Distributed-Lag Linear Structural Equation Modelling with the R Package dlsem

Alessandro Magrini
Dep. Economics and Management
University of Pisa, Italy
<alessandro.magrini@for.unipi.it>

`dlsem` version 1.8 – 13 June 2017

Contents

1 Introduction **1**

2 Theory **1**

3 Installation **4**

4 Illustrative example **4**

4.1 Specification of the model code 5
4.2 Specification of control options 5
4.3 Estimation 6
4.4 Assessment and decomposition of causal effects 9

1 Introduction

Package `dlsem` implements inference functionalities for distributed-lag linear structural equation models with constrained lag shapes (DLSEMs: Magrini *et al.*, 2016). Endpoint-constrained quadratic, quadratic decreasing and gamma lag shapes are available. DLSEMs allow to perform dynamic causal inference, that is to assess and disentangle causal effects at different time lags. The vignette is structured as follows. In Section 2, theory on distributed-lag linear structural equation models with constrained lag shapes is presented. In Section 3, instructions for the installation of the `dlsem` packages are provided. In Section 4, the practical use of `dlsem` is illustrated through a fictitious impact assessment problem.

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 effect on the response variable 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:

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

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 $(a_j + b_j)/2$. 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 $\frac{\delta_j}{(\delta_j-1)\log(\lambda_j)}$. 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 regression 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 regression coefficent, but values a_j and b_j can be computed through numerical approximation.

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 lowest Akaike Information Criterion (Akaike, 1974)[1].

Structural equation models (SEMs) have a long history starting with the contribution of Wright (1934) and further elaborated by Pearl (2000) in the context of causal inference. The basic feature of a SEM is a directed acyclic graph (DAG, see Pearl, 2000, pages 12 and on). In a DAG, variables are represented by nodes and directed edges may connect pairs of variables 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:

$$p(V_1, \ldots, V_m) = \prod_{j=1}^{J} p(V_j \mid \Pi_j) \tag{6}$$

[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.

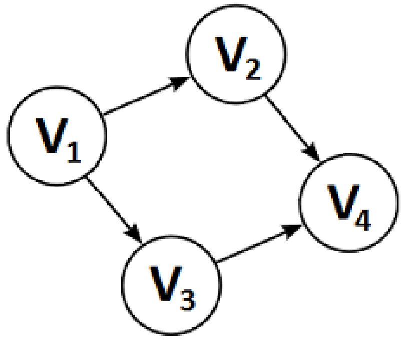

Figure 1: The directed acyclic graph of a SEM. The regression model applied to variable V_1 has no covariates, the regression models applied to variables V_2 and V_3 have V_1 as covariate, the regression model applied to variable V_4 has V_2 and V_3 as covariates.

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 on). A SEM is defined by a specification of $p(V_j \mid \Pi_j)$ for $j = 1, \ldots, J$. In a linear parametric formulation (linear SEM), $p(V_j \mid \Pi_j)$ is the linear regression model where V_j is the response variable and its parents in the DAG are the covariates.

In the Pearl's framework, the DAG has a causal interpretation, and a causal effect is associated to each edge, directed path or couple of nodes to represent expected changes induced by an intervention (Pearl, 2000, Chapter 5.3; Pearl, 2012). For a linear SEM:

- 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.

Thus, a linear SEM can be employed to assess and decompose the average change in the value of any variable induced by an intervention provoking a unit variation in the value of any other variable. The causal effect of a variable on another is termed *overall* causal effect, the causal effect associated 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.

A distributed-lag linear structural equation model (DLSEM) is a SEM where each regression model is a distributed-lag linear regression model with constrained lag shapes. For a DLSEM, the DAG does not explicitly include time lags, and an edge connecting two nodes implies that there is at least one time lag where the coefficient of the variable represented by the parent node in the regression model of the variable represented by the child node is non-zero. A DLSEM can be exploited to assess and decompose 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.

