Distributed-lag linear structural equation models in R: the `dlsem` package

Alessandro Magrini
Dep. Statistics, Computer Science, Applications
University of Florence, Italy
<magrini@disia.unifi.it>

`dlsem` version 2.4.6 – 23 March 2020

Contents

1 Introduction

Structural causal models (SCMs, Pearl, 2000) are a mathematical framework describing the behaviour of a multivariate system, and represent one of the prevalent methodologies for causal inference in contemporary applied sciences. Markovian SCMs are a special case where the joint probability distribution of the considered variables can be factored according to a directed acyclic graph. Distributed-lag linear structural equation models (DLSEMs) are Markovian SCMs, where each factor of the joint probability distribution is a distributed-lag linear regression with constrained lag shapes. They were firstly introduced in the context of lag exposure assessment (Magrini, 2018), then applied to impact assessment of research expenditure in Agriculture (Magrini *et al.*, 2019). DLSEMs account for temporal delays in the dependence relationships among the variables through a single parameter per covariate, thus allowing to perform dynamic causal inference in a feasible fashion.

Package `dlsem` implements inference functionalities for DLSEMs with several types of constrained lag shapes (Magrini , 2020). Currently, endpoint-constrained quadratic ('ecq'), quadratic decreasing ('qd'), linearly decreasing ('ld') and gamma ('gam') lag shapes are available.

This vignette is structured as follows. In Section 2, theory on the DLSEM is presented. In Section 3, instructions for the installation of the `dlsem` package are provided. In Section 4, the practical use of `dlsem` is illustrated through a simple impact assessment problem.

2 Theory

Distributed-lag linear regression Let Y be a response variable and $X_1, \ldots, X_p$ be the covariates, with y_t and x_{jt}, respectively, the value of Y and of X_j at time t. Under the hypothesis that time is discrete and that both Y and $X_1, \ldots, X_p$ are stationary time series, 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:

$$y_t = \beta_0 + \sum_{j=1}^{J} \sum_{l=0}^{L_j} \beta_{j,l} \; x_{j,t-l} + \epsilon_t \tag{1}$$

where $x_{j,t-l}$ is the value of the j-th covariate at l time lags before t, and ϵ_t is the random error at time t uncorrelated with the covariates and with ϵ_k, $\forall k \neq t$. 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.

Least squares can be used to consistently estimate the lag shapes [1], but, since time series data are likely to be serially correlated and heteroskedastic, a good practice is to apply the Heteroskedasticity and Autocorrelation Consistent (HAC) correction for the covariance matrix of least squares estimators (Newey & West, 1978).

The model in Formula 1 has the disadvantage that a parameter is required for each lagged instance of a covariate, and lagged instances of the same covariate tend to be highly correlated. The Almon's polynomial lag shape (Almon, 1965) overcomes these limitations by forcing the coefficients for lagged instances of the same covariate to follow a polynomial of order Q:

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

For instance, for $q = 2$ we have that $\beta_{j,l} = \phi_{j,0} + \phi_{j,1} l + \phi_{j,2} l^2$. Unfortunately, the Almon's polynomial lag shape may show multiple modes and coefficients with different signs, thus entailing problems of interpretation. Type II constrained lag shapes (Magrini , 2020) overcome this issue. They include 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:

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

the *linearly decreasing* lag shape:

$$\beta_{j,l} = \begin{cases} \theta_j \frac{b_j + 1 - l}{b_j + 1 - a_j} & a_j \leq l \leq b_j \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

and the *gamma* lag shape:

$$\beta_{j,l} = \theta_j (l+1)^{\frac{a_j}{1-a_j}} b_j^l \left[\left(\frac{a_j}{(a_j - 1) \log b_j} \right)^{\frac{a_j}{1-a_j}} b_j^{\frac{a_j}{(a_j-1)\log b_j} - 1} \right]^{-1} \tag{6}$$
$$0 < a_j < 1 \qquad 0 < b_j < 1$$

The endpoint-constrained quadratic lag shape is zero for a time lag $l < a_j$ or $l > b_j$, 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

[1] Note that stationarity of the time series is required to obtain consistent least squares estimation (Granger and Newbold, 1974). If some of them contains a stochastic trend (unit root), a reasonable procedure is to sequentially apply differencing to all variables until they are made stationary according to a unit root test. Sequential testing is required for interpretation, as it allows to have the same order of differencing for all variables.

to value 0 at lag $b_j + 1$ according to a quadratic function. The linearly decreasing lag shape is a linear version of the quadratic one. The gamma lag shape is positively skewed with mode equal to θ_j at lag $\frac{a_j}{(a_j-1)\log b_j}$.

