Available online at www.sciencedirect.com
SCIENCE DIRECT®

ELSEVIER

Journal of Econometrics 132 (2006) 379–407

JOURNAL OF Econometrics

www.elsevier.com/locate/jeconom

Non-causality in bivariate binary time series

Rocco Mosconi[a,*], Raffaello Seri[b]

[a]Dipartimento di Ingegneria Gestionale, Politecnico di Milano, P.za Leonardo da Vinci 32, 20133 Milano, Italy
[b]Dipartimento di Economia, Università degli Studi dell'Insubria, via Ravasi 2, 21100 Varese, Italy

Available online 15 April 2005

Abstract

In this paper we develop a dynamic discrete-time bivariate probit model, in which the conditions for Granger non-causality can be represented and tested. The conditions for simultaneous independence are also worked out. The model is extended in order to allow for covariates, representing individual as well as time heterogeneity. The proposed model can be estimated by Maximum Likelihood. Granger non-causality and simultaneous independence can be tested by Likelihood Ratio or Wald tests. A specialized version of the model, aimed at testing Granger non-causality with bivariate discrete-time survival data is also discussed. The proposed tests are illustrated in two empirical applications.

JEL classification: C50; C33; C35

Keywords: Granger non-causality; Multivariate choice models; Markov chains; Multivariate survival analysis

1. Introduction

The epistemological status of the statistical-probabilistic notion of causality based on predictability is still a matter of profound controversy among philosophers and methodologists (see Basmann, 1988, pp. 96–98; Geweke, 1984; Swamy and von zur

*Corresponding author. Tel.: +39 0223992747; fax: +39 02700423151.
E-mail addresses: rocco.mosconi@polimi.it (R. Mosconi), raffaello.seri@uninsubria.it (R. Seri).

0304-4076/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.jeconom.2005.02.005

Muehlen, 1988). This notion fits, in a probabilistic sense, the two key aspects of the Hume theory of causation: the systematic conjunction of cause and effect, and the time precedence of the cause with respect to the effect. Nonetheless, it fails to account for what probably is the deepest, though empirically less helpful, aspect, i.e., the idea that the cause "forces" or "produces" the effect. Despite these limitations, the notion of causality based on predictability proved to be a valuable tool for applied research thanks to its operational usefulness in the construction, estimation, interpretation and application of econometric models.

In Economics, the notion of non-causality has mainly been used to model macroeconomic variables, and hence one single realization of the processes involved in the analysis is usually assumed, such processes are assumed to be continuous and exogenous processes are seldom included in the information set. In this framework, non-causality is usually tested assuming that the process of interest belongs to the class of Vector ARIMA processes. On the other hand, in microeconometric applications, where the variables are often qualitative and where longitudinal data are usually available, the VARIMA framework is not appropriate, and including covariates to account for individual heterogeneity becomes an essential aspect of modelling. Therefore, a set of ad hoc tools has to be developed in order to make the usual definitions of non-causality operational.

In this paper, we discuss non-causality analysis in a bivariate discrete-time binary process,[1] in a panel data setting, possibly allowing for covariates. This case has received a certain attention, but no general technique has been proposed in order to deal with it. In the econometric literature, the case of Markov chains without covariates is discussed in Chamberlain (1982), Bouissou et al. (1986) and in Gouriéroux et al. (1987). Furthermore, in the sociometric literature, models for studying interdependencies and causality relations between transitions are so relevant that Stolnitz (1983) advocated their construction and use as one of the five great challenges of demographic research. In particular, the special case of two-wave bivariate binary models is often called the "sixteen-fold table problem" and has received attention since Lazarsfeld (1948, see also McCullough, 1978). Specific models for the study of the interaction between two binary choice process have been proposed by Yamaguchi (1990); Lillard (1993) and Petersen (1995). In particular, Yamaguchi (1990) is the only one dealing explicitly with causality testing, but he adopts a definition different from Granger non-causality and hardly acceptable. As a general remark, it seems that one of the critical points in these papers is the modelling of simultaneity between transitions, since these models are often based on a non-stated assumption of strong simultaneous independence.

The paper is organized as follows. Section 2 illustrates in a very general setting the probability aspects of the definitions of non-causality based on predictability. Next, under the maintained assumption that the process of interest is a Markov chain with stationary transition probabilities, and that the information set is restricted to the

[1] The case of continuous-time counting processes, fully observed or subject to grouping and censoring, is addressed in the literature (Schweder, 1970; Aalen et al., 1980; Tuma, 1980; Winship, 1986; Aalen, 1987; Florens and Fougère, 1996).

history of the process. Section 3 shows how non-causality analysis may be performed in a dynamic bivariate probit model. Section 4 extends the simple dynamic probit model illustrated in Section 3 in two directions. First, the assumption of stationary transition probabilities is dropped, allowing the transition probabilities to depend on covariates. Then, the first-order Markov assumption is also relaxed, allowing for more complex dynamic structures. Section 5 shows under which conditions the Maximum Likelihood estimates of the parameters of the proposed models, as well as the Likelihood Ratio tests for hypotheses on such parameters, display the usual asymptotic properties. Possible problems in finite samples are also illustrated. Section 6 shows how the proposed analysis does specialize when one is interested in a specific four states Markov chain, corresponding to discrete-time bivariate survival data. Sections 7 and 8 illustrate the proposed methodology, using respectively data about marriage and fertility timing in a sample of 266 American women and about the adoption of two interrelated technologies by 552 Italian metalworking plants. Section 9 concludes.

2. Some preliminary definitions

In a general setting (see Florens and Fougère, 1996), a mathematically rigorous definition of non-causality based on predictability requires the specification of the stochastic process to be predicted, the available information set, and the reduced information set. Although several generalizations exist, we will briefly review here the concept of discrete-time one step ahead strong non-causality (the terminology is drawn from Florens and Fougère, 1996). Here *one step ahead* (as opposed to *global*) is referred to the prediction horizon, whereas *strong* (as opposed to *weak*) means that the focus is on predicting the whole distribution, rather than just the mean. Notice that Granger's (1969) original definition is stated in terms of the mean. Chamberlain (1982) and Florens and Mouchart (1982) propose the definition involving the whole distribution (see also Granger, 1988).

Let $\{Y_t = (Y_t^1, Y_t^2),\ t \in I \subseteq \mathbb{N} = \{1, 2, \ldots\}\}$, or $\{Y_t\}$ for short,[2] be a discrete-time stochastic process on a probability space $(\Omega, \mathscr{A}, \mathbb{P})$. The statistical problem of non-causality is to test whether $\mathbb{P}$ satisfies non-causality conditions. The available information is described by the *filtration* $\{\mathscr{F}_t, t \in I\} = \{\mathscr{F}_t\}$. For simplicity, we will assume here that $\{\mathscr{F}_t\}$ is the *canonical filtration* associated with the stochastic process $\{(Y_t, X_t)\} = \{(Y_t^1, Y_t^2, X_t)\}$,[3] where $\{Y_t^1\}$, $\{Y_t^2\}$ and $\{X_t\}$ may either be scalar or vector processes. The reduced information set is represented by the canonical filtrations $\{\mathscr{G}_t^1\} = \{\sigma\{(Y_s^1, X_s), 1 \leqslant s \leqslant t\}\}$ and $\{\mathscr{G}_t^2\} = \{\sigma\{(Y_s^2, X_s), 1 \leqslant s \leqslant t\}\}$. Let then $\{\mathscr{Y}_t^1\}$, $\{\mathscr{Y}_t^2\}$

[2] The following notation is used through the paper: $\{Z_t\}$ denotes a stochastic process, Z_t being the value of the process at time t; $\{z_t\}$ and z_t represent the corresponding realizations. Moreover, $\mathbb{P}\{z_t | w_t\}$ is adopted as a short notation for $\mathbb{P}\{Z_t = z_t | W_t = w_t\}$. The equality between random variables is always to be intended almost surely.

[3] The *canonical* (or *self exciting*) filtration associated with the process $\{Z_t\}$ defined on $(\Omega, \mathscr{A}, P)$ is a family $\{\mathscr{F}_t\}$ of sub-σ-fields of $\mathscr{A}$, where $\mathscr{F}_t = \sigma\{Z_s, 1 \leqslant s \leqslant t\}$. Intuitively, $\mathscr{F}_t$ represents the history of $\{Z_t\}$ up to time t.

and $\{\mathscr{Y}_t\}$ be the canonical filtrations associated with the processes $\{Y_t^1\}$, $\{Y_t^2\}$ and $\{Y_t\}$, respectively. Notice that $\mathscr{Y}_t^1 \subseteq \mathscr{G}_t^1 \subseteq \mathscr{F}_t$, $\forall t \in I$, and similarly $\mathscr{Y}_t^2 \subseteq \mathscr{G}_t^2 \subseteq \mathscr{F}_t$, $\forall t \in I$.

In the paper, we will adopt the following definitions, stated in terms of conditional independence of sub-σ-fields of $\mathscr{A}$ (see Florens and Mouchart, 1982, Appendix, for the relevant results about conditional independence).