A causal effect evaluated at a single lag is 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 1.8.

4 Illustrative example

The practical use of package `dlsem` is illustrated through a fictitious impact assessment problem, aiming at testing 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 proposed problem is shown in Figure 2. The analysis will be conducted on the dataset industry, containing 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
 1st Qu.:105.07   1st Qu.: 87.88   1st Qu.:  7536
```

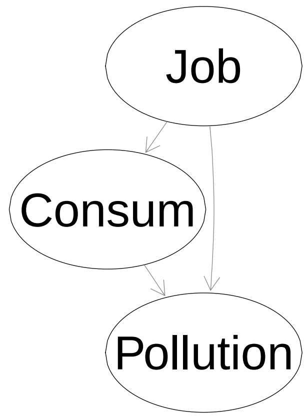

Figure 2: The DAG for the industrial development problem. 'Job': number of job positions in industry. 'Consum': private consumption index. 'Pollution': amount of greenhouse gas emissions.

```
 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 is meant to include the formal specification of the regression models in the DLSEM. 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( ), qdec( ) and gamma( ) can be employed to specify, respectively, an endpoint-constrained quadratic, a quadratic decreasing or a gamma lag shape. Operators quec( ) and qdec( ) have three arguments: the name of the variable to which the lag shape is applied, the minimum lag with a non-zero coefficient (a_j), and the maximum lag with a non-zero coefficient (b_j). Operator gamma( ) has three arguments: the name of the variable to which the lag shape is applied, parameter δ_j and parameter λ_j. If none of these two operators is applied to a variable, it is assumed that the coefficient associated to that variable is 0 for time lags greater than 0 (no lag). The group factor and exogenous variables must not be specified in the model code (see Subsection 4.3). The regression model for variables with no covariates besides the group factor and exogenous variables can be omitted from the model code (here, we could omit the regression model for the number of job positions). In this problem, an endpoint-constrained quadratic lag shape between 0 and 15 time lags is assumed for all variables:

```
> mycode <- 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 must be a named list containing one or more among several components. The key component is adapt, a named vector of logical values where each value must refer to one response

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

variable and indicates whether values a_j and b_j for each lag shape in the regression model of that variable must be selected on the basis of the best fit to data, instead of employing the ones specified in the model code. If adaption is requested for a regression model, three further components are taken into account: max.gestation, max.lead, min.width and sign. Each of these three components is a named list, where each component of the list must refer to one response variable and must be a named vector including, respectively, the maximum gestation lag, the maximum lead lag, the minimum lag width and the sign (either '+' for non-negative, or '-' for non-positive) of the coefficients of one or more covariates. In this problem, adaptation of lag shapes is performed for all regression models with the following constraints: (i) maximum gestation lag of 3 years, (ii) maximum lead lag of 15 years, (iii) minimum lag width of 5 years, (iv) all coefficients with non-negative sign

```
> mycontrol <- 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="+"))
+   )
```

4.3 Estimation

Once the model code and control options are specified, the structural model can be estimated from data using the command dlsem( ). The user can indicate a group factor 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. By indicating exogenous variables, they will be included as non-lagged covariates in each regression model, in order to eliminate spurious effects due to differences between the levels of the group factor. 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. Before 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 using the Expectation-Maximization algorithm (Dempster *et al.*, 1977). In this problem, the region is indicated as the group factor, while population and gross domestic product are indicated as exogenous variables. Also, we request the logarithmic transformation and provide control options to argument control:

```
> mod0 <- dlsem(mycode,group="Region",exogenous=c("Population","GDP"),
+   data=industry,control=mycontrol,log=T)
Checking stationarity...
 Order 1 differentiation performed
Start estimation...
 Estimating regression model 1/3 (Job)
 Estimating regression model 2/3 (Consum)
 Estimating regression model 3/3 (Pollution)
Estimation completed
```

After estimating the structural model, the user can display the DAG where each edge is coloured according to the sign of its causal effect (green for non-negative, red for non-positive). The result is shown in Figure 3: the group factor and exogenous variables are omitted from the DAG.

