

# Fitting directed acyclic graphs with latent nodes as finite mixtures models, with application to education transmission

Antonio Forcina · Salvatore Modica

Received: date / Accepted: date

**Abstract** This paper describes an efficient EM algorithm for maximum likelihood estimation of a system of nonlinear structural equations corresponding to a directed acyclic graph model that can contain an arbitrary number of latent variables. The endogenous variables in the model must be categorical, while the exogenous variables may be arbitrary. The models discussed in this paper are an extended version of finite mixture models suitable for causal inference. An application to the problem of education transmission is presented as an illustration.

**Keywords** Extended latent class models · mixture models · structural equations · causal inference

62H17 91B15 62P20

## 1 Introduction

Structural equation models (SEM) are defined by a system of nonlinear equations specifying which variables have a direct causal effect on each endogenous variable in the system. A recursive non parametric SEM is equivalent to a directed acyclic graph (DAG) and, also, to a set of conditional independence statements. Pearl (1995) has shown that, under certain conditions, (the *back-door* and the *front-door* criteria) causal effects can be estimated from the frequency distribution of the observed variables; these conditions are, however, rather restrictive and are difficult to combine with statistical modeling assumptions. In this paper we restrict attention to models where the full joint distribution of observed and latent variables is identified and we describe an efficient algorithm for maximum likelihood estimation; certain routines of this algorithm may also be used to compute natural direct causal effects, Pearl (2010).

---

A. Forcina  
Dip. Economia, Finanza e Statistica, Univ. of Perugia, Italy  
Tel.: +39075 5855230  
Fax: +123-45-678910  
E-mail: forcina@stat.unipg.it

S. Modica  
Dip. di Scienze Economiche, Aziendali e Finanziarie, Univ. of Palermo, Italy

The class of models considered in this paper may be seen as an extension of latent class models in the sense that observable variables need not be independent conditionally on the latent ones. In addition, an observable variable may have a direct effect on a latent one and a latent variable may have a direct effect on another latent variable which is conceptually distinct. These models are not entirely new, for example Hagenaars (2002) has considered an application to a social science context of a model which is a special case of those considered here. The class of mixture models considered by Alfò and Trovato (2011) may be seen as a special case of those studied here, relative to the dependence structure; a more detailed discussion will be given in section 2.2.

We present an application in the context of education transmission, a much debated issue in Econometrics and Labor Economics. In order to assess the causal effect of the education of the parents on that of their child, one needs to control for the latent endowments of the parents and that of the child, which are likely to be strongly associated. The approach we propose is based on estimating a recursive system of structural equations where the natural endowment of parents and child are treated as two latent endogenous variables; this, we believe, provides an innovative contribution to the existing literature on the subject which we review briefly in Section 5.

The class of models studied in this paper are defined in section 2 where we examine the relationship with related models. The computation of maximum likelihood estimates and their implementation are discussed in section 3, an approach to the evaluation of causal effects is presented in section 4 and the application to education transmission is presented in section 5.

## 2 A class of semi-parametric structural equation models

We recall, following Pearl (2000), that a non parametric recursive structural equation model is a system of equations in the variables  $Z_1, \dots, Z_n$

$$Z_i = f_i(pa_i, \varepsilon_i), \quad i = 1, \dots, n \quad (1)$$

where  $pa_i$  is the subset of variables which are assumed to be the direct causes of  $Z_i$ , these are usually called *parents*, and  $\varepsilon_1, \dots, \varepsilon_n$  is a set of independent background exogenous variables which account for all residual effects. The fact that the system is recursive implies that, if  $Z_h$  is a parent of  $Z_i$ ; then  $h < i$ . The system is non-parametric in the sense that the distribution of the  $\varepsilon_i$  and the form of the functions  $f_i$  do not need to be specified. Such a system is equivalent to a causal DAG where endogenous variables are represented by nodes and there is an arrow from  $Z_h$  to  $Z_i$  if  $Z_h$  is a direct cause of  $Z_i$ , that is if  $Z_h \in pa_i$ . A convenient property of causal DAGs is that the joint distribution may be factorized into the product of the conditional distribution of each node given its parents. A DAG can contain one or more latent nodes, for example in the case of education transmission discussed in section 5, the unobservable endowments of the parents and that of the child are supposed to affect the educational achievements of the latter.