Definition 2.1. Strong one step ahead Granger non-causality: $\{Y_t^2\}$ *does not strongly cause* $\{Y_t^1\}$ *one step ahead, given* $\{\mathscr{G}_{t-1}^1\}$, briefly $Y^1 \nleftarrow Y^2$, if

$$\mathscr{Y}_t^1 \perp\!\!\!\perp \mathscr{Y}_{t-1}^2 \mid \mathscr{G}_{t-1}^1 \quad \forall t \in I. \tag{1}$$

Similarly, $\{Y_t^1\}$ *does not strongly cause* $\{Y_t^2\}$ *one step ahead, given* $\{\mathscr{G}_{t-1}^2\}$, briefly $Y^1 \nrightarrow Y^2$, if

$$\mathscr{Y}_t^2 \perp\!\!\!\perp \mathscr{Y}_{t-1}^1 \mid \mathscr{G}_{t-1}^2 \quad \forall t \in I. \tag{2}$$

Definition 2.2. Strong simultaneous independence: $\{Y_t^1\}$ and $\{Y_t^2\}$ are *strongly simultaneously independent given* $\{\mathscr{F}_{t-1}\}$, briefly $Y^1 \nleftrightarrow Y^2$, if

$$\mathscr{Y}_t^1 \perp\!\!\!\perp \mathscr{Y}_t^2 \mid \mathscr{F}_{t-1} \quad \forall t \in I. \tag{3}$$

Notice that the term *simultaneous* in the latter definition has exactly the same meaning as *instantaneous* in Geweke (1984) and Granger (1988). A different term is suggested here since Florens and Fougère (1996) observe that one step ahead non-causality in discrete time has an analogue in continuous time when the time distance between "cause" and "effect" goes to zero, a circumstance that they define as *instantaneous causality*. Therefore, they use instantaneous as a synonym for discrete-time *one step ahead causality*, while they do not give any definition similar to (3). Moreover, for the simultaneous condition (3), the term dependence is proposed instead of causality (as in Granger, 1988) or feedback (as in Geweke, 1984), since the notion is completely a-directional in nature (not even bi-directional).

In the macroeconometric literature, in order to make such general definition operational, $\{Y_t\}$ is assumed to be a continuous process belonging to the class of Vector ARIMA processes, exogenous processes $\{X_t\}$ are seldom included in the information set (so that $\mathscr{G}_t^1$ and $\mathscr{G}_t^2$ do coincide with $\mathscr{Y}_t^1$ and $\mathscr{Y}_t^2$, respectively), and one single realization of $\{Y_t\}$ is observed. Conversely in a microeconometric discrete choice setting, we will assume that N individual realizations $(i = 1, \ldots, N)$ of the process are observed, with $t = 1, \ldots, T$. As we will see, depending on the dynamic structure of the model, N may have to be large with respect to T, but if very simple dynamic structures are assumed, a small N, or even $N = 1$, can be enough if T is large. Notice that, in microeconometric settings, $\{X_t\}$ is needed to model individual heterogeneity, and may well include some time-fixed variables.

In our framework, at any time $t \in \{1, \ldots, T\}$, the state space of $Y_t = (Y_t^1, Y_t^2)$ is given by the following states: $\{(0,0), (1,0), (0,1), (1,1)\}$. Basically the model can be represented by the diagram in Fig. 1, where each box represents one of the four states where the process could belong at time $(t-1)$, and the arrows represent the transitions which may occur at time t.

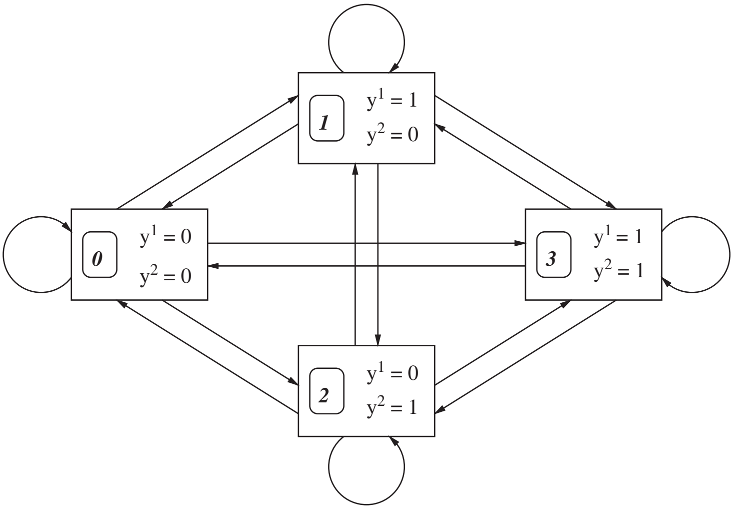

Fig. 1. State-transition diagram for a binary bivariate Markov model.

Let us illustrate how definitions (1)–(2)–(3) can be made operational by applying them to a precise stochastic process and information set. To make the simplest possible example, let us restrict the information set to the canonical filtration associated with $\{Y_t\}$, and furthermore make the assumption that $\{Y_t\}$ is a *first-order Markov process* (or *Markov chain*), so that $\mathbb{P}\{y_t|y_{t-1},\dots,y_1\} = \mathbb{P}\{y_t|y_{t-1}\}$. The most restrictive definition of Markov process requires that the transition probabilities do not vary over time. More specifically, under this assumption the process is defined a Markov chain with *stationary transition probabilities*. Notice that this alone does exclude any impact of covariates on the transition probabilities. In this simplified framework, the definitions given above specialize as follows.

Definition 2.3. Strong one step ahead non-causality for a Markov chain with stationary transition probabilities: Y_{t-1}^2 *does not strongly cause* Y_t^1 *one step ahead, given* Y_{t-1}^1, if[4]

$$\mathbb{P}\{y_t^1|y_{t-1}\} = \mathbb{P}\{y_t^1|y_{t-1}^1\} \quad \forall t \in \{2,\dots,T\}. \tag{4}$$

Similarly, Y_{t-1}^1 *does not strongly cause* Y_t^2 *one step ahead, given* Y_{t-1}^2 if

$$\mathbb{P}\{y_t^2|y_{t-1}\} = \mathbb{P}\{y_t^2|y_{t-1}^2\} \quad \forall t \in \{2,\dots,T\}. \tag{5}$$

Definition 2.4. Strong simultaneous independence for a Markov chain with stationary transition probabilities: Y_t^1 and Y_t^2 are *strongly simultaneously*

[4] The equivalence between (1) and (4) in this framework comes immediately by noticing that, under the Markov assumption and the assumption that the information set $\mathscr{F}_{t-1}$ coincides with $\mathscr{Y}_{t-1}$, the conditional independence statement (1) implies $\mathbb{P}\{y_t^1, y_{t-1}^2|y_{t-1}^1\} = \mathbb{P}\{y_t^1|y_{t-1}^1\} \cdot \mathbb{P}\{y_{t-1}^2|y_{t-1}^1\}$, $\forall t \in \{1,\dots,T\}$, which in turn implies (4). The same argument holds for the other definitions.

independent given Y_{t-1} if

$$\mathbb{P}\{y_t|y_{t-1}\} = \mathbb{P}\{y_t^1|y_{t-1}\} \cdot \mathbb{P}\{y_t^2|y_{t-1}\} \quad \forall t \in \{2, \ldots, T\} \tag{6}$$

or equivalently

$$\mathbb{P}\{y_t^1|y_t^2, y_{t-1}\} = \mathbb{P}\{y_t^1|y_{t-1}\} \quad \forall t \in \{2, \ldots, T\}$$

or equivalently

$$\mathbb{P}\{y_t^2|y_t^1, y_{t-1}\} = \mathbb{P}\{y_t^2|y_{t-1}\} \quad \forall t \in \{2, \ldots, T\}.$$

The appropriate statistical model where these conditions can be tested is the joint distribution of Y_t given Y_{t-1}. Granger non-causality conditions involve only the marginal distributions of Y_t^1 and Y_t^2 (conditional on Y_{t-1}), whereas testing for simultaneous independence requires the joint distribution to be fully specified, and compared to the product of the marginal distributions. Notice that, since Y_{t-1}, as well as Y_t, can belong to a finite set of four states, the most general model representing $\mathbb{P}\{y_t|y_{t-1}\}$ involves 16 parameters, corresponding to the transition probabilities from each of the states in $(t-1)$ to each of the states in t (or some one-to-one transformation of the transition probabilities). More precisely, since the sum of the transition probabilities for transitions starting from each of the states is equal to 1, only 12 parameters are enough to describe the conditional distribution completely.

3. A Markov dynamic bivariate probit model for homogeneous population

Essentially, the type of data set in which we want to check for non-causality consists in observations on the choices of N individuals facing two interacting binary choices in discrete time. It seems therefore natural to use, as a statistical model, a dynamic version of a bivariate discrete choice model. In this section, we introduce a dynamic bivariate probit model, derived using a latent regression approach.[5] Alternative specifications of the logit type have been considered, but they have been discarded since writing non-causality constraints seems much more difficult: this will be briefly discussed at the end of this section.[6]

In this section, the following assumptions are maintained:

- the population is homogeneous (no covariates are introduced);
- the process is first-order Markov (all the information from the history of the process which is relevant for the transition probabilities in t is represented by the state of the process in $(t-1)$).

[5] The univariate static probit model is well known. The first attempt to extend the model in a multivariate direction based on a latent regression approach is due to Ashford and Sowden (1970). Dynamic versions of the univariate probit model are discussed, for example, in Heckman (1978, 1981).

[6] We refer to Amemiya (1981) for a survey of multivariate binary regression models.

These simplifying assumptions will both be relaxed in Section 4.

In order to use the bivariate probit setting to represent the distribution of $Y_{i,t} = (Y^1_{i,t}, Y^2_{i,t})^\mathsf{T}$ conditionally on the state of the system in $(t-1)$, it is convenient to remark that the state in $(t-1)$ can be defined by the four dimensional function of $y_{i,t-1}$:

$$s_{i,t-1} = (1, y^1_{i,t-1}, y^2_{i,t-1}, y^1_{i,t-1}y^2_{i,t-1})^\mathsf{T}.$$

In fact, $s_{i,t-1}$ is an invertible linear transformation of

$$s^*_{i,t-1} = [(1-y^1_{i,t-1})(1-y^2_{i,t-1}), y^1_{i,t-1}(1-y^2_{i,t-1}), (1-y^1_{i,t-1})y^2_{i,t-1}, y^1_{i,t-1}y^2_{i,t-1}]^\mathsf{T},$$

where $s^*_{i,t-1}$ involves four mutually exclusive dummies representing the four states of the process in $(t-1)$.[7] The reason for using $s_{i,t-1}$ instead of $s^*_{i,t-1}$ (or $y_{i,t-1}$ directly) to describe the state in $(t-1)$ is that, by doing so, the non-causality restrictions are more easily written and interpreted.

