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 2.1 – 04 January 2018

Contents

1 Introduction

Structural causal models (SCMs, Pearl, 2000, Chapter 5) consist in the simultaneous application of regression models to a set of variables, and, in a parametric linear formulation (linear SCMs), represent one of the prevalent methodologies for causal inference in contemporary applied sciences. *Distributed-lag linear structural equation models* (DLSEMs) are an extension of linear SCMs, where each factor of the joint probability distribution is a distributed-lag linear regression with constrained lag shapes (Judge *et al.*, 1985, Chapters 9-10). DLSEMs account for temporal delays in the dependence relationships among domain variables and allow to assess causal effects at different time lags. As such, they are suitable to investigate the effect of an external impulse on a multidimensional system through time. Econometrics and epidemiology are two of the main fields of application for DLSEMs.

Package `dlsem` implements inference functionalities for DLSEMs with several types of constrained lag shapes. 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. Section 5 includes considerations on future development of the package.

2 Theory

Lagged instances of one or more covariates can 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 P:

$$\beta_{j,l} = \sum_{p=0}^{P} \phi_p l^p \tag{2}$$

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. Package `dlsem` includes 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:

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

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 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 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 can be computed through numerical approximation.

For these three lag shapes it holds:

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

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 can 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 Akaike Information Criterion (AIC, Akaike, 1974) or the Bayesian Information Criterion (BIC, Schwarz, 1978)[1].

Structural causal models (SCMs) were developed by Pearl in the context of causal inference. They are rooted to path analysis (Wright, 1934) and simultaneous equation models (Haavelmo, 1943; Koopmans *et al.*, 1950). The basic feature of a SCM is a directed acyclic graph (DAG, see Pearl, 2000, pages 12 and following). In a DAG, variables are represented by nodes and directed edges may connect pairs of nodes without creating directed cycles (Figure 1). If a variable receives an edge from another variable, the latter is called *parent* of the former. A DAG encodes a factorization of the joint probability distribution:

$$p(V_1, \ldots, V_m) = \prod_{j=1}^{J} p(V_j \mid \Pi_j) \tag{7}$$

where Π_j is the set of parents of variable V_j. As such, if some pairs of variables are not connected by an edge, the DAG implies a set of conditional independence statements (Pearl, 2000, pages 16 and following). A SCM is defined by a specification of $p(V_j \mid \Pi_j)$ for $j = 1, \ldots, J$. In a linear parametric formulation (linear SCM), $p(V_j \mid \Pi_j)$ is the linear regression model where V_j is the response variable and its parents in the DAG are the covariates.

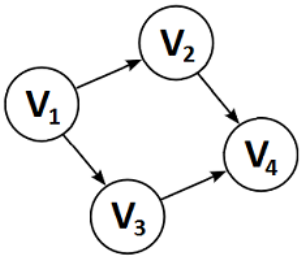

Figure 1: An example of directed acyclic graph.

The DAG of a SCM has a causal interpretation, and a causal effect is associated to each edge, directed path or couple of nodes according to the following definitions (Pearl, 2000, Section 5.3; Pearl, 2012):

- the causal effect associated to each edge in the DAG is the coefficient of the variable represented by the node originating the edge in the regression model of the variable represented by the node receiving the edge;
- the causal effect associated to a directed path is the product of the causal effects associated to each edge in the path;
- the causal effect of a variable on another is the sum of the causal effects associated to each directed path connecting the nodes representing the two variables.

In this view, each causal effect in a linear SCM represents the average change in the value of a variable induced by an intervention provoking a unit variation in the value of another variable. The causal effect of a variable on another is termed *overall* causal effect, the causal effect associated

[1] Neither the response variable nor the covariates must contain a trend in order to obtain unbiased estimates (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.

to a directed path made by a single edge is called *direct* effect, while the causal effects associated to the other directed paths are denoted as *indirect* effects.