The methodology described in this paper is applicable when endogenous variables, observed or latent, are categorical. Our models differ from non parametric SEM because, when a variable is assumed to depend on two or more other variables, we allow some of these effects to be additive on a logit scale appropriate to the nature of the response variable under consideration. Essentially, logits of type *reference category* or *adjacent* are more appropriate when response categories are not ordered, logits of type *global* are preferable when

response categories are ordered and logits of type *continuation* are more suitable when response categories correspond to survival or achievements, see Colombi and Forcina (2001) for a detailed discussion. If  $Z_i$  has categories coded as  $0, 1, \dots, c_i - 1$ , the  $i$ th structural equation has  $c_i - 1$  components, one for each logit of  $Z_i$  and, in the special case when the effects of its parents are additive, the  $h$ th logit ( $h = 1, \dots, c_i - 1$ ) may be written as

$$\lambda_{ih} = \sum_{l=1}^h \beta_{i0l} + \sum_{Z_j \in pa_i} \sum_{l=1}^{c_j-1} \beta_{ijl} I(Z_j \geq l), \quad (2)$$

where  $I(Z_j \geq l)$  is the indicator function. Note that we have used the incremental coding for the  $\beta$ s, this means that, for instance  $\beta_{i0h}$  is the difference in the intercepts of the  $h$  and  $h - 1$  logits for  $Z_i$ . The reconstruction formulas for the case of logits global and adjacent, the only types used in this paper, are given below

$$(g) : P(Z_i = h) = \frac{\exp(\lambda_{ih})}{1 + \exp(\lambda_{ih})} - \frac{\exp(\lambda_{i,h-1})}{1 + \exp(\lambda_{i,h-1})}$$

$$(a) : P(Z_i = h) = \frac{\exp(\sum_{l=1}^h \lambda_{il})}{1 + \sum_{h=1}^{c_i-1} \exp(\sum_{l=1}^h \lambda_{il})}$$

From the software point of view, any model of our class is determined by the following specifications:

- An ordered list of the endogenous variables such that, if there is an arrow from  $Z_i$  to  $Z_j$ , then  $Z_i$  comes before  $Z_j$ ;
- A binary indicator specifying which variables, among the endogenous ones, are latent;
- For each endogenous variable, the list of its parents;
- For each node, the corresponding link function; this is determined by the number of categories of the node variable and the type of logit (adjacent, global, continuation) which determines how its conditional distribution is parameterized;
- For each endogenous variable, a regression model which specifies how its logits depend on the parents and, possibly, on additional exogenous variables measured at the level of statistical units; this is determined by a design matrix for each response variable .

## 2.1 Identifiability

Identifiability results for latent class models under conditional independence are by now well established. Recent results by Allman et al (2009) can handle several extended latent class models where certain subsets of the observable variables may be associated conditionally to the latent. Though, to our knowledge, no results are available to determine whether a general DAG with an arbitrary number of latent variables is identifiable, the numerical method described by Forcina (2008) can be used to determine whether a given model is locally identifiable with very high probability everywhere in the parameter space; this approach was used in the application. Essentially, the method samples points from the parameter space and checks whether the jacobian matrix obtained by differentiating the log-linear parameters of the saturated model for the joint distribution of the observable variables with respect to the actual parameters of the model is well away from being singular.