Each individual i has to make two binary choices at time t, i.e., to choose the value of the binary bivariate vector $Y_{i,t}$. The latent regression approach assumes that the individual will choose $Y^1_{i,t} = 1$ when a latent continuous random variable $Y^{1*}_{i,t}$ crosses a threshold which, with no loss of generality, is set equal to zero. The same holds for $Y^2_{i,t}$. In the current framework, the distribution of the latent variables is assumed to depend on the choice made in $(t-1)$. The latent regression is therefore given by:

$$\begin{aligned} y^{1*}_{i,t} &= \beta_1^\mathsf{T} s_{i,t-1} + \varepsilon^1_{i,t}, \\ y^{2*}_{i,t} &= \beta_2^\mathsf{T} s_{i,t-1} + \varepsilon^2_{i,t}. \end{aligned} \tag{7}$$

As usual in the multivariate probit setting, a standardized bivariate normal distribution is then assumed for $\varepsilon_{i,t} = (\varepsilon^1_{i,t}, \varepsilon^2_{i,t})$:

$$\varepsilon_{i,t} = \begin{pmatrix} \varepsilon^1_{i,t} \\ \varepsilon^2_{i,t} \end{pmatrix} \sim \text{iid}\mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho_{i,t} \\ \rho_{i,t} & 1 \end{bmatrix} \right), \tag{8}$$

where the correlation $\rho_{i,t}$ is assumed to depend on the state $s_{i,t-1}$ as follows:

$$\rho_{i,t} = \frac{2\exp(\gamma^\mathsf{T} s_{i,t-1})}{1+\exp(\gamma^\mathsf{T} s_{i,t-1})} - 1. \tag{9}$$

The logit-type functional form in (9) is chosen so as to bound the correlation coefficient between -1 and 1 and is known as z-transformation (see Fisher, 1921; Hotelling, 1953): other choices are possible (see e.g. Morimune, 1979). Standardizing the variances to be equal to 1 is needed for identification purposes, and implies no loss of generality. Notice that the assumption that $\varepsilon_{i,t}$ is independently distributed

[7] Obviously, $s_{i,t-1} = Qs^*_{i,t-1}$, with

$$Q = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

matches perfectly the Markov assumption, since failure of this condition means that there is some information left in the history of the process after conditioning on $s_{i,t-1}$.

A compact notation for the latent regression is

$$y_{i,t}^* = B^{\mathsf{T}} s_{i,t-1} + \varepsilon_{i,t}, \quad \varepsilon_{i,t} \sim \mathrm{iid}\mathcal{N}(0, R),$$

where $y_{i,t}^* = (y_{i,t}^{1*}, y_{i,t}^{2*})^{\mathsf{T}}, \varepsilon_{i,t} = (\varepsilon_{i,t}^1, \varepsilon_{i,t}^2)^{\mathsf{T}},\ B = [\beta_1, \beta_2]$ and $R = \begin{bmatrix} 1 & \rho_{i,t} \\ \rho_{i,t} & 1 \end{bmatrix}$.

The distribution of $Y_{i,t}$ can be elegantly written by introducing the diagonal matrix $D_{y_{i,t}} = 2\,\mathrm{diag}(y_{i,t}) - I_2$. It is easily computed that

$$\mathbb{P}(y_{i,t}|y_{i,t-1}) = \mathbb{P}(D_{y_{i,t}}(B^{\mathsf{T}} s_{i,t-1} + \varepsilon_{i,t}) > 0) = \mathbb{P}(-D_{y_{i,t}} \varepsilon_{i,t} < D_{y_{i,t}} B^{\mathsf{T}} s_{i,t-1}).$$

Our normality assumption for $\varepsilon_{i,t}$ implies that

$$\mathbb{P}(Y_{i,t} = y_{i,t}|y_{i,t-1}) = \Phi_2(D_{y_{i,t}} B^{\mathsf{T}} s_{i,t-1}; 0, D_{y_{i,t}} R D_{y_{i,t}}^{\mathsf{T}}), \tag{10}$$

where $\Phi_2(\cdot; \mu, \Sigma)$ denotes the integrated bivariate normal with mean μ and covariance matrix Σ.

Notice that B is a 4×2 matrix, while γ is a 4-dimensional vector; therefore, as a whole, the distribution (10) depends on 12 parameters freely varying in $\mathbb{R}^{12}$; such parameters can be easily shown to be a bijective transformation of the probabilities associated to the transitions illustrated in Fig. 1. Notice that the marginal distribution of $Y_{i,t}^1$ and $Y_{i,t}^2$ (given $y_{i,t-1}$) is given by

$$\mathbb{P}\{y_{i,t}^1|y_{i,t-1}\} = \Phi_1((2y_{i,t}^1 - 1)\beta_1^{\mathsf{T}} s_{i,t-1}; 0, 1), \tag{11}$$

$$\mathbb{P}\{y_{i,t}^2|y_{i,t-1}\} = \Phi_1((2y_{i,t}^2 - 1)\beta_2^{\mathsf{T}} s_{i,t-1}; 0, 1). \tag{12}$$

The conditions for strong one step ahead non-causality and strong simultaneous independence are easily stated as restrictions on the parameter space of (10):

$$\mathsf{H}_{1 \nleftarrow 2}\ (Y^1 \nleftarrow Y^2):\ \beta_1 = H_1 \varphi_1, \tag{13}$$

$$\mathsf{H}_{1 \nrightarrow 2}\ (Y^1 \nrightarrow Y^2):\ \beta_2 = H_2 \varphi_2, \tag{14}$$

$$\mathsf{H}_{1 \nLeftrightarrow 2}\ (Y^1 \nLeftrightarrow Y^2):\ \gamma = 0, \tag{15}$$

where

$$H_1 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}, \quad H_2 = \begin{bmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}. \tag{16}$$

Under $\mathsf{H}_{1 \nleftarrow 2}$, y_{t-1}^2 and $y_{t-1}^1 y_{t-1}^2$ are excluded from (11), so that $\mathbb{P}\{y_{i,t}^1|y_{i,t-1}\} = \mathbb{P}\{y_{i,t}^1|y_{i,t-1}^1\}$. Similarly, under $\mathsf{H}_{1 \nrightarrow 2}$, y_{t-1}^1 and $y_{t-1}^1 y_{t-1}^2$ are excluded from (12), so that $\mathbb{P}\{y_{i,t}^2|y_{i,t-1}\} = \mathbb{P}\{y_{i,t}^2|y_{i,t-1}^2\}$. Finally, under $\mathsf{H}_{1 \nLeftrightarrow 2}$, $\rho_{i,t}$ is equal to zero, and hence the joint distribution (10) factors out in the product of the marginal distributions (11) and (12).

To assess the degree of generality of the proposed probit model, it seems important to compare it with alternative specifications found in the literature, mainly belonging to the logit class. We consider two different approaches:

- Logistic model: apparently it would be simple to modify the bivariate dynamic probit model to get a bivariate dynamic logit model, by replacing the bivariate normal with a bivariate generalization of the logistic distribution (see e.g., Amemiya, 1981, p. 1531). However, the available generalizations, proposed in Gumbel (1961), have the disadvantage that the correlation is constrained to be equal to 0.5 for the Gumbel's type A bivariate logistic distribution and between $-\pi^2/3$ and $\pi^2/3$ for the Gumbel's type B bivariate logistic distribution (see Johnson and Kotz, 1972, pp. 291–294). This lack of flexibility in the distribution of the errors in the latent regression implies a restriction on the degree of simultaneous dependence of the observed process, and is therefore highly undesirable in general, and even more given the purpose of this paper.
- Log-linear model: this model is described, for example, in Morimune (1979, p. 957) and Amemiya (1981, p. 1528), and it is not derived using the latent regression approach. Unlike the previous one, it does not impose any restriction on the degree of correlation of the process. However, while the condition for strong simultaneous independence is easily written in the log-linear model, the conditions for strong one step ahead non-causality are not. In fact, strong one step ahead non-causality implies a rather complex non-linear constraint on the parameters.[8] This extra complexity does not seem justified in the absence of covariates, since in this case the bivariate log-linear model is completely equivalent to the bivariate probit model, both being one to one reparameterizations of the transition probabilities: therefore the maximized likelihood will be the same for the two models. Notice however that the equivalence fails when covariates are introduced, since the way covariates affect the transition probabilities is obviously different in the probit and logit models. It might therefore be useful, before the non-causality analysis is carried on, to test for the appropriateness of the probit specification, see for example Morimune (1979) or Murphy (1994). If the preliminary analysis selects the logit specification, the present approach needs to be adapted: a complete analysis of the logit specification is beyond the scope of this paper.

4. Introducing covariates and relaxing the first-order Markov hypothesis

In this section, the model presented in Section 3 will be extended in two directions. First we will relax the assumption of stationary transition probabilities by introducing covariates, in order to account for individual and/or time heterogeneity.

[8] Morimune (1979, p. 958) briefly illustrates a modified version of the logistic model where the marginal distributions are standard univariate logits. In this model, non-causality constraints are linear, but simultaneous independence implies a non-linear constraint.

This will be done under the first-order Markov assumption. Then we will drop the first-order Markov assumption to allow for more complex dynamics: this will be done in the absence of covariates. Relaxing both hypotheses is straightforward and left to the reader. Some general remarks about possible extensions of the model conclude the section.

4.1. Extending the information set

The information set available to predict Y_t is now enlarged to $\mathscr{F}_{t-1} = \mathscr{Y}_{t-1} \vee \mathscr{X}_{t-1}$.[9] It is important to remark that $\mathscr{X}_{t-1}$ could be replaced by $\mathscr{X}_t$, since this is simply a matter of time translation which, from a mathematical perspective, is completely irrelevant in the following discussion. However, for the economic interpretation of the results, if $\mathscr{F}_{t-1}$ contains information on covariates observed at time t (and not yet at time $t-1$), then this requires that these covariates are "valid conditioning variables", in the sense that the distribution of Y_t given the past and current X_t represents the economic mechanism that we want to analyse. This valid conditioning requirement is violated when Y_t and X_t are "jointly dependent", i.e., when X_t is an endogenous regressor in the usual econometric sense.