Distributed-lag linear structural equation models (DLSEMs) are linear SCMs where each factor of the probability distribution in Equation 7 is a distributed-lag linear regression model (Judge *et al.*, 1985, Chapters 9-10). Time lags are not explicitly shown in the DAG of a DLSEM, and an edge connects two nodes, say $V_j \longrightarrow V_i$, if and only if there is at least one time lag where the coefficient of variable V_j in the regression model of variable V_i is non-zero. A DLSEM can be exploited to assess the causal effect of any variable to another at different time lags by extending the rules above:

- The causal effect associated to each edge in the DAG at lag k is represented by the coefficient at lag k of the variable represented by the parent node in the regression model of the variable represented by the child node.
- The causal effect associated to a directed path at lag k is computed as follows:
  1. denote the number of edges in the path as p;
  2. enumerate all the possible p-uples of lags, one lag for each of the p edges, such that their sum is equal to k;
  3. for each p-uple of lags:
     - for each lag in the p-uple, compute the coefficient associated to the corresponding edge at that lag;
     - compute the product of all these coefficients;
  4. sum all these products.
- The causal effect of a variable on another at lag k is represented by the sum of the causal effects at lag k associated to each directed path connecting the two variables.

These rules define a causal effect evaluated at a single lag, denoted as *instantaneous* causal effect. The *cumulative* causal effect at a prespecified lag, say k, is obtained by summing all the instantaneous causal effects for each lag up to k.

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")
```

The latest version of `dlsem` is 2.1.

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.

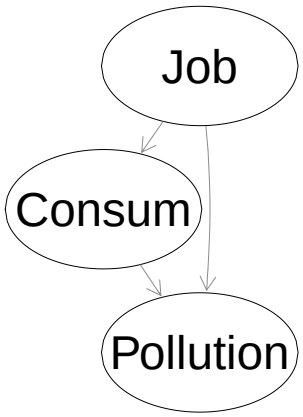

Figure 2: The DAG for the industrial development problem. 'Job': number of job positions in industry. 'Consum': private consumption index. 'Pollution': amount of greenhouse gas emissions.

```
> 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
```

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`$(\cdot)$, `qdec`$(\cdot)$ and `gamma`$(\cdot)$ can be employed to specify, respectively, an endpoint-constrained quadratic, a quadratic decreasing or a gamma lag shape. Operators `quec`$(\cdot)$ and `qdec`$(\cdot)$ have three arguments: the name of the covariate to which the lag shape is applied, the gestation lag (a_j) and the lead lag (b_j). Operator `gamma`$(\cdot)$ has three arguments: the name of the covariate to which the lag shape is applied, parameter δ_j and 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 specification of regression models with no endogenous covariates may be omitted from the model code (for example, one could avoid to specify the regression model 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~quec(Job,0,15),
+   Pollution~quec(Job,0,15)+quec(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. Control options 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.

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;

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

- `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()
```

In alternative, 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`$(\cdot)$. The user can 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 model, 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 model, in order to eliminate cross-sectional spurious effects. Each exogenous variable can be either qualitative or quantitative and its coefficient in each regression model is 0 for time lags greater than 0 (no lag). The user can 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, differentiation is performed until the hypothesis of unit root is rejected by the Augmented Dickey-Fuller test for all quantitative variables[3], and missing values are imputed with their conditional mean using the Expectation-Maximization algorithm (Dempster *et al.*, 1977)[4]. 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 model 1/3 (Job)
 Estimating regression model 2/3 (Consum)
 Estimating regression model 3/3 (Pollution)
Estimation completed
```

The results of command `dlsem`$(\cdot)$ 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 response variable;
- `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 of parameters a_j, b_j and θ_j $(j = 1, \ldots, J)$ for each endogenous variable, together with goodness of fit indices:

```
> summary(indus.mod)

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 0.1006394 0.01783725 5.642089 4.589874e-08 ***

Response: Pollution
       a b     theta  se(theta)  t value     Pr(>|t|)
Job    1 8 0.1048006 0.03008457 3.483532 5.989626e-04 ***
Consum 1 6 0.2320105 0.03660783 6.337729 1.339514e-09 ***
```

[3] If the group factor is specified, the panel version of the Augmented Dickey-Fuller test proposed by Levin *et al.* (2002) is used instead.

[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.

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

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 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 full summary of parameter estimation, including the group-specific intercepts and the coefficients of exogenous variables, can be obtained by applying the `summary` method to component `estimate` of the `dlsem` object:

```
> lapply(indus.mod$estimate,summary)

$Job

Call:
lm(formula = Job ~ Region + Population + GDP, data = industry)

Residuals:
      Min        1Q    Median        3Q       Max
-0.035183 -0.008863  0.000619  0.008844  0.035491

