Austrian Journal of Statistics
January 2019, Volume 48, 14–42.
http://www.ajs.or.at/
doi:10.17713/ajs.v48i2.777

Distributed-lag Linear Structural Equation Models in R: The dlsem Package

Alessandro Magrini
Dep. Statistics, Computer Science, Applications
University of Florence, Italy

Abstract

In this paper, an extension of linear Markovian structural causal models is introduced, called distributed-lag linear structural equation models (DLSEMs), where each factor of the joint probability distribution is a distributed-lag linear regression with constrained lag shapes. DLSEMs account for temporal delays in the dependence relationships among the variables and allow to assess dynamic causal effects. As such, they represent a suitable methodology to investigate the effect of an external impulse on a multidimensional system through time. In this paper, we present the dlsem package for R implementing inference functionalities for DLSEMs. The use of the package is illustrated through an example on simulated data and a real-world application aiming at assessing the impact of agricultural research expenditure on multiple dimensions in Europe.

Keywords: constrained lag shapes, directed acyclic graphs, dynamic causal inference, probabilistic graphical models, time series.

1. Introduction

Structural causal models (SCMs, Pearl 2000, Chapter 5) represent one of the prevalent methodologies for causal inference in contemporary applied sciences. In particular, a Markovian SCM is such that a directed acyclic graph (DAG) encodes causal relationships among the variables, which also implies a factorization of the joint probability distribution according to conditional independence relationships. In a linear parametric formulation (linear Markovian SCM), each factor of the joint probability distribution is a linear regression model, and a causal effect is associated to each edge, directed path or couple of nodes in the DAG to represent average changes in the value of a variable induced by an intervention provoking a unit variation in the value of another variable. A linear Markovian SCM can be extended by letting each factor of the joint probability distribution be a dynamic linear model, in order to account for temporal variations in the dependence relationships among the variables.

When the objective is forecasting, a large variety of dynamic linear models is available. For example, linear regression with polynomial lag shapes (Almon 1965) is a common feature of dynamic predictive models. Mixed-data sampling (MIDAS) regression (Andreou, Ghysels, and Kourtellos 2007; Ghysels, Sinko, and Valkanov 2007), implemented in the R package midasr (Ghysels, Kvedaras, and Zemlys 2016), is a broader class of time series models dealing with data sampled at different frequencies. The R package dlnm (Gasparrini 2011) contains functionalities for linear and non-linear regression models with spline lag-shapes. However, the assessment of causal effects in a dynamic context requires to take into account prior knowledge on lag shapes, and a predictive model may not be adequate without introducing appropriate mathematical constraints. For instance, the effect of an intervention has typically the same sign, or at least is null, at all the possible time lags, but neither the Almon's nor a spline lag shape can be directly applied to this purpose, as they cannot avoid regression coefficients to have different signs.

Constrained lag shapes (Judge, Griffiths, Hill, Lutkepohl, and Lee 1985, Chapters 9-10) overcome such problem, because regression coefficients are non-zero only outside a pre-specified interval of time lags, and within that interval they are allowed to rise from value zero to a maximum before declining again to zero, or to simply decrease from a maximum value to zero. Also, parameter estimation is straightforward, because, given the interval of time lags, the maximum value can be estimated using ordinary least squares. Thus, if prior knowledge on the interval is available, it can be directly exploited, while if the interval is not known a-priori, it may be determined by model selection across all the possible ones.

In this paper, we introduce an extension of linear Markovian SCMs, called *distributed-lag linear structural equation models* (DLSEMs), where each factor of the joint probability distribution is a distributed-lag linear regression with constrained lag shapes. They were introduced for the first time by Magrini (2018) in the context of lag exposure assessment. DLSEMs represent a suitable methodology to investigate the effect of an external impulse on a multidimensional system through time.

The aim of this paper is to illustrate the dlsem package, which implements inference functionalities for DLSEMs in R. The paper is structured as follows. In Section 2, theory on DLSEMs is presented. In Section 3, instructions for the installation of the dlsem package are provided. In Section 4, the practical use of the dlsem package is illustrated through an example on simulated data. In Section 5, the dlsem package is applied to address a real-world application aiming at assessing the impact of agricultural research expenditure on multiple dimensions in Europe. Section 6 includes conclusive remarks and considerations on future development.

2. Theory

In this section, theory on distributed-lag linear regression and on structural causal models is provided (Subsections 2.1 and 2.2), then distributed-lag linear structural equation models are presented (Subsection 2.3).

2.1. Distributed-lag linear regression

Lagged instances of one or more covariates may be included in the linear regression model to account for temporal delays in their influence on the response:

$$y_t = \beta_0 + \sum_{j=1}^{J} \sum_{l=0}^{L_j} \beta_{j,l}\ x_{j,t-l} + \epsilon_t \qquad \epsilon_t \sim \mathrm{N}(0, \sigma^2) \tag{1}$$

where y_t is the value of the response variable at time t and $x_{j,t-l}$ is the value of the j-th covariate at l time lags before t. The set $(\beta_{j,0}, \beta_{j,1}, \ldots, \beta_{j,L_j})$ is denoted as the *lag shape* of the j-th covariate and represents its regression coefficient (in the remainder, simply 'coefficient') at different time lags.

Parameter estimation is inefficient because lagged instances of the same covariate are typically highly correlated. The Almon's polynomial lag shape (Almon 1965) is a well-known solution to this problem, where coefficients for lagged instances of a covariate are forced to follow a polynomial of order Q:

$$\beta_{j,l} = \begin{cases} \phi_{j,0} & l = 0 \\ \sum_{q=0}^{Q} \phi_{j,q} l^q & \text{otherwise} \end{cases} \tag{2}$$

Unfortunately, the Almon's polynomial lag shape may show multiple modes and coefficients with different signs, thus entailing problems of interpretation. Constrained lag shapes (Judge *et al.* 1985, Chapters 9-10) overcome this deficiency. Some examples are the *endpoint-constrained quadratic* lag shape:

$$\beta_{j,l} = \begin{cases} \theta_j \left[-\frac{4}{(b_j-a_j+2)^2} l^2 + \frac{4(a_j+b_j)}{(b_j-a_j+2)^2} l - \frac{4(a_j-1)(b_j+1)}{(b_j-a_j+2)^2} \right] & a_j \leq l \leq b_j \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

the *quadratic decreasing* lag shape:

$$\beta_{j,l} = \begin{cases} \theta_j \frac{l^2 - 2(b_j+1)l + (b_j+1)^2}{(b_j-a_j+1)^2} & a_j \leq l \leq b_j \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

and the *gamma* lag shape:

$$\beta_{j,l} = \theta_j (l+1)^{\frac{\delta}{1-\delta}} \lambda_j^l \left[\left(\frac{\delta_j}{(\delta_j - 1)\log(\lambda_j)} \right)^{\frac{\delta_j}{1-\delta_j}} \lambda_j^{\frac{\delta_j}{(\delta_j-1)\log(\lambda_j)} - 1} \right]^{-1} \tag{5}$$
$$0 < \delta_j < 1 \qquad 0 < \lambda_j < 1$$

The endpoint-constrained quadratic lag shape is zero for a lag $l \leq a_j - 1$ or $l \geq b_j + 1$, and symmetric with mode equal to θ_j at lag $(a_j + b_j)/2$. The quadratic decreasing lag shape decreases from value θ_j at lag a_j to value 0 at lag $b_j + 1$ according to a quadratic function. The gamma lag shape is positively skewed with mode equal to θ_j at lag $\frac{\delta_j}{(\delta_j-1)\log(\lambda_j)}$. Value a_j is denoted as the *gestation lag*, value b_j as the *lead lag*, and value $b_j - a_j$ as the *lag width*. A static coefficient (no lag shape) is obtained if $a_j = b_j = 0$. Since it is not expressed as a function of a_j and b_j, the gamma lag shape cannot reduce to a static coefficient, but the corresponding values of a_j and b_j may be computed through numerical approximation from the values of δ_j and λ_j.

For these three lag shapes it holds:

$$\begin{aligned} \beta_{j,l} > 0 \Longleftrightarrow \theta_j > 0 \\ \beta_{j,l} < 0 \Longleftrightarrow \theta_j < 0 \end{aligned} \qquad \forall \; a_j \leq l \leq b_j \tag{6}$$

and we refer to the *lag sign* as the sign of parameter θ_j.

A linear regression model with constrained lag shapes is linear in parameters $\beta_0, \theta_1, \ldots, \theta_J$, provided that the values of $a_1, \ldots, a_J, b_1, \ldots, b_J$ are known. Thus, one may use ordinary least squares to estimate parameters $\beta_0, \theta_1, \ldots, \theta_J$ for several models with different values of $a_1, \ldots, a_J, b_1, \ldots, b_J$, and then select the one with the minimum value of the Bayesian Information Criterion (BIC, Schwarz 1978)[1]. A heuristic algorithm is shown below.

```
0. Let $J$ be the number of covariates and $T$ the greatest time lag under consideration. Initialize selected as an empty vector. Set candidates $= \{X_1, \ldots, X_J\}$. For $j = 1, \ldots, J$, set $a_j = 0$ and $b_j = 0$.

1. Repeat until candidates is not empty:

    a) initialize fitting as an empty vector. Set cand.temp = candidates;
```

[1] Alternatively, the Akaike Information Criterion (AIC, Akaike 1974) may be used.

- b) repeat until cand.temp is not empty:
  - b1) determine j such that the first element in cand.temp is X_j. Initialize fit.temp as an empty matrix with $T+1$ rows and $T+1$ columns;
  - b2) for $s_1 = 0, \ldots, T$ and for $s_2 = 0, \ldots, T$:
    - b2.1) set $a_j = s_1$, $b_j = s_2$ and $\mathsf{cov.temp} = \{\mathsf{selected} \cup X_j\}$;
    - b2.2) use ordinary least squares to estimate parameters $\beta_0, \theta_1, \ldots, \theta_J$ for the model with covariates cov.temp, gestation lags $\{a_k : \ X_k \in \mathsf{cov.temp}\}$ and lead lags $\{b_k : \ X_k \in \mathsf{cov.temp}\}$. Compute the BIC for this model and insert it into $\mathsf{fit.temp}(s_1, s_2)$;
  - b3) determine k_1 and k_2 such that $\mathsf{fit.temp}(k_1, k_2)$ is the best value in fit.temp. Insert $\mathsf{fit.temp}(k_1, k_2)$ into fitting. Set $a_j = k_1$ and $b_j = k_2$. Remove X_j from cand.temp;
- c) determine X_k such that the k-th value in fitting is the minimum one. Insert X_k into selected and remove X_k from candidates.

2. Use ordinary least squares to estimate parameters $\beta_0, \theta_1, \ldots, \theta_J$ for the model with covariates in selected, gestation lags $a_1, \ldots, a_J$ and lead lags $b_1, \ldots, b_J$.

Note that neither the response variable nor the covariates must contain unit root in order to obtain unbiased estimates with ordinary least squares (Granger and Newbold 1974). A reasonable procedure is to sequentially apply differentiation to all variables until the Augmented Dickey-Fuller test (Dickey and Fuller 1981) rejects the hypothesis of unit root for all of them.