Let us first maintain the first-order Markov assumption, and for notational simplicity let us also assume, without loss of generality, that all the information in $\mathscr{X}_{t-1}$ which is relevant for the transition probabilities in t is given by X_{t-1}.

Extending model (10) so that the transition probabilities depend on x_{t-1} can be easily done replacing $s_{i,t-1}$ with

$$z^*_{i,t-1} = [s^{\mathsf{T}}_{i,t-1}, x^{*\mathsf{T}}_{i,t-1}]^{\mathsf{T}}, \tag{17}$$

where $x^*_{i,t}$ is the part of $x_{i,t}$ which is linearly independent of $s_{i,t}$ (typically, if $x_{i,t}$ includes the constant, it has to be dropped to avoid perfect collinearity with $s_{i,t}$). If we denote by k the dimension of $x_{i,t}$, and by k^* the dimension of $x^*_{i,t}$, then B and γ will be now of dimension $(4+k^*) \times 2$ and $(4+k^*) \times 1$. It is important to point out that this way to include the covariates amounts to assuming that the impact on the transition probabilities is the same irrespective of $s_{i,t-1}$, so that the effect of the covariates is the same whatever state the individual belongs to in $(t-1)$. A more general model, allowing for interaction among the covariates and the state of the process in $(t-1)$, i.e., $s_{i,t-1}$, ensues from using in (10), instead of $s_{i,t-1}$,

$$z_{i,t-1} = s_{i,t-1} \otimes x_{i,t-1}. \tag{18}$$

Notice that, in this case, B and γ will be of dimension $4k \times 2$ and $4k \times 1$, so that many more parameters have to be estimated. Conforming to a similar tradition in log-linear models, we will refer to the model deriving from (18) as *saturated* model, while the model deriving from (17) will be referred to as *unsaturated.* Notice that the unsaturated model is nested in the saturated one, and therefore the decision about which one is convenient for describing the data may be empirically based, for

[9] Let $\mathscr{M}_1$ and $\mathscr{M}_2$ be σ-fields. $\mathscr{M}_1 \vee \mathscr{M}_2$ denotes the σ-field generated by $\mathscr{M}_1 \cup \mathscr{M}_2$. Hence $\{\mathscr{F}_t\} = \{\mathscr{Y}_t \vee \mathscr{X}_t\}$ corresponds to the canonical filtration associated to $\{(Y_t, X_t)\}$.

example, on likelihood ratio tests. A simple example may help understanding the difference between the two models. Assume that each individual i belongs, at any time t, to either one or the other of two mutually exclusive and exhaustive classes C_1 and C_2. Define

$$D^1_{i,t} = \mathbb{1}_{\{\text{individual } i \in C_1 \text{ at time } t\}}, \quad D^2_{i,t} = \mathbb{1}_{\{\text{individual } i \in C_2 \text{ at time } t\}},$$

so that $D^1_{i,t} + D^2_{i,t} = 1$. Let $X_{i,t} = (D^1_{i,t}, D^2_{i,t})^{\mathsf{T}}$. In this setting, one may take $x^*_{i,t} = d^1_{i,t}$, and therefore $z^*_{i,t-1}$ and $z_{i,t-1}$ are respectively defined by (17) and (18). The most striking difference between the saturated and unsaturated model in this case is that, in the unsaturated model, $X_{i,t}$ (i.e., belonging to class C_1 or C_2) has the same impact (positive or negative or none) on the probability of $Y^1_{i,t}$ and $Y^2_{i,t}$ irrespective of the state in $(t-1)$. Conversely, in the saturated model, $X_{i,t}$ may have, say, a positive effect on the probability of $Y^1_{i,t}$ if $Y_{i,t-1} = (0,0)^{\mathsf{T}}$, and no effect on the probability of $Y^1_{i,t}$ if $Y_{i,t-1} = (1,0)^{\mathsf{T}}$.

The conditions for Granger non-causality in the presence of covariates are formally identical to (13) and (14), but the restriction matrices are defined as follows for the unsaturated model:

$$H^*_1 = \begin{bmatrix} H_1 & 0 \\ 0 & I_{k^*} \end{bmatrix}, \quad H^*_2 = \begin{bmatrix} H_2 & 0 \\ 0 & I_{k^*} \end{bmatrix},$$

while, for the saturated model the matrices are $H^*_1 = I_k \otimes H_1$ and $H^*_2 = I_k \otimes H_2$. It is easily checked that these restrictions matrices exclude all the regressors involving y^2_{t-1} from $\mathbb{P}\{y^1_{i,t} | y_{i,t-1}, x_{i,t-1}\}$, and all the regressors involving y^1_{t-1} from $\mathbb{P}\{y^2_{i,t} | y_{i,t-1}, x_{i,t-1}\}$. As for the simultaneous independence condition (15), it remains unchanged, since $\rho_{i,t}$ must be identically equal to zero for all (i,t) in order to factor out the joint distribution (10) into the product of the marginal (11) and (12), which requires that $\rho_{i,t}$ does not depend on covariates.

So far we have discussed observed heterogeneity. However, in microeconomic applications it is customary to assume also some degree of unobserved heterogeneity, by introducing fixed or random effects. Extending the notion of fixed and random effects to dynamic discrete choice models is extremely difficult, due to the non-linearity of such models. Even in the univariate case there is no completely satisfactory solution to the problem. A discussion may be found in Chamberlain (1984), Maddala (1987), Wooldridge (2002) and Honoré (2002). Formally, the fixed effects dynamic probit model might be easily extended to our multivariate setting by adding individual dummies to $s_{i,t}$. However, it is easily demonstrated that, due to the incidental parameters problem, maximum likelihood estimates in this model are consistent in general only when $T \to \infty$ (even if N is fixed). Conversely, when T is small, the random effects model seems convenient. The random effects model may be extended in the bivariate case by introducing a random variable indexed with i in each of the equations in (7). If a parametric distribution (e.g., a bivariate normal distribution) is chosen for the random effects, they can be integrated out as in the univariate case. As illustrated in Wooldridge (2002), it may be convenient to condition the distribution of the unobserved individual effects upon the first observation $y_{i,1}$ and the explanatory variables: this approach leads to manageable estimators that are consistent when $N \to \infty$ and T is fixed. Notice that strict exogeneity of the covariates, which is not necessary in general for non-causality analysis, seems to be needed in the presence of random effects (see Wooldridge, 2000).

4.2. Relaxing the first-order Markov assumption

Let us now relax the first-order Markov assumption. This will be done in the spirit of AutoRegressive models: models based on ARMA-like extensions, like those developed in the univariate case by Heckman (1978, 1981), will not be discussed here. For the sake of simplicity, we go back to the assumption that the information set available in t is $\mathcal{Y}_{t-1}$ (no covariates). Consider first the case where the relevant information for the transition probabilities is given by the last two states visited by an individual, rather than the last one only. There are therefore 16 possible paths followed in $(t-2)$ and $(t-1)$, at the end of which the individual may choose among four states. Hence, the most general model one may use to describe $\mathbb{P}\{y_{i,t}|y_{i,t-1}, y_{i,t-2}\}$ requires $16 \times (4-1) = 48$ transition probabilities.[10]

This model can be written in the form (10) by replacing $s_{i,t-1}$ by $s^2_{i,t-1} = s_{i,t-1} \otimes s_{i,t-2}$, i.e., using the saturated model with $x_{i,t-1} = s_{i,t-2}$. To generalize to the case in which the last ℓ states visited are relevant for the transition probabilities, then $s^{\ell}_{i,t-1} = s_{i,t-1} \otimes s_{i,t-2} \otimes \cdots \otimes s_{i,t-\ell}$ has to be used in (10) instead of $s_{i,t-1}$. It seems natural to refer to this model as bivariate Probit Vector AutoRegressive model of order ℓ, or PVAR(ℓ). We will call PVARX(ℓ) the model in which exogenous covariates are also included. Notice that the number of parameters does increase very rapidly, since B and γ will be of dimension $4^{\ell} \times 2$ and $4^{\ell} \times 1$. The dynamic structure of the process may be simplified by using the unsaturated model rather than the saturated one, which would dramatically reduce the number of parameters to $3 \times (3\ell + 1)$, although the interpretation of the ensuing model is unclear. A further simplification could be based on the following underlying latent regression:

$$y^*_{i,t} = \mu + \sum_{j=1}^{\ell} A_j y_{i,t-j} + \varepsilon_{it}, \tag{19}$$

where $y^*_{i,t} = (y^{1*}_{i,t}, y^{2*}_{i,t})$, A_j $(j = 1, \ldots, \ell)$ are 2×2 parameter matrices, μ is a 2×1 parameter vector, and $\varepsilon_{i,t}$ is the vector defined in (8) with $\rho_{i,t} = \rho$. The total number of parameters is further reduced to $4\ell + 3$. Although this model resembles the usual VAR closely, the left-hand side involves the latent variabes y^*, while the right-hand side involves the binary variables y. In principle, a proper VAR in terms of the latent variables, like

$$y^*_{i,t} = \mu^* + \sum_{j=1}^{\ell} A^*_j y^*_{i,t-j} + \varepsilon^*_{it} \tag{20}$$

[10] It is easily shown that this second-order Markov 4 states model can be rewritten as a first-order Markov 16 states model, where 192 out of the 16^2 transition probabilities are set to zero, while 16 transition probabilities can be written as linear functions of the remaining 48.

could be used. The univariate version of such model is introduced and discussed in Heckman (1981) and Grether and Maddala(1982). Dueker (2001) and Dueker and Wesche (2001) generalize the model to a VAR with one latent variable and several observed variables, and show how their model can be estimated using MCMC techniques; multivariate extensions with two or more binary variables do not seem to be thoroughly discussed yet. The interpretation of this type of model usually found in the literature is that the propensities to choose $y_{i,t}^1 = 1$ and $y_{i,t}^2 = 1$ depend on past propensities rather than past choices. Notice that while model (19) is nested in (10), (20) is not, and therefore the two models are non-nested in each other. More precisely, it is highlighted for example in Maddala (1987) that models like (20) imply that transition probabilities depend on the entire history of the observed process $y_{i,t}$: therefore, under (20) $y_{i,t}$ is not a Markov chain. Of course, non-causality analysis within models (10)–(19) and (20) could give conflicting results. Therefore, a proper discussion of model (20), as well as the development of tests for comparing such model with (10)–(19) would be interesting, but are beyond the scope of this paper.