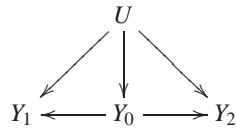
Typical modeling restrictions that might be used to achieve identifiability are assumptions of additivity within a given link function, like, for example, a multivariate logistic

function. Continuous covariates may be included as exogenous variables; these are the variables determined outside the system so that there is no equation that describes their behavior. Clearly, when continuous covariates are available, a linear regression model within the assumed link function must be used.

## 2.2 Discussion

An interesting instance of the models described above was used informally by Hagenaars (2002) as an extended latent class model. It may be interesting to note that, while in a basic latent class model the parameters which determine the marginal distribution of the latent are somehow separate from those which determine the conditional distribution of the responses, in the general context described here, in principle, any node of the DAG may correspond to a latent variable and, if there is a latent node  $Z_i$  which has no parents, its marginal distribution is determined by the  $\beta_{i0h}$ , the intercept parameters for the adjacent logits, whose number equals the number of latent categories minus 1.

A different, but closely related literature is that based on finite mixture models, like those developed, for instance, in Alfò and Trovato (2011) where a selection variable and two or more response variables are assumed to depend on a multivariate continuous latent distribution. However, when the underlying distribution is approximated with a discrete distribution with  $K$  support points, the resulting model is equivalent to a DAG model with a single discrete latent variable, say  $U$ ; the special case where there are two responses  $Y_1, Y_2$  and a selection variable  $Y_0$  is displayed in the DAG below



It is worth noting that the true multivariate nature of the underlying latent, once turned into a discrete one, should show up in the values of the estimated intercept parameters  $\beta_{ijl}$ , where  $i$  indexes the response variable,  $j$  the latent and  $l$  the category of the latent; the fact that  $\beta_{ijl}$  is positive (or negative) for all  $i, l$  indicates that the underlying latent is essentially uni-dimensional.

## 3 Maximum likelihood estimation

Under the assumption that, conditionally on exogenous covariates, the joint distribution of the variables (both the observable and the latent ones) in the DAG is multinomial, any identifiable model may be fitted by an EM algorithm. In the E-step we update the hypothetical latent distribution on the basis of the posterior probabilities that the subjects with a given observed response profile belong to each possible latent configuration and in the M-step we maximize the multinomial likelihood of the latent distribution.

In spite of the rather complex framework, the E-step has the familiar form of the product of the observed frequencies times the estimated posterior probabilities. Let  $\hat{\pi}_{h|j}(i)$  denote the probability of belonging to latent configuration  $h$  conditionally on having observed configuration  $j$  for the  $i$ th unit, where  $j$  and  $h$  denote, respectively, a given cell of the observed and

latent frequency table; let also  $N_j(i)$  denotes the observed frequency in cell  $j$  for the  $i$ th unit, the reconstructed frequency table is given by

$$M_{j,h}(i) = N_j(i) \frac{\hat{\pi}_{h|j}(i)}{\sum_h \hat{\pi}_{h|j}(i)}.$$

Due to the recursive nature of this class of models, the M-step may be performed by maximizing the conditional likelihood of each endogenous variable conditionally on its parents and on exogenous variables. An efficient algorithm for fitting these generalized logistic models is described in Evans and Forcina (2012), section 4.

Though the theory required to implement the EM algorithm to our models is straightforward, the difficulty lies in setting up a software that can perform these tasks efficiently having as input a general DAG with an arbitrary number of latent variables. Essentially, in the E-step we first need to compute the marginal probability distribution of the observed variables and then expand this back into the joint distribution while, in the M-step, we first need to compute, for each node, the conditional distribution of the response variable given its parents and, at the end, reconstruct the joint distribution recursively. The basic idea is to arrange probabilities and frequencies in lexicographic order so that the categories of  $Z_j$  run faster than those of  $Z_i$  if  $j > i$ . Marginal distributions are computed by first rearranging entries into a two-way table where the variables to be retained are by column and then summing across rows. Expansion of a smaller table into a larger one is performed first by replicating each entry a number of times equal to the number of cells of the omitted variables and then rearranging entries according to the original ordering of variables. Rearrangement of cells are performed by suitable indices which are constructed before starting the algorithm. The MATLAB functions that implement the EM algorithm on a general DAG will be made available as supplementary material.