2.2. Structural causal models

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

- $\boldsymbol{V} = \{V_1, \ldots, V_J\}$ is a set of endogenous variables;
- $\Omega_{\boldsymbol{V}} = \Omega_{V_1} \times \ldots \times \Omega_{V_J}$ is the cartesian product of the domains of variables in $\boldsymbol{V}$;
- $\boldsymbol{U} = \{U_1, \ldots, U_K\}$ is a set of unobserved variables;
- $\Omega_{\boldsymbol{U}} = \Omega_{U_1} \times \ldots \times \Omega_{U_K}$ is the cartesian product of the domains of variables in $\boldsymbol{U}$;
- $\boldsymbol{f} : \ \Omega_{\boldsymbol{V}} \times \Omega_{\boldsymbol{U}} \longrightarrow \Omega_{\boldsymbol{V}}$ is a measurable function;
- $\mathbb{P}_{\boldsymbol{U}}$ is a probability measure on $\Omega_{\boldsymbol{U}}$.

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

$$p(v_1, \ldots, v_J) = \prod_{j=1}^{J} p(v_j \mid \Pi_j = \pi_j) \tag{7}$$

where Π_j is the set of variables in $\boldsymbol{V}$ such that, for $j > 1$, V_j is independent of variables in $\{V_1, \ldots, V_{j-1}\} \setminus \Pi_j$, given variables in Π_j. This means that the joint probability distribution of variables in $\boldsymbol{V}$ can be factored according to conditional independence relationships holding among them disregarding variables in $\boldsymbol{U}$. Pearl (2000, pages 12 and following) shows that these conditional independence relationships are encoded into a directed acyclic graph (DAG)

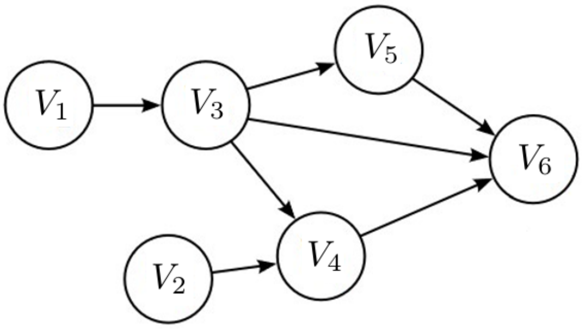

Figure 1: An example of directed acyclic graph.

such that Π_j is the parent set of V_j, $\forall\ j = 1,\ldots,J$. For example, in the Markovian SCM associated to the DAG in Figure 1, it holds:

$$p(v_1,v_2,v_3,v_4,v_5,v_6) = p(v_1)\ p(v_2)\ p(v_3 \mid v_1)\ p(v_4 \mid v_2,v_3)\ p(v_5 \mid v_3)\ p(v_6 \mid v_3,v_4,v_5) \quad (8)$$

and, by way of illustration, V_6 is independent of V_1 and V_2 given V_3, V_4 and V_5.

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

$$p(v_1,\ldots,v_J \mid \mathrm{do}(V_i = v_i)) = \prod_{j\neq i} p(v_j \mid \pi_j)\ |_{V_i=v_i} \quad (9)$$

where $|_{V_i=v_i}$ indicates that $p(v_i \mid \pi_i)$ is replaced by value v_i. This formula, called *truncated factorization* (Pearl 2000, Section 3.2), allows to compute the effect of an intervention from the (pre-intervention) distribution in Formula 7, that is to predict the effect of an intervention from non-experimental (observational) data. In a Markovian SCM, the effect of $\mathrm{do}(V_i = v_i)$ on V_j, called *causal effect* of V_i on V_j, is given by the following expression (see Pearl 2000, page 70 and following):

$$p(V_j = v_j \mid \mathrm{do}(V_i = v_i)) = \sum_{\pi_i} p(V_j = v_j \mid V_i = v_i, \Pi_i = \pi_i) p(\Pi_i = \pi_i) \quad (10)$$

where Π_i is the parent set of V_i.

In a linear parametric formulation (linear Markovian SCMs), each factor $p(v_j \mid \pi_j)$ of the joint probability distribution in Formula 7 is the linear regression model where V_j is the response variable and variables in Π_j are the covariates. For example, in the linear Markovian SCM associated to the DAG in Figure 1, $p(v_4 \mid v_2, v_3)$ 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 of Formula 10, as shown in the following paragraphs.

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

$$\beta_{j\,|\,i} := \Delta\mathrm{E}(V_j \mid \Delta V_i = 1,\, \Delta V_{k:\ V_k \in \{\Pi_j \setminus V_i\}} = 0) \quad (11)$$

Expression 11 is a special case of Expression 10, 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 variables in Π_i but not in Π_j are independent of V_j conditionally to variables in Π_j (see Formula 7), we can conclude that $\beta_{j|i}$ represents the average effect of $\mathrm{do}(\Delta V_i = 1)$ on V_j:

$$\beta_{j\,|\,i} := \Delta\mathrm{E}(V_j \mid \Delta V_i = 1,\, \Delta V_{k:\ V_k \in \{\Pi_j \setminus V_i\}} = 0) = \Delta\mathrm{E}(V_j \mid \mathrm{do}(\Delta V_i = 1); < V_i, V_j >) \quad (12)$$

which is called *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 12 is associated to the edge $< V_i, V_j >$. For example, in the linear Markovian SCM associated to the DAG in Figure 1, $\beta_{4|3}$ represents the change in the expected value of V_4 given a unit variation of V_3, at constant value of V_2, equating 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 $\text{do}(\Delta V_i = 1)$ influences the expected value of V_j independently through each directed path connecting V_i to V_j, for an *overall* causal effect equal to the sum of the causal effects associated to each of these paths:

$$\Delta\text{E}(V_j \mid \text{do}(\Delta V_i = 1)) := \sum_{<V_{d_0},\ldots,V_{d_m}>:\ d_0=i \wedge d_m=j} \Delta\text{E}(V_j \mid \text{do}(\Delta V_i = 1); <V_{d_0},\ldots,V_{d_m}>) \quad (13)$$

where $\Delta\text{E}(V_j \mid \text{do}(\Delta V_i = 1); < V_{d_0},\ldots,V_{d_m}>)$ is the causal effect of $\text{do}(\Delta V_i = 1)$ on V_j associated to the directed path $< V_{d_0},\ldots,V_{d_m} >$ ($d_0 = i$ and $d_m = j$) connecting V_i to V_j, denoted as the *pathwise* causal effect of V_i on V_j through $< V_{d_0},\ldots,V_{d_m} >$.

A pathwise causal effect associated to an edge (direct causal effect) can be computed using Formula 12. Instead, a pathwise causal effect associated to a multi-edge directed path, also referred as *indirect* causal effect, can be computed through the product of the regression coefficients associated to each edge in the path (see, for example, Wright 1934):

$$\begin{aligned}\Delta\text{E}(V_j \mid \text{do}(\Delta V_i = 1); < V_i,\ldots,V_j >) &:= \prod_{k:\ V_k\in<V_i,\ldots,V_j>\wedge k\neq i} \Delta\text{E}(V_k \mid \text{do}(\Delta V_{k-1} = 1); < V_{k-1}, V_k >) = \\ &= \prod_{k:\ V_k\in<V_i,\ldots,V_j>\wedge k\neq i} \beta_{k|k-1}\end{aligned} \quad (14)$$

Note that Formula 2.2 is a generalization of Formula 12. 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 to 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 $\Delta\text{E}(V_6 \mid \text{do}(\Delta V_3 = 1))$, is equal to $\beta_{6|3} + \beta_{4|3} \cdot \beta_{6|4} + \beta_{5|3} \cdot \beta_{6|5}$.

2.3. Distributed-lag linear structural equation models

Delayed dependence relationships among the variables may be taken into account in a Markovian SCM by specifying each factor of the joint probability distribution in Formula 7 equal to the distributed-lag linear regression in Formula 1. We refer to this Markovian SCM as *distributed-lag linear structural equation model* (DLSEM). The definition of causal effects at different time lags in a DLSEM is provided in the following paragraphs.

Direct causal effects Let $b^{(l)}_{j|i}$ be the coefficient of V_i at lag l in the regression model of V_j. Such coefficient equates the *direct* causal effect of V_i on V_j at lag l:

$$\Delta\text{E}^{(l)}(V_j \mid \text{do}(\Delta V_i = 1); < V_i, V_j >) := b^{(l)}_{j|i} \quad (15)$$

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

$$e_{(q_1,\ldots,q_m)}(< V_{d_0},\ldots,V_{d_m} >; d_0 = i, d_m = j) = \prod_{k=1}^{m} b^{(q_k)}_{d_k|d_{k-1}} \quad (16)$$

we obtain one of the possible causal effects of V_i on V_j through $< V_{d_0}, \ldots, V_{d_m} >$ at lag l. Thus, the *indirect* causal effect of V_i on V_j through $< V_{d_0}, \ldots, V_{d_m} >$ ($d_0 = i$ and $d_m = j$) at lag l is equal to the sum of all the causal effects that can be obtained from Formula 16:

$$\Delta \mathrm{E}^{(l)}(V_j \mid \mathrm{do}(\Delta V_i = 1); < V_{d_0}, \ldots, V_{d_m} >, d_0 = i, d_m = j) := \sum_{(q_1,\ldots,q_m) \in Q_m^{(l)}} \prod_{k=1}^{m} b_{d_k|d_{k-1}}^{(q_k)} \tag{17}$$

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

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

The DAG of a DLSEM includes all the possible temporal instances of each variable in $\boldsymbol{V}$, but it 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.

A DLSEM is a special case of dynamic Bayesian network (Murphy 2002), where endogenous variables follow the Gaussian distribution and lag shapes are possibly constrained to predefined functional forms.

3. Installation

Before installing dlsem, you must have installed R version 2.1.0 or higher, which is freely available at http://www.r-project.org/.

To install the dlsem package, type the following in the R command prompt:

```
> install.packages("dlsem")
```

and R will automatically install the package to your system from CRAN. In order to keep your copy of dlsem up to date, use the command:

```
> update.packages("dlsem")
```

All the results shown in this paper are obtained using R 3.4.3 with the dlsem package version 2.3. The official web page of dlsem is: https://cran.r-project.org/web/packages/dlsem/.

4. Example on simulated data

In this section, the practical use of the dlsem package is illustrated through a simple impact assessment problem referred as “industrial development problem”. The objective is to test whether the influence through time of the number of job positions in industry (proxy of the industrial development, label: Job) on the amount of greenhouse gas emissions (proxy of pollution, label: Pollution) is direct and/or mediated by the amount of private consumption (label: Consum). The static representation of the DAG for the industrial development problem is shown in Figure 2. The analysis will be conducted on the dataset industry, containing simulated data for 10 imaginary regions in the period 1983-2015.

