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]
a.r.martins@ua.pt
Manuel Scotto[2]
manuel.scotto@tecnico.ulisboa.pt
Sónia Gouveia[1]
sonia.gouveia@ua.pt

[1] Institute of Electronics and Informatics Engineering of Aveiro
University of Aveiro, PT

[2] Center for Computational and Stochastic Mathematics and Department of Mathematics, IST
University of Lisbon, PT

Abstract

A comparison between strategies aiming at optimal lag selection for covariates in INGARCH models, in the context of the analysis of the association between air quality and daily number of respiratory hospital admissions in Portugal is presented. To this end, a block-forward (BF) approach is developed for automatic selection of covariates. Then, two strategies are used for optimal lag selection: (i) fixed lag (FL) approach, with optimal lag being selected as the one which maximises the cross-correlation between the covariate and the daily admissions; and (ii) changeable lag (CL) approach, with optimal lag being selected as that minimising the AIC criterion among all candidate lags. Results show that CL models have more significant covariates and lower AIC values than FL models. The coefficients of covariates simultaneously in FL and CL models are similar, despite having different optimal lags. Hence, the lag selection strategy has an impact on model fitting, which cannot be neglected.

1 Introduction

This study considers the INteger Generalised AutoRegressive Conditional Heteroskedastic (INGARCH) processes to model the association between hospital admissions and air quality. These have an ARMA-like structure, though the data generating mechanism is analogous to that of a GARCH model in the sense that the conditional mean recursively depends on the past conditional means and on the past observations [2, 3]. The INGARCH formulation incorporates link/transformation functions [8], to deal with negative serial correlation [4] and, time-varying covariates [5].
Model construction with covariates demands optimal criteria for covariate selection. The importance of such criteria is evident, as model performance can be improved by ignoring irrelevant covariates and, by considering the relevant covariates at optimal lags. These criteria should also address collinearity, as a strong association among covariates may obscure their relationship with the response and may lead to computational instability in model estimation. This paper introduces a novel method for optimal selection of time-varying covariates - the block-forward (BF). Briefly, covariates expected to induce the same effect on the response are included in one block. For each block, the significant covariate leading to the lowest Akaike Information Criteria (AIC) model is included in the model, as long as the covariates already in the model remain significant.
In time series context, the association between a response and a predictor are usually lagged. As an example, it is well-known that the maximal association between air pollution and hospital admissions may be delayed up to 7 days [7]. Traditionally, the optimal response/predictor lag is evaluated from the absolute cross-correlation function (CCF), previously to model construction. However, this procedure does not consider the possible associations among covariates. Thus, optimal lag choice in the process of covariate selection (and not *a priori*) is a promising approach, as different lagged versions of the same predictor can be thought of as a block of collinear covariates. Thus, we aim at the comparison of two strategies for lag selection, one based on the traditional CCF criterion (fixed lag, FL) and another considering the optimal AIC criterion among several candidate lags (changeable lag, CL), using the BF covariate selection method.

2 Data & Methods

2.1 Data

This study included the analysis of time series of air quality ($PM_{2.5}$, PM_{10}, NO_x, NO_2, CO, O_3 and SO_2), of temperature and of daily counts of hospital admissions (due to respiratory causes) during the 2005-2017 period. Figure 1 shows an example of a hospital admission time series, which clearly exhibits an annual seasonal pattern.
The spatial matching of air quality, temperature and hospital admissions was based on a 20km influence circumference centered around each air quality monitoring station. Hourly air quality data at 58 monitoring stations 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/`). Daily series were obtained from the maximum daily values, when at least 75% of hourly observations were available at a given day, otherwise were obtained through 1-NN imputation. Temperature series were matched to each spatial location based on their geographical proximity (measured with euclidean distance). All hospital admissions episodes registered in Portugal (2005-2017) were provided by Administração Central do Sistema de Saúde (`http://www.acss.min-saude.pt`). For each spatial location, the daily number of hospital admissions due to respiratory causes was recorded as the count of episodes connected with respiratory system diseases' (ICD-9:460–519 and ICD-10:J00-J99) from patients with address within the 20km influence circumference.

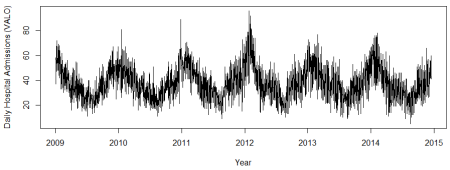

Figure 1: Hospital Admission time series at Valongo, Portugal.