To start the algorithm, an initial E-step is performed by assuming the the posterior probabilities  $\pi_{h|j}(i)$  are uniform, except for a small random perturbation. In the initial M-step a one-step ahead logistic model is fitted and estimates are adjusted to smooth possibly large absolute values. In this way an initial estimate of the latent distribution is obtained. With some expertise, the models described in this paper could also be fitted with the LG-Syntax module described by Vermunt and Magidson (2008)

The methodology described by Bartolucci and Forcina (2006), section 3.3, was used to compute standard errors of the parameter estimates from the estimate of the expected information matrix. The idea is to collect all parameters into the vector  $\beta$ , to compute the score vector of the log-likelihood for the observed distribution by the chain rule and the information matrix as follows

$$s = \frac{\partial L(\beta)}{\partial \gamma} \frac{\partial \gamma}{\partial \theta'} \frac{\partial \theta}{\partial \beta'}, \quad F = E(ss'/n),$$

where  $\gamma$  is the vector of log-linear parameters for the saturated log-linear model of the observed distribution,  $\theta$  is the vector of log-linear parameters for the latent distribution and  $F$  is the expected information matrix. The extension of this procedure to a general DAG model is a rather complex task which is handled by specific routines which exploit the rearrangement indices mentioned above.

#### 4 Evaluation of causal effects

In this paper we formulate the questions of interest and compute appropriate answers within the formal language developed by J. Pearl (see for example Pearl 2000, Chapter 3) which

we summarize briefly below. It may be useful to note that, in Pearl’s framework, the joint distribution of the observed variables in the DAG is assumed to be known, or estimated from observed frequencies, and the formal language is required to evaluate causal effects by taking into proper account the causal relations described by the DAG. The fact that certain variables are or are not endogenous, is irrelevant when we estimate the statistical model, as long as the conditional independencies implied by the DAG are true. However, while in a non-parametric context certain causal effects may not be estimable from the joint distribution of the observable variables, in our semi-parametric framework, once the statistical model is identifiable, any causal effect of interest may be easily computed from the estimated latent distribution.

In a structural equation model, see equation (1), we may evaluate the causal effect of a subset of variables  $X = (Z_i)_{i \in I}$  on  $Y = (Z_j)_{j \in J}$ , with  $J$  disjoint from  $I$ , by first applying to the “do operator”

$$P(y | do(x)) = \sum_{i \notin I \cup J} P(z_1, \dots, z_n | do(x)),$$

this is equivalent to determine the distribution that would arise if we could perform an idealized experiment where the variables in  $X$  were randomized. Once the intervention distribution has been constructed, we need to choose how to compare distributions of  $Y$  for different values of  $x$ : the two most obvious alternatives are differences or ratios of the relevant probabilities. Because in the application we deal with ordered categorical distributions, we simply compute the ratio of the corresponding survival probabilities.

#### 4.1 Direct effects

In a complex DAG causal effects may act through several different pathways, and we may be interested in assessing the effects that act along certain specific paths. Consider, for instance, the model described in Table 1 presented in section 5. There,  $S_p$  (parents’ education) affects  $S_c$  (child education) directly, or by affecting  $U_c$  (child latent endowment) or  $Y$  (family income) which, in turn, affect  $S_c$ . The effect of  $U_p$  (parents’ latent endowment) travels through many channels, but we would mainly be interested in its effect on  $S_c$  while observed family backgrounds is held fixed, to capture the effect of natural inheritance, that is the path from  $U_p$  to  $S_c$  going through  $U_c$ .