For the endpoint-constrained quadratic, quadratic decreasing and linearly decreasing lag shapes, a_j represents the *gestation lag*, b_j the *lead lag*, and $b_j - a_j$ the *lag width* (a static lag shape is obtained if $a_j = b_j = 0$). Gestation lag, lead lag and lag width are not explicit in a gamma lag shape, but they can be approximated numerically from parameters a_j and b_j. For these constrained lag shapes, it holds:

$$\begin{array}{l} \beta_{j,l} > 0 \Longleftrightarrow \theta_j > 0 \\ \beta_{j,l} < 0 \Longleftrightarrow \theta_j < 0 \end{array} \qquad \forall l:\ a_j \leq l \leq b_j \tag{7}$$

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

A linear regression with these 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 fit several regressions with different values of $a_1, \ldots, a_J, b_1, \ldots, b_J$, and select the one with the minimum residual sum of squares (see Magrini , 2020 for details).

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 *et al.*, 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:

$$p(v_1, \ldots, v_J) = \prod_{j=1}^{J} p(v_j \mid \Pi_j = \pi_j) \tag{8}$$

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) 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:

$$p(v_1, v_2, v_3, v_4) = p(v_1)\ p(v_2 \mid v_1)\ p(v_3 \mid v_1)\ p(v_4 \mid v_2, v_3) \tag{9}$$

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

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:

$$p(v_1, \ldots, v_J \mid \mathrm{do}(V_i = v_i)) = \prod_{j \neq i} p(v_j \mid \pi_j)\ |_{V_i = v_i} \tag{10}$$

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 8, that is to predict such effect from non-experimental (observational) data. In a

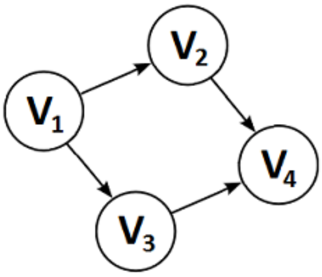

Figure 1: An example of directed acyclic graph.

Markovian SCM, the effect of $\text{do}(V_i = v_i)$ 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):

$$p(V_j = v_j \mid \text{do}(V_i = v_i)) = \sum_{\pi_i} p(V_j = v_j \mid V_i = v_i, \Pi_i = \pi_i) p(\Pi_i = \pi_i) \tag{11}$$

where Π_i is the parent set of V_i.

In a linear parametric formulation of SCMs (linear Markovian SCMs), each factor $p(v_j \mid \pi_j)$ of the joint probability distribution in Formula 8 is the linear regression 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 where V_4 is the response variable and V_2 and V_3 are the covariates.

As shown by Magrini, 2018, the computation of causal effects in a linear Markovian SCM involves the coefficients of the regressions only, without the need of Formula 11. Let $\text{do}(aV_i = 1)$ be an intervention changing the value of V_i by a unit. Under such intervention:

- if V_i is parent of V_j, the *direct* causal effect of V_i on V_j is equal to the coefficient of V_i in the regression of V_j;
- the causal effect of V_i on V_j through a multi-edge directed path $< V_i, \ldots, V_j >$ connecting V_i to V_j, called *indirect* causal effect of V_i on V_j through $< V_i, \ldots, V_j >$, is equal to (see, for example, Wright, 1934):

$$e(< V_i, \ldots, V_j >) = \prod_{k:\ V_k \in < V_i, \ldots, V_j > \wedge k \neq i} \beta_{k|k-1} \tag{12}$$

  where $\beta_{k|k-1}$ is the coefficient of V_{k-1} in the regression of V_k. The direct causal effect of V_i on V_j and all the indirect causal effects of V_i on V_j are denoted as *pathwise* causal effects of V_i on V_j;
- the *overall* causal effect of V_i on V_j is equal to the sum of all the pathwise causal effects of V_i on V_j.

For example, in the linear Markovian SCM associated to the DAG in Figure 1, there are two directed paths connecting V_1 to V_4: $< V_1, V_2, V_4 >$ with pathwise causal effect $\beta_{2|1} \cdot \beta_{4|2}$, and $< V_1, V_3, V_4 >$ with pathwise causal effect $\beta_{3|1} \cdot \beta_{4|3}$. Thus, the overall causal effect of V_1 on V_4 is equal to $\beta_{2|1} \cdot \beta_{4|2} + \beta_{3|1} \cdot \beta_{4|3}$.