```
> plot(mod0)
```

All edges result statistically significant, providing evidence that the influence of industrial development on pollution is both direct and mediated by private consumption.

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

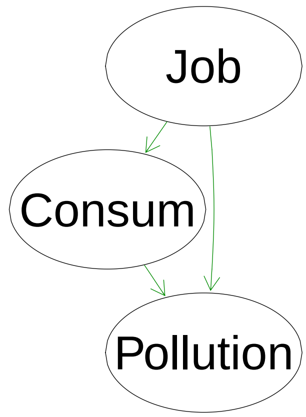

Figure 3: The DAG where each edge is coloured with respect to the sign of its causal effect. Green: non-negative causal effect. Red: non-positive causal effect. Grey: not statistically significant causal effect (no such edges here).

The user can also request the summary of estimation:

```
> summary(mod0)
$Job

Call:
lm(formula = Job ~ Region + Population + GDP, data = `structure(list(Region = structure(c(1L, 1L, 1L, 1L, 1L, 1

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 + quec(Job, 0, 5) + Population +
    GDP, data = `structure(list(Region = structure(c(1L, 1L, 1L, 1L, 1L, 1L, 1L, `)
```

```
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
Job         0.100639   0.017837   5.642 4.59e-08 ***
Population  0.839726   0.307290   2.733 0.006736 **
GDP        -0.816565   0.027103 -30.128  < 2e-16 ***
---
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 + quec(Job, 1, 8) + quec(Consum,
    1, 6) + Population + GDP, data = `structure(list(Region = structure(c(1L, 1L, 1L, 1L, 1L, 1L, 1L, `)

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 ***
Job         0.104801   0.030085   3.484 0.000599 ***
Consum      0.232011   0.036608   6.338 1.34e-09 ***
Population -0.533564   0.322472  -1.655 0.099457 .
GDP         0.134247   0.029659   4.526 9.91e-06 ***
---
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 summary of estimation returns estimates of parameters θ_j ($j = 1, \ldots, J$). Instead, the command edgeCoeff( ) can be used to obtain estimates and confidence intervals of coefficients at the relevant time lags $\beta_{j,l}$ $(j = 1, \ldots, J;\ l = 0, 1, \ldots)$:

```
> edgeCoeff(mod0)
$`0`
                   estimate lower 95% upper 95%
Consum~Job       0.04929275 0.0321693 0.0664162
Pollution~Job    0.00000000 0.0000000 0.0000000
Pollution~Consum 0.00000000 0.0000000 0.0000000

$`1`
                   estimate  lower 95%  upper 95%
Consum~Job       0.08215458 0.05361550 0.11069366
Pollution~Job    0.04140270 0.01810801 0.06469739
Pollution~Consum 0.11363780 0.07849493 0.14878066

$`2`
                   estimate  lower 95% upper 95%
Consum~Job       0.09858550 0.06433860 0.1328324
Pollution~Job    0.07245472 0.03168901 0.1132204
Pollution~Consum 0.18939633 0.13082488 0.2479678

$`3`
                   estimate  lower 95% upper 95%
Consum~Job       0.09858550 0.06433860 0.1328324
Pollution~Job    0.09315607 0.04074302 0.1455691
Pollution~Consum 0.22727559 0.15698986 0.2975613

$`4`
                   estimate  lower 95% upper 95%
Consum~Job       0.08215458 0.05361550 0.1106937
Pollution~Job    0.10350674 0.04527002 0.1617435
Pollution~Consum 0.22727559 0.15698986 0.2975613

$`5`
                   estimate  lower 95% upper 95%
Consum~Job       0.04929275 0.03216930 0.0664162
Pollution~Job    0.10350674 0.04527002 0.1617435
Pollution~Consum 0.18939633 0.13082488 0.2479678

$`6`
                   estimate  lower 95% upper 95%
Consum~Job       0.00000000 0.00000000 0.0000000
Pollution~Job    0.09315607 0.04074302 0.1455691
Pollution~Consum 0.11363780 0.07849493 0.1487807

$`7`
                   estimate  lower 95% upper 95%
Consum~Job       0.00000000 0.00000000 0.0000000
Pollution~Job    0.07245472 0.03168901 0.1132204
Pollution~Consum 0.00000000 0.00000000 0.0000000

$`8`
                  estimate  lower 95%  upper 95%
Consum~Job       0.0000000 0.00000000 0.00000000
Pollution~Job    0.0414027 0.01810801 0.06469739
Pollution~Consum 0.0000000 0.00000000 0.00000000
```