Effects exerted through specific paths are called ‘direct effects’. In the literature different definitions of direct effects have been considered; the one used in our application is the ‘natural direct effect’ which is defined as follows (see for example Pearl (2000) Definition 4.5.1 or Pearl (2010) section 6.1.3). Suppose we are interested in the causal effect of a set of variables  $X$  on  $Y$  exerted through all paths except those going through a set of mediating variables  $M = (Z_i)_{i \in K}$ , with  $K$  disjoint from  $I, J$ . Then, first we compute the intervention distribution obtained by setting  $X = x$  and  $M = m$

$$P(y | do(x), do(m)) = \sum_{i \notin I \cup J \cup K} P(z_1, \dots, z_n | do(x), do(m));$$

the effect of  $M$  is then averaged out, with weights provided by the distribution of  $M$  when  $X$  is set to its reference category by intervention.

Computation of direct effects requires the computation of several intervention distributions, a task that is similar to the one implemented within the EM algorithm described above to reconstruct the joint distribution. In practice, the basic ingredients are the DAG structure and, for each node, the estimated conditional distribution given its parents. Then, nodes are processed one at a time to reconstruct the required intervention distribution.

## 5 Application to Education transmission

The question of assessing the effect of raising the education of the parents by policy intervention on the education of their children is difficult because the answer depend on the extent to which the association between parents' and children's education is due to the transmission of unobservable endowments across generations.

### 5.1 Background and Literature

For simplicity, consider the very simple model in the four variables  $S^p, U^p$  and  $S^c, U^c$ , which denote schooling and unobservable endowments respectively for parents and child, while  $\varepsilon^p, \varepsilon^c$  are exogenous errors and assume that

$$S^c = f(S^p, U^c, \varepsilon^c) \quad (3)$$

$$U^c = g(S^p, U^p, \varepsilon^p). \quad (4)$$

This model says that a child's education depends on her own endowment and her parents' education, and in turn the child's endowment depends on her parents' schooling and endowment. Under this model the observed association between  $S^p$  and  $S^c$  is partly due to the effect of endowment on schooling within each generation combined with the transmission effect from  $U^p$  to  $U^c$ . Thus the stronger the endowment transmission effect the weaker the scope of education policy. One could substitute from equation (4) into (3) to get the reduced form equation

$$S^c = f(S^p, U^p, \varepsilon) \quad (5)$$

which requires controlling only for parents' endowment. Three main approaches in this direction have been pursued. Behrman and Rosenzweig (2002) take differences between subjects with twin mothers, having adjusted for assortative mating in order to control for differences between education of fathers; Plug (2004) uses data on adoptees under the assumption that there should be no endowment transmission, although, as noted by Holmlund et al (2011), association may be induced by selective placement of adoptees. Finally, Black et al (2005) analyze a dataset where differences in parent's education was exogenously induced by reforms in municipal schooling laws which they used as an instrument. For a critical assessment see Holmlund et al (2011) who apply the three methods to a single data set and show that they produce conflicting results.

Alternatively one could estimate equation (3) in isolation, which requires controlling only for the child's endowment. By fitting a much more complex version of (3), Cameron and Heckman (1998) address the issue of how the family background affects the probability of transition from one grade of education to the next. Though their model resembles (3) the heterogeneity is assumed independent from the observed covariates, so it could be interpreted as the component of  $U^c$  which is not determined by family background.

The variable  $U^p$ , named *family endowment*, is essentially identified by the variables it affects, so it is meant to capture the family environment in which children grow up. It is in principle a cross classification of various characteristics of the family, but in practice it turns out to be naturally ordered in a scale of 'quality'. The child's unobservable  $U^c$  is identified mainly through cognitive and non-cognitive test scores, so it is not to be interpreted as strictly reflecting an individual intrinsic endowment; it is rather a mixture of this and other unobservables like motivation and acquired knowledge useful for schooling advancement.



## 5.2 Data

We use data from the National Child Development Survey (NCDS), produced by a UK cohort study targeting the population born in the UK between the 3rd to the 9th of March 1958. Individuals were surveyed at different stages of their life and information on their schooling achievement, various tests results and family background was collected. A complete description of the data is available at

<http://www.esds.ac.uk/longitudinal/access/ncds>.