Distributed-lag linear structural equation models Distributed-lag linear structural equation models (DLSEMs) are Markovian SCMs where each factor of the joint probability distribution in Formula 8 is a distributed-lag linear regression with constrained lag shapes. They were firstly introduced in the context of lag exposure assessment (Magrini, 2018), then applied to impact assessment of research expenditure in Agriculture (Magrini *et al.*, 2019). The DAG of a DLSEM would involve all the possible temporal instances of each variable in $\boldsymbol{V}$. Here, for simplicity, a static DAG is still used for a DLSEM, where the 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 of variable V_j is non-zero. Causal effects at different time lags in a DLSEM are defined as follows:

- if V_i is parent of V_j, the *direct* causal effect of V_i on V_j at lag l is equal to the coefficient of V_i at lag l in the regression of V_j;
- 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 $Q_m^{(l)}$ be the set of all the possible ordered m-uples of time lags such that their sum is equal to l. The *indirect* causal effect of V_i on V_j through such path at lag l is equal to:

$$e(< V_{d_0}, \ldots, V_{d_m} >; d_0 = i, d_m = j) = \sum_{(q_1,\ldots,q_m)\in Q_m^{(l)}} \prod_{k=1}^{m} b_{d_k|d_{k-1},q_k} \tag{13}$$

  where $b_{d_k|d_{k-1},q_k}$ is the coefficient of $V_{d_{k-1}}$ at lag q_k in the regression of V_{d_k};
- the *overall* causal effect of V_i on V_j at lag l is equal to the sum of all the pathwise causal effects of V_i on V_j at lag 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.

3 Installation

Before installing `dlsem`, you must have installed R version 3.5.0 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")
```

The latest version of `dlsem` is 2.4.6.

4 Illustrative example

The practical use of package `dlsem` is illustrated through a simple impact assessment problem denoted as "industrial development problem". The objective is to test whether the influence through time of the number job positions in industry (proxy of the industrial development) on the amount of greenhouse gas emissions (proxy of pollution) is direct and/or mediated by the amount of private consumption. 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.

```
> 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
```

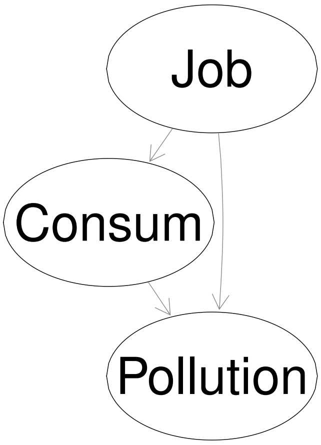

Figure 2: The DAG for the industrial development problem. ‘Job’: number of job positions in industry. ‘Consum’: private consumption index. ‘Pollution’: amount of greenhouse gas emissions.

```
 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  
```

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 regressions. The variables for which a regression is specified are called *endogenous* variables. The other variables are referred as *exogenous* variables (not to be confused with the unobserved disturbances).

The model code must be a list of formulas, one for each regression. In each formula, the response and the covariates must be quantitative variables[2], and operators `ecq`(·), `qd`(·), `ld`(·) and `gam`(·)[3] may be employed to specify, respectively, an endpoint-constrained quadratic, a quadratic decreasing, a linearly decreasing or a gamma lag shape. Operators `ecq`(·), `qd`(·), `ld`(·) and `gam`(·) have three arguments: the name of the covariate to which the lag shape is applied, and the two shape parameters a_j and b_j (see Magrini , 2020 for details).

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 specification of regressions with no endogenous covariates may be omitted from the model code (for example, one could avoid to specify the regression for the number of job positions). In this problem, all lag shapes are assumed to be endpoint-constrained quadratic lag shapes between 0 and 15 time lags:

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

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

[3] The operators `ecq`(·), `qd`(·) and `gam`(·) replace the old operators `quec.lag`(·), `qdec.lag`(·) and `gamma.lag`(·). If an old operator is employed, it is automatically replaced by the new one and a warning is returned.

4.2 Specification of control options

The second step to build a DLSEM with the `dlsem` package is the specification of control options. Control options are distinguished into global (applied to all the regressions) and local (regression-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;
- `min.gestation`: the minimum gestation lag for all lag shapes. If not provided, it is taken as equal to 0;
- `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 may be either `'+'` for positive or `'-'` for negative. If not provided, adaptation will disregard the lag sign.

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 of that variable;
- `min.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 of the endogenous variable above and include the minimum gestation lag for its lag shape;
- `max.gestation`: the same as `min.gestation`, with the exception that the named vector must include the maximum gestation lag;
- `max.lead`: the same as `min.gestation`, with the exception that the named vector must include the maximum lead lag;
- `min.width`: the same as `min.gestation`, with the exception that the named vector must include the minimum lag width;
- `sign`: the same as `min.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 may be expressed in several ways. The most simple solution is to specify only global control options, as the constraints hold for all the regressions:

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

In alternative, one may specify only local control options, by repeating them for each regression:

