sciendo

DOI: 10.2478/bile-2018-0012

Biometrical Letters
Vol. 55 (2018), No. 2, 179-195

Linear Markovian models for lag exposure assessment

Alessandro Magrini

Department of Statistics, Computer Science, Applications – University of Florence, Italy
alessandro.magrini@unifi.it

SUMMARY

Linear regression with temporally delayed covariates (distributed-lag linear regression) is a standard approach to lag exposure assessment, but it is limited to a single biomarker of interest and cannot provide insights on the relationships holding among the pathogen exposures, thus precluding the assessment of causal effects in a general context. In this paper, to overcome these limitations, distributed-lag linear regression is applied to Markovian structural causal models. Dynamic causal effects are defined as a function of regression coefficients at different time lags. The proposed methodology is illustrated using a simple lag exposure assessment problem.

Key words: directed acyclic graph; distributed-lag linear regression; dynamic causal inference; structural causal models; polynomial lag shape.

1. Introduction

Lag exposure assessment has the aim of investigating the effect of several pathogen exposures on a biomarker over time. Linear regression with temporally delayed (lagged) pathogen exposures, known as *distributed-lag linear regression*, is a standard approach to address such problems. Distributed-lag linear regression was proposed for the first time in the econometric field (Koyck, 1954; Solow, 1960; Almon, 1965), but recently it has received increasing attention for lag exposure assessment problems (Schwartz, 2000; Zanobetti et al., 2000; Martins et al., 2006; Welty et al., 2009; Gasparrini and Leone, 2014).

Unfortunately, distributed-lag linear regression is limited to a single biomarker of interest and cannot provide insights on the relationships holding among the pathogen exposures. These limitations may preclude the assessment

of causal effects, as they generally depend on the causal structure relating all the variables of interest (biomarkers and exposures), and not simply on the relationships between each biomarker and the exposures by which it is directly influenced. Structural causal models (SCMs; Pearl, 2000), and in particular linear Markovian SCMs, allow a recursive application of linear regression, thus appearing to be a natural extension of distributed-lag linear regression to a multivariate domain.

In this paper, distributed-lag linear regression is applied to Markovian structural causal models in order to obtain a methodology for lag exposure assessment in a multivariate domain. Our proposal, compared with existing methods, allows one to consider several different biomarkers and to study the relationships holding between each pathogen exposure and biomarker, as well as among the pathogen exposures. Thus, it is possible to assess dynamic causal effects whatever the causal structure relating the variables of interest. Several rules are provided to compute these from regression coefficients at different time lags.

This paper is structured as follows. In section 2, distributed-lag linear regression is introduced and discussed in the context of lag exposure assessment. In section 3, existing theory on SCMs is summarized, with special emphasis on linear Markovian SCMs. In section 4, a methodology is detailed for the use of linear Markovian SCMs for lag exposure assessment in a multivariate domain. In section 5, the proposed methodology is applied to a biometric problem. Section 6 contains concluding remarks and considerations on the future development of the proposed methodology.

2. Lag exposure in the linear regression model

Suppose that we wish to investigate the effect of a single pathogen exposure on a biomarker over time. Let X_t be the measurement of the pathogen exposure at time t and Y_t be the measurement of the biomarker at time t.

By assuming that time is a discrete variable, and that the influence of X on Y does not depend on time but only on the temporal distance (lag) between the exposure and the observation, the influence of X on Y may be modeled using a linear regression model including temporally delayed (lagged) instances of X:

$$y_t = \beta_0 + \sum_{l=0}^{L} \beta_l x_{t-l} + \varepsilon_t \quad (1)$$

where ε_t is the random error at time t, uncorrelated with X and with $\varepsilon_0, \dots, \varepsilon_{t-1}$. In this model, the observed value of Y at time t depends not only on the value of X at time t, but also on all values of X since L time instants before t. In particular, the value of Y at time t is expected to increase by β_l for a unitary increase in the value of X at time $t - l$, for any t. Equivalently, as Y is exposed to a unitary increase in the value of X, the value of Y is expected to increase by β_l after l instants (time lags).

Suppose now that there is a vector of p pathogens, say $\boldsymbol{X} = (X_1, \dots, X_p)$, rather than a single one. In this case, the model becomes:

$$y_t = \beta_0 + \sum_{i=1}^{p} \sum_{l=0}^{L_i} \beta_{i,l} x_{i,t-l} + \varepsilon_t \quad (2)$$

where ε_t is the random error at time t, uncorrelated with the variables in $\boldsymbol{X}$ and with $\varepsilon_0, \dots, \varepsilon_{t-1}$. In this new formulation, the expected value of Y at time t is expected to increase by $\beta_{i,l}$ for a unitary increase in the value of X_i at time $t - l$, given constant values of the variables in $\boldsymbol{X}$ besides X_i.

The set $\boldsymbol{\beta}_i = (\beta_{i,0}, \beta_{i,1}, \dots, \beta_{i,L_i})$ is called the *lag shape* of the covariate X_i and represents its influence on Y at different time lags. Note that the case where a covariate X_i has only a static influence on the response Y is obtained by setting $L_i = 0$. The case where $L_i = 0$ for all i coincides with the classical linear regression model.