Some variables are inherently discrete (notably schooling level) while others would be more naturally described as continuous, like income and test scores. Because the finite mixture model approach used in this paper can be applied only when all endogenous variables are categorical, continuous variables were turned into discrete. Though clearly a continuous variable contains more information relative to a discrete approximation, there are two reasons why a model based on categorical variables may involve less parametric restrictions than one based on the original continuous measurements. First, a continuous variable used as explanatory in a regression model implies linearity unless additional polynomial terms are introduced; instead, once it has been transformed into a set of dummy variables corresponding to discrete categories, it can capture patterns of non linearity in a non parametric way. Models involving a continuous variable as response are usually based on the rather restrictive assumption of normality while, when used as categorical, the discrete distribution is assumed to be multinomial, that is completely unrestricted, at least in the first stage.

The original sample contains 18560 observations, but more than 80% have at least a missing entry. Incompleteness is scattered across many variables included in the survey. The subsample of complete data which we analyze amounts to almost 3000 subjects, 1471 males (sons) and 1330 females (daughters). The marginal distributions of the summary statistics for the most relevant variables in the complete-case sub-sample do not differ significantly from the same distributions in the whole sample, but we cannot really exclude selection bias. Our main dependent variable  $S^c$  is the amount of education achieved by each individual, which takes four levels: no qualification, O-level, A-level and higher education.

Children are tested at the age of 7 and 11 for mathematics, reading and non-cognitive skills, and again at 16 for math and reading, and we use the test scores for identification of the unobservable endowment. More specifically, after taking principal components (which in all cases explain no less than 90% of the total variance) for math and reading we combine scores at 7 and 11 into two ordered variables:  $EM$  and  $ER$ . Math and reading scores at 16 are coded in two additional variables  $LM$  and  $LR$ . For *non-cognitive skills* (available at ages 7 and 11), principal components yields two factors; these were averaged and then dichotomized at the median into the binary variable  $NC$ .

Parents' schooling is defined as the age at which they left school (12 to 21 years); for each parent we extract a three level variable corresponding to significant educational steps: leaving up to 14 years of age; after 14 but not later than 16; after 16; these are called  $S^m$ ,  $S^f$  for mother and father respectively. As usual there are many missing data on family income; to alleviate the problem, since few mothers in the dataset have an income, we neglect mother's income (thus avoiding to drop data with missing mother's income) and concentrate on fathers'. We group their income in three categories into the ordered variable  $Y$ .

The NCDS contains also information on parents' interest in their children's education, as reported by teachers; this turns out to be an important variable; it should measure the amount of effort or concern, and, perhaps, is related to the value that family gives to the



child's education. Parents' interest is originally classified into as many as 5 categories; we extract three binary parents' interest variables,  $I^7$ ,  $I^{11}$ ,  $I^{16}$ .

## 6 The Model

We estimate a system of equations which is an extended and highly complex version of (3) and (4); because of its complexity, it is convenient to summarize the basic features of the model in table 1 below where, for each variable in the DAG we give the number of categories, the type of logit ( $g$  for global and  $a$  for adjacent) and the list of parents. Note that

**Table 1** Description of the model

$i$	$Z_i$	n.cat.	logit	$pa_i$
1	$U^p$	3	$a$	-
2	$I^7$	2	$g$	$U^p$
3	$I^{11}$	2	$g$	$U^p$
4	$I^{16}$	2	$g$	$U^p$
5	$S^m$	3	$g$	$U^p$
6	$S^f$	3	$g$	$U^p, S^m$
7	$Y$	3	$g$	$U^p, S^m, S^f$
8	$U^c$	3	$a$	$U^p, I^7, I^{11}, I^{16}, S^m, S^f$
9	$EM$	3	$g$	$U^c$
10	$LM$	3	$g$	$U^c, EM$
11	$ER$	3	$g$	$U^c$
12	$LR$	3	$g$	$U^c, ER$
13	$NC$	2	$a$	$U^c$
14	$s^c$	4	$g$	$S^m, S^f, Y, U^c$