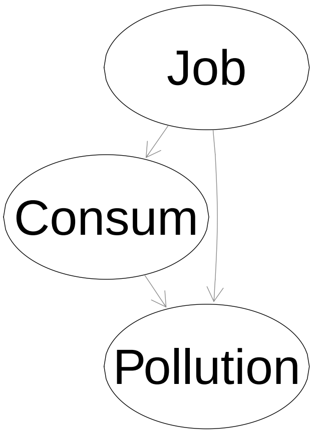

Figure 2: The static representation of the DAG for the industrial development problem. 'Job': number of job positions in industry. 'Consum': private consumption index. 'Pollution': amount of greenhouse gas emissions.

```
> data(industry)
> summary(industry)

     Region         Year        Population            GDP
 1      : 32   Min.   :1983   Min.   : 4771649   Min.   :  97119
 2      : 32   1st Qu.:1991   1st Qu.: 8310737   1st Qu.: 186783
 3      : 32   Median :1998   Median :25381874   Median : 463942
 4      : 32   Mean   :1998   Mean   :32368547   Mean   : 727735
 5      : 32   3rd Qu.:2006   3rd Qu.:56273337   3rd Qu.:1307044
 6      : 32   Max.   :2014   Max.   :78308254   Max.   :1883702
 (Other):128
      Job             Consum          Pollution
 Min.   : 34.77   Min.   : 37.35   Min.   :  3161
 1st Qu.:105.07   1st Qu.: 87.88   1st Qu.:  7536
 Median :137.03   Median :108.47   Median : 25320
 Mean   :127.61   Mean   :108.17   Mean   : 32202
 3rd Qu.:152.68   3rd Qu.:124.85   3rd Qu.: 47109
 Max.   :200.83   Max.   :211.16   Max.   :101441
```

Since data are grouped into regions, we inspect minimum and maximum values by region:

```
> by(industry[,-c(1:2)],industry$Region,function(x){apply(x,2,quantile,prob=c(0,1))})

industry$Region: 1
     Population    GDP       Job   Consum Pollution
0%      9819005 175519  83.41447 112.1351  6957.444
100%    9973483 190379 200.83461 126.4740  7794.559
--------------------------------------------------------------------------------
industry$Region: 2
     Population     GDP       Job   Consum Pollution
0%     77532689 1688751  89.04076 53.35230  49290.65
100%   78308254 1883702 146.64997 88.71258  67572.55
--------------------------------------------------------------------------------
industry$Region: 3
     Population      GDP      Job   Consum Pollution
0%      5097638 153685.9 116.4307 108.5695  9689.909
100%    5136994 170871.8 183.5543 154.0917 11167.293
--------------------------------------------------------------------------------
industry$Region: 4
     Population      GDP      Job   Consum Pollution
```

```
0%     36713363 545880.2 137.0105 100.5134  37810.26
100%   37391776 597517.1 168.2146 140.6425  54913.93
--------------------------------------------------------------------------------
industry$Region: 5
    Population       GDP      Job   Consum Pollution
0%      4771649  97119.23 127.3341 102.2382  5440.151
100%    4791872 112493.15 171.6086 144.7355  6439.347
--------------------------------------------------------------------------------
industry$Region: 6
    Population     GDP      Job   Consum Pollution
0%     54035307 1257045 104.8938 104.4684  86609.89
100%   55146391 1356454 171.4953 197.3554 101440.95
--------------------------------------------------------------------------------
industry$Region: 7
    Population     GDP      Job   Consum Pollution
0%     56248555 1105558 146.3254 103.1810  33807.57
100%   56666469 1218726 171.5073 211.1634  42752.34
--------------------------------------------------------------------------------
industry$Region: 8
    Population      GDP       Job   Consum Pollution
0%     13954409 353391.7  60.57895 37.34921  10317.25
100%   14050385 382004.6 141.52760 86.21321  16831.72
--------------------------------------------------------------------------------
industry$Region: 9
    Population      GDP       Job   Consum Pollution
0%      8281398 215630.6  34.77417 81.35061  3160.522
100%    8353897 241041.2 152.80416 93.08245  7895.383
--------------------------------------------------------------------------------
industry$Region: 10
    Population     GDP       Job   Consum Pollution
0%     56127386 1222594  85.95371  89.0765  41684.86
100%   56608406 1353481 183.57437 111.8265  48170.81
```

These summaries highlight a relevant heterogeneity between regions. As shown in the remainder, the `dlsem` package can take into account a grouping factor, thus explaining the variance due to differences among several groups. Furthermore, it can be noted that regions have the same number of measurements:

```
> table(industry$Region)

 1  2  3  4  5  6  7  8  9 10
32 32 32 32 32 32 32 32 32 32
```

This feature is not essential for the application of DLSEMs: in general, the sample size is not required to be equal for all groups.

4.1. Specification of the model code

The first step to build a DLSEM with the `dlsem` package is the definition of the model code, which includes the formal specification of the regression models. The variables for which a regression model is specified are called *endogenous* variables. The other variables are referred as *exogenous* variables.

The model code must be a list of formulas, one for each regression model. In each formula, the response and the covariates must be quantitative variables[2], and operators `quec.lag`$(\cdot)$, `qdec.lag`$(\cdot)$ and `gamma.lag`$(\cdot)$ can be employed to specify, respectively, an endpoint-constrained quadratic, a quadratic decreasing or a gamma lag shape. Operators `quec.lag`$(\cdot)$ and `qdec.lag`$(\cdot)$ have three mandatory arguments:

[2] Qualitative variables can be included only as exogenous variables, as described in Subsection 4.3.

- the name of the covariate to which the lag shape is applied,
- the gestation lag (a_j),
- the lead lag (b_j).

Operator `gamma.lag`$(\cdot)$ has three mandatory arguments:

- the name of the covariate to which the lag shape is applied,
- parameter δ_j,
- parameter λ_j.

If none of these two operators is applied to a covariate, it is assumed that its coefficient is equal to 0 for time lags greater than 0 (no lag shape). The group factor and exogenous variables must not appear in the model code (see Subsection 4.3 for the way to include them).

The following code specifies all lag shapes as endpoint-constrained quadratic lag shapes between 0 and 15 time lags:

```
> indus.code <- list(
+   Job ~ 1,
+   Consum ~ quec.lag(Job,0,15),
+   Pollution ~ quec.lag(Job,0,15)+quec.lag(Consum,0,15)
+   )
```

The formula of regression models with no endogenous covariates may be omitted from the model code. For example, the following code (where the formula specifying the regression model for variable `Job` is omitted) is equivalent to the previous one:

```
> indus.code <- list(
+   Consum ~ quec.lag(Job,0,15),
+   Pollution ~ quec.lag(Job,0,15)+quec.lag(Consum,0,15)
+   )
```

4.2. Specification of control options

The second step to build a DLSEM with the `dlsem` package is the specification of control options, which are distinguished into global (applied to all regression models) and local (model-specific) options.

Global control options must be a named list with one or more of the following components:

- `adapt`: a logical value indicating if adaptation of lag shapes must be performed, that is parameters of lag shapes must be chosen on the basis of fit to data. Default is `FALSE`, meaning no adaptation;
- `max.gestation`: the maximum gestation lag for all lag shapes. If not provided, it is taken as equal to `max.lead` (see below);
- `max.lead`: the maximum lead lag for all lag shapes. If not provided, it is computed accordingly to the sample size;
- `min.width`: the minimum lag width for all lag shapes. It cannot be greater than `max.lead`. If not provided, it is taken as 0;
- `sign`: the lag sign for all lag shapes, that can be either `'+'` for positive or `'-'` for negative. If not provided, adaptation will disregard the lag sign;

- `selection`: the criterion to be used for the adaptation of lag shapes, that can be one among `'bic'` (the default) and `'aic'` to minimise BIC or AIC, respectively.

Local control options must be a named list containing one or more among the following components:

- `adapt`: a named vector of logical values, where each component must have the name of one endogenous variable and indicate if adaptation of lag shapes must be performed for the regression model of that variable;
- `max.gestation`: a named list. Each component of the list must have the name of one endogenous variable and be a named vector. Each component of the named vector must have the name of one covariate in the regression model of the endogenous variable above and include the maximum gestation lag for its lag shape;
- `max.lead`: the same as `max.gestation`, with the exception that the named vector must include the maximum lead lag;
- `min.width`: the same as `max.gestation`, with the exception that the named vector must include the minimum lag width;
- `sign`: the same as `max.gestation`, with the exception that the named vector must include the lag sign (either `'+'` for positive or `'-'` for negative).

Local control options have no default values, and global ones are applied in their absence. If some local control options conflict with global ones, only the former are applied.

Suppose that one wants to perform adaptation with the following constraints for all lag shapes: (i) maximum gestation lag of 3 years, (ii) maximum lead lag of 15 years, (iii) minimum lag width of 5 years, (iv) positive lag sign. Control options for these constraints can be expressed in several ways. The most simple solution is to specify only global control options, as the constraints hold for all regression models:

```
> indus.global <- list(adapt=T,max.gestation=3,max.lead=15,min.width=5,sign="+")
> indus.local <- list()
```

Alternatively, one may specify only local control options by repeating them for each regression model:

```
> indus.global <- list()
> indus.local <- list(
+   adapt=c(Consum=T,Pollution=T),
+   max.gestation=list(Consum=c(Job=3),Pollution=c(Job=3,Consum=3)),
+   max.lead=list(Consum=c(Job=15),Pollution=c(Job=15,Consum=15)),
+   min.width=list(Consum=c(Job=5),Pollution=c(Job=5,Consum=5)),
+   sign=list(Consum=c(Job="+"),Pollution=c(Job="+",Consum="+"))
+   )
```

or both local and global control options:

```
> indus.global <- list(adapt=T,min.width=5)
> indus.local <- list(
+   max.gestation=list(Consum=c(Job=3),Pollution=c(Job=3,Consum=3)),
+   max.lead=list(Consum=c(Job=15),Pollution=c(Job=15,Consum=15)),
+   sign=list(Consum=c(Job="+"),Pollution=c(Job="+",Consum="+"))
+   )
```

4.3. Parameter estimation

Once the model code and control options are specified, parameter estimation can be performed using the command `dlsem`(·). The main arguments of the command include:

- model.code, the first argument of the command, requiring the model code in the format defined in Subsection 4.1;
- group, accepting the name of a single group factor (optional). By indicating the group factor, one intercept for each level of the group factor will be estimated in each regression model.
- exogenous, accepting the name of exogenous variables (optional). Exogenous variables can be either qualitative or quantitative, and will be included in each regression model with no lag shape;
- log, a logical value indicating whether logarithmic transformation must be applied to all strictly positive quantitative variables (default is FALSE). If log is set to TRUE, the coefficient of each strictly positive quantitative covariate is (approximatively) interpreted as an elasticity, that is as an expected percentage increase in the value of the response variable for 1% increase in the value of the covariate[3];
- data, requiring an object of class data.frame containing data;
- global.control and local.control, accepting global and local options, respectively, in the format defined in Subsection 4.2 (optional).

Before parameter estimation, differentiation is performed until the hypothesis of unit root is rejected by the Augmented Dickey-Fuller test for all quantitative variables. If the group factor is specified, the panel version of the Augmented Dickey-Fuller test proposed by Levin, Lin, and Chub (2002) is used instead. Also, missing values, if present, are imputed with their conditional mean using the Expectation-Maximization algorithm (Dempster, Laird, and Rubin 1977)[4]. For further arguments controlling differentiation and imputation options, see the documentation of command dlsem(·) by typing ?dlsem.