The autoregressive lag shape may be further considered, and the distributed-lag linear regression becomes:

$$y_t = \beta_0 + \sum_{h=1}^{H} \phi_h y_{t-h} + \sum_{i=1}^{p} \sum_{l=0}^{L_i} \beta_{i,l} x_{i,t-l} + \varepsilon_t \tag{3}$$

In the remainder of this paper, a distributed-lag linear regression model with response variable Y and covariates $X_1, \dots, X_p$ will be indicated with the following notation:

$$Y \sim \mathcal{L}(Y; H) + \mathcal{L}(X_1; L_1) + \cdots + \mathcal{L}(X_p; L_p) \tag{4}$$

where $\mathcal{L}(\cdot\,; L)$ denotes a lag shape of length L.

3. Structural causal models

Structural causal models (SCMs) were developed by Pearl (2000) in the context of causal inference. They are rooted in path analysis (Wright, 1934) and simultaneous equation models (Haavelmo, 1943; Koopmans et al., 1950). An SCM consists of a tuple $\{\boldsymbol{V}, \boldsymbol{U}, \Omega_{\boldsymbol{V}}, \Omega_{\boldsymbol{U}}, \boldsymbol{f}, \wp_{\boldsymbol{U}}\}$, where:

$\boldsymbol{V} = (V_1, \dots, V_J)$ is a set of endogenous variables;

$\Omega_{\boldsymbol{V}} = \Omega_{V_1} \times \dots \times \Omega_{V_J}$ is the Cartesian product of the domains of variables in V;

$\boldsymbol{U} = (U_1, \dots, U_K)$ is a set of unobserved variables;

$\Omega_{\boldsymbol{U}} = \Omega_{U_1} \times \dots \times \Omega_{U_K}$ is the Cartesian product of the domains of variables in U;

$\boldsymbol{f}: \Omega_{\boldsymbol{V}} \times \Omega_{\boldsymbol{U}} \to \Omega_{\boldsymbol{V}}$ is a measurable function;

$\wp_{\boldsymbol{U}}$ is a probability measure on $\Omega_{\boldsymbol{U}}$.

Markovian SCMs (Pearl, 2000, Chapter 3) are a special case where $\boldsymbol{f}$ is acyclic and the variables in $\boldsymbol{U}$ are mutually independent. In a Markovian SCM, the following factorization of the joint probability distribution of variables in $\boldsymbol{V}$ holds:

$$p(v_1, \dots v_J) = \prod_{j=1}^{J} p(v_j | \Pi_j = \pi_j) \tag{5}$$

where Π_j is the set of variables in $\boldsymbol{V}$ such that, for $j > 1$, V_j is independent of the variables in $\{V_1, \dots V_{j-1}\} \setminus \Pi_j$, given the variables in Π_j. This means that the joint probability distribution of the variables in $\boldsymbol{V}$ can be factored according to conditional independence relationships holding among them, disregarding the

variables in $\boldsymbol{U}$. Pearl (2000, page 12 and following) shows that these conditional independence relationships are encoded into a directed acyclic graph (DAG) such that Π_j is the parent set of V_j, $\forall j = 1, \dots J$. For example, in the Markovian SCM associated with the DAG in Figure 1, it holds that

$$\begin{aligned} &p(v_1, v_2, v_3, v_4, v_5, v_6) \\ &= p(v_1)p(v_2)p(v_3|v_1)p(v_4|v_2, v_3)p(v_5|v_3)p(v_6|v_3, v_4, v_5) \end{aligned} \tag{6}$$

and, for example, V_6 is independent of V_1 and V_2 given V_3, V_4 and V_5.

Let do($V_i = v_i$) denote an intervention setting the value of V_i to v_i. Then, in a Markovian SCM it holds that

$$p\big(v_1, \dots v_J|\text{do}(V_i = v_i)\big) = \prod_{j \neq i} p(v_j|\Pi_j = \pi_j)\big|_{V_i = v_i} \tag{7}$$

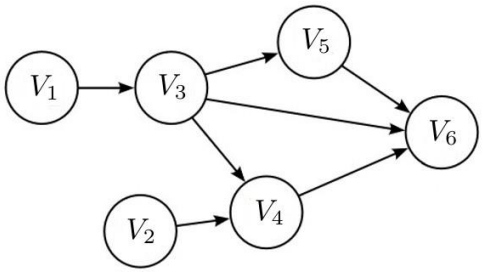

Figure 1. An example of a directed acyclic graph

where $\cdot|_{V_i = v_i}$ indicates that V_i is replaced by the value v_i. This formula, called *truncated factorization* (Pearl, 2000, section 3.2), allows one to compute the effect of an intervention from the (pre-intervention) distribution in formula (5), that is, to predict the effect of an intervention from non-experimental (observational) data. In a Markovian SCM, the effect of $\text{do}(V_i = v_i)$ on V_j, called the *causal effect* of V_i on V_j, is given by the following expression (see Pearl, 2000, page 70 and following):

$$\begin{aligned} &p\big(V_j = v_j|\text{do}(V_i = v_i)\big) \\ &\qquad = \prod_{\pi_i} p(V_j = v_j|V_i = v_i, \Pi_i = \pi_i)p(\Pi_i = \pi_i) \end{aligned} \tag{8}$$

where Π_i is the parent set of V_i.