there is an arrow from  $S^m$  to  $S^f$  to account for assortative mating. In the fitted model the dependence of each node on its parents is assumed linear on the appropriate logit link transformation. In particular, because all observable variables in the system are naturally ordered, we use cumulative (or 'global') logits. The levels of unobservable variables, instead, are assumed to correspond to unordered qualitative types, so we use adjacent logits. Separate models were fitted for daughters and sons to account for gender effects.

### 6.1 Main Estimation Results

In Table 2 we display some of the most relevant parameter estimates from different structural equations included in the model which we fitted to data on sons and daughters separately. First of all note that all the  $\beta_{i1h}$  parameters are negative and usually significant, this indicates that the three parents' latent class may be ordered from best to worst; the only exception are the  $\beta_{81h}$ s which, being positive, indicate that the child's latent class may also be ordered from best to worst. This is in agreement with the fact that all the other  $\beta_{8jh}$  are negative, indicating that increasing rearing efforts and higher education on the parents' side are positively associated with an improvement of the latent endowment of the child. The few displayed parameters from the equations for early and late score in math indicate that better endowed children get better score and that performances are correlated in time. Finally, for the educational achievements, the displayed estimates confirm that more endowed children get higher

**Table 2** Parameter estimates and standard errors(se) for sons and daughters

	Daughters		Sons	
	coeff	se	coeff	se
$\beta_{411}$	-1.4133	0.3007	-1.3860	0.2721
$\beta_{412}$	-2.7674	0.3951	-2.2284	0.2546
$\beta_{511}$	-2.5027	0.2776	-2.4058	0.2285
$\beta_{512}$	-0.7036	0.1750	-0.6301	0.1437
$\beta_{611}$	-2.9051	0.3717	-5.7254	0.5329
$\beta_{612}$	-0.2087	0.2082	-0.4647	0.1749
$\beta_{811}$	1.2421	0.6426	0.1621	0.5733
$\beta_{812}$	2.8051	0.5843	1.4549	0.4474
$\beta_{841}$	-0.5668	0.2332	-0.8648	0.1737
$\beta_{852}$	-0.6575	0.2957	-0.8067	0.2468
$\beta_{862}$	-0.0092	0.3903	-0.6447	0.4554
$\beta_{981}$	-2.8296	0.1876	-2.8735	0.2047
$\beta_{982}$	-2.6018	0.1278	-2.5633	0.1204
$\beta_{10.81}$	-3.1387	0.2905	-3.0296	0.2333
$\beta_{10.82}$	-2.0644	0.2211	-1.5633	0.2147
$\beta_{10.91}$	0.3476	0.1984	0.5046	0.1842
$\beta_{14.51}$	0.1585	0.1491	-0.0599	0.1422
$\beta_{14.52}$	0.3622	0.2009	-0.1078	0.2000
$\beta_{14.61}$	0.0818	0.1539	0.3573	0.1464
$\beta_{14.62}$	0.1118	0.2050	0.6453	0.2129
$\beta_{14.81}$	-1.8843	0.1831	-3.0141	0.2135
$\beta_{14.82}$	-2.8394	0.2107	-1.7979	0.2177

achievements. The parameter estimates for the association with the education of the parents are less obvious to interpret: essentially we see that while the association with the education of the father is positive and significant for the son, the association for the daughter is close to 0 and smaller than the association with the education of the mother. These results may be interpreted as indicating a possible gender (or role) effect which may act either as pressure from the related parent or as an effort of emulation. A more specific interpretation of these results within the context of causal inference is described in the next section.