In the following code, the region is indicated as the group factor, population and gross domestic product are indicated as exogenous variables, the logarithmic transformation is requested, and both global and local control options specified in Subsection 4.2 are provided:

```
> indus.mod <- dlsem(indus.code,group="Region",exogenous=c("Population","GDP"),
+   data=industry,global.control=indus.global,local.control=indus.local,log=T)

Checking stationarity...
 Order 1 differentiation performed
Starting estimation...
 Estimating regression model 1/3 (Job)
 Estimating regression model 2/3 (Consum)
 Estimating regression model 3/3 (Pollution)
Estimation completed
```

Messages inform that differentiation was applied one time in order to achieve stationarity of all time series. No mention to imputation was made, meaning that data are complete. The results of command dlsem(·) is an object of class dlsem. Among the components of dlsem objects, we highlight:

- estimate: a list of objects of class lm, one for each endogenous variable;

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

[4] Imputation of missing values is performed after eventual logarithmic transformation and differentiation by assuming group-specific means and time-invariant covariance matrix. Qualitative variables cannot contain missing values. Each quantitative variable must have at least 3 observed values if the group factor is not specified, otherwise it must have at least 3 observed values per group.

- `model.code`: the model code after eventual adaptation;
- `data.used`: data after eventual logarithmic transformation and differentiation.

The `summary` method for class `dlsem` returns the summary of the estimation:

```
> summary(indus.mod)

ENDOGENOUS PART

Response: Job
-

Response: Consum
                              Estimate Std. Error  t value     Pr(>|t|)
quec.lag(Job, 0, 5, Region) 0.1006394 0.01783725 5.642089 4.589874e-08 ***

Response: Pollution
                                 Estimate Std. Error  t value     Pr(>|t|)
quec.lag(Job, 1, 8, Region)    0.1048006 0.03008457 3.483532 5.989626e-04 ***
quec.lag(Consum, 1, 6, Region) 0.2320105 0.03660783 6.337729 1.339514e-09 ***


EXOGENOUS PART

Response: Job
            Estimate Std. Error   t value      Pr(>|t|)
Population -2.015755 0.36919466  -5.45987  1.004944e-07 ***
GDP        -1.274005 0.03253314 -39.16023 1.591909e-119 ***

Response: Consum
             Estimate Std. Error    t value     Pr(>|t|)
Population  0.8397265 0.30729012   2.732683 6.735972e-03  **
GDP        -0.8165645 0.02710312 -30.128064 1.096637e-84 ***

Response: Pollution
             Estimate Std. Error   t value     Pr(>|t|)
Population -0.5335639 0.32247211 -1.654605 9.945701e-02   .
GDP         0.1342472 0.02965881  4.526384 9.908715e-06 ***


INTERCEPTS

Response: Job
             Estimate  Std. Error    t value     Pr(>|t|)
Region1  -0.027108664 0.002403134 -11.280545 8.189303e-25 ***
Region2  -0.014868387 0.002401561  -6.191135 1.975106e-09 ***
Region3  -0.014228172 0.002401629  -5.924383 8.639991e-09 ***
Region4  -0.005320298 0.002403060  -2.213968 2.758788e-02   *
Region5  -0.008833821 0.002401537  -3.678402 2.784066e-04 ***
Region6  -0.015622725 0.002401342  -6.505831 3.260886e-10 ***
Region7  -0.005154175 0.002401605  -2.146138 3.266936e-02   *
Region8  -0.027052095 0.002401793 -11.263293 9.395308e-25 ***
Region9  -0.046951445 0.002402163 -19.545484 2.514703e-55 ***
Region10 -0.023440072 0.002402647  -9.755938 1.077582e-19 ***

Response: Consum
             Estimate  Std. Error   t value     Pr(>|t|)
Region1   0.013228135 0.003105034  4.260222 2.905842e-05 ***
Region2  -0.009181367 0.002452433 -3.743779 2.255585e-04 ***
Region3   0.014910423 0.002369592  6.292400 1.413274e-09 ***
```

```
Region4   0.012261936 0.002143643  5.720139 3.065699e-08 ***
Region5   0.012591239 0.002189363  5.751097 2.609354e-08 ***
Region6   0.027006345 0.002425256 11.135464 1.319250e-23 ***
Region7   0.023946916 0.002133839 11.222454 6.881615e-24 ***
Region8  -0.014297098 0.003061892 -4.669367 4.962066e-06 ***
Region9   0.019452657 0.004455213  4.366268 1.860065e-05 ***
Region10  0.003490765 0.002834166  1.231673 2.192426e-01

Response: Pollution
              Estimate  Std. Error    t value     Pr(>|t|)
Region1   0.0181034164 0.005671781  3.1918397 1.624344e-03  **
Region2   0.0166945282 0.002993763  5.5764369 7.290975e-08 ***
Region3   0.0008710423 0.004745009  0.1835702 8.545228e-01
Region4   0.0038743955 0.003341414  1.1595078 2.475294e-01
Region5  -0.0047651007 0.003653564 -1.3042336 1.935420e-01
Region6  -0.0138551604 0.006254540 -2.2152164 2.778958e-02   *
Region7  -0.0133904268 0.004809974 -2.7838873 5.847590e-03  **
Region8   0.0294218841 0.004102569  7.1715751 1.164801e-11 ***
Region9   0.0029735558 0.008691559  0.3421200 7.325933e-01
Region10  0.0171095625 0.004253094  4.0228508 7.951079e-05 ***


ERRORS
               sigma  df
Job       0.01337002 298
Consum    0.01076881 247
Pollution 0.01112485 216


GOODNESS OF FIT

R-squared: 0.8609
AIC: -4786.373
BIC: -4636.377
```

We see that the number of job positions in industry (`Job`) significantly influences, on one hand, the amount of private consumption (`Consum`) from 0 to 5 years and, on the other hand, the amount of greenhouse gas emissions (`Pollution`) from 1 to 8 years, while the amount of private consumption (`Consum`) significantly influences the amount of greenhouse gas emissions (`Pollution`) from 1 to 6 years. This result provides evidence that the influence of industrial development on pollution is both direct and mediated by private consumption.

The `plot` method for class `dlsem` displays the DAG of the model in the static representation, where each edge is coloured with respect to the sign of the estimated causal effect (green: positive, red: negative, light gray: not statistically significant):

```
> plot(indus.mod)
```

Note that the DAG includes only the endogenous variables. Argument `conf` in the `plot` method controls the confidence level, which is equal to 0.95 by default. Here, a statistically significant and positive causal effect is associated to each edge, thus all of them are shown in green (Figure 3). Colours can be suppressed by setting option `style` to `1` in the `plot` method (default is `style=2`). Instead, by setting option `style` to `0`, all edges are shown in black disregarding statistical significance of causal effects (see Figure 2).

4.4. Assessment of causal effects

After parameter estimation is performed by means of command `dlsem(·)`, the command `causalEff(·)` can be used on the resulting object of class `dlsem` to compute all the path-

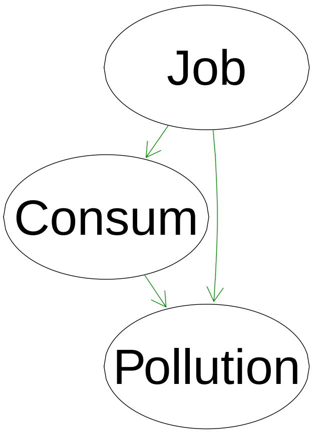

Figure 3: The static representation of the DAG for the industrial development problem, where each edge is coloured with respect to the sign of the estimated causal effect. Green: positive causal effect. Red: negative causal effect. Light grey: not statistically significant causal effect (no such edges here).

wise causal lag shapes and the overall one connecting two variables. The main arguments of command `causalEff`(·) include:

- `x`, the first argument, requiring an object of class `dlsem`;
- `from`, requiring the name of one or more starting variables, that is the variables generating the causal effect;
- `to`, requiring the name of the ending variable, that is the variable receiving the causal effect;
- `lag`, accepting specific time lags at which the causal effect must be computed. If no values are provided, all the relevant time lags are considered.
- `cumul`, a logical value indicating whether the cumulative causal effect must be returned. Default is `FALSE`.

Only exogenous variables can be indicated as starting or ending variables. Note that, due to the properties of the multiple linear regression model, causal effects are net of the influence of the group factor and exogenous variables.

The following code returns the cumulative causal effect (by path and the overall one) of the number of job positions on the amount of greenhouse gas emissions:

```
> causalEff(indus.mod,from="Job",to="Pollution",cumul=T)

$`Job*Consum*Pollution`
      estimate   std. err.   lower 95%   upper 95%
0  0.000000000 0.000000000 0.000000000 0.000000000
1  0.005601519 0.001338424 0.002978257 0.008224781
2  0.024273250 0.003426876 0.017556697 0.030989803
3  0.062239103 0.006316396 0.049859194 0.074619011
4  0.121988641 0.009697230 0.102982419 0.140994863
5  0.200409911 0.013132357 0.174670963 0.226148858
6  0.287544654 0.016155815 0.255879838 0.319209470
7  0.365965924 0.018423705 0.329856126 0.402075722
8  0.425715462 0.019838657 0.386832409 0.464598516
9  0.463681315 0.020535961 0.423431571 0.503931059
10 0.482353046 0.020776857 0.441631154 0.523074937
```

```
11 0.487954565 0.020819922 0.447148267 0.528760863
12 0.487954565 0.020819922 0.447148267 0.528760863

$`Job*Pollution`
    estimate  std. err.  lower 95%  upper 95%
0  0.0000000 0.00000000 0.00000000 0.00000000
1  0.0414027 0.01188526 0.01810801 0.06469739
2  0.1138574 0.02395552 0.06690547 0.16080937
3  0.2070135 0.03590255 0.13664579 0.27738119
4  0.3105202 0.04660327 0.21917950 0.40186096
5  0.4140270 0.05526967 0.30570041 0.52235354
6  0.5071830 0.06139921 0.38684280 0.62752328
7  0.5796378 0.06482646 0.45258023 0.70669529
8  0.6210405 0.06590698 0.49186516 0.75021576
9  0.6210405 0.06590698 0.49186516 0.75021576
10 0.6210405 0.06590698 0.49186516 0.75021576
11 0.6210405 0.06590698 0.49186516 0.75021576
12 0.6210405 0.06590698 0.49186516 0.75021576

$overall
     estimate  std. err.  lower 95%  upper 95%
0  0.00000000 0.00000000 0.00000000 0.00000000
1  0.04700422 0.01322369 0.02108627 0.07292217
2  0.13813067 0.02736157 0.08450297 0.19175836
3  0.26925259 0.04213927 0.18666113 0.35184405
4  0.43250887 0.05612473 0.32250642 0.54251132
5  0.61443689 0.06809948 0.48096437 0.74790940
6  0.79472770 0.07710063 0.64361323 0.94584216
7  0.94560369 0.08260701 0.78369692 1.10751046
8  1.04675592 0.08481875 0.88051424 1.21299761
9  1.08472178 0.08498455 0.91815513 1.25128843
10 1.10339351 0.08504308 0.93671214 1.27007488
11 1.10899503 0.08505361 0.94229301 1.27569704
12 1.10899503 0.08505361 0.94229301 1.27569704
```

The output of command `causalEff`$(\cdot)$ is a list of matrices including point estimates, standard errors[5] and asymptotic confidence intervals for all the pathwise causal lag shapes and the overall one connecting the starting variables to the ending variable.