Linear Markovian structural causal models

In a linear parametric formulation of SCMs (linear Markovian SCMs), each factor $p(v_j|\Pi_j = \pi_j)$ of the joint probability distribution in formula (5) is the linear regression model where V_j is the response variable and the variables in Π_j are the covariates. For example, in the linear Markovian SCM associated with the DAG in Figure 1, $p(v_4|v_2, v_3)$ is the linear regression model where V_4 is the response variable and V_2 and V_3 are the covariates.

In a linear Markovian SCM, the computation of causal effects involves the coefficients of the regression models only, without the need for formula (8), as shown in the following paragraphs. The regression coefficient notation used in section 2 is modified to include subscripts with both the response variable and the covariate, separated by a vertical pipe. For instance, $\beta_{j|i}$ indicates the coefficient of V_i in the regression model of V_j.

Direct causal effects. The coefficient of V_i in the regression model of V_j, say $\beta_{j|i}$, represents the expected value of V_j given a unit variation of V_i and given constant values of the parents of V_j besides V_i:

$$\beta_{j|i} \coloneqq \Delta\mathrm{E}(V_j|\Delta V_i = 1, \Delta V_{k:V_k \in \{\Pi_j \setminus \mathrm{V}_i\}} = 0) \tag{9}$$

Expression (9) is a special case of (8), where the intervention is $\mathrm{do}(\Delta V_i = 1)$ and the conditioning set is $\{\Pi_j \setminus V_i\}$ instead of Π_i. Since the variables in Π_i but not in Π_j are independent of V_j conditionally on the variables in Π_j (see formula (5)), we can conclude that $\beta_{j|i}$ represents the average effect of $\mathrm{do}(\Delta V_i = 1)$ on V_j:

$$\begin{aligned} \beta_{j|i} &\coloneqq \Delta\mathrm{E}\left(V_j \middle| \Delta V_i = 1, \Delta V_{k:V_k \in \{\Pi_j \setminus \mathrm{V}_i\}} = 0\right) \\ &= \Delta\mathrm{E}(V_j|\mathrm{do}(\Delta V_i = 1); < V_i, V_j >) \end{aligned} \tag{10}$$

which is called the *direct* causal effect of V_i on V_j. The notation $\Delta\mathrm{E}(V_j|\mathrm{do}(\Delta V_i = 1); < V_i, V_j >)$ emphasizes that the causal effect in formula (10) is associated with the edge $< V_i, V_j >$. For example, in the linear Markovian

SCM associated with the DAG in Figure 1, $\beta_{4|3}$ represents the expected value of V_4 given a unit variation of V_3 and given a constant value of V_2, equating to the direct causal effect of V_3 on V_4.

Indirect causal effects and the overall causal effect. Suppose that there exists more than one directed path connecting variable V_i to variable V_j. In this case, it is straightforward to show that the intervention $\mathrm{do}(\Delta V_i = 1)$ influences the expected value of V_j independently along each directed path connecting V_i to V_j, for an *overall* causal effect equal to the sum of the causal effects associated with each of these paths:

$$\begin{aligned}&\Delta\mathrm{E}\Big(V_j\Big|\mathrm{do}(\Delta V_i = 1)\Big)\\&= \prod_{<V_{d_0},\ldots,V_{d_m}>:\, d_0=i\wedge d_{m=j}} \Delta\mathrm{E}(V_j|\mathrm{do}(\Delta V_i = 1); < V_{d_0}, \ldots, V_{d_m} >)\end{aligned} \tag{11}$$

where $\Delta\mathrm{E}(V_j|\mathrm{do}(\Delta V_i = 1); < V_{d_0}, \ldots, V_{d_m}>)$ is the causal effect of $\mathrm{do}(\Delta V_i = 1)$ on V_j associated with the directed path $< V_{d_0}, \ldots, V_{d_m} >$ ($d_0 = i \wedge d_m = j$) connecting V_i to V_j, denoted as the pathwise causal effect of V_i on V_j along $< V_{d_0}, \ldots, V_{d_m} >$.