For the unrestricted $\mathsf{PVAR}(\ell)$ model, the Granger non-causality conditions are formally identical to (13) and (14), but the restriction matrices are defined as $H_1^* = H_1^{\otimes \ell}$ and $H_2^* = H_2^{\otimes \ell}$ where $A^{\otimes \ell}$ is the Kronecker product of ℓ copies of A. In this case, the restrictions matrices exclude all the regressors involving $y_{i,t-1}^2, \ldots, y_{i,t-\ell}^2$ from $\mathbb{P}\{y_{i,t}^1 | y_{i,t-1}, \ldots, y_{i,t-\ell}\}$, and all the regressors involving $y_{i,t-1}^1, \ldots, y_{i,t-\ell}^1$ from $\mathbb{P}\{y_{i,t}^2 | y_{i,t-1}, \ldots, y_{i,t-\ell}\}$. Again, the simultaneous independence condition (15) remains unchanged. The restriction matrices for the restricted versions of the PVAR, as well as those needed for the PVARX may be obtained accordingly.

Notice that some of the covariates can be deterministic functions of the past values of $y_{i,t}$. In this case, the coefficients of these variables have to be included in the test for non-causality. This extension, while destroying the Markovian character of the model, allows for considering much more general forms of dependence on the past (see e.g., Heckman, 1978, 1981).

5. Estimation and testing

The purpose of this section is to discuss the properties of the parameter estimates in model (10), as well as the properties of the tests for the hypotheses (13), (14) and (15). Some hints will also be given about the generalizations illustrated in Section 4. We will discuss the asymptotic properties of Maximum Likelihood estimates and LR tests, although several other standard procedures for estimating and testing may be used. Some finite sample results will also be illustrated.

5.1. Inference in the homogeneous model

We assume that each individual i, $i = 1, \ldots, N$, is observed at each time t during a period of known length $t = 1, \ldots, T$; the extension to the case of unbalanced panel in which every individual i is observed for a length T_i is straightforward if we suppose that T_i is a stopping time with respect to the filtration $\{\mathscr{F}_t\}$. The log-likelihood conditional on the first observation of each cross-sectional unit can be written in compact form as:[11]

$$\ln \mathsf{L}_{NT}(\theta) = \sum_{i=1}^{N} \sum_{t=2}^{T} \ln \Phi_2 \left(D_{y_{i,t}} B s_{i,t-1}; 0, D_{y_{i,t}} \begin{bmatrix} 1 & \rho_{i,t} \\ \rho_{i,t} & 1 \end{bmatrix} D_{y_{i,t}}^{\mathsf{T}} \right), \tag{21}$$

where θ is the parameter vector composed of B and γ and $\rho_{i,t}$ is given by (9).

Let us discuss the asymptotics involved, by considering three cases:

- $T \to \infty$, N finite,
- T finite, $N \to \infty$,
- $T \to \infty$, $N \to \infty$.

To keep the notation simple, notice that the first-order Markov model for homogeneous population can be written compactly, since it is equivalent to a Markov model for the univariate process $\{U_t\}$ which, at any time t, takes on values in a finite state-space $\mathscr{U} = \{0, 1, 2, 3\}$,[12] with a stationary (or *time-homogeneous*) transition probability matrix $P = (P_{hk})$, $(h, k) \in \mathscr{U} \times \mathscr{U}$; P is clearly a stochastic matrix, that is, $P_{hk} \geqslant 0 \; \forall h, k$, and $\sum_{k \in \mathscr{U}} P_{hk} = 1$. Moreover we define the nth power of P, $P^n = (P_{n,hk})$, where $P_{n,hk} = \mathbb{P}\{U_{t+n} = k | U_t = h\}$. The relevant asymptotic theory for homogeneous Markov chains is presented in Anderson and Goodman (1957) for $N \to \infty$ and in Billingsley (1961) for $T \to \infty$. However, we state some results in the following that allow a unified treatment for N and $T \to \infty$.

It is intuitive that to ensure consistent and asymptotically normal estimates of the transition probabilities what is needed is that *all* the transitions whose probabilities have to be estimated (i.e., are not known) *can* be observed infinitely many times as T and/or N go to infinity. Through a Delta Method expansion, this will imply corresponding asymptotic results for the estimators of θ (for a related result for Markov chains, see Theorem 4.1 in Billingsley, 1961). Thus, LR and Wald tests are asymptotically χ^2 distributed under the null hypothesis.

A.1. (necessary condition) Each state with at least one unknown exiting transition probability must be visited infinitely often with probability 1 as either T or N or both go to infinity.

A.2. (sufficient condition) Infinitely many of the individuals who have reached each state with at least one unknown exiting transition probability must be observed for at least one time period in that state.

[11] We do not make any attempt to use the first observation of each cross-sectional unit. It is in fact extremely difficult, especially when heterogeneity is introduced, to make a distributional assumption on the first observation which is consistent with the dynamic multivariate probit model; it is also unclear whether such consistence is needed or not. We believe that introducing a rather arbitrary and hardly testable assumption is not justified by the possible gain in efficiency. Coherently, if the model is extended by introducing random effects, we suggest to follow Wooldridge (2002) in that respect.

[12] The elements of $\mathscr{U}$ correspond element-wise with the elements of $\mathscr{Y}$, defined as $\mathscr{Y} = \{(0,0), (0,1), (1,0), (1,1)\}$, representing the state space of the process $\{Y_t\}$ at any time t.

In the following we will enunciate some results about the conditions on P under which both the necessary and sufficient conditions are fulfilled. Let us first state the condition when $T \to \infty$ with N fixed.[13]

Lemma 5.1. *Assume that P is such that each state with at least one unknown exiting transition probability is persistent. Then Conditions* A.1. *and* A.2. *are fulfilled for $T \to \infty$ for any $N \geqslant 1$.*

The proof of the part related to Condition A.1. is in Billingsley (1995), Theorem 8.2. Notice that each state of an irreducible Markov chain with finite state space is persistent (Billingsley, 1995, Example 8.7). Therefore in our case, this ensures that ML estimates show the usual asymptotic properties when $T \to \infty$ for any $N \geqslant 1$, finite or infinite.

A similar result can be obtained when $N \to \infty$, but it depends on the initial conditions of the process. These are defined by a vector of *initial probabilities* $p = (p_h)$, representing, $\forall h \in \mathcal{U}$, the probability of being in state h in $t = 1$. Of course, $p_h \geqslant 0$, $\forall h \in \mathcal{U}$, and $\sum_{h \in \mathcal{U}} p_h = 1$.

Lemma 5.2. *Assume that P and p are such that there exists at least a $h_k \in \mathcal{U}$ and a finite integer n_k such that $p_{h_k} P_{n_k,h_k k} > 0$ for each state $k \in \mathcal{U}$ with at least one unknown exiting transition probability. Then*:

(i) *Condition* A.1. *is fulfilled if $N \to \infty$ for any $T >$* $\max\{\bar{n}_k, k \in \mathcal{U}\}$, *where $\bar{n}_k$ is, for each k, the minimum n_k such that $p_{h_k} P_{n_k,h_k k} > 0$;*
(ii) *Condition* A.2. *is fulfilled if $T > T^* = \max\{\bar{n}_k, k \in \mathcal{U}\} + 1$.*

Notice that if $p_h \neq 0$ $\forall h \in \mathcal{U}$, and the chain is irreducible, then the conditions on p and P are met, and moreover $T^* = 1$, so that the ML estimates show the usual asymptotic properties when $N \to \infty$ for $T > 1$, be it bounded or not. As a whole, the conditions that $p_h \neq 0$, for every $h \in \mathcal{U}$ and that P is irreducible are sufficient, although non-necessary, for consistency and asymptotic normality of the Maximum Likelihood estimator with either $T \to \infty$ or $N \to \infty$, where in the latter case T must be at least equal to 2.

Extension to higher-order Markovian models such as the PVAR(ℓ) are straightforward, since it is always possible to rewrite a bivariate PVAR(ℓ) model as a finite state-space Markov chain with 4^ℓ states, and hence the conditions stated in Propositions 5.1 and 5.2 apply to the transition probabilities matrix and initial probabilities vector corresponding to the new Markov chain.

Extension to unobserved heterogeneity modelled by fixed effects implies that the transition probability matrix is different across individuals. Therefore, the results for $T \to \infty$ remain valid (with more severe finite sample problems) provided that the conditions on P hold for every individual. Conversely, the results for $N \to \infty$ with T fixed cannot be extended due to the incidental parameters problem.

[13] For definitions of irreducibility and persistence, see Billingsley (1995, Section 8).

Let us briefly address some problems possibly arising in finite samples. In this case, the transition probabilities are estimated essentially as the ratio of the number of cases where the transition occurred (say, N_{hk}) over the number of cases where it could have occurred (say, N_h). In any finite sample, the distribution of N_{hk} given N_h will be Binomial, and will hence converge in distribution to the Normal as $N_h \to \infty$ (see Anderson and Goodman, 1957; Billingsley, 1961; and Gouriéroux, 2000, Chapter 6). Propositions 5.1 and 5.2 state conditions which ensure divergence of N_h with either N (Proposition 5.1) or T (Proposition 5.2), and convergence to normality of the distribution of the ratio $\frac{N_{hk}}{N_h}$. Notice however that, for states which have been visited few times (N_h small), the distribution of N_{hk} given N_h may be very far from normality, especially when the true transition probability P_{hk} is close to zero or one, which will give highly skewed distributions. In any case, in order to get an idea of how reliable the asymptotic distribution can be, it is convenient to check the number of observations *in each* state with at least one unknown exiting transition probability. In fact, even if $N \times T$ is large, the information about some of the transitions may be quite poor.