Since the logarithmic transformation was applied to all quantitative variables, the resulting causal effects are interpreted as elasticities, that is, for a 1% of job positions more, greenhouse gas emissions are expected to grow by 0.61% after 5 years and by 1.11% after 10 years. The influence ends after 11 years, as the cumulative causal effects at lags 11 and 12 are equal.

A pathwise or an overall causal lag shape can be displayed using the command `lagPlot`$(\cdot)$. For instance, one may display the causal lag shape associated to each path connecting the number of job positions to the amount of greenhouse gas emissions (asterisks separate the name of variables in a path):

```
> lagPlot(indus.mod,path="Job*Pollution")
> lagPlot(indus.mod,path="Job*Consum*Pollution")
```

while the following code displays the the overall causal lag shape of the number of job positions on the amount of greenhouse gas emissions:

[5] A pathwise causal effect composed by the direct causal effects $\kappa_1, \dots, \kappa_m$ is equal to $\prod_{i=1}^m \kappa_i$, and the estimates of $\kappa_1, \dots, \kappa_m$, say $\hat{\kappa}_1, \dots, \hat{\kappa}_m$, are each other independent as they refer to different regression models. Thus, it holds:

$$\mathrm{Var}\left[\widehat{\prod_{i=1}^m \kappa_i}\right] = \prod_{i=1}^m \left(\hat{\kappa}_i^2 + \mathrm{Var}[\hat{\kappa}_i]\right) - \prod_{i=1}^m \hat{\kappa}_i^2$$

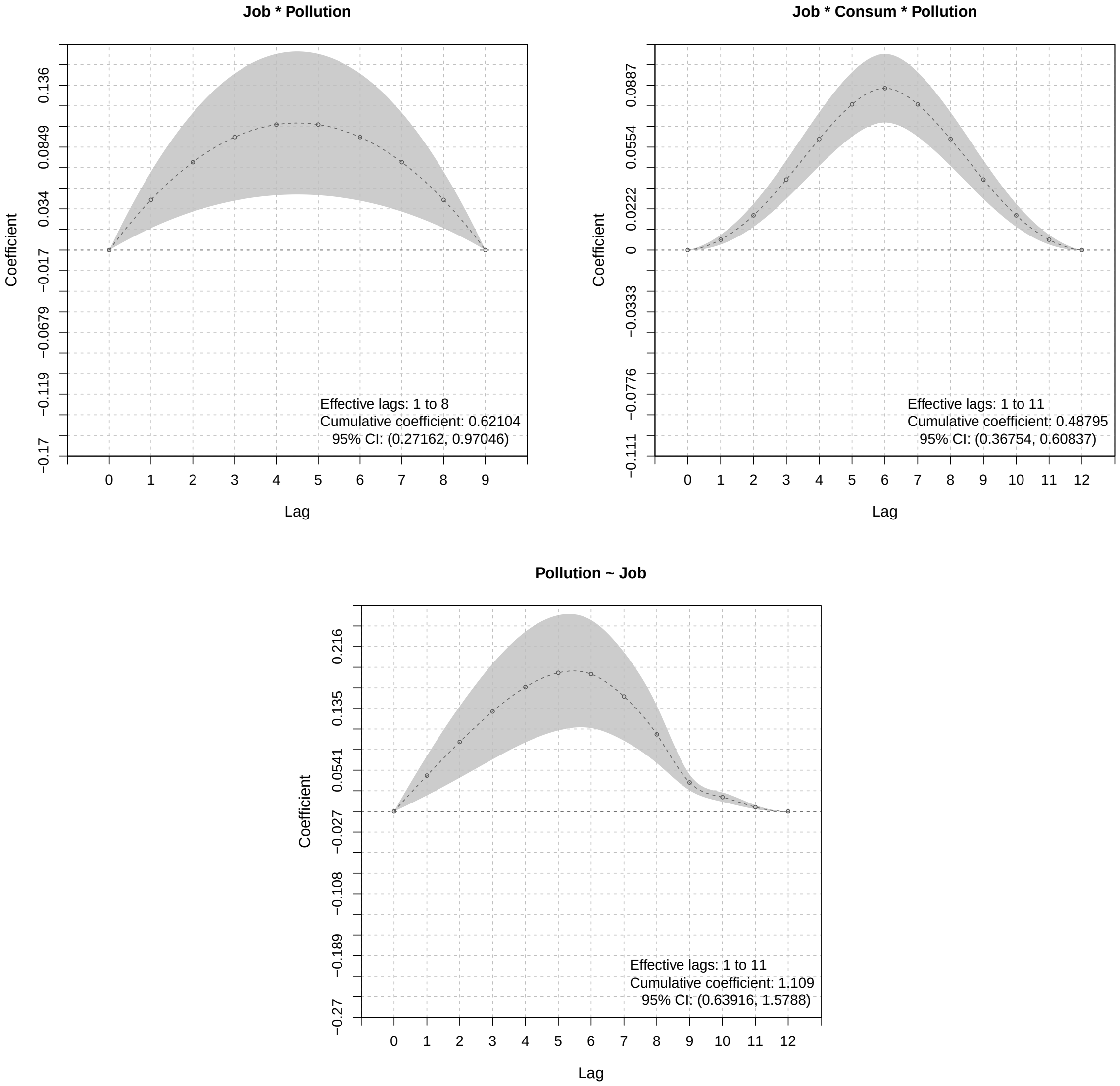

Figure 4: The pathwise causal lag shapes (upper panels) and the overall one (lower panel) connecting `Job` and `Pollution`. 95% asymptotic confidence intervals are shown in grey.

```
> lagPlot(indus.mod,from="Job",to="Pollution")
```

The resulting graphics are shown in Figure 4. Note that a multi-edge pathwise causal lag shape is a mixture of different lag shapes, thus it may show an irregular aspect, like it is the case of the overall causal lag shape displayed in the lower panel of Figure **??**.

4.5. Model comparison

We now fit two alternative models for the industrial development problem, such that all lag shapes are quadratic decreasing and gamma lag shapes, respectively:

```
> # model 2: quadratic decreasing lag shapes
> indus.code_2 <- list(
>   Job ~ 1,
>   Consum~qdec.lag(Job,0,15),
>   Pollution~qdec.lag(Job,0,15)+qdec.lag(Consum,0,15)
>   )
> indus.mod_2 <- dlsem(indus.code_2,group="Region",exogenous=c("Population","GDP"),
>   data=industry,global.control=indus.global,local.control=indus.local,log=T)
> summary(indus.mod_2)
```

```
A distributed-lag linear structural equation model
 Group factor: Region (10 groups)
 Exogenous variables: Population, GDP

Response: Job
-

Response: Consum
    a b     theta  se(theta)  t value     Pr(>|t|)
Job 0 5 0.1057272 0.02883474 3.666659 0.0003008825 ***

Response: Pollution
       a  b      theta  se(theta)  t value     Pr(>|t|)
Job    2 15 0.22363345 0.03182028 7.028016 7.426072e-11 ***
Consum 0  5 0.07433732 0.05778413 1.286466 2.003167e-01

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-squared: 0.8547,  AIC: -4278.3,  BIC: -4133.748

> # model 3: gamma lag shapes
> indus.code_3 <- list(
>   Job ~ 1,
>   Consum~gamma.lag(Job,0.5,0.5),
>   Pollution~gamma.lag(Job,0.5,0.5)+gamma.lag(Consum,0.5,0.5)
>   )
> indus.mod_3 <- dlsem(indus.code_3,group="Region",exogenous=c("Population","GDP"),
>   data=industry,global.control=indus.global,local.control=indus.local,log=T)
> summary(indus.mod_3)

A distributed-lag linear structural equation model
 Group factor: Region (10 groups)
 Exogenous variables: Population, GDP

Response: Job
-

Response: Consum
    a b    theta se(theta)  t value     Pr(>|t|)
Job 0 5 0.213074 0.0620565 3.433548 0.0006942689 ***

Response: Pollution
       a  b     theta  se(theta)   t value     Pr(>|t|)
Job    2 12 0.3322248 0.02637051 12.598346 1.030076e-26 ***
Consum 0  5 0.0931422 0.03770555  2.470251 1.440266e-02   *

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-squared: 0.8563,  AIC: -4620.154,  BIC: -4471.726
```

We see that the three models provide different results. The `BIC` method for class `dlsem` can be used to compare them according to the BIC:

```
> lapply(list(QUEC=indus.mod,QDEC=indus.mod_2,GAMMA=indus.mod_3),BIC)

$QUEC
      Job    Consum Pollution (overall)
-1733.060 -1553.811 -1349.505 -4636.377

$QDEC
      Job     Consum  Pollution  (overall)
```

```
-1733.060 -1536.0729  -864.6145 -4133.7476

$GAMMA
      Job    Consum Pollution (overall)
-1733.060 -1605.498 -1133.168 -4471.726
```

The model with endpoint-constrained quadratic lag shapes has the best fit according to the BIC. Note that the fit for variable `Job` is constant in each model because it has no endogenous covariates.

5. A real-world application

In this section, we illustrate the use of the **dlsem** package to address a real-world application aiming at assessing the impact of agricultural research expenditure on multiple dimensions in Europe. We refer to this problem as "agricultural research problem". The dataset `agres` is used in the illustration, which contains official data for EU 15 countries in the period 1980-2014:

```
> data(agres)
> names(agres)

 [1] "COUNTRY"        "YEAR"           "GDP"            "EMPL_AGR"
 [5] "UAA"            "PATENT_OTHER"   "GBAORD_AGR"     "BERD_AGR"
 [9] "RD_EDU_AGR"     "EU_PRO_AGR"     "PATENT_AGR"     "TFPC"
[13] "EPC"            "GVA_AGR"        "PPI_AGR"        "ENTR_INCOME_AGR"
[17] "ENERGY_RENEW"   "INCOME_RURAL"   "UNEMPL_RURAL"   "HEALTH_RURAL"
```

Dataset `agres` includes the code of the country (`COUNTRY`), the year (`YEAR`) and several other variables. We consider a subset of them in order to cover four interrelated causal levels:

- *context level*, including factors determining differences among countries and directly influencing the other levels;
- *investment level*, including factors contributing to fund agricultural research;
- *research level*, including indicators of the agricultural research activity;
- *impact level*: including variables affecting economic conditions of producers and consumers.

The considered variables for each level are:

- *context level*: gross domestic product (international dollars, label: `GDP`), persons employed in Agriculture (count, label: `EMPL_AGR`), utilized agricultural area (hectares, label: `UAA`), mechanical, chemical and environment-related patent applications (count, label: `PATENT_OTHER`);
- *investment level*: government research expenditure (million euro PPS, label: `GBAORD_AGR`), business enterprise research expenditure (million euro PPS, label: `BERD_AGR`);
- *research level*: agricultural researchers with tertiary education (count, label: `RD_EDU_AGR`), agricultural patent applications (count, label: `PATENT_AGR`);
- *impact level*: net entrepreneurial income of Agriculture (index 2005=100, label: `ENTR_INCOME_AGR`), producer price of agricultural output (index 2005=100, label: `PPI_AGR`).