A pathwise causal effect associated with an edge (direct causal effect) can be computed using formula (10). A pathwise causal effect associated with a multi-edged directed path, also referred to as an *indirect* causal effect, can be computed from the product of the regression coefficients associated with each edge in the path (see, for example, Wright, 1934):

$$\begin{aligned}&\Delta\mathrm{E}\big(V_j\big|\mathrm{do}(\Delta V_i = 1); < V_i, \ldots, V_j >\big):\\&= \prod_{k:V_k\in<V_i,\ldots,V_j>\wedge k\neq i} \Delta\mathrm{E}(V_k|\mathrm{do}(\Delta V_{k-1} = 1); < V_{k-1}, V_k >)\\&= \prod_{k:V_k\in<V_i,\ldots,V_j>\wedge k\neq i} \beta_{k|k-1}\end{aligned} \tag{12}$$

Note that formula (12) is a generalization of (10). In this view, it is clear that both direct and indirect causal effects belong to the class of pathwise causal effects. For example, in the linear Markovian SCM associated with the DAG in

Figure 1, there are three directed paths connecting V_3 to V_6: $< V_3, V_6 >$ with pathwise (direct) causal effect $\beta_{6|3}$, $< V_3, V_4, V_6 >$ with pathwise (indirect) causal effect $\beta_{4|3} \cdot \beta_{6|4}$, and $< V_3, V_5, V_6 >$ with pathwise (indirect) causal effect $\beta_{5|3} \cdot \beta_{6|5}$. Thus, the overall causal effect of V_3 on V_6, namely $\Delta\mathrm{E}(V_6|\mathrm{do}(\Delta V_3 = 1))$, is equal to $\beta_{6|3} + \beta_{4|3} \cdot \beta_{6|4} + \beta_{5|3} \cdot \beta_{6|5}$.

4. Lag exposure in linear Markovian structural causal models

Lag exposure may be taken into account in Markovian SCMs by specifying each factor of the joint probability distribution in formula (5), equal to the distributed-lag linear regression in formula (3). We refer to this family of Markovian SCMs as *distributed-lag linear structural causal models* (DLSCMs). The DAG of a DLSCM includes all of the possible temporal instances of each variable in $\boldsymbol{V}$. Figure 2 shows the DAG of several DLSCMs on two variables X and Y. The definition of causal effects at different time lags in a DLSCM is provided in the following paragraphs.

Direct causal effects. Let $\beta_{j|i,l}$ be the coefficient of V_i at lag l in the regression model of V_j. This coefficient equates to the *direct* causal effect of V_i on V_j at lag l:

$$\Delta\mathrm{E}_l\big(V_j\big|\mathrm{do}(\Delta V_i = 1); < V_i, V_j >\big) = \beta_{j|i,l} \tag{13}$$

Indirect causal effects. Let $< V_{d_0}, \dots, V_{d_m} >$, with $d_0 = i \wedge d_m = j$, be a directed path composed of m edges connecting V_i to V_j, and $Q_m^{(l)}$ be the set of all possible ordered m-tuples of time lags such that their sum is equal to l. If we compute the m direct causal effects associated with each edge in $< V_{d_0}, \dots, V_{d_m} >$ at one of the m-tuples in $Q_m^{(l)}$, say $(q_1, \dots, q_m)$, and multiply them:

$$e_{(q_1,\dots,q_m)}(< V_{d_0}, \dots, V_{d_m} >; d_0 = i, d_m = j) = \prod_{k=1}^{m} \beta_{d_k|d_{k-1},q_k} \tag{14}$$

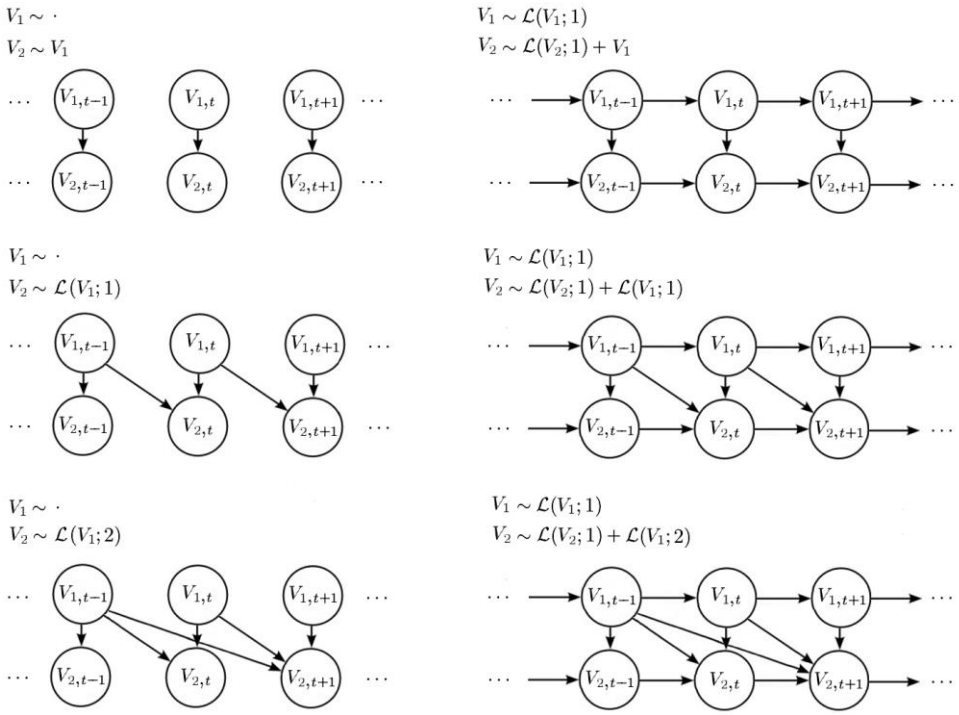

Figure 2. The DAG of several DLSCMs on two variables $\boldsymbol{V_1}$ and $\boldsymbol{V_2}$