5.2. Inference in the model with covariates

Let us now discuss the conditions for consistency and asymptotic normality in the model with covariates. Since the cross-section case (in which T is finite and $N \to \infty$) appears as a modification of the theory for iid random variables, we consider only the time series case, in which $N = 1$ and $T \to \infty$, and we remove any reference to the index i. We rely on the *partial likelihood* concept as developed by Cox (1975), Wong (1986) and Fokianos and Kedem (1998). The last reference is particularly relevant for our purposes, since the asymptotic properties of a generalized linear model for categorical data are studied in the context of partial likelihood.

Let $\{\mathscr{F}_t\}$ be a filtration and $\{Y_t\}$ be a time series such that Y_t is $\mathscr{F}_t$-measurable for any t. Usually, we will take $\mathscr{F}_t = \sigma\{(Y_s, X_s), 1 \leqslant s \leqslant t\}$. With some notational imprecision, let us denote by $f(y_t|\mathscr{F}_{t-1};\theta)$ the density of Y_t given $\mathscr{F}_{t-1}$, where θ is a parameter vector. The partial likelihood of θ based on $\{Y_t, \mathscr{F}_t\}$, is given by

$$\mathsf{L}_T(\theta) = \prod_{t=2}^{T} f(y_t|\mathscr{F}_{t-1};\theta). \tag{22}$$

The only conditions for writing this form of likelihood are that $\{\mathscr{F}_t\}$ is a filtration and Y_t is adapted to $\mathscr{F}_t$ for any t. If $\{Y_t\}$ does not Granger cause $\{X_t\}$ (or equivalently $\{X_t\}$ is strictly exogenous for $\{Y_t\}$, see Chamberlain, 1982), then the partial likelihood is equivalent to the classical *conditional likelihood* defined by $f(y_2,\ldots,y_T|x_1,\ldots,x_{T-1};\theta)$.[14] However, the properties of inference based on partial likelihood do not depend on strict exogeneity of $\{X_t\}$ in any manner.

[14] For an example, see Kaufmann (1987), where the same model of Fokianos and Kedem (1998), is studied.

The following assumptions allow for deriving the asymptotic properties of the estimates of θ, say $\hat{\theta}_T$, obtained by maximizing (22). In the following, Z_t is a vector of regressors representing the relevant information in $\mathscr{F}_t$; in other words, Z_t is a vector of functions of current and past values of Y_t and X_t. Examples are given in (17) or (18), but nothing compels the process $\{Y_t\}$ to be Markovian of any order since complex functions of the past of $\{Y_t\}$ can also be included in $\{Z_t\}$ provided they respect the conditions given below.

B.1. The statistical model is described by the conditional distribution of Y_t given Z_{t-1} (see the models described in Section 4), for $\theta \in \Theta$, where Θ is an open and bounded subset of a suitable Euclidean space.
B.2. The true conditional probability of Y_t given Z_{t-1} is obtained for a value $\theta_0 \in \Theta$.
B.3. $\sum_{t=1}^{T} z_t z_t^{\mathsf{T}}$ is almost surely non-singular for any T large enough.
B.4. For any t, Z_t lies in a non-random compact subset Γ of a suitable Euclidean space $\mathbb{R}^d$.
B.5. The empirical cdf of the data $\{z_1, \ldots, z_{T-1}\}$ converges almost surely to a cdf F; moreover, F is such that $\int_{\mathbb{R}^d} zz^{\mathsf{T}}\,\mathrm{d}F(z)$ is positive definite.

We will not prove rigorously the results: the proof of Fokianos and Kedem (1998) can be adapted to deal with our model too. However, we discuss the assumptions in some detail.

Notice that the compactness of Γ (B.4.) and the boundedness of θ_0 (B.1.) imply that all the transition probabilities $\mathbb{P}(Y_t = y_t | Z_{t-1}; \theta_0)$ are bounded away from 0 and 1.[15] This guarantees the fulfillment of Conditions A.1. and A.2. given in the previous subsection for the model with no covariates.

The boundedness of Θ (B.1.) and Γ (B.4.) ensures that the score and the Hessian are bounded. Moreover, the score evaluated at θ_0 is a square-integrable zero mean martingale (see Wong, 1986; Fokianos and Kedem, 1998): through the LLN and CLT for martingales, this entails that $\sqrt{T} \cdot \frac{\partial \ln \mathrm{L}_T(\theta)}{\partial \theta}|_{\theta=\theta_0}$ converges to a centered Gaussian random vector. The non-singularity of $\sum_{t=1}^{T} z_t z_t^{\mathsf{T}}$ in B.3. guarantees that the Hessian is well-defined and invertible in finite samples, while B.5. ensures that the Hessian is also invertible in the limit. Assumption B.5. entails the convergence of functions of $\{Z_t\}$ of the form $T^{-1}\sum_{t=1}^{T} f(z_t)$, such as the Hessian, to a non-random limit and is quite standard in the econometric literature (see Amemiya, 1973, p. 999).[16]

Under these assumptions, the probability that a locally unique maximum likelihood estimator exists converges to 1, and there exists a sequence of maximum partial likelihood estimators $\hat{\theta}_T$ which is consistent and asymptotically normal:

$$\sqrt{T}(\hat{\theta}_T - \theta_0) \xrightarrow{\mathscr{D}} \mathscr{N}(0, \mathfrak{I}^{-1}(\theta_0)).$$

[15] In the Markov case with covariates, the chain, even if non-stationary, is irreducible.

[16] This is an asymptotic mean ergodicity condition, allowing for a certain degree of non-stationarity (see Gray and Kieffer, 1980).

The covariance matrix can be defined in three alternative ways. As usually, the *Hessian* $H_T(\theta)$ is given by $H_T(\theta) = -\frac{1}{T}\frac{\partial^2 \ln \mathsf{L}_T(\theta)}{\partial\theta\partial\theta^{\mathsf{T}}}$. The *conditional information matrix* $G_T(\theta)$ is given by

$$G_T(\theta) = \frac{1}{T}\sum_{t=1}^{T} \mathrm{Cov}_{\theta_0}\left[\frac{\partial \ln \mathsf{L}_t(\theta)}{\partial\theta} - \frac{\partial \ln \mathsf{L}_{t-1}(\theta)}{\partial\theta}\middle|\mathscr{F}_{t-1}\right]$$
$$= \frac{1}{T}\sum_{t=1}^{T} \mathrm{Cov}_{\theta_0}\left[\frac{\partial \ln \mathbb{P}\{Y_t|Z_{t-1};\theta\}}{\partial\theta}\middle|Z_{t-1}\right].$$

The *unconditional information matrix* is $F_T(\theta) = \mathbb{E}_{\theta_0}[G_T(\theta)]$. Under the assumptions, the three expressions for the covariance matrix coincide asymptotically:

$$F_T(\hat{\theta}_T) = G_T(\hat{\theta}_T) + o_{\mathbb{P}}(1) = H_T(\hat{\theta}_T) + o_{\mathbb{P}}(1) \to \mathfrak{I}(\theta_0).$$

The asymptotic theory of the usual tests (Wald, LM and LR) is standard.

As concerns asymptotic efficiency, an assumption of weak exogeneity of the covariate process $\{X_t\}$ with respect to θ is obviously needed (see Engle et al., 1983, p. 277–278), since otherwise efficient estimates of θ would require the complete specification of the marginal density of $\{X_t\}$. More precise remarks on the asymptotic efficiency of the estimator can be found in Wong (1986), for the general case, and in Slud and Kedem (1994), for the case of a logistic autoregression.

Also the Eicker–White sandwich matrix can in principle be used:

$$H_T(\hat{\theta}_T)^{-1}\hat{G}_T(\hat{\theta}_T)\cdot H_T(\hat{\theta}_T)^{-1}.$$

This is reasonable when the PVAR is not supposed to be the true model and causality testing is just intended to be an exploratory tool. In this case, more robustness against stronger forms of dependence can be obtained using a HAC (heteroskedasticity and autocorrelation consistent) estimator of $G_T(\hat{\theta}_T)$ (see Davidson, 2000, Section 9.4.3, and references therein).

6. Non-causality with survival data

Special cases of the model discussed in Sections 3 and 4 can be obtained when some of the transition probabilities are set to 0. In the following section we will deal with the case of survival models,[17] i.e., models in which the states with $Y_t^j = 0$ are not accessible from the states with $Y_{t-1}^j = 1$, $j = \{1, 2\}$; this implies that every decision with respect to a variable Y_t^j is in a certain sense irreversible. This implies that 7 out of 12 transitions illustrated in Fig. 1 have zero probability, leaving only five unrestricted transition probabilities, as illustrated in Fig. 2.

It is convenient to consider first state 0 (corresponding to $Y_t = (0,0)^{\mathsf{T}}$) from which all states can be reached; therefore the choice can be modelled through a bivariate

[17] A standard reference for survival models is Kalbfleisch and Prentice (1980). A counting process perspective on these models is in Andersen et al. (1993). A review of multivariate survival models is Hougaard (1987).