For further details, see the documentation of the agres dataset by typing ?agres.

In order to highlight the opportunity to include qualitative exogenous variables, we consider an additional variable in the context level, that is a dummy indicating whether the Decoupling policy implemented in 2005 is in vigour or not. Such variable can be defined with the following code:

```
> agres$POLICY <- factor(1*(agres$YEAR>=2005))
> levels(agres$POLICY) <- c("no","yes")
```

It is important to indicate qualitative variables with numerical labels as factors, otherwise they will be interpreted as quantitative variables. With the following code we define the four causal levels and request the summary of the considered data:

```
> context.var <- c("GDP","EMPL_AGR","UAA","PATENT_OTHER","POLICY")
> investment.var <- c("GBAORD_AGR","BERD_AGR")
> research.var <- c("RD_EDU_AGR","PATENT_AGR")
> impact.var <-  c("ENTR_INCOME_AGR","PPI_AGR")
> all.var <- c(context.var,investment.var,research.var,impact.var)
> summary(agres[,c("COUNTRY",all.var)])
```

```
    COUNTRY          GDP              EMPL_AGR             UAA
 AT      : 25   Min.   :  83844   Min.   :  53000   Min.   : 1300000
 BL      : 25   1st Qu.: 217173   1st Qu.: 126375   1st Qu.: 2675000
 DE      : 25   Median : 356676   Median : 317250   Median : 4076000
 DK      : 25   Mean   : 877300   Mean   : 480384   Mean   : 9996738
 EL      : 25   3rd Qu.:1656223   3rd Qu.: 788050   3rd Qu.:16958000
 ES      : 25   Max.   :3161940   Max.   :1903150   Max.   :30593000
 (Other):200    NA's   :14                          NA's   :14
  PATENT_OTHER    POLICY      GBAORD_AGR        BERD_AGR         RD_EDU_AGR
 Min.   :   1.5   no :210   Min.   : 13.77   Min.   :  0.257   Min.   :  2367
 1st Qu.: 128.1   yes:140   1st Qu.: 39.91   1st Qu.:  1.692   1st Qu.:  5797
 Median : 370.9             Median : 81.97   Median : 12.243   Median : 12621
 Mean   : 908.9             Mean   :162.26   Mean   : 28.209   Mean   : 24137
 3rd Qu.: 975.0             3rd Qu.:274.44   3rd Qu.: 57.947   3rd Qu.: 32686
 Max.   :7692.9             Max.   :792.50   Max.   :160.057   Max.   :140091
 NA's   :14                 NA's   :12       NA's   :200       NA's   :111
   PATENT_AGR       ENTR_INCOME_AGR     PPI_AGR
 Min.   :  0.3333   Min.   : 43.00   Min.   : 60.36
 1st Qu.:  8.7641   1st Qu.: 91.05   1st Qu.: 97.78
 Median : 28.7500   Median :113.35   Median :102.77
 Mean   : 59.4942   Mean   :119.06   Mean   :107.44
 3rd Qu.: 73.5277   3rd Qu.:139.11   3rd Qu.:113.93
 Max.   :549.9620   Max.   :272.00   Max.   :191.60
 NA's   :21         NA's   :4        NA's   :23
```

In order to verify the presence of heterogeneity between groups (here, countries), we inspect the summary of government research expenditure, gross domestic product and utilized agricultural area for some countries:

```
> cou.ind <- which(agres[,"COUNTRY"] %in% c("AT","DK","FR","IT"))
> by(agres[cou.ind,c("GBAORD_AGR","GDP","UAA")],factor(agres[cou.ind,"COUNTRY"]),summary)
```

```
agres[cou.ind, "COUNTRY"]: AT
   GBAORD_AGR          GDP               UAA
 Min.   :27.77   Min.   :223054   Min.   :3154000
 1st Qu.:32.28   1st Qu.:253438   1st Qu.:3226000
 Median :34.36   Median :295257   Median :3368500
 Mean   :34.26   Mean   :291552   Mean   :3333208
 3rd Qu.:35.81   3rd Qu.:330722   3rd Qu.:3427500
 Max.   :41.37   Max.   :349515   Max.   :3519000  
                 NA's   :1        NA's   :1        
-------------------------------------------------------------
agres[cou.ind, "COUNTRY"]: DK
   GBAORD_AGR          GDP              UAA         
 Min.   : 39.80   Min.   :190030   Min.   :2609000  
 1st Qu.: 50.71   1st Qu.:217949   1st Qu.:2646750  
 Median : 58.96   Median :250066   Median :2674000  
 Mean   : 61.79   Mean   :241123   Mean   :2683875  
 3rd Qu.: 67.96   3rd Qu.:265359   3rd Qu.:2711500  
 Max.   :113.86   Max.   :276868   Max.   :2788000  
                  NA's   :1        NA's   :1        
-------------------------------------------------------------
agres[cou.ind, "COUNTRY"]: FR
   GBAORD_AGR         GDP               UAA          
 Min.   :198.9   Min.   :1650030   Min.   :28774000  
 1st Qu.:277.5   1st Qu.:1777188   1st Qu.:29227000  
 Median :316.7   Median :2081240   Median :29736000  
 Mean   :341.9   Mean   :2031360   Mean   :29684292  
 3rd Qu.:406.3   3rd Qu.:2264428   3rd Qu.:30109500  
 Max.   :541.9   Max.   :2351940   Max.   :30593000  
 NA's   :1       NA's   :1         NA's   :1         
-------------------------------------------------------------
agres[cou.ind, "COUNTRY"]: IT
   GBAORD_AGR         GDP               UAA          
 Min.   :119.9   Min.   :1501700   Min.   :13630000  
 1st Qu.:155.9   1st Qu.:1617625   1st Qu.:14296750  
 Median :223.8   Median :1781445   Median :15303000  
 Mean   :241.4   Mean   :1728950   Mean   :15044833  
 3rd Qu.:312.4   3rd Qu.:1827698   3rd Qu.:15653250  
 Max.   :436.4   Max.   :1918530   Max.   :16840000  
 NA's   :3       NA's   :1         NA's   :1         
```

From these summaries it seems that propensity to research investments is heterogeneous across countries: those with higher gross domestic product and utilized agricultural area tend to expend more for agricultural research.

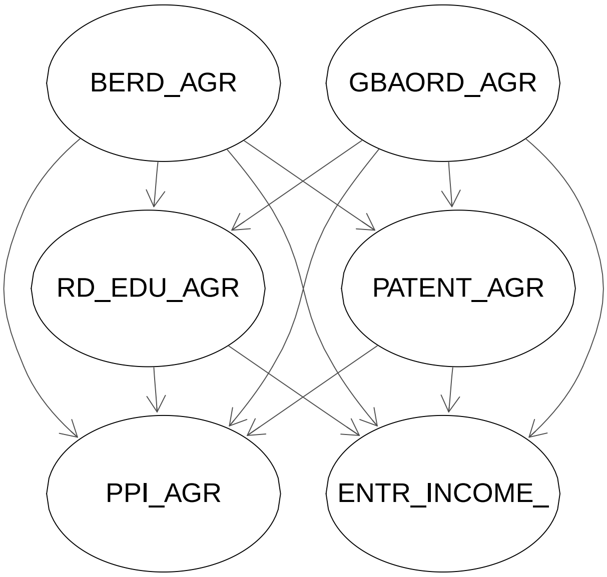

Figure 5: The static representation of the DAG for the agricultural research problem.

We hypothesize the DAG for the agricultural research problem on the basis of the assumption of conditional independence within the same level, and on the following causal order: investment level, research level, impact level (Figure 5). Also, we consider the variables in the context level as exogenous variables, and assume endpoint-constrained quadratic lag shapes for all covariates. On these grounds, the model code can be defined as follows:

```
> auxcode <- c(
+   paste(investment.var,"~1",sep=""),
+   paste(research.var,"~",paste("quec.lag(",investment.var,",0,20)",
+     collapse="+",sep=""),sep=""),
+   paste(impact.var,"~",paste("quec.lag(",c(investment.var,research.var),",0,20)",
+     collapse="+",sep=""),sep="")
+   )
> agres.code <- list()
> for(i in 1:length(auxcode)) {
+   agres.code[[i]] <- formula(auxcode[i])
+   }
```

Constraints for the adaptation of lag shapes are the following: (i) 3 years of maximum gestation lag for all lag shapes; (ii) 5 years of minimum lag width for all lag shapes; (iii) 15 years of maximum lag lead for the lag shapes in the regression model of variables in the research level; (iv) 20 years of maximum lag lead for the lag shapes in the regression model of variables in the impact level; (v) negative lag sign for the lag shapes in the regression model of producer price of agricultural output (PPI_AGR), and positive lag sign for all other lag shapes. The last constraint is motivated by the fact that lower values of producer price typically indicate an improvement of economic conditions. The following code defines these constraints efficiently:

```
> agres.global <- list(adapt=T,max.gestation=3,min.width=5,max.lead=20,sign="+")
> auxcon1 <- rep(15,length(investment.var))
> names(auxcon1) <- investment.var
> auxcon2 <- rep("-",length(investment.var)+length(research.var))
> names(auxcon2) <- c(investment.var,research.var)
> agres.local <- list(max.lead=list(RD_EDU_AGR=auxcon1,PATENT_AGR=auxcon1),
+   sign=list(PPI_AGR=auxcon2))
```

Finally, we request the logarithmic transformation for all quantitative variables in order to interpret coefficients as elasticities. The code performing parameter estimation is the following:

```
> agres.mod <- dlsem(agres.code,group="COUNTRY",exogenous=context.var,data=agres,
+   global.control=agres.global,local.control=agres.local,log=T)
```

```
Logarithm not applied to variables: POLICY
Checking stationarity...
 Order 1 differentiation performed
EM converged after 32 iterations. Log-likelihood: 2372.762
Estimating regression model 1/6 (GBAORD_AGR)
Estimating regression model 2/6 (BERD_AGR)
Estimating regression model 3/6 (RD_EDU_AGR)
Estimating regression model 4/6 (PATENT_AGR)
Estimating regression model 5/6 (ENTR_INCOME_AGR)
Estimating regression model 6/6 (PPI_AGR)
Estimation completed
```

Parameter estimation takes a couple of minutes. Messages inform that logarithmic transformation was not applied to POLICY as it is a qualitative variable, and that differentiation was applied one time in order to achieve stationarity of all time series. Also, the Expectation-Maximization algorithm was run, as data contain missing values.