we obtain one of the possible causal effects of V_i on V_j along $< V_{d_0}, \dots, V_{d_m} >$ at lag l. Thus, the *indirect* causal effect of V_i on V_j along $< V_{d_0}, \dots, V_{d_m} > (d_0 = i \wedge d_m = j)$ at lag l is equal to the sum of all of the causal effects that can be obtained from formula (14):

$$\begin{aligned} &\Delta \mathrm{E}_l\left(V_j \middle| \mathrm{do}(\Delta V_i = 1); < V_{d_0}, \dots, V_{d_m} >, d_0 = i, d_m = j\right) \\ &= \sum_{(q_1,\dots,q_m) \in Q_m^{(l)}} \prod_{k=1}^{m} \beta_{d_k | d_{k-1}, q_k} \end{aligned} \tag{15}$$

Table 1 shows the addenda of the indirect causal effect of V_1 on V_6 along $< V_1, V_3, V_4, V_6 >$ at lag 3, that is $\Delta \mathrm{E}_3(V_6 | \mathrm{do}(\Delta V_1 = 1); < V_1, V_3, V_4, V_6 >)$, in the linear Markovian SCM associated with the DAG in Figure 1.

Table 1. Addenda of the indirect causal effect of $\boldsymbol{V_1}$ on $\boldsymbol{V_6}$ along $< \boldsymbol{V_1}, \boldsymbol{V_3}, \boldsymbol{V_4}, \boldsymbol{V_6} >$ ($\boldsymbol{m}$=3) at lag 3 in the linear Markovian SCM associated with the DAG in Figure 1. Their sum equates to $\Delta\mathbf{E_3}(\boldsymbol{V_6}|\mathbf{do}(\Delta\boldsymbol{V_1} = \mathbf{1}); < \boldsymbol{V_1}, \boldsymbol{V_3}, \boldsymbol{V_4}, \boldsymbol{V_6} >$

$(q_1, q_2, q_3) \in Q_3^{(3)}$	$e_{(q_1,q_2,q_3)}(< V_1, V_3, V_4, V_6 >)$
(0,0,3)	$\beta_{3\|1,0} \cdot \beta_{4\|3,0} \cdot \beta_{6\|4,3}$
(0,1,2)	$\beta_{3\|1,0} \cdot \beta_{4\|3,1} \cdot \beta_{6\|4,2}$
(0,2,1)	$\beta_{3\|1,0} \cdot \beta_{4\|3,2} \cdot \beta_{6\|4,1}$
(0,3,0)	$\beta_{3\|1,0} \cdot \beta_{4\|3,3} \cdot \beta_{6\|4,0}$
(1,0,2)	$\beta_{3\|1,1} \cdot \beta_{4\|3,0} \cdot \beta_{6\|4,2}$
(1,1,1)	$\beta_{3\|1,1} \cdot \beta_{4\|3,1} \cdot \beta_{6\|4,1}$
(1,2,0)	$\beta_{3\|1,1} \cdot \beta_{4\|3,2} \cdot \beta_{6\|4,0}$
(2,0,1)	$\beta_{3\|1,2} \cdot \beta_{4\|3,0} \cdot \beta_{6\|4,1}$
(2,1,0)	$\beta_{3\|1,2} \cdot \beta_{4\|3,1} \cdot \beta_{6\|4,0}$
(3,0,0)	$\beta_{3\|1,3} \cdot \beta_{4\|3,0} \cdot \beta_{6\|4,0}$

Overall causal effects. The *overall* causal effect of V_i on V_j at lag l, say $\Delta\mathrm{E}_l(V_j|\mathrm{do}(\Delta V_i = 1))$, is represented by the sum of the pathwise causal effects at lag l associated with each directed path connecting V_i to V_j.

The causal effects just defined are evaluated at a single time lag. The *cumulative* causal effect at a pre-specified time lag, say l, is obtained by summing all causal effects at each time lag up to l. A *pathwise causal lag shape* is the set of causal effects associated with a path at different time lags. An *overall causal lag shape* is the set of the overall causal effects of a variable on another at different time lags.

In the following subsections, several constraints on the lag shapes that may fit with prior knowledge on the phenomenon of interest are introduced, and a static representation of the DAG of a DLSCM is proposed.

4.1. Constraining the lag shapes

The practical application of the model in formula (3) critically depends on the relationship among the coefficients composing each lag shape. From a theoretical point of view, the lag shape of a covariate should have a regular form. For

example, the effect of an exposure may be small at first, then it may reach a peak before diminishing to zero after some time lags. In this view, a model with no constraints on the lag shapes may not fit with prior knowledge on the phenomenon of interest and thus may be difficult to interpret. For simplicity, the index of the response variable is omitted from the regression coefficient notation; thus the same notation as in section 2 is used.