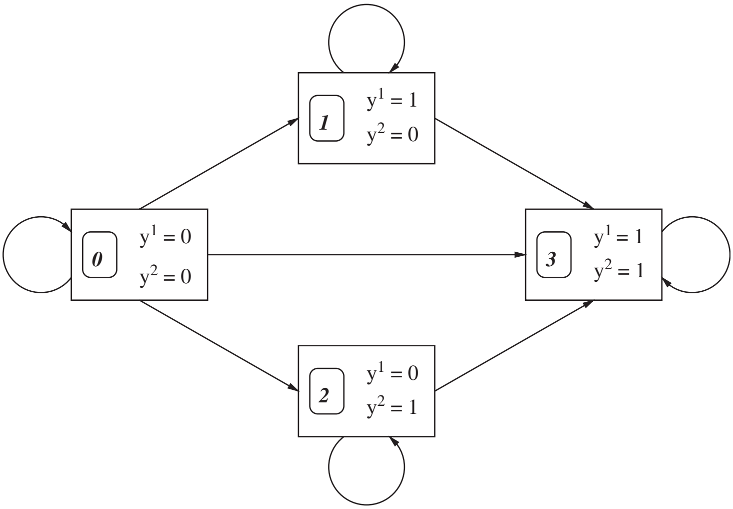

Fig. 2. State-transition diagram for a bivariate survival model.

probit model:

$$\mathbb{P}\{y_{i,t}|Y_{i,t-1}=(0,0)^{\mathsf{T}}\}=\Phi_2\left(D_{y_{i,t}}\begin{bmatrix}\beta_{10}\\ \beta_{20}\end{bmatrix};0,D_{y_{i,t}}\begin{bmatrix}1 & \rho\\ \rho & 1\end{bmatrix}D^{\mathsf{T}}_{y_{i,t}}\right).$$

We have eliminated the subscript from ρ since no other correlation coefficient is present in the model. It may be convenient to reparameterize the model using $\rho=\frac{\exp(\gamma_0)-1}{\exp(\gamma_0)+1}$ so that all parameters vary in $\mathbb{R}$.

When considering transitions from state 1 (that is $Y_{i,t-1}=(1,0)^{\mathsf{T}}$), it is important to remark that the bivariate distribution $\mathbb{P}\{y_{i,t}|Y_{i,t-1}=(1,0)^{\mathsf{T}}\}$ collapses into its marginal $\mathbb{P}\{y^2_{i,t}|Y_{i,t-1}=(1,0)^{\mathsf{T}}\}$, since $\mathbb{P}\{Y^1_{i,t}=1|Y_{i,t-1}=(1,0)^{\mathsf{T}}\}=1$. A similar argument holds from state 2. Therefore we may write:

$$\mathbb{P}\{y^1_{i,t}|Y_{i,t-1}=(0,1)^{\mathsf{T}}\}=\Phi_1((2y^1_{i,t}-1)(\beta_{10}+\beta_{11});0,1),$$
$$\mathbb{P}\{y^2_{i,t}|Y_{i,t-1}=(1,0)^{\mathsf{T}}\}=\Phi_1((2y^2_{i,t}-1)(\beta_{20}+\beta_{21});0,1).$$

As a whole, the model includes five parameters ($\beta_{10},\beta_{11},\beta_{20},\beta_{21}$ and γ_0), freely varying in $\mathbb{R}^5$, which can be shown to be bijective transformations of the transition probabilities.

Conditions for simultaneous independence and Granger non-causality can be easily adapted from those introduced in Section 3, and take the particularly simple form:

$$\mathsf{H}_{1\nleftarrow 2}\ (Y^1\nleftarrow Y^2):\ \beta_{11}=0,$$
$$\mathsf{H}_{1\nrightarrow 2}\ (Y^1\nrightarrow Y^2):\ \beta_{21}=0,$$
$$\mathsf{H}_{1\nLeftrightarrow 2}\ (Y^1\nLeftrightarrow Y^2):\ \gamma_0=0.$$

Also the introduction of exogenous variables can be accounted for by paralleling the solutions presented in Section 4. Assume that the exogenous variables are represented by a k-dimensional vector $x_{i,t}$ (including a constant): unsaturated versions of the models may be obtained by replacing β_{10}, β_{20} and γ_0 by linear functions of $x_{i,t}$, so that the total number of parameters becomes $3k+2$, while saturated versions may be obtained by replacing β_{11} and β_{21} also by linear functions of $x_{i,t}$ (the number of parameters is raised to $5k$). Of course, non-causality analysis may be carried on in these models by setting to zero all parameters replacing β_{11} (for $\mathsf{H}_{1\nleftarrow 2}$), β_{21} (for $\mathsf{H}_{1\nrightarrow 2}$) or γ_0 (for $\mathsf{H}_{1\nLeftrightarrow 2}$).

Extending the model by relaxing the first-order Markov assumption requires some caution. In fact, in this case conditioning on the lagged endogenous variables is meaningful only when the individual is in state 1 or 2. Moreover, if individual i is in state 1 at time t, then $y^2_{i,t-j}=0$ for any $j\geqslant 0$, and therefore only lagged $y^1_{i,t}$ need to be introduced to describe the path followed by the individual to reach state 1. A similar argument holds for state 2. The implication is that, in order to increase the dynamics up to lag ℓ, one should simply replace β_{11} by a linear function of $(y^1_{i,t-1}, y^1_{i,t-2}, \ldots, y^1_{i,t-\ell})$ and β_{21} by a linear function of $(y^2_{i,t-1}, y^2_{i,t-2}, \ldots, y^2_{i,t-\ell})$; the total number of parameters becomes then $2\ell+3$. Also in this case, non-causality analysis may be carried on by setting to zero all parameters replacing β_{11} (for $\mathsf{H}_{1\nleftarrow 2}$), β_{21} (for $\mathsf{H}_{1\nrightarrow 2}$) or γ_0 (for $\mathsf{H}_{1\nLeftrightarrow 2}$). A more parsimonious representation could be obtained by replacing β_{11} (and β_{21}) by a function of the time spent in state 1 (and 2). Of course, it is possible to develop models of order ℓ with exogenous variables: unsaturated versions of the model are obtained by replacing β_{10}, β_{20} and γ_0 only by linear functions of $x_{i,t}$, while saturated versions require that all the parameters describing the dynamics are also replaced by linear functions of $x_{i,t}$.

As for the asymptotic properties of the estimates and tests, notice that for this model, even in the case of homogeneous populations, Proposition 5.1 does not hold, since this model can be represented as a Markov chain with an absorbing state corresponding to state 3. This means that $N\to\infty$ is needed in order to ensure consistency and asymptotic normality of the estimates.

7. An illustrative example for panel data

The model developed in Section 4 will be employed here to analyze the relationship between marital status and the decision to have children. The analysis is only meant to be an illustration of the methodology rather than a serious attempt to model the decision process.[18] We use data from the well-known PSID database.[19]

[18] Some papers have studied the dynamical interdependencies between marital duration and fertility (Koo and Janowitz, 1983; Waite and Lillard, 1991; Lillard and Waite, 1993; Lillard, 1993).

[19] The Panel Study of Income Dynamics (PSID), begun in 1968, is a longitudinal study of a representative sample of U.S. individuals (men, women, and children) and the family units in which they reside. The study is conducted at the Survey Research Center, Institute for Social Research, University of Michigan. Information about the original 1968 sample individuals and their current co-residents (spouses, co-habitors, children, and anyone else living with them) is collected each year.

Starting from all women appearing in the database from 1968 to 1993 ($t = 1, \ldots, 26$), subjects with missing covariates have been eliminated, leading to a sample of 266 individuals. The variables have been elaborated in order to yield a certain uniformity over time. The data are available from the authors upon request, together with a precise description of how they have been obtained starting from PSID variables.

Therefore the following variables have been selected:

- $y^1_{i,t}$: the variable is set to 1 if individual i gave birth to (at least) one child during the year t, 0 otherwise;
- $y^2_{i,t}$: this variable is set to 1 if individual i was married during the year t, 0 otherwise.

The causal relationship between $y^1_{i,t}$ and $y^2_{i,t}$ will be analyzed conditioning on the vector $x_{i,t}$ including the following covariates:

- $age_{i,t}$: age of individual i in year t,
- $age^2_{i,t}$: age of individual i in year t squared,
- $income_{i,t}$: income of individual i in year t,
- $hours_{i,t}$: hours worked by individual i in year t,
- $edu_{i,t}$: years of school completed by individual i in year t.

As discussed in Section 4, in order to have an economically meaningful interpretation of the results of non-causality analysis, simultaneous regressors should not be endogenous. In this application, this means that they should not be part of the same optimization problem leading to the decision about marriage and maternity. Actually, *income* and *hours* might be regarded as endogenous in this sense. However, for illustrative purpose, we maintain the exogeneity assumption, since the alternative would complicate the example significantly. In fact, simply replacing $income_{i,t}$ and $hours_{i,t}$ with $income_{i,t-1}$ and $hours_{i,t-1}$ would not be a convincing solution, since two equations for $income_{i,t}$ and $hours_{i,t}$ should also be added. The resulting system would be four dimensional, involving binary, continuous and count variables: explicitly writing and testing non-causality restrictions in such a system is possible in principle, but beyond the scope of this paper.[20]

The number of observations in each of the four states described in Fig. 1 is given in Table 1, which shows that we have few observations in state $(y^1, y^2) = (1, 0)$.

The first step in our analysis consists in determining the maximum lag of the PVAR model, and whether the saturated or unsaturated version of the model is more appropriate for the data. We will refer to the saturated and unsaturated models of order j by S_j and U_j, respectively. To reduce the dimension of the parameter space, we have considered here restricted versions of the models, where $\rho_{i,t}$ is a function of lagged $s_{i,t}$, but not of $x_{i,t}$. Notice that U_j is nested in S_j and U_{j+1}, while S_j is nested in

[20] See Dufour and Renault (1998) for the problems involved in testing for non-causality between two variables in higher dimensional systems.