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
Region1    -0.027109   0.002403 -11.281  < 2e-16 ***
Region2    -0.014868   0.002402  -6.191 1.98e-09 ***
Region3    -0.014228   0.002402  -5.924 8.64e-09 ***
Region4    -0.005320   0.002403  -2.214 0.027588 *
Region5    -0.008834   0.002402  -3.678 0.000278 ***
Region6    -0.015623   0.002401  -6.506 3.26e-10 ***
Region7    -0.005154   0.002402  -2.146 0.032669 *
Region8    -0.027052   0.002402 -11.263  < 2e-16 ***
Region9    -0.046951   0.002402 -19.545  < 2e-16 ***
Region10   -0.023440   0.002403  -9.756  < 2e-16 ***
Population -2.015755   0.369195  -5.460 1.00e-07 ***
GDP        -1.274005   0.032533 -39.160  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.01337 on 298 degrees of freedom
  (10 observations deleted due to missingness)
Multiple R-squared:  0.8903,        Adjusted R-squared:  0.8859
F-statistic: 201.5 on 12 and 298 DF,  p-value: < 2.2e-16


$Consum

Call:
lm(formula = Consum ~ Region + Population + GDP + quec(Job, 0,
    5), data = industry)

Residuals:
```

```
       Min         1Q     Median         3Q        Max
-0.0275870 -0.0066042 -0.0001772  0.0074214  0.0263515

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
Region1     0.013228   0.003105   4.260 2.91e-05 ***
Region2    -0.009181   0.002452  -3.744 0.000226 ***
Region3     0.014910   0.002370   6.292 1.41e-09 ***
Region4     0.012262   0.002144   5.720 3.07e-08 ***
Region5     0.012591   0.002189   5.751 2.61e-08 ***
Region6     0.027006   0.002425  11.135  < 2e-16 ***
Region7     0.023947   0.002134  11.222  < 2e-16 ***
Region8    -0.014297   0.003062  -4.669 4.96e-06 ***
Region9     0.019453   0.004455   4.366 1.86e-05 ***
Region10    0.003491   0.002834   1.232 0.219243
Population  0.839726   0.307290   2.733 0.006736 **
GDP        -0.816565   0.027103 -30.128  < 2e-16 ***
Job         0.100639   0.017837   5.642 4.59e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.01077 on 247 degrees of freedom
  (60 observations deleted due to missingness)
Multiple R-squared:  0.8575,        Adjusted R-squared:   0.85
F-statistic: 114.4 on 13 and 247 DF,  p-value: < 2.2e-16


$Pollution

Call:
lm(formula = Pollution ~ Region + Population + GDP + quec(Job,
    1, 8) + quec(Consum, 1, 6), data = industry)

Residuals:
      Min        1Q    Median        3Q       Max
-0.026978 -0.007834  0.000029  0.006816  0.033939

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
Region1     0.018103   0.005672   3.192 0.001624 **
Region2     0.016695   0.002994   5.576 7.29e-08 ***
Region3     0.000871   0.004745   0.184 0.854523
Region4     0.003874   0.003341   1.160 0.247529
Region5    -0.004765   0.003654  -1.304 0.193542
Region6    -0.013855   0.006254  -2.215 0.027790 *
Region7    -0.013390   0.004810  -2.784 0.005848 **
Region8     0.029422   0.004103   7.172 1.16e-11 ***
Region9     0.002974   0.008692   0.342 0.732593
Region10    0.017110   0.004253   4.023 7.95e-05 ***
Population -0.533564   0.322472  -1.655 0.099457 .
GDP         0.134247   0.029659   4.526 9.91e-06 ***
Job         0.104801   0.030085   3.484 0.000599 ***
Consum      0.232011   0.036608   6.338 1.34e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.01112 on 216 degrees of freedom
  (90 observations deleted due to missingness)
```

```
Multiple R-squared:  0.7177,        Adjusted R-squared:  0.6994
F-statistic: 39.22 on 14 and 216 DF,  p-value: < 2.2e-16
```

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)
```

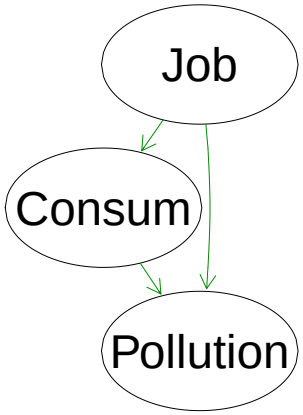

Figure 3: The DAG where each edge is coloured with respect to the sign of the estimated causal effect. Green: positive causal effect. Red: negative causal effect. Grey: not statistically significant causal effect (no such edges here).