Almon's polynomial lag shape (Almon, 1965) constrains coefficients to be polynomials of order *Q*:

$$\beta_{i,l} = \begin{cases} \varphi_{i,0} & l = 0 \\ \sum_{q=0}^{Q} \varphi_{i,q} l^q & \text{otherwise} \end{cases} \tag{16}$$

For instance, for $Q = 2$ we have $\beta_{i,l} = \varphi_{i,0} + \varphi_{i,1} l + \varphi_{i,2} l^2$. Almon's polynomial lag shape reduces the number of parameters required to represent the lag shape of a covariate, but multiple modes and coefficients with different signs may occur, thus problems of interpretation may still arise.

The *endpoint-constrained quadratic* (ECQ) lag shape (Andrews and Fair, 1992):

$$\beta_{i,l} = \begin{cases} \theta_i \left[-\frac{4}{(b_i - a_i + 2)^2} l^2 + \frac{4(a_i + b_i)}{(b_i - a_i + 2)^2} l - \frac{4(a_i - 1)(b_i + 1)}{(b_i - a_i + 2)^2} \right], \\ 0, \end{cases} \tag{17}$$

denoted $\mathrm{ECQ}(\cdot\,; \theta_i, a_i, b_i)$, overcomes the limitation of Almon's lag shape, as it is zero for a time lag $l < a_i$ or $l > b_i$, and symmetric with mode equal to θ_i at lag $(a_i + b_i)/2$.

The quadratic decreasing (QD) lag shape:

$$\beta_{i,l} = \begin{cases} \theta_i \dfrac{l^2 - 2(b_i + 1)l + (b_i + 1)^2}{(b_i - a_i + 1)^2}, & a_i \le l \le b_i \\ 0, & \text{otherwise} \end{cases} \tag{18}$$

denoted $\mathrm{QD}(\cdot\,; \theta_i, a_i, b_i)$, is a truncated version of the ECQ, which decreases from value θ_i at lag a_i to value 0 at lag $b_i + 1$.

The *gamma* lag shape (Schmidt, 1974):

$$\beta_{i,l} = \theta_i (l+1)^{\frac{\delta_i}{1-\delta_i}} \lambda_i^l \left[\left(\frac{\delta_i}{(\delta_i - 1)\log\lambda_i} \right)^{\frac{\delta_i}{1-\delta_i}} \lambda_i^{\frac{\delta_i}{(\delta_i-1)\log\lambda_i} - 1} \right]^{-1} \qquad (19)$$
$$0 < \delta_i < 1 \qquad 0 < \lambda_i < 1$$

denoted $\mathrm{G}(\cdot\,; \theta_i, \delta_i, \lambda_i)$, is positively skewed with mode equal to θ_i at lag $\delta_i[(\delta_i - 1)\log\lambda_i]^{-1}$.

The value a_i is called the *gestation lag*, the value b_i the *lead lag*, and the value $b_i - a_i$ the *lag width*. Note that the ECQ and the QD lag shapes degenerate to a static coefficient if $a_i = b_i = 0$. The gamma lag shape cannot reduce to a static coefficient, but the corresponding values of a_i and b_i may be computed from the values of δ_i and λ_i by numerical approximation. The ECQ, QD and gamma lag shapes have the following property:

$$\begin{matrix} \beta_{i,l} > 0 \Leftrightarrow \theta_i > 0 \\ \beta_{i,l} < 0 \Leftrightarrow \theta_i < 0 \end{matrix} \quad \forall l: a_i \leq l \leq b_i \qquad (20)$$

which is referred to as *monotonicity*.

4.2. Static representation

The DAG of a DLSCM may be represented in a static version for more clarity. For example, only a single temporal instance for each variable is represented, and an edge $< V_i, V_j >$ exists if and only if there exists at least one time lag where the coefficient of variable V_i in the regression model of variable V_j is non-zero. Figure 3 shows the DAG of the following DLSCM represented in static form:

$$\begin{cases} A \sim \cdot \\ B \sim \mathrm{ECQ}(A; 0.0918, 0, 4) \\ C \sim \mathrm{ECQ}(A; 0.1161, 2, 6) + \mathrm{ECQ}(B; 0.1922, 1, 5) \end{cases} \qquad (21)$$

Whenever the EQC, QD or gamma lag shape is applied, the sign of the parameter θ_i can be associated with the corresponding direct causal effect, and thus with the corresponding edge in the DAG, due to the monotonicity property. For instance, a positive sign can be associated with each direct causal effect in the DLSCM in formula (21) and with each edge in its DAG as shown in Figure 3.

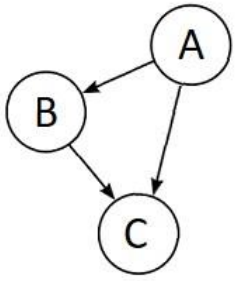

Figure 3. The DAG of the DLSCM in formula (21) represented in static form

5. Application to a simple lag exposure assessment problem

Consider the DLSCM in formula (21), where *A* and *B* are the irradiation (log becquerel) from two distinct radioactive pathogens, and *C* is the equivalent dose (log gray) absorbed by a subject exposed to them. Time is expressed in hours. This model postulates that:

- the irradiation from the first pathogen enforces the irradiation from the second according to an ECQ lag shape persisting from 0 to 4 hours;
- the equivalent dose depends on the irradiation from the first and the second pathogen according to, respectively, an ECQ lag shape persisting from 2 to 6 hours and an ECQ lag shape persisting from 1 to 5 hours.