The summary of parameter estimation can be obtained using the summary method. Here, we display the summary for the endogenous part only by accessing to the component endogenous:

```
> summary(agres.mod)$endogenous

Response: GBAORD_AGR
-

Response: BERD_AGR
-

Response: RD_EDU_AGR
           a  b     theta  se(theta)   t value    Pr(>|t|)
GBAORD_AGR 3 15 0.1462298 0.15397605 0.9496919 0.344429088
BERD_AGR   0  6 0.1402433 0.05310359 2.6409389 0.009517316 **

Response: PATENT_AGR
           a  b      theta se(theta)    t value Pr(>|t|)
GBAORD_AGR 0 15  0.2409959 0.1997313  1.2066005 0.230272
BERD_AGR   3 15 -0.0810642 0.1198558 -0.6763478 0.500293

Response: ENTR_INCOME_AGR
           a  b      theta se(theta)   t value    Pr(>|t|)
GBAORD_AGR 0  5 0.02623935 0.2183972 0.1201451 0.905075394
BERD_AGR   1  7 0.02470520 0.1197374 0.2063281 0.837765194
RD_EDU_AGR 2 20 0.83906765 0.5254023 1.5970003 0.119519563
PATENT_AGR 0 12 1.27776437 0.4259429 2.9998491 0.005027301 **

Response: PPI_AGR
           a  b       theta  se(theta)   t value     Pr(>|t|)
GBAORD_AGR 1 12 -0.06422218 0.05124818 -1.253160 0.2186972630
BERD_AGR   3 20 -0.14556419 0.03645437 -3.993052 0.0003303346 ***
RD_EDU_AGR 0  5 -0.05753030 0.01772700 -3.245349 0.0026353764  **
PATENT_AGR 1  6 -0.09521903 0.04245053 -2.243058 0.0315197334   *
```

Instead, goodness of fit statistics can be displayed by accessing to component gof:

```
> summary(agres.mod)$gof

       Rsq        AIC        BIC
  0.205822  28.395596 445.977607
```

The static representation of the DAG with coloured edges is shown in Figure 6:

```
> plot(agres.mod)
```

Results show that business enterprise research expenditure (BERD_AGR) influences producer price (PPI_AGR) both directly from 3 to 20 years, and indirectly through the number of researchers with tertiary education (RD_EDU_AGR) from 0 to 5 years. Producer price (PPI_AGR) is also influenced by the number of patent applications (PATENT_AGR) from 1 to 6 years, independently of business enterprise research expenditure (BERD_AGR), which also influences entrepreneurial income (ENTR_INCOME_AGR) from 0 to 12 years.

The pathwise causal lag shapes and the overall one connecting BERD_AGR and PPI_AGR can be displayed by means of the following code (results shown in Figure 7):

```
> lagPlot(agres.mod,path="BERD_AGR*PPI_AGR")
> lagPlot(agres.mod,path="BERD_AGR*RD_EDU_AGR*PPI_AGR")
> lagPlot(agres.mod,from="BERD_AGR",to="PPI_AGR")
```

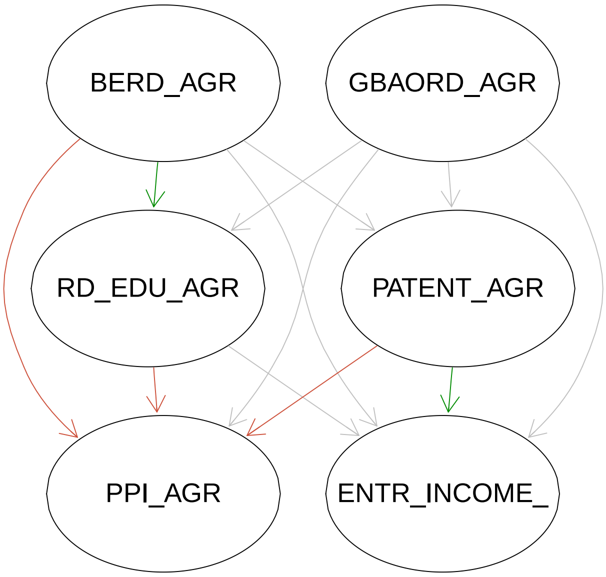

Figure 6: The static representation of the DAG for the agricultural research problem (endpoint-constrained lag shapes) where each edge is coloured with respect to the sign of the estimated causal effect. Green, red and light grey indicate positive, negative and not statistically significant causal effects, respectively.

We see that the path connecting `BERD_AGR` to `PPI_AGR` passing through `RD_EDU_AGR` is composed by two edges with causal effects of different signs. However, the absolute value of the negative one is greater, thus the causal effect associated to the path results negative (upper right panel of Figure 7). As a consequence, the cumulative overall causal effect is monotonically negative (lower panel of Figure 7). We can conclude that for a 1% of business enterprise research expenditure more, producer price is expected to increase by 2.03% after 20 years:

```
> causalEff(agres.mod,from="BERD_AGR",to="PPI_AGR",lag=20,cumul=T)$overall

    estimate lower 95% upper 95%
20 -2.032343 -2.277942 -1.786745
```

As a further step, we fit a second model using gamma lag shapes instead of endpoint-constrained quadratic ones:

```
> auxcode_2 <- c(paste(investment.var,"~1",sep=""),
+   paste(research.var,"~",paste("gamma.lag(",investment.var,",0.5,0.5)",
+     collapse="+",sep=""),sep=""),
+   paste(impact.var,"~",paste("gamma.lag(",c(investment.var,research.var),",0.5,0.5)",
+     collapse="+",sep=""),sep=""))
> agres.code_2 <- list()
> for(i in 1:length(auxcode_2)) {
+   agres.code_2[[i]] <- formula(auxcode_2[i])
+   }
> agres.mod_2 <- dlsem(agres.code_2,group="COUNTRY",exogenous=context.var,data=agres,
+   global.control=agres.global,local.control=agres.local,log=T)

Logarithm not applied to variables: POLICY
Checking stationarity...
 Order 1 differentiation performed
EM converged after 32 iterations. Log-likelihood: 2372.762
Estimating regression model 1/6 (GBAORD_AGR)
Estimating regression model 2/6 (BERD_AGR)
```

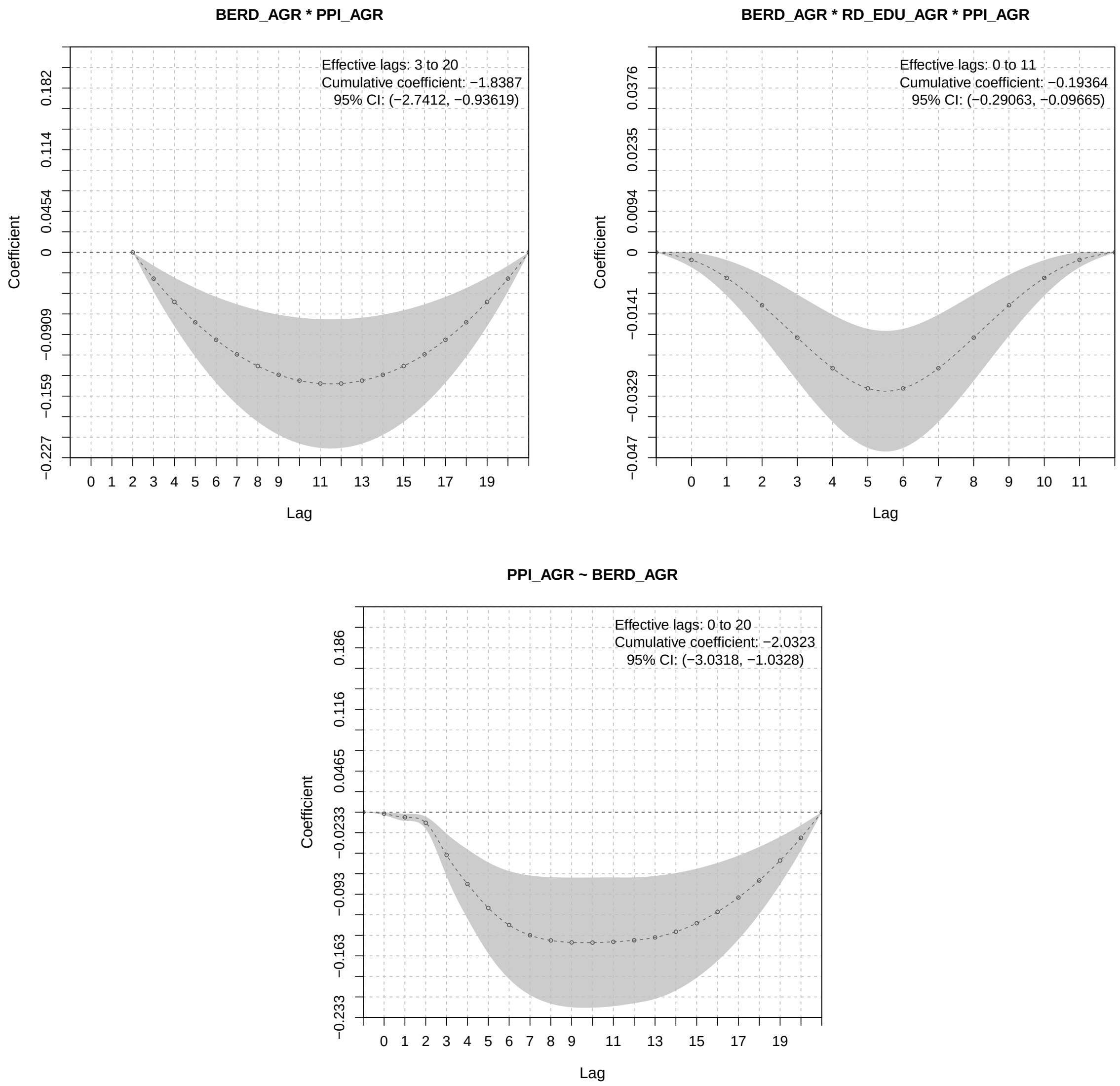

Figure 7: The pathwise causal lag shapes (upper panels) and the overall one (lower panel) connecting `BERD_AGR` to `PPI_AGR`. 95% asymptotic confidence intervals are shown in grey.

```
Estimating regression model 3/6 (RD_EDU_AGR)
Estimating regression model 4/6 (PATENT_AGR)
Estimating regression model 5/6 (ENTR_INCOME_AGR)
Estimating regression model 6/6 (PPI_AGR)
Estimation completed
```

Parameter estimation with gamma lag shapes is longer and in this case takes a dozen of minutes. Imputation and differentiation is obviously unchanged. Results with gamma lag shapes are quite different from those with endpoint-constrained quadratic ones:

```
> summary(agres.mod_2)$endogenous

Response: GBAORD_AGR
-

Response: BERD_AGR
-

Response: RD_EDU_AGR
           a b      theta se(theta)   t value     Pr(>|t|)
```

```
GBAORD_AGR 0 5 0.03513002 0.1543571 0.2275892 8.201452e-01
BERD_AGR   0 5 0.78183701 0.1185141 6.5969981 2.350165e-10 ***

Response: PATENT_AGR
           a  b      theta se(theta)   t value   Pr(>|t|)
GBAORD_AGR 0 15 0.31217999 0.1802083 1.7323291 0.08578436 .
BERD_AGR   1  6 0.02856498 0.1018026 0.2805917 0.77950677

Response: ENTR_INCOME_AGR
           a  b       theta se(theta)    t value   Pr(>|t|)
GBAORD_AGR 0  5  0.09161682 0.3597409  0.2546745 0.80006248
BERD_AGR   0 18 -0.33275458 0.1326417 -2.5086732 0.01554989 *
RD_EDU_AGR 2 15  0.53828995 0.2395878  2.2467339 0.02929268 *
PATENT_AGR 0 20  0.36390869 0.1558521  2.3349617 0.02377460 *

Response: PPI_AGR
           a  b        theta  se(theta)    t value   Pr(>|t|)
GBAORD_AGR 0 12 -0.049777487 0.02186578 -2.2765021 0.02461862  *
BERD_AGR   0  5  0.061170793 0.03929708  1.5566243 0.12223777
RD_EDU_AGR 0 15 -0.066418855 0.02109762 -3.1481683 0.00208097 **
PATENT_AGR 1  6 -0.006869514 0.02424436 -0.2833448 0.77740903
```