2.2 INGARCH models

The INGARCH process (Y_t) assumes that the conditional distribution of Y_t is Negative Binomial i.e.,

$$Y_t|\mathcal{F}_{t-1} \sim NB(\lambda_t, \phi), \tag{1}$$

where $\lambda_t := E(Y_t|\mathcal{F}_{t-1})$ and $\phi \in (0,\infty)$ represents the dispersion parameter. Note that $Var(Y_t|\mathcal{F}_{t-1}) = \lambda_t + \lambda_t^2/\phi$ so, the limiting case $\phi \to \infty$ is the Poisson distribution with parameter λ_t. In this formulation,

$$\mathcal{F}_{t-1} := \sigma(\, Y_s, \boldsymbol{X}_{s+1},\ s \le t-1\,) \tag{2}$$

expresses the joint history of the process (up to time $t-1$) and covariates (up to and including time t). Also, the conditional expectation λ_t satisfies the recursion

$$g(\lambda_t) = \beta_0 + \sum_{k=1}^{p} \beta_k\, \tilde{g}(Y_{t-k}) + \sum_{\ell=1}^{q} \alpha_\ell\, g(\lambda_{t-\ell}) + \boldsymbol{\eta}^T\, \boldsymbol{X}_t, \tag{3}$$

where p and q are the INGARCH model orders, $\beta_0 > 0, \beta_k \ge 0$, $\alpha_\ell \ge 0$, $\forall_{k,\ell}$ and $\sum_{k=1}^{p} \beta_k + \sum_{\ell=1}^{q} \alpha_\ell < 1$. The latter condition ensures the stationary of the INGARCH process. Also, the link function $g : \mathbb{R}^+ \to \mathbb{R}$ and the transformation function $\tilde{g} : \mathbb{N}_0 \to \mathbb{R}$ were set as the natural logarithm function, to easily accommodate covariates into the model [5]. Finally, $\boldsymbol{X}_t = (X_{t,1}, \ldots, X_{t,r})^T$ is a time-varying r-dimensional covariate vector for each time t and $\boldsymbol{\eta} := (\eta_1, ..., \eta_r)^T$ is the parameter vector of the covariates coefficients. The estimation of INGARCH coefficients require a fixed order p and q. Optimal (p,q) pairs were chosen by AIC minimisation, varying from 0 to 7 in order to accommodate several INGARCH-like structures and include terms related with the presence of weekly seasonality.

2.3 Block-Forward and optimal lag selection for covariates

The block-forward (BF) selection method allows the automatic selection of significant covariates in $\boldsymbol{X}_t$. In the conventional forward method, e.g. used in linear regression, covariates are sequentially added to the model according to their statistical significance. In the BF method, the covariates are organised in blocks, where each block includes the covariates that are expected to induce a similar effect on Y_t. Consequently, the covariates in the same block are also expected to be correlated. For each block, the significant covariate leading to the lowest AIC model enters the model, as long as the other covariates remain significant (at 5% significance level). The order of the blocks is presented in Fig. 2 and reflects the current knowledge on the effect of temperature and air pollutants on hospital admissions [1]. In the BF implementation, two approaches were considered in the computation of the optimal lag between each covariate and Y_t. The fixed lag (FL) approach considers the covariate lag as that maximising the absolute values of the sample cross-correlation between the covariate and Y_t. And, the changeable lag (CL) approach that selects the optimal lag in which the BF conditions for a covariate to enter the INGARCH model are optimised. In practice, the implementation of FL and CL approaches is quite similar: while the FL approach considers the same number of candidates and covariates in one block, the CL approach considers that the number of candidates in one block is equal to the number of covariates in that block times the number of lags to be tested (in this case 8, from 0 to 7). Taking the example of block 2, FL approach tests up to 2 candidates to enter the model while CL approach will test up to 16 candidates. Note that, in both approaches, one candidate per block is selected at most.

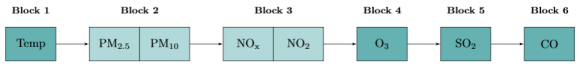

Figure 2: Blocks of covariates in the block-forward approach.

3 Results

The constructed INGARCH models were compared with respect to the number of selected covariates, the corresponding coefficients estimates and the chosen lags. Figure 3 shows the number of selected covariates out of the available for both approaches. Overall, CL models select more covariates than FL models. As an instance, temperature is selected in 54/58 CL models compared to 41/58 in FL models. Also, air quality covariates are more often selected in CL than in FL models. The median number of covariates included in the FL and CL models is, respectively, 2 and 3 covariates. Overall, both approaches show that air quality covariates are significantly associated with daily hospital admissions, beyond the well-established effect of temperature [6].

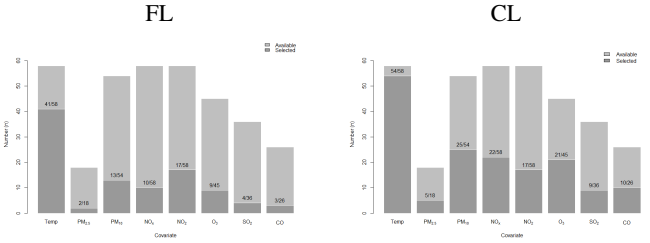

Figure 3: Barplot of the number of selected (dark grey) over the number of available (light grey) covariates for the 58 spatial locations analysed.