```
> 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 user may indicate a single group factor (just one) to argument `group` and one or more exogenous variables to argument `exogenous`. By indicating the group factor, one intercept for each level of the group factor will be estimated in each regression, in order to explain the variability due to differences between groups. By indicating exogenous variables, they will be included as non-lagged covariates in each regression, in order to eliminate cross-sectional spurious effects. Each exogenous variable may be either qualitative or quantitative and its coefficient in each regression is 0 for time lags greater than 0 (no lag shape). The user may decide to apply the logarithmic transformation to all strictly positive quantitative variables by setting argument `log` to `TRUE`, in order to interpret each coefficient as an elasticity (percentage increase in the value of the response variable for 1% increase in the value of a covariate). Before parameter estimation, differencing is sequentially applied until all the time series are made stationary. By default, the Kwiatkowski-Phillips-Schmidt-Shin (KPSS, Kwiatkowski *et al.*, 1950) test is performed if the number of periods is less than 100, otherwise the Augmented Dickey-Fuller (ADF, Dickey and Fuller, 1981) test is used. [4], and each missing value is replaced by its conditional mean computed through the Expectation-Maximization algorithm (Dempster *et al.*, 1977)[5]. The HAC correction of the covariance matrix of least squares estimators is applied by default (it can be disabled by setting argument `hac` to `FALSE`). In this problem, the region is indicated as the group factor, while population and gross domestic product are indicated as exogenous variables. Also, the logarithmic transformation is requested, and global and local control options are provided to arguments `global.control` and `local.control`,respectively:

```
> 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 1/3 (Job)
Estimating regression 2/3 (Consum) ... 5%
Estimating regression 2/3 (Consum) ... 45%
Estimating regression 2/3 (Consum) ... 50%
Estimating regression 2/3 (Consum) ... 55%
Estimating regression 2/3 (Consum) ... 95%
Estimating regression 2/3 (Consum) ... 100%
```

[4] If the group factor is specified, group-specific p-values are combined according to the method proposed by (Demetrescu *et al.*, 2006).

[5] Qualitative variables cannot contain missing values.

```
Estimating regression 3/3 (Pollution) ... 5%
Estimating regression 3/3 (Pollution) ... 10%
Estimating regression 3/3 (Pollution) ... 15%
Estimating regression 3/3 (Pollution) ... 20%
Estimating regression 3/3 (Pollution) ... 25%
Estimating regression 3/3 (Pollution) ... 25%
Estimating regression 3/3 (Pollution) ... 30%
Estimating regression 3/3 (Pollution) ... 35%
Estimating regression 3/3 (Pollution) ... 40%
Estimating regression 3/3 (Pollution) ... 45%
Estimating regression 3/3 (Pollution) ... 50%
Estimating regression 3/3 (Pollution) ... 55%
Estimating regression 3/3 (Pollution) ... 60%
Estimating regression 3/3 (Pollution) ... 65%
Estimating regression 3/3 (Pollution) ... 70%
Estimating regression 3/3 (Pollution) ... 75%
Estimating regression 3/3 (Pollution) ... 75%
Estimating regression 3/3 (Pollution) ... 80%
Estimating regression 3/3 (Pollution) ... 85%
Estimating regression 3/3 (Pollution) ... 90%
Estimating regression 3/3 (Pollution) ... 95%
Estimating regression 3/3 (Pollution) ... 100%
Estimation completed
```

The results of command `dlsem`(·) is an object of class `dlsem`. Among the components of `dlsem` objects, we found:

- `estimate`: a list of objects of class `lm`, one for each regression;
- `call`: a list containing the call for each regression after eventual adaptation of lag shapes;
- `model.code`: the model code after eventual adaptation of lag shapes;
- `data`: data after eventual logarithmic transformation and differencing, which were used in the estimation.

