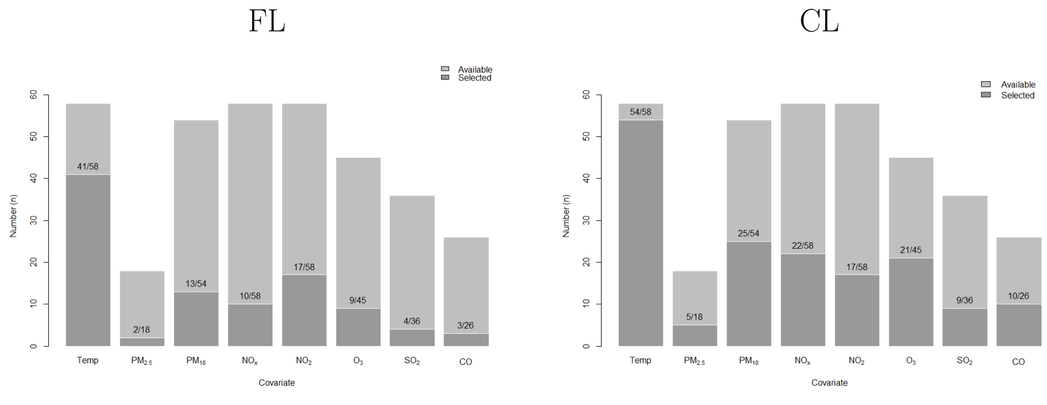

Fig. 4: Barplot of the number of selected (dark grey) over the number of available (light grey) covariates for the 58 spatial locations analysed.

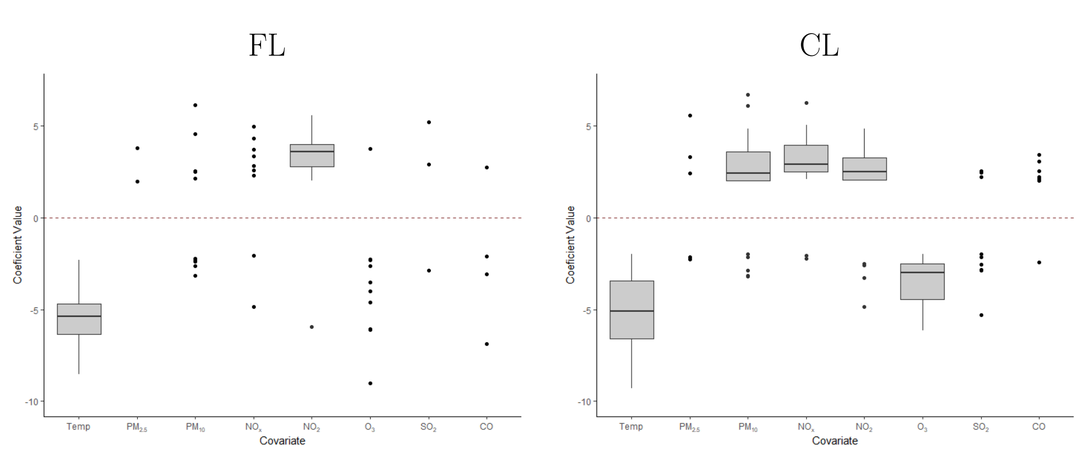

Fig. 5: Distribution of the scaled coefficients at the 58 spatial locations. Boxplots are shown when there are at least 15 locations.

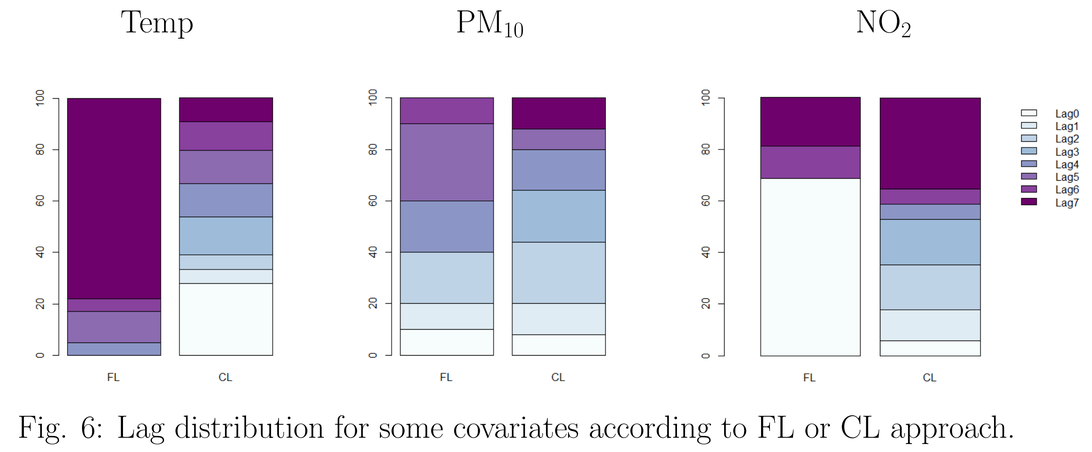

Fig. 6: Lag distribution for some covariates according to FL or CL approach.

- CL approach includes more covariates in models than FL approach
- Both approaches result in similar estimated coefficients
- The lag distribution varies considerably for some covariates (e.g., Temp) depending on the approach used
- CL models have, on average, lower (<20 units) AIC values compared to FL models

Conclusion

Lag selection strategy has an impact on model fitting, which cannot be neglected. Overall, CL approach includes more covariates in models than FL approach. Despite being computationally more demanding, CL approach tunes in the choice of the lag for each covariate, by accounting for the dependence among covariates.