## 6.2 Estimated direct effects

The results are presented in Table 3, where estimates for sons and daughters are considered separately. The comparisons are expressed as ratios of survival probabilities, so, for example the upper-left value of 1.3640 says that the probability that a girl reaches education level at least 1 when  $R = (1, 1, 1)$  is 1.3640 times larger than when  $R = (0, 0, 0)$ . The effect of parents' education is calculated excluding the income path, so it includes the indirect effect exerted via child's endowment.

**Table 3** Causal effects on  $S^c$ 

	Daughters		Sons	
	$S^c > 0$	$S^c > 2$	$S^c > 0$	$S^c > 2$
<i>rearing efforts</i> $R = (I^7, I^{11}, I^{16})$				
from min to max	1.3640	1.8120	1.7379	2.7337
<i>Separately for the three components</i>				
$R^7$ from 0 to 1	1.0398	1.0733	1.0780	1.1263
$R^{11}$ from 0 to 1	1.1761	1.3554	1.2824	1.5182
$R^{16}$ from 0 to 1	1.1023	1.1943	1.2373	1.4215
<i>Mother's schooling</i> $S^m$				
from 0 to 1	0.9383	0.9144	0.9893	0.9699
from 1 to 2	1.2005	1.5480	1.1840	1.2741
from min to max	1.1264	1.4155	1.1713	1.2357
<i>Father's schooling</i> $S^f$				
from 0 to 1	1.0593	1.1330	1.1746	1.4384
from 1 to 2	1.0259	1.0748	1.3195	1.9446
from min to max	1.0867	1.2177	1.5499	2.7971
<i>Income</i> $Y$				
from 0 to 1	1.0275	1.0773	1.0071	1.0178
from 1 to 2	1.0546	1.1596	1.0720	1.1844
from min to max	1.0836	1.2493	1.0796	1.2055

## References

- Alfö M, Trovato G (2011) A finite mixture models for multivariate counts under endogenous selectivity. *Statistical Computations* 21:185–202
- Allman ES, Matias C, Rhodes JA (2009) Identifiability of parameters in latent structure models with many observed variables. *The Annals of Statistics* 37:3099–3132
- Bartolucci F, Forcina A (2006) A class of latent marginal models for capture-recapture data with continuous covariates. *Journal of the American Statistical Association* 101:786–794
- Behrman JR, Rosenzweig M (2002) Does increasing women's schooling increases schooling of the next generation? *American Economic Review* 92:323–334
- Black SE, Devenaux PJ, Kjell GS (2005) Why the apple doesn't fall far: understanding intergenerational transmission of human capital. *American Economic Review* 95:437–449
- Colombi R, Forcina A (2001) Marginal regression models for the analysis of positive association of ordinal response variables. *Biometrika* 88(4):1001–1019
- Evans R, Forcina A (2012) Two algorithms for fitting constrained marginal models. Tech. rep., <http://arxiv.org/>
- Forcina A (2008) Identifiability of extended latent class models with individual covariates. *Computational statistics and data analysis* 52:5263–5268
- Hagenaars JA (2002) *Applied latent class analysis*, Cambridge University press, chap 9, pp 234 – 286
- Holmlund H, Lindahl M, Plug E (2011) The causal effect of parents' schooling on children's schooling: A comparison of estimation methods. *Journal of Economic Literature* 49:615–651
- Pearl J (1995) Causal diagrams for empirical research. *Biometrika* 82:669–688
- Pearl J (2000) *Causality*. Cambridge University Press
- Pearl J (2010) The mediation formula: a guide to the assessment of causal pathways in non linear models. Tech. rep., Dept. of Computer Science, Univ. of California, Los Angeles
- Plug E (2004) Estimating the effect of mother's schooling using a sample of adoptees. *The American Economic Review* 94:358–368
- Vermunt J, Magidson J (2008) *LG-Syntax user guide: manual for Latent Gold 4.5 syntax module*. Statistical Innovations inc, Belmont, MA