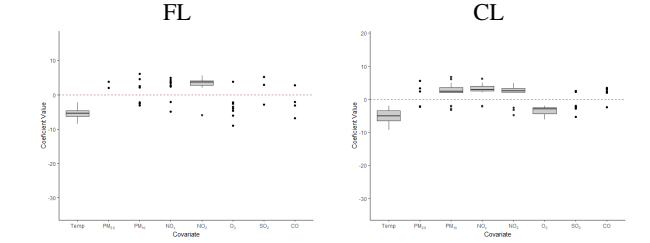

Figure 4: Distribution of the scaled coefficients at the 58 spatial locations. Boxplots are shown when there are at least 15 locations.

Lag	Temp		PM_{10}		NO_2		O_3		CO	
	FL	CL	FL	CL	FL	CL	FL	CL	FL	CL
0	0.0	27.8	10.0	8.0	68.8	5.9	10.0	19.0	0.0	0.0
1	0.0	5.6	10.0	12.0	0.0	11.8	10.0	19.0	0.0	0.0
2	0.0	5.6	20.0	24.0	0.0	17.6	0.0	9.5	0.0	20.0
3	0.0	14.8	0.0	20.0	0.0	17.6	0.0	23.8	0.0	20.0
4	4.9	13.0	20.0	16.0	0.0	5.9	0.0	9.5	0.0	30.0
5	12.2	13.0	30.0	8.0	0.0	0.0	0.0	4.8	0.0	10.0
6	4.9	11.1	10.0	0.0	12.5	5.9	0.0	9.5	50.0	10.0
7	78.0	9.3	0.0	12.0	18.8	35.3	80.0	4.8	50.0	10.0
Total (%)	100	100	100	100	100.0	100	100	100	100	100
Total (N)	41	54	10	25	16	17	10	21	4	10

Table 1: Distribution of the chosen lags according to FL and CL approach.

Figure 4 displays the distribution of the estimated scaled coefficients (i.e. coefficient divided by its standard error) for each covariate. Temperature and O_3 are negatively associated with respiratory hospital admissions, whereas the remaining air pollutants are, in general, positively associated. The magnitude of the coefficients and, the overall direction of the association are similar for both approaches. Hence, there is no major impact on the quantification of the covariate effect between approaches.
Table 1 shows the distribution of the chosen lags for some of the covariates analysed (Temp, PM_{10}, NO_2, O_3 and CO) according to each approach. There is some variability in the proportion of selected lags depending on the approach. For instance, while lag 7 is the preferred for Temp (78.0%) in the FL approach, lag 0 (27.8%) and lag 3 (14.8%) are the most frequently chosen in the CL approach. It is worthy to mention that CL models have, on average, lower AIC (< 20 units). Recall that the AIC criterion is a trade-off between information and number of covariates in a model (where increased number of covariates is penalised). Thus, the information of the covariates added pay-off the increase in complexity.

4 Conclusion

Despite the CL approach choosing more variables and different lags, the coefficients estimates remain similar for the covariates between approaches. However, the AIC of CL models is lower than that of FL models, indicating that the amount of information introduced by the additional variables in CL models pays-off the increased number of variables. Thus, tuning the lag during covariate selection is more advantageous as it increases the model performance. The trade-off is that the CL approach is computationally more demanding as both the covariates and their lagged versions are tested in the BF algorithm, which is an important aspect to consider when performing such analysis. Either way, an adequate modelling strategy is essential to assist in hospital planning and resources management and, ultimately, to contribute to better health/environmental policies.

References

[1] N. M. Ab, A. N. Aizuddin, and R. Hod. Effect of air pollution and hospital admission: a systematic review. *Annals of Global Health*, 84 (4):670, 2018.

[2] R. Ferland, A. Latour, and D. Oraichi. Integer-valued GARCH process. *Journal of Time Series Analysis*, 27(6):923–942, 2006.

[3] A. Heinen. Modelling time series count data: an autoregressive conditional Poisson model. *Social Science Research Network*, 2003.

[4] M. Ispány, V. A. Reisen, G. C. Franco, et al. On generalized additive models with dependent time series covariates. In *International Work-Conference on Time Series Analysis*, pages 289–308. Springer, 2017.

[5] T. Liboschik, K. Fokianos, and R. Fried. tscount: An R Package for Analysis of Count Time Series Following Generalized Linear Models. *Journal of Statistical Software*, 82(5):1–51, 2017.

[6] E. Martínez-Solanas and X. Basagaña. Temporal changes in the effects of ambient temperatures on hospital admissions in spain. *PloS one*, 14(6):e0218262, 2019.

[7] A. Slama, A. Śliwczyński, J. Woźnica, et al. Impact of air pollution on hospital admissions with a focus on respiratory diseases: a time-series multi-city analysis. *Environmental Science and Pollution Research*, 26(17):16998–17009, 2019.

[8] D. Tjøstheim. Some recent theory for autoregressive count time series. *Test*, 21(3):413–438, 2012.