4.4 Assessment and decomposition of causal effects

Causal effects can be computed using the command causalEff( ). The user must specify one or more starting variables (argument from) and the ending variable (argument to). Optionally, specific time lags at which causal effects must be computed can be provided to argument lag, otherwise all the relevant ones are considered. Also, the user can choose whether instantaneous (argument cumul set to FALSE, the default) or cumulative (argument cumul set to TRUE) causal effects must be returned. Here, the cumulative causal effect of the number of job positions on the amount of greenhouse gas emissions is requested at time lags 0, 5, 10, 15 and 20:

```
> causalEff(mod0,from="Job",to="Pollution",lag=seq(0,20,by=5),cumul=T)
```

```
$`Job*Consum*Pollution`
    estimate lower 95% upper 95%
0  0.0000000 0.0000000 0.0000000
5  0.2004099 0.1494260 0.2513939
10 0.4823530 0.3645648 0.6001413
15 0.4879546 0.3675431 0.6083661
20 0.4879546 0.3675431 0.6083661

$`Job*Pollution`
    estimate lower 95% upper 95%
0  0.0000000 0.0000000 0.0000000
5  0.4140270 0.1810801 0.6469739
10 0.6210405 0.2716201 0.9704608
15 0.6210405 0.2716201 0.9704608
20 0.6210405 0.2716201 0.9704608

$overall
    estimate lower 95% upper 95%
0  0.0000000 0.0000000 0.0000000
5  0.6144369 0.3305060 0.8983677
10 1.1033935 0.6361849 1.5706021
15 1.1089950 0.6391632 1.5788269
20 1.1089950 0.6391632 1.5788269
```

The output of command causalEff is a list of matrices, each containing estimates and confidence intervals of the causal effect associated to each path connecting the starting variables to the ending variable at the requested time lags. Also, estimates and confidence intervals of the overall causal effect is shown in the component named overall.

Since the logarithmic trasformation was applied to all quantitative variables, causal effects above are interpreted as elasticities, that is, for a 1% of job positions more, greenhouse gas emissions are expected to grow by 1.31% after 20 years. Actually, the effect ends before 15 years, as the cumulative causal effects after 15 and 20 years are equal. The time lag up to which the effect is non-zero can be found by running command causalEff without providing a value to argument lag:

```
> causalEff(mod0,from="Job",to="Pollution",cumul=T)$overall
```

```
     estimate  lower 95%  upper 95%
0  0.00000000 0.00000000 0.00000000
1  0.04700422 0.02108627 0.07292217
2  0.13813067 0.06526392 0.21099741
3  0.26925259 0.13357327 0.40493191
4  0.43250887 0.22417152 0.64084623
5  0.61443689 0.33050605 0.89836772
6  0.79472770 0.43994019 1.14951521
7  0.94560369 0.53269372 1.35851366
8  1.04675592 0.59612995 1.49738190
9  1.08472178 0.62369628 1.54574727
10 1.10339351 0.63618492 1.57060209
11 1.10899503 0.63916318 1.57882687
12 1.10899503 0.63916318 1.57882687
```

The estimated lag shape associated to a path or to an overall causal effect can be displayed using the command lagPlot( ). For instance, we can display the lag shape associated to each path connecting the number of job positions to the amount of greenhouse gas emissions:

```
> lagPlot(mod0,path="Job*Pollution")
> lagPlot(mod0,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(mod0,from="Job",to="Pollution")
```

The resulting graphics are shown in Figure 4.