Thus, the influence of the first pathogen (A) on the equivalent dose (C) is both direct (along $< A, C >$) and indirect (along $< A, B, C >$).

The overall causal effect of *A* on *C* (Figure 4 c) is the sum of all of the possible pathwise causal effects from *A* to *C* (formula (11)):

- the direct causal effect of *A* on *C*, persisting from 1 to 7 hours (Figure 4 a);
- the (indirect) causal effect along < *A*,*B*,*C* > (Figure 4 b), composed of the direct causal effect of *A* on *B* and the direct causal effect of *B* on *C*. Since the former persists from 0 to 4 hours and the latter from 1 to 5 hours, the causal effect of *A* on *C* along <*A*,*B*,*C* > is found to persist from 0 to 9 hours (formula (15)).

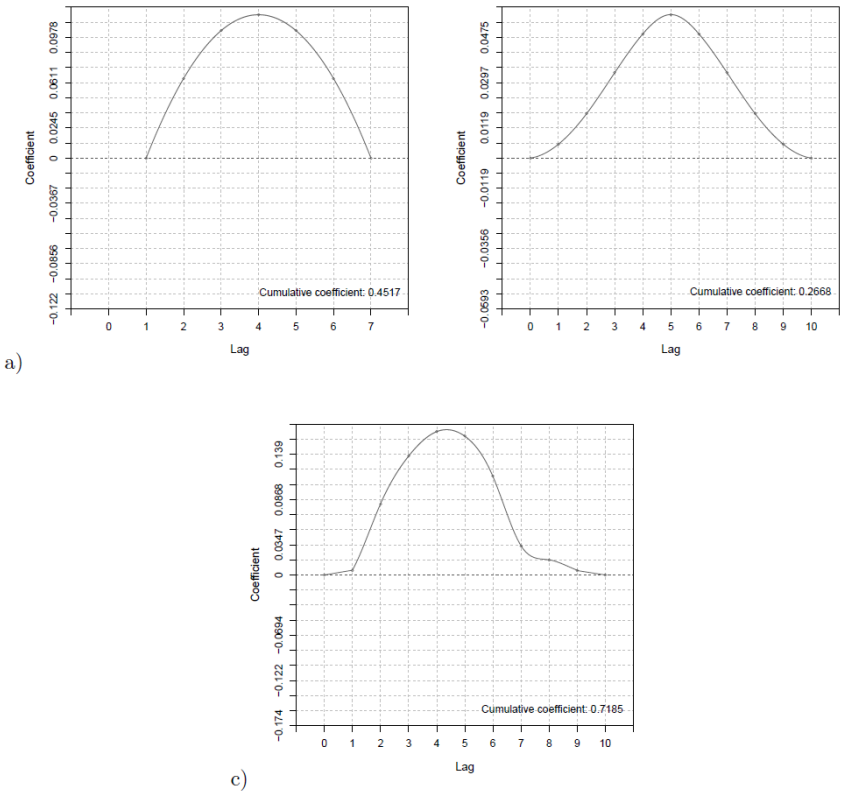

Figure 4. Decomposition of the causal effect of $\boldsymbol{A}$ on $\boldsymbol{C}$. a) The direct causal effect of $\boldsymbol{A}$ on $\boldsymbol{C}$. b) The causal effect of $\boldsymbol{A}$ on $\boldsymbol{C}$ along $< \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C} >$ (indirect causal effect). c) The overall causal effect of $\boldsymbol{A}$ on $\boldsymbol{C}$

Note that the lag shapes of the indirect pathwise causal effects and of the overall effect constitute a mixture of ECQ lag shapes, thus they may have an irregular character.

Being composed of two pathwise causal effects persisting from 1 to 7 hours and from 0 to 9 hours, the overall causal effect of *A* on *C* persists from 0 to 9 hours (Figure 4 c). By summing the overall causal effects up to 9 time lags, a cumulative overall causal effect equal to 0.72 is obtained. Since the logarithmic scale is used, we can conclude that, for a 1% increase in irradiation from pathogen *A*, the equivalent dose absorbed by the subject is expected to increase by approximately 0.72% after 9 hours. [The true expected growth rate for the

response variable due to a 1% increase in the value of a covariate with coefficient κ is equal to 1.01^{κ}, which corresponds to a percentage increase of $(1.01^{\kappa} - 1) \cdot 100$. The approximation $(1.01^{\kappa} - 1) \cdot 100 \approx \kappa$ proposed here is reasonable for $|\kappa| < 10$.]