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|)
ecq(Job, 0, 5, Region) 0.1006394 0.01679967 5.990556 7.347355e-09 ***

Response: Pollution
                             Estimate Std. Error  t value     Pr(>|t|)
ecq(Job, 1, 8, Region)    0.08925573 0.03120937 2.859902 4.653304e-03  **
ecq(Consum, 1, 7, Region) 0.24416273 0.03430066 7.118310 1.593832e-11 ***


EXOGENOUS PART

Response: Job
            Estimate Std. Error   t value     Pr(>|t|)
Population -2.015755 0.39488913  -5.10461 5.918596e-07 ***
GDP        -1.274005 0.03999778 -31.85189 6.015402e-98 ***
```

```
Response: Consum
             Estimate Std. Error    t value     Pr(>|t|)
Population  0.8397265 0.20162453   4.164803 4.309595e-05 ***
GDP        -0.8165645 0.02990941 -27.301257 1.500457e-76 ***

Response: Pollution
             Estimate Std. Error   t value     Pr(>|t|)
Population -0.5444401 0.35131433 -1.549723 1.226718e-01
GDP         0.1467766 0.02815773  5.212656 4.343876e-07 ***


INTERCEPTS

Response: Job
             Estimate  Std. Error    t value     Pr(>|t|)
Region1  -0.027108664 0.002471484 -10.968577 9.695829e-24 ***
Region2  -0.014868387 0.002299389  -6.466235 4.104727e-10 ***
Region3  -0.014228172 0.003020166  -4.711056 3.784960e-06 ***
Region4  -0.005320298 0.003022567  -1.760192 7.940111e-02   .
Region5  -0.008833821 0.002249825  -3.926448 1.071505e-04 ***
Region6  -0.015622725 0.002392959  -6.528622 2.855024e-10 ***
Region7  -0.005154175 0.001800131  -2.863223 4.491035e-03  **
Region8  -0.027052095 0.002431495 -11.125706 2.802035e-24 ***
Region9  -0.046951445 0.002368957 -19.819459 2.396982e-56 ***
Region10 -0.023440072 0.002406224  -9.741433 1.200852e-19 ***

Response: Consum
             Estimate  Std. Error   t value     Pr(>|t|)
Region1   0.013228135 0.002838281  4.660614 5.159972e-06 ***
Region2  -0.009181367 0.002046187 -4.487061 1.107575e-05 ***
Region3   0.014910423 0.002424204  6.150648 3.086339e-09 ***
Region4   0.012261936 0.001834972  6.682356 1.550222e-10 ***
Region5   0.012591239 0.002755083  4.570184 7.703658e-06 ***
Region6   0.027006345 0.001908162 14.153070 1.060383e-33 ***
Region7   0.023946916 0.002205349 10.858561 1.035773e-22 ***
Region8  -0.014297098 0.003110621 -4.596220 6.868376e-06 ***
Region9   0.019452657 0.004320357  4.502558 1.035492e-05 ***
Region10  0.003490765 0.002953700  1.181828 2.384106e-01

Response: Pollution
              Estimate  Std. Error    t value     Pr(>|t|)
Region1   0.0156171437 0.006110882  2.5556285 1.128676e-02   *
Region2   0.0189615505 0.003138285  6.0420110 6.588110e-09 ***
Region3  -0.0028332829 0.005241024 -0.5405971 5.893423e-01
Region4   0.0019955283 0.003391747  0.5883481 5.569133e-01
Region5  -0.0074913183 0.003436561 -2.1798880 3.034664e-02   *
Region6  -0.0197953108 0.006507258 -3.0420356 2.640840e-03  **
Region7  -0.0182342490 0.004709758 -3.8715897 1.432683e-04 ***
Region8   0.0328607464 0.004644322  7.0754664 2.049059e-11 ***
Region9  -0.0007537518 0.008850955 -0.0851605 9.322127e-01
Region10  0.0163606362 0.004130296  3.9611295 1.013163e-04 ***


ERRORS
           Std. Dev.  df       Rsq
Job       0.01337002 298 0.8902711
Consum    0.01076881 247 0.8575233
Pollution 0.01101466 216 0.7232624
```

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 4 time lags and, on the other hand, the amount of greenhouse gas emissions (`Pollution`) from 2 to 6 time lags, while the amount of private consumption (`Consum`) significantly influences the amount of greenhouse gas emissions (`Pollution`) from 1 to 5 time lags. 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 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)
```

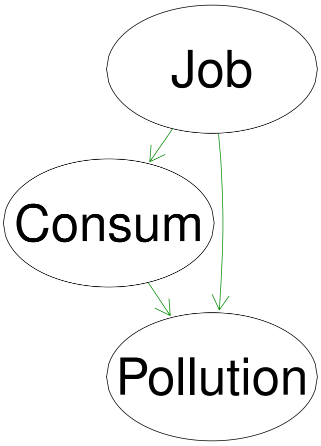

Figure 3: The DAG where each edge is coloured with respect to the sign of the estimated causal effect. Green: positive causal effect. Red: negative causal effect. Grey: not statistically significant causal effect (no such edges here).