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 the causal effect associated to any edge, directed path or couple of nodes at different time lags. 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 lags at which the causal effect must be computed can be provided to argument `lag`, otherwise all the relevant ones are considered. Also, the user can choose whether the instantaneous (argument `cumul` set to `FALSE`, the default) or the cumulative (argument `cumul` set to `TRUE`) causal effect must be returned. 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 can be obtained by means of the following code:

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

$`Job*Consum*Pollution`
      estimate   lower 95%   upper 95%
0  0.000000000 0.000000000 0.000000000
1  0.005601519 0.002978257 0.008224781
```

```
2  0.024273250 0.017556697 0.030989803
3  0.062239103 0.049859194 0.074619011
4  0.121988641 0.102982419 0.140994863
5  0.200409911 0.174670963 0.226148858
6  0.287544654 0.255879838 0.319209470
7  0.365965924 0.329856126 0.402075722
8  0.425715462 0.386832409 0.464598516
9  0.463681315 0.423431571 0.503931059
10 0.482353046 0.441631154 0.523074937
11 0.487954565 0.447148267 0.528760863
12 0.487954565 0.447148267 0.528760863

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

$overall
     estimate  lower 95%  upper 95%
0  0.00000000 0.00000000 0.00000000
1  0.04700422 0.02108627 0.07292217
2  0.13813067 0.08450297 0.19175836
3  0.26925259 0.18666113 0.35184405
4  0.43250887 0.32250642 0.54251132
5  0.61443689 0.48096437 0.74790940
6  0.79472770 0.64361323 0.94584216
7  0.94560369 0.78369692 1.10751046
8  1.04675592 0.88051424 1.21299761
9  1.08472178 0.91815513 1.25128843
10 1.10339351 0.93671214 1.27007488
11 1.10899503 0.94229301 1.27569704
12 1.10899503 0.94229301 1.27569704
```

The output of command `causalEff`$(\cdot)$ is a list of matrices, each containing estimate and confidence interval for the causal effect associated to each path connecting the starting variables to the ending variable at the requested time lags. Also, estimate and confidence interval for the overall causal effect are contained in the component named `overall`.

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.

The estimated lag shape associated to a path or to an overall causal effect can be displayed using the command `lagPlot`$(\cdot)$. For instance, one can display the 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 lag shape associated to the overall causal effect 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 the lag shape associated to a path including more than one edge is a mixture of constrained lag shapes, thus it may show an irregular aspect, like it is the case of the overall causal effect displayed in the lower panel of Figure 4.

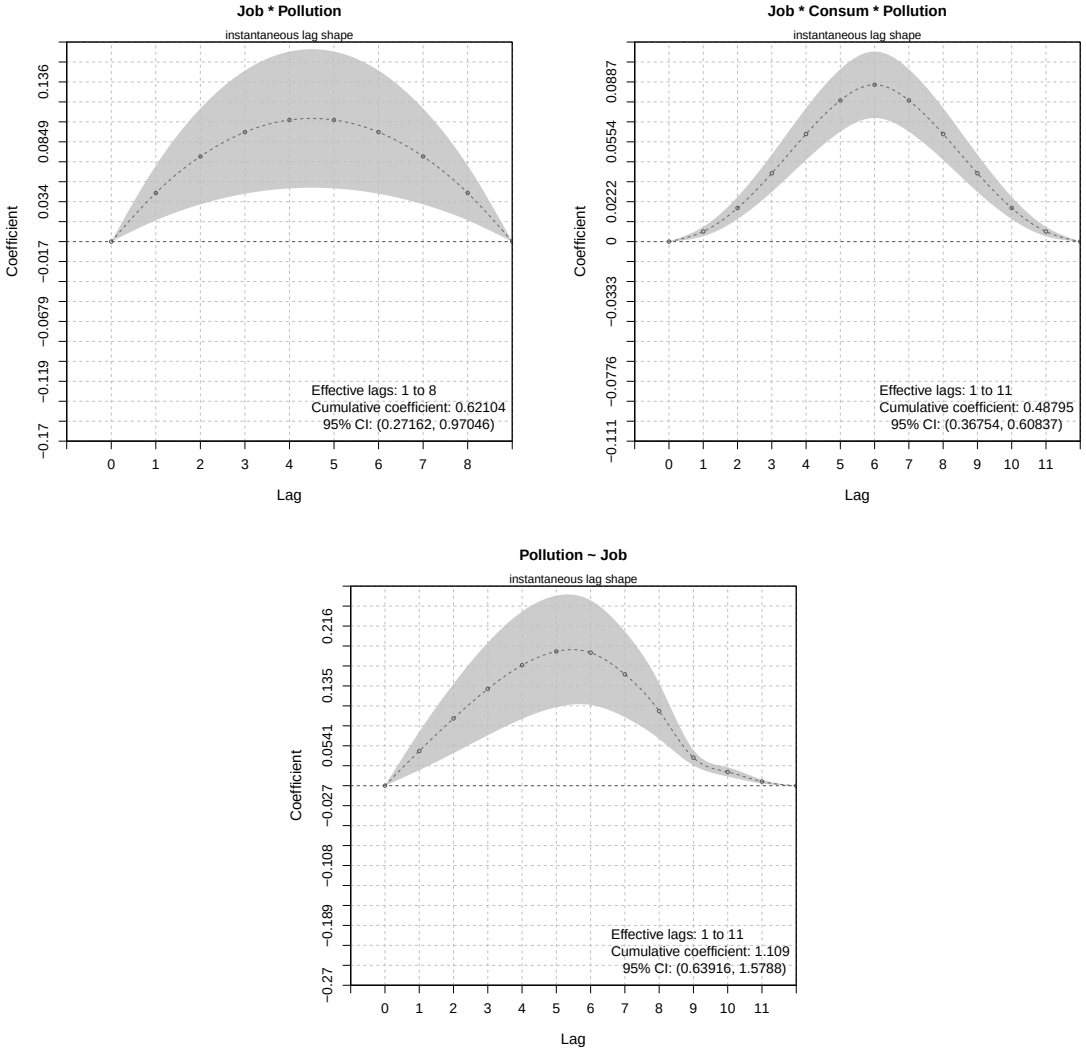