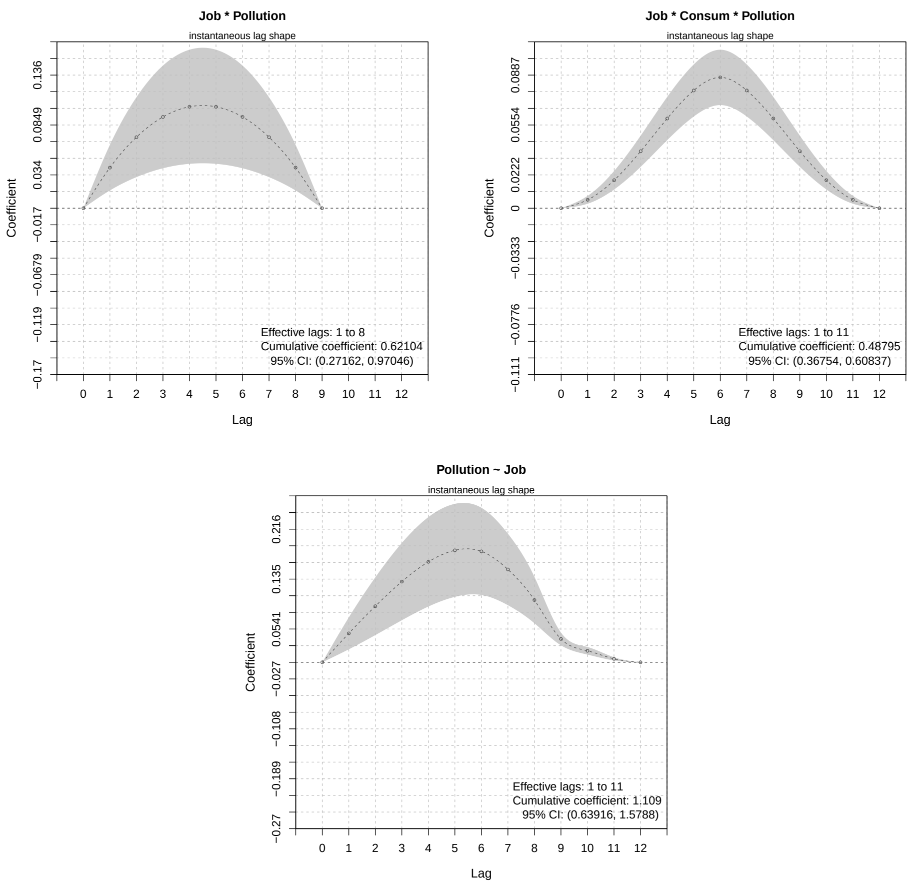

Figure 4: The estimated lag shape associated to each path connecting the number of job positions to the amount of greenhouse gas emissions (upper panels) and to the overall causal effect (lower panel). 95% confidence intervals are shown in grey.

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.

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.

A. Magrini, F. Bartolini, A. Coli, and B. Pacini (2016). Distributed-Lag Structural Equation Modelling: An Application to Impact Assessment of Research Activity on European Agriculture. *Proceedings of the 48th Meeting of the Italian Statistical Society*, 8-10 June 2016, Salerno, IT.

J. Pearl (2012). The Causal Foundations of Structural Equation Modelling. In: R. H. Hoyle (ed.), Handbook of Structural Equation Modelling, Chapter 5. Guilford Press, New York, US-NY.

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.