The result is shown in Figure 3. Note that the DAG includes only the endogenous variables.

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 pathwise causal lag shapes and the overall one connecting two variables. The main arguments of command `causalEff`(·) include the name of one or more variables generating the causal effect (argument `from`), and the name of the variable receiving the causal effect (argument `to`). Optionally, specific time to which computation should be focused may be provided to argument `lag`, otherwise the whole lag shapes will be considered. Cumulative causal effects may be returned by setting the argument `cumul` to `TRUE`. 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 cumulative causal effect of the number of job positions on the amount of greenhouse gas emissions may be obtained by means of the following code:

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

$`Job*Consum*Pollution`
     estimate   std. err.   lower 95%   upper 95%
0  0.00000000 0.000000000 0.000000000 0.000000000
1  0.00526551 0.001155426 0.003000917 0.007530104
2  0.02306795 0.002994438 0.017198958 0.028936941
3  0.05992652 0.005597756 0.048955120 0.070897922
4  0.11935156 0.008739051 0.102223340 0.136479789
5  0.20008939 0.012074665 0.176423479 0.223755297
6  0.29486857 0.015211890 0.265053814 0.324683329
```

```
7  0.38964776 0.017804653 0.354751276 0.424544234
8  0.47038558 0.019657879 0.431856843 0.508914313
9  0.52981062 0.020771816 0.489098611 0.570522633
10 0.56666919 0.021303440 0.524915219 0.608423168
11 0.58447163 0.021481811 0.542368057 0.626575209
12 0.58973714 0.021512862 0.547572709 0.631901577
13 0.58973714 0.021512862 0.547572709 0.631901577

$`Job*Pollution`
     estimate  std. err.  lower 95%  upper 95%
0  0.00000000 0.00000000 0.00000000 0.00000000
1  0.03526152 0.01232963 0.01109590 0.05942715
2  0.09696919 0.02485116 0.04826181 0.14567656
3  0.17630761 0.03724486 0.10330902 0.24930620
4  0.26446142 0.04834566 0.16970566 0.35921717
5  0.35261522 0.05733608 0.24023857 0.46499187
6  0.43195365 0.06369478 0.30711416 0.55679313
7  0.49366131 0.06725017 0.36185339 0.62546923
8  0.52892283 0.06837109 0.39491796 0.66292770
9  0.52892283 0.06837109 0.39491796 0.66292770
10 0.52892283 0.06837109 0.39491796 0.66292770
11 0.52892283 0.06837109 0.39491796 0.66292770
12 0.52892283 0.06837109 0.39491796 0.66292770
13 0.52892283 0.06837109 0.39491796 0.66292770

$overall
     estimate  std. err.  lower 95%  upper 95%
0  0.00000000 0.00000000 0.00000000 0.00000000
1  0.04052703 0.01348505 0.01409681 0.06695725
2  0.12003714 0.02782539 0.06550036 0.17457391
3  0.23623413 0.04276247 0.15242122 0.32004704
4  0.38381298 0.05689906 0.27229288 0.49533308
5  0.55270461 0.06907040 0.41732911 0.68808011
6  0.72682222 0.07835349 0.57325220 0.88039224
7  0.88330906 0.08420037 0.71827937 1.04833876
8  0.99930841 0.08669842 0.82938263 1.16923419
9  1.05873345 0.08695776 0.88829937 1.22916753
10 1.09559203 0.08708628 0.92490605 1.26627800
11 1.11339446 0.08713009 0.94262263 1.28416630
12 1.11865997 0.08713775 0.94787313 1.28944682
13 1.11865997 0.08713775 0.94787313 1.28944682
```

The output of command `causalEff`(·) is a list of matrices including point estimates 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 11 and 12 years are equal.

A pathwise or an overall causal lag shape can be displayed using the command `lagPlot`(·). 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:

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

or the overall causal lag shape of the number of job positions on the amount of greenhouse gas emissions:

```
> 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.