Figure 4: The estimated lag shape associated to each path connecting the number of job positions to the amount of greenhouse gas emissions (upper panels) and to the overall causal effect (lower panel). 95% confidence intervals are shown in grey.

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(Job,0,15),
+   Pollution~qdec(Job,0,15)+qdec(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)

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

> 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(Job,0.5,0.5),
+   Pollution~gamma(Job,0.5,0.5)+gamma(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)

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

> 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. Methods `AIC` and `BIC` for class `dlsem` can be used to compare them according to AIC and BIC, respectively:

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

$QUEC
      Job    Consum Pollution (overall)
-1781.636 -1603.661 -1401.076 -4786.373

$QDEC
       Job     Consum  Pollution  (overall)
-1781.6357 -1585.9224  -910.7421 -4278.3002

$GAMMA
      Job    Consum Pollution (overall)
-1781.636 -1655.876 -1182.643 -4620.154

> 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)
-1733.0602 -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 both AIC and BIC. Note that the fit for variable `Job` is constant in each model because it has no endogenous covariates.

5 Future development

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 complete search over all the possible models is infeasible for most real-world applications, a heuristic search is currently implemented. Further development of the package may be directed towards the improvement of the search strategy.

Grouped data are currently managed through fixed effects estimation. In the future, random effects estimation may be implemented 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.

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

References

H. Akaike (1974). A New Look at the Statistical Identification Model. *IEEE Transactions on Automatic Control*, 19: 716-723.

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

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.

G. G. Judge, W. E. Griffiths, R. C. Hill, H. Lutkepohl, and T. C. Lee (1985). The Theory and Practice of Econometrics. John Wiley & Sons, 2nd ed., New York, US-NY.

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.

A. Levin, C. Lin, and C. J. Chub (2002). Unit Root Tests in Panel Data: Asymptotic and Finite-Sample Properties. *Journal of Econometrics*, 108: 1-24.

J. Pearl (2012). The Causal Foundations of Structural Equation Modelling. In: R. H. Hoyle (ed.), Handbook of Structural Equation Modelling, pages 68-91. Guilford Press, New York, US-NY.

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

G. Schwarz (1978). Estimating the Dimension of a Model. *Annals of Statistics*, 6, 461-464.

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