The static representation of the DAG with coloured edges is shown in Figure 8:

```
> plot(agres.mod)
```

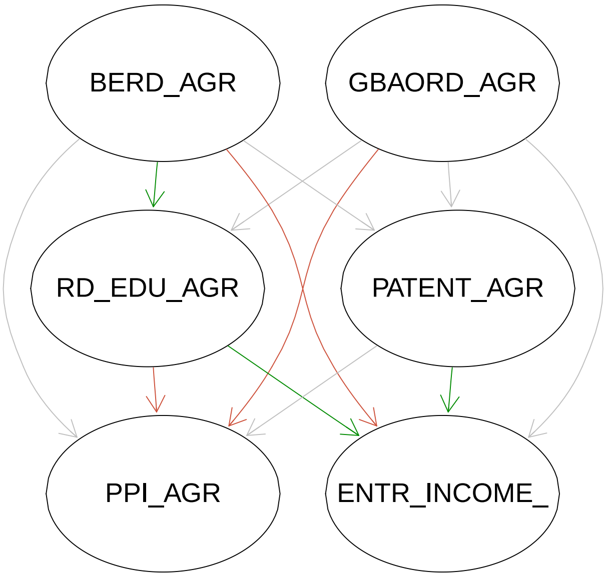

Figure 8: The static representation of the DAG for the agricultural research problem (gamma lag shapes) where each edge is coloured with respect to the sign of the estimated causal effect. Green, red and light grey indicate positive, negative and not statistically significant causal effects, respectively.

According to the BIC, the model with gamma lag shapes should be preferred:

```
> lapply(list(QUEC=agres.mod,GAMMA=agres.mod_2),BIC)

$QUEC
     GBAORD_AGR        BERD_AGR      RD_EDU_AGR      PATENT_AGR ENTR_INCOME_AGR
       27.98495       178.69873       185.28630       248.85848        70.89533
```

```
        PPI_AGR       (overall)
     -135.99300       575.73078

$GAMMA
     GBAORD_AGR        BERD_AGR      RD_EDU_AGR      PATENT_AGR ENTR_INCOME_AGR
       27.98495       178.69873       160.36565       285.71881        84.81348
        PPI_AGR       (overall)
     -270.89223       466.68938
```

6. Conclusions and future development

Distributed-lag linear structural equation models (DLSEMs) allow to study the dynamic causal path generated by an external impulse into a multi-dimensional system, thanks to a variety of lag shapes characterized by flexibility and simplicity of estimation. In the real-world application illustrated in this paper, DLSEMs are applied to an impact assessment problem in the field of agricultural economics, but they appear as a suitable solution in several other scientific domains where the effect of an external impulse is studied through time. For instance, epidemiologic problems may be addressed, like the analysis of the dynamic effect of pollutants on human health (Martins, Pereira, Lin, Santos, Prioli, do Carmo Luiz, Saldiva, and Ferreira Braga 2006; Steinvil, Fireman, Kordova-Biezuner, Cohen, Shapira, Berliner, and Rogowski 2009). Also, viticultural-oenological domains can be dynamically characterized by focusing, for example, on grape maturation (Magrini, Di Blasi, and Stefanini 2017) and vinification (Stefanini and Pantani 2013; Magrini, Pantani, Bartolini, and Stefanini 2016). Lag shapes included in the package may represent a large number of real-world lag structures: unimodal symmetric (with the endpoint-constrained quadratic lag shape), unimodal asymmetric (with the gamma lag shape) and skewned ones (with the quadratic decreasing lag shape). Nevertheless, additional lag shapes with further specific features may be added in future.

Parameter estimation in DLSEM cannot be performed in a single step unless gestation and lead lags are all known. Since the number of possible models rises exponentially as the number of covariates and time lags increases, complete search is infeasible for most real-world applications, thus a heuristic search was implemented. Further development of the package may be directed towards the improvement of the search strategy.

The main limitation of DLSEMs relies in the sample size to achieve efficient estimates. Likewise all lag models proposed in the literature, DLSEMs require to drop away a number of statistical units in order to estimate as many regression coefficients at different time lags. Thus, long time series are typically required to efficiently estimate wide lag structures. An extensive simulation study to assess the potential of the method as a function of the sample size is a natural continuation of the research on this topic.

DLSEMs are a special case of dynamic Bayesian networks (Murphy 2002), where endogenous variables follow the Gaussian distribution and lag shapes are possibly constrained to predefined functional forms. An important functionality of DLSEMs implemented in the `dlsem` package is the opportunity to condition on exogenous variables of any type, either categorical or continuous.

The current implementation of the package deals with grouped data through fixed effects estimation. Feature releases may include random effects estimation to enhance inference whenever the considered groups are a subset of the possible ones, or covariates with values constant within groups (second-level covariates) are involved.

Acknowledgements

This research was partially supported by the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement n°609448 (IMPRESA project, `http://www.impresa-project.eu`), and by the University of Florence (Florence, Italy) under 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 valuable comments that helped improving the presentation of our work. The authors declare that there are no conflicts of interest.

References

Akaike H (1974). "A New Look at the Statistical Identification Model." *IEEE Transactions on Automatic Control*, **19**, 716–723. `doi:10.1109/TAC.1974.1100705`.

Almon S (1965). "The Distributed Lag between Capital Appropriations and Net Expenditures." *Econometrica*, **33**, 178–196. `doi:10.2307/1911894`.

Andreou E, Ghysels E, Kourtellos A (2007). "Regression Models with Mixed Sampling Frequencies." *Journal of Econometrics*, **158**, 246–261. `doi:10.1016/j.jeconom.2010.01.004`.

Dempster AP, Laird NM, Rubin DB (1977). "Maximum Likelihood from Incomplete Data via the EM Algorithm." *Journal of the Royal Statistical Society, Series B*, **39**(1), 1–38. `doi:10.1.1.133.4884`.

Dickey DA, Fuller WA (1981). "Likelihood Ratio Statistics for Autoregressive Time Series with a Unit Root." *Econometrica*, **49**, 1057–1072. `doi:10.2307/1912517`.

Gasparrini A (2011). "Distributed Lag Linear and Non-Linear Models in R: The Package `dnlm`." *Journal of Statistical Software*, **43**(8), 1–20. `doi:10.18637/jss.v043.i08`.

Ghysels E, Kvedaras V, Zemlys V (2016). "Mixed Frequency Data Sampling Regression Models: The R Package `midasr`." *Journal of Statistical Software*, **72**(4), 1–35. `doi:10.18637/jss.v072.i04`.

Ghysels E, Sinko A, Valkanov R (2007). "MIDAS Regressions: Further Results and New Directions." *Econometric Reviews*, **26**(1), 53–90. `doi:10.1080/07474930600972467`.

Granger CWJ, Newbold P (1974). "Spurious Regressions in Econometrics." *Journal of Econometrics*, **2**(2), 111–120. `doi:10.1016/0304-4076(74)90034-7`.

Haavelmo T (1943). "The Statistical Implications of a System of Simultaneous Equations." *Econometrica*, **11**(1), 1–12. `doi:0012-9682(194301)11:1`.

Judge GG, Griffiths WE, Hill RC, Lutkepohl H, Lee TC (1985). *The Theory and Practice of Econometrics.* 2nd edition. John Wiley & Sons, New York, US-NY.

Koopmans TC, Rubin H, Leipnik RB (1950). "Measuring the Equation Systems of Dynamic Economics." In TC Koopmans (ed.), *Statistical Inference in Dynamic Economic Models*, pp. 53–237. John Wiley & Sons, New York, US-NY.

Levin A, Lin C, Chub CJ (2002). "Unit Root Tests in Panel Data: Asymptotic and Finite-Sample Properties." *Journal of Econometrics*, **108**, 1–24. `doi:10.1016/S0304-4076(01)00098-7`.

Magrini A (2018). "Linear Markovian Models for Lag Exposure Assessment." *Biometrical Letters*, **55**(2), 179–195. doi:10.2478/bile-2018-0012.

Magrini A, Di Blasi S, Stefanini FM (2017). "A Conditional Linear Gaussian Network to Assess the Impact of Several Agronomic Settings on the Quality of Tuscan Sangiovese Grapes." *Biometrical Letters*, **54**(1), 25–42. doi:10.1515/bile-2017-0002.

Magrini A, Pantani OL, Bartolini AB, Stefanini FM (2016). "On Prefermentative Maceration Techniques: Statistical Analysis of Sensory Descriptors in Sangiovese Wine." *Biometrical Letters*, **53**(1), 1–20. doi:10.1515/bile-2016-0001.

Martins LC, Pereira LAA, Lin CA, Santos UP, Prioli G, do Carmo Luiz O, Saldiva PHN, Ferreira Braga AL (2006). "The Effects of Air Pollution on Cardiovascular Diseases: Lag Structures." *Revista de Saúde Pública*, **40**(4). doi:10.1590/S0034-89102006000500018.

Murphy K (2002). "Dynamic Bayesian Networks: Representation, Inference and Learning." PhD Thesis, Computer Science Division, UC Berkeley, US-CA.

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

Schwarz G (1978). "Estimating the Dimension of a Model." *Annals of Statistics*, **6**, 461–464. doi:10.1214/aos/1176344136.

Stefanini FM, Pantani OL (2013). "A Bayesian Model to Compare Vinification Procedures." *Biometrical Letters*, **50**(2), 61–80. doi:10.2478/bile-2013-0018.

Steinvil A, Fireman E, Kordova-Biezuner L, Cohen M, Shapira I, Berliner S, Rogowski O (2009). "Environmental Air Pollution Has Decremental Effects on Pulmonary Function Test Parameters up to One Week after Exposure." *American Journal of Medical Science*, **338**(4), 273–279. doi:10.1097/MAJ.0b013e3181adb3ed.

Wright S (1934). "The Method of Path Coefficients." *Annals of Mathematical Statistics*, **5**(3), 161–215. doi:10.1214/aoms/1177732676.

Affiliation:

Alessandro Magrini
Department of Statistics, Computer Science, Applications
University of Florence, Italy
E-mail: alessandro.magrini@unifi.it
ORCID: orcid.org/0000-0002-7278-5332

Austrian Journal of Statistics
published by the Austrian Society of Statistics
Volume 48
January 2019

http://www.ajs.or.at/
http://www.osg.or.at/
Submitted: 2018-01-03
Accepted: 2018-06-23