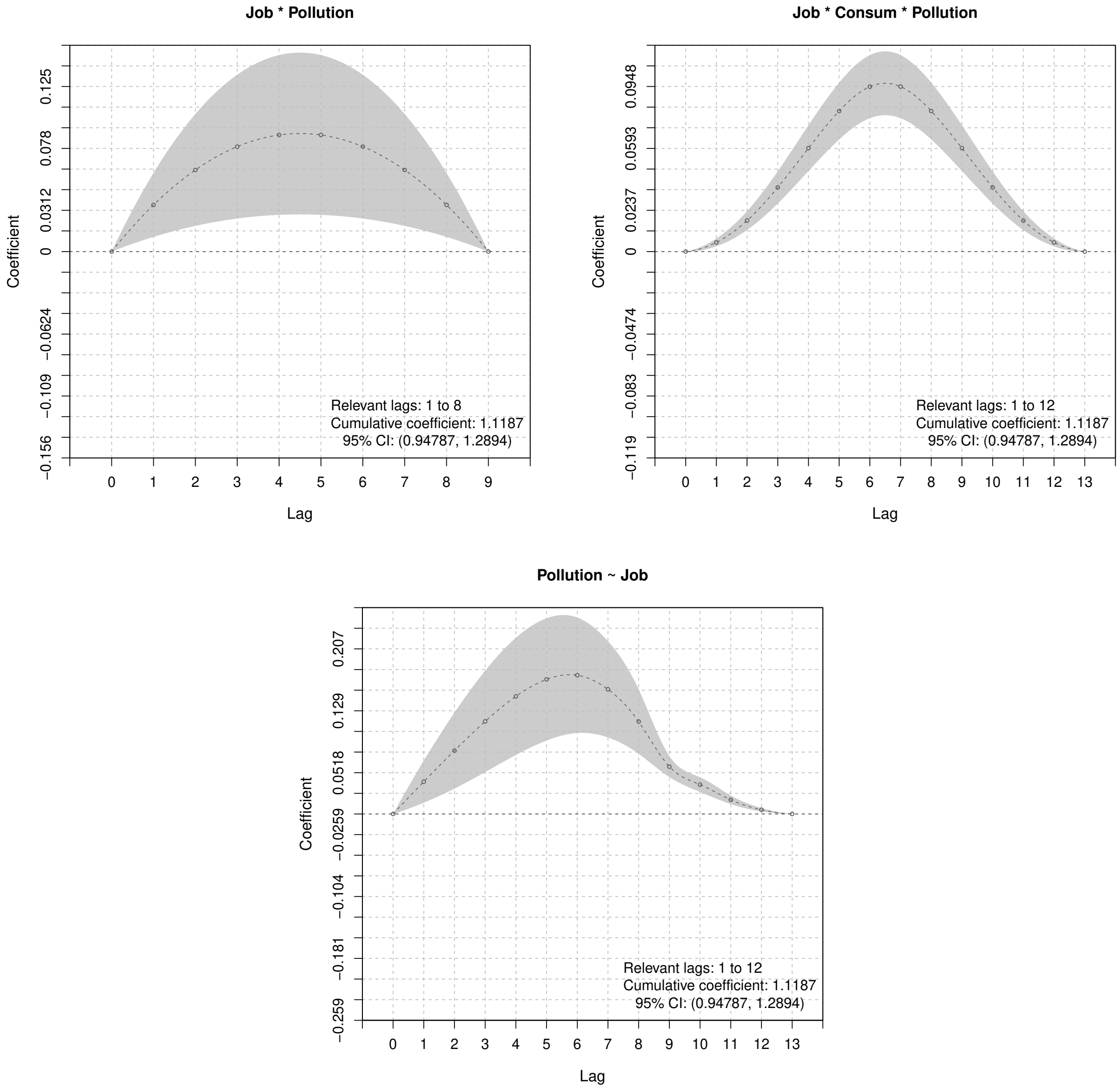

Figure 4: The pathwise causal lag shapes (upper panels) and the overall one (lower panel) connecting the number of job positions to the amount of greenhouse gas emissions. 95% asymptotic confidence intervals are shown in grey.

4.5 Comparison among alternative models

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~qd(Job,0,15),
+   Pollution~qd(Job,0,15)+qd(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,quiet=T)
> summary(indus.mod_2)$endogenous

$Job
NULL

$Consum
                       Estimate Std. Error  t value     Pr(>|t|)
qd(Job, 0, 7, Region) 0.1044609 0.02495916 4.185271 4.073917e-05 ***

$Pollution
                           Estimate Std. Error   t value     Pr(>|t|)
qd(Job, 2, 12, Region)    0.1441207 0.03359200  4.290328 2.938223e-05 ***
qd(Consum, 3, 10, Region) 0.3247139 0.02850865 11.390015 7.186589e-23 ***