In the traditional approach, a distributed-lag linear regression model applied to this problem would have considered the influence of both A and B on C, but the influence of A on B would have been disregarded. Thus, the relationship between the two exposures and the biomarker would have been assessed, but it would not have been possible to compute the causal effect of A on C, as this also depends on the disregarded relationship between A and B. Note that similar considerations would hold if A was an exposure and B and C were biomarkers, or if more exposures and/or biomarkers were considered.

6. Concluding remarks

We have shown that distributed-lag linear regression combined with Markovian structural causal models allows one to perform lag exposure assessment in a multivariate domain. Existing methods focus on one regression model at a time, but the proposed methodology allows one to consider several different biomarkers and to study the relationships holding between each pathogen exposure and biomarker, as well as among the pathogen exposures. This makes it possible, for the first time, to assess dynamic causal effects whatever the causal structure relating the variables of interest.

In principle, the proposed methodology supports unconstrained and constrained lag shapes of any type. We have presented three of the possible constrained lag shapes that may represent the most common real-world lag structures: unimodal symmetric, unimodal asymmetric and skewed.

Future work will be directed towards the development of a procedure for estimating model parameters. At first glance, ordinary least squares estimation could be recursively applied provided that the time series are stationary. However, the estimation of distributed-lag linear regression with the constrained

lag shapes presented in this paper cannot be performed in a single step unless all gestation and lead lags are known. Since the number of possible models grows exponentially as the number of covariates and time lags increases, the development of a heuristic search within the parameter estimation procedure appears necessary.

The presence of unit roots in data is a challenging issue for time series models (Granger and Newbold, 1974), and so it is for the proposed methodology. Currently available solutions for unit roots, such as differentiation and autocorrelated errors, may be implemented in the parameter estimation procedure.

The present contribution could be extended to grouped data through the introduction of random effects in the distributed-lag linear regression models. In general, any extension that could be applied to linear regression may be applied to the proposed methodology.

Particular attention has been paid to defining dynamic causal effects as a function of regression coefficients at different time lags. Thus, only linear relationships have been considered for the moment. Nevertheless, future work may include the extension of the theory developed in this paper to specific classes of non-linear distributed-lag regression models (Gasparrini et al., 2017).

Acknowledgements

This work was partially supported by the University of Florence (Italy) funding framework Progetto strategico di ricerca di base per l'anno 2015, grant Disegno e analisi di studi sperimentali e osservazionali per le decisioni in ambito epidemiologico, socio-economico, ambientale e tecnologico. We thank the anonymous reviewers for their valuable comments that helped improve the presentation of our work.

REFERENCES

Almon S. (1965): The Distributed Lag between Capital Appropriations and Net Expenditures. Econometrica 33: 178–196.

Andrews W.K., Fair R.C. (1992): Estimation of Polynomial Distributed Lags and Leads with End Point Constraints. Journal of Econometrics 53: 123–139.

Gasparrini A., Leone M. (2014): Attributable Risk from Distributed Lag Models. BMC Medical Research Methodology 14(1): 14–55.

Gasparrini A., Scheipl F., Armstrong B., Kenward M.G. (2017): A Penalized Framework for Distributed Lag Non-Linear Models. Biometrics 73(3): 938–948.

Granger C.W.J., Newbold P. (1974): Spurious Regressions in Econometrics. Journal of Econometrics 2(2): 111–120.

Haavelmo T. (1943): The Statistical Implications of a System of Simultaneous Equations. Econome,rica 1(1): 1–12.

Koopmans T.C., Rubin H., Leipnik R.B. (1950): Measuring the Equation Systems of Dynamic Economics. In T. C. Koopmans (ed.), Statistical Inference in Dynamic Economic Models, pages 53–237. John Wiley & Sons, New York, US-NY.

Koyck L.M. (1954): Distributed Lags and Investment Analysis. North-Holland, Amsterdam, NL.

Martins L.C., Pereira L.A.A., Lin C.A., Santos U.P., Prioli G., do Carmo Luiz O., Saldiva P.H.N., Ferreira Braga A.L. (2006): The effects of air pollution on cardiovascular diseases: Lag structures. Revista de Sau´de Pu´blica 40(4). doi: 10.1590/S0034-89102006000500018.

Pearl J. (2000): Causality: Models, Reasoning, and Inference. Cambridge University Press, Cambridge, UK.

Schmidt P. (1974): A Modification of the Almon Distributed Lag. Journal of the American Statistical Association 69: 679-681.

Schwartz J. (2000): The Distributed Lag between Air Pollution and Daily Deaths. Epidemiology 11(3): 320–326.

Solow R.M. (1960): On a Family of Lag Distributions. Econometrica 28: 393–406.

Welty L.J., Peng R.D., Zeger S.L., Dominici F. (2009): Bayesian Distributed Lag Models: Estimating Effects of Particulate Matter Air Pollution on Daily Mortality. Biometrics 65(1): 282–291.

Wright S. (1934): The Method of Path Coefficients. Annals of Mathematical Statistics 5(3):161–215.

Zanobetti A., Wand M.P., Schwartz J., Ryan L.M. (2000): Generalized Additive Distributed Lag Models: Quantifying Mortality Displacement. Biostatistics 1(3): 279–292.