> # model 3: linearly decreasing lag shapes
> indus.code_3 <- list(
+   Job ~ 1,
+   Consum~ld(Job,0.5,0.5),
+   Pollution~ld(Job,0.5,0.5)+ld(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,quiet=T)
> summary(indus.mod_3)$endogenous

$Job
NULL

$Consum
                       Estimate Std. Error  t value     Pr(>|t|)
ld(Job, 0, 5, Region) 0.1113044  0.0210713 5.282273 2.800099e-07 ***

$Pollution
                          Estimate Std. Error  t value     Pr(>|t|)
ld(Job, 3, 9, Region)    0.1276649 0.02158717 5.913925 1.371993e-08 ***
ld(Consum, 2, 8, Region) 0.2589867 0.03597007 7.200060 1.106564e-11 ***

> # model 4: gamma lag shapes
> indus.code_4 <- list(
+   Job ~ 1,
+   Consum~gam(Job,0.5,0.5),
+   Pollution~gam(Job,0.5,0.5)+gam(Consum,0.5,0.5)
+   )
> indus.mod_4 <- dlsem(indus.code_4,group="Region",exogenous=c("Population","GDP"),
+   data=industry,global.control=indus.global,local.control=indus.local,log=T,quiet=T)
> summary(indus.mod_4)$endogenous

$Job
NULL

$Consum
                              Estimate Std. Error  t value   Pr(>|t|)
gam(Job, 0.85, 0.2, Region) 0.06797784  0.0291622 2.331026 0.02102363 *

$Pollution
                                Estimate Std. Error   t value     Pr(>|t|)
gam(Job, 0.95, 0.05, Region)   0.1186122 0.02744747  4.321425 2.852175e-05 ***
gam(Consum, 0.9, 0.15, Region) 0.3496251 0.03333223 10.489099 1.556773e-19 ***
```

Here the option `quiet` was set to `TRUE` to suppress messages on the estimation progress. We see that the four models provide different results. Method `compareModels` can be used to compare them according to information criteria:

```
> compareModels(list(indus.mod,indus.mod_2,indus.mod_3, indus.mod_4))
```

```
    logLik  p      AIC      BIC
1 -209.9538 39 497.9076 617.8394
2 -210.0674 39 498.1347 618.0665
3 -210.0282 39 498.0564 617.9881
4 -209.9326 39 497.8651 617.7969
```

The model with endpoint-constrained quadratic lag shapes has the lowest value of each information criterion, and thus the best fit to data. Note that information criteria for variable `Job` are the same in each model because it has no endogenous covariates.

5 Final remarks

Lag shapes included in the package may represent a large number of real-world lag structures, nevertheless new lag shapes with further specific features may be added in the future. The same holds for further functionalities for linear models for time series data.

Please, do not hesitate to contact me for questions, feedbacks or bug reports.

References

S. Almon (1965). The Distributed Lag between Capital Appropriations and Net Expenditures. *Econometrica*, 33, 178-196.

M. Demetrescu, U. Hassler, and A. Tarcolea. Combining Significance of Correlated Statistics with Application to Panel Data. *Oxford Bulletin of Economics and Statistics*, 68(5), 647-663.

A. P. Dempster, N. M. Laird, and D. B. Rubin (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society*, Series B, 39(1): 1-38.

D. A. Dickey, and W. A. Fuller (1981). Likelihood Ratio Statistics for Autoregressive Time Series with a Unit Root. *Econometrica*,49: 1057-1072.

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

T. Haavelmo (1943). The Statistical Implications of a System of Simultaneous Equations. *Econometrica*, 11(1): 1-12.

T. C. Koopmans, H. Rubin, and R. B. Leipnik (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.

D. Kwiatkowski, P. C. B. Phillips, P. Schmidt and Y. Shin (1992). Testing the Null Hypothesis of Stationarity against the Alternative of a Unit Root. *Journal of Econometrics*, 54(1-3): 159-178.

A. Magrini (2020). A Family of Theory-Based Lag Shapes for Distributed-Lag Linear Regression. To be appeared on *Italian Journal of Applied Statistics*.

A. Magrini, F. Bartolini, A. Coli, B. Pacini (2019). A Structural Equation Model to Assess the Impact of Agricultural Research Expenditure on Multiple Dimensions. *Quality and Quantity*, 53(4): 2063-2080.

A. Magrini (2018). Linear Markovian Models for Lag Exposure Assessment. *Biometrical Letters*, 55(2): 179-195.

W. K. Newey, and K. D. West (1978). A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix. *Econometrica*, 55(3), 703-708.

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

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