Table 1
Number of observations in each state

	$y^1_{i,t}=0$	$y^1_{i,t}=1$
$y^2_{i,t}=0$	3274	151
$y^2_{i,t}=1$	2977	403

Table 2
Information criteria for the estimated models

Models	L	p	BIC
U_1	−2980.79	22	−3077.87
U_2	−2962.69	31	−3099.48
S_1	−2894.40	52	−3123.86
S_2	−2828.62	208	−3746.47

S_{j+1} but non-nested in any of the unsaturated models. Therefore, there is no natural ordering of the models, and this suggests to use information criteria for model selection. For $j=1,2$, we have computed the BIC criterion (Bayesian Information Criterion, due to Schwarz, 1978) for both the unsaturated and saturated models. The preferred model is selected by maximizing $\mathsf{BIC}=\ln(\mathsf{L})-\frac{p}{2}\ln M$, where L is the likelihood,[21] M is the number of observation (in our case, $M=6805$) and p is the number of parameters. Table 2 shows that model U_1 is preferred according to BIC. Estimates of model U_1 are given in Table 3.

Within this model, the non-causality relations between the two processes $\{Y^1_t\}$ and $\{Y^2_t\}$ are tested through Wald tests, whose results are displayed in Table 4. The results show the following:[22]

- the hypothesis $\mathsf{H}_{1\nleftarrow 2}$, concerning the non-causality of Y^2 towards Y^1 is strongly rejected: a marital relation seems to increase significantly the probability of having a child, as common sense suggests.
- the hypothesis $\mathsf{H}_{1\nrightarrow 2}$, concerning the non-causality of Y^1 towards Y^2 is accepted. Hence, fertility timing does not seem to have any impact on the marriage and divorce decisions of American women.
- the hypothesis $\mathsf{H}_{1\nLeftrightarrow 2}$, concerning the simultaneous independence between Y^2 and Y^1 is rejected. By using (9) to compute $\rho_{i,t}$, and noticing that the only significant

[21] Estimation has been performed using GAUSS-386i and Ox Professional 3.0 (see Doornik, 1999). The covariance matrix of the estimates has been calculated through the cross-product of first derivatives.

[22] We have also checked whether the results of non-causality analysis are affected by the selected model. In this example, the results are exactly the same irrespective of the lag order and the choice about saturation.

Table 3
Estimates of model U_1

Usable observations		6805	Degrees of freedom		6783
Function value		−2980.793			
Variable		Coeff.	Std error	T-stat	Signif.
Children bearing equation ($\mathbb{P}\{y_{i,t}^1 \vert y_{i,t-1}, x_{i,t}\}$)					
1	CONST	−3.871	0.417	−9.27	0.00
2	Y1_1	−0.309	0.263	−1.17	0.24
3	Y2_1	0.561	0.0641	8.75	0.00
4	Y1_1*Y2_1	−0.360	0.288	−1.25	0.21
5	AGE	0.226	0.0329	6.89	0.00
6	AGESQ	−0.526	0.0594	−8.85	0.00
7	INCOME	0.687	0.585	1.17	0.24
8	HOURS	−9.480	3.615	−2.62	0.01
9	EDU	0.0141	0.00786	1.79	0.07
Marriage equation ($\mathbb{P}\{y_{i,t}^2 \vert y_{i,t-1}, x_{i,t}\}$)					
1	CONST	−2.432	0.421	−5.77	0.00
2	Y1_1	−0.170	0.293	−0.58	0.56
3	Y2_1	3.320	0.0719	46.18	0.0
4	Y1_1*Y2_1	−0.0270	0.327	−0.08	0.93
5	AGE	0.0348	0.0320	1.09	0.28
6	AGESQ	−0.110	0.0574	−1.92	0.05
7	INCOME	0.0220	0.628	0.03	0.97
8	HOURS	1.427	3.831	0.37	0.71
9	EDU	0.0830	0.00921	9.01	0.00
Correlation					
1	CONST	0.644	0.136	4.73	0.00
2	Y1_1	−2.171	433.675	−0.00	0.99
3	Y2_1	−0.389	0.238	−1.63	0.10
4	Y1_1*Y2_1	1.026	433.677	0.00	1.00

Table 4
Causality testing through Wald tests

Hypothesis	χ^2	DoF	Signif.
$\mathsf{H}_{1 \nleftarrow 2}$ ($Y^1 \nleftarrow Y^2$)	76.62	2	0.00
$\mathsf{H}_{1 \nrightarrow 2}$ ($Y^1 \nrightarrow Y^2$)	2.10	2	0.35
$\mathsf{H}_{1 \nLeftrightarrow 2}$ ($Y^1 \nLeftrightarrow Y^2$)	27.51	4	0.00

coefficient in γ is the constant, the correlation among the latent variables seems positive (around 0.3) from all states, maybe somewhat lower from states with $Y_t^2 = 1$ (married women).

8. An illustrative example for survival data

To illustrate the model developed in Section 6, we investigate the causal relationship between the adoption of two technologies introduced in the 70s in the Italian metalworking industry.[23] The dataset involves survival data, namely the spell of non-adoption for two related technologies in a sample of Italian plants, and therefore the analysis described in Section 6 will be performed. The two technologies considered are Computer Aided Design or Manufacturing (CADCAM), which will be labelled by 1, and Flexible Manufacturing Systems (FMS) which will be labelled by 2. Both technologies are originated from the Flexible Automation (FA) paradigm and therefore they are expected to display significative interactions.

Data on the diffusion of FA within the Italian metalworking industry are provided by the FLAUTO database, developed at Politecnico di Milano. Our sample includes 552 plants. CADCAM and FMS have been introduced in Italy around 1970, hence the observation window is assumed to begin in this year and calendar time t is set to 0 in 1969. Since the dataset originates from a retrospective survey carried on in 1989, the observed adoption time never exceeds $T = 20$.

For each plant $i = 1, \ldots, 552$, we observe the year of adoption of both technologies, say t_i^1 and t_i^2. To fit these survival type data into our framework, it is convenient to transform them as follows:

$$\begin{aligned} y_{i,t}^1 &= \mathbb{1}_{\{\text{plant } i \text{ adopts CADCAM at time } t\}} \quad t = t_i^E, \ldots, \min[t_i^1, T], \\ y_{i,t}^1 &= \mathbb{1}_{\{\text{plant } i \text{ adopts FMS at time } t\}} \quad t = t_i^E, \ldots, \min[t_i^2, T], \end{aligned}$$

where t_i^E is the year plant i enters the sector. Notice that about 30% of the plants enter after 1970. The data are right-censored because the firms are not observed after 1989, so that t_i^1 and t_i^2 never exceed $T = 20$. However, this kind of censoring involves no bias, since the censoring time may be regarded as a Markov time with respect to the filtration generated by the process. The number of observations in each of the three relevant states described in Fig. 2 is given in Table 5, which shows that we have few observations in state $(y^1, y^2) = (0, 1)$.

Let us now introduce the covariates, i.e., the vector $x_{i,t}$ in our notation. The only time-invariant covariate considered is the size of the plant expressed in thousands of employees at June 1989. In addition, conforming to Colombo and Mosconi (1995), we consider two different time scales, by using both the calendar time t, and the duration of non-adoption $\tau_{i,t} = t - \max(0, t_i^E)$. For plants that entered the sector before 1970, the two time scales coincide. We expect calendar time to reflect phenomena which do not depend on the existence of the firm, notably price and/or performance changes of the technologies over time, epidemic effects, and other time varying factors. Instead, the duration of non-adoption captures effects related to the

[23] This section is inspired by the study carried on in Colombo and Mosconi (1995). We simplify the economic analysis by introducing only a small subset of the covariates used there. Therefore, this should be considered only as an illustration of the methods introduced in the previous sections rather than a contribution to the economic debate. On the other hand, the econometric approach to non-causality analysis is made more rigorous here.

Table 5
Number of observations in each state

	$y^1_{i,t}=0$	$y^1_{i,t}=1$
$y^2_{i,t}=0$	8795	880
$y^2_{i,t}=1$	178	==

Table 6
Estimates of the unsaturated model

Usable observations		9853	Degrees of freedom		9842
Function value		−1475.488			
Variable		Coeff.	Std error	T-stat	Signif.
CADCAM equation ($\mathbb{P}\{y^1_{i,t}\|y_{i,t-1},x_{i,t}\}$)					
1	CONST	−3.798	0.129	−29.51	0.00
2	Y2_1	0.469	0.125	3.76	0.00
3	SIZE	0.285	0.043	6.64	0.00
4	t	0.139	0.011	12.48	0.00
5	$\tau_{i,t}$	−0.003	0.009	−0.30	0.38
FMS equation ($\mathbb{P}\{y^2_{i,t}\|y_{i,t-1},x_{i,t}\}$)					
1	CONST	−3.562	0.192	−18.56	0.00
2	Y1_1	0.198	0.109	1.81	0.03
3	SIZE	0.252	0.057	4.41	0.00
4	t	0.082	0.018	4.61	0.00
5	$\tau_{i,t}$	−0.006	0.014	−0.41	0.34
Correlation					
1	CORRELATION	0.246	0.081	3.05	0.00

existence of the firm, such as learning phenomena. The collinearity-like problems created by these variables are discussed in Colombo and Mosconi (1995).

As a whole, the covariates introduced in the model are $x_{i,t}=(size_i,t,\tau_{i,t})^{\mathsf{T}}$; for expositional purposes, $\rho_{i,t}$ is assumed not to depend on $x_{i,t}$ and it is not z-transformed. Model S_1 has 17 parameters, while model U_1, which is nested in S_1, has 11 parameters. The χ^2_6 likelihood ratio test takes on value 2.563, and therefore model U_1 is selected and it will be used for non-causality analysis. Estimates of model U_1 are reported in Table 6. These estimates might suffer some bias due to the omission of several variables that have proved significant in previous studies: however, the sign, magnitude and significance of the coefficients of *size* and t resemble other studies, while duration of non-adoption is properly signed but insignificant.

Based on Table 6, Wald type non-causality tests may be done by simply analyzing the t-test for the parameters β_{12}, β_{21} and ρ. β_{11} is positive and significant (the hypothesis of Granger non-causality is rejected), which suggest a positive effect of adoption of FMS on the following adoption of CADCAM: this has a simple economic interpretation and confirms the presence of an interaction of the two technologies. On the other hand, β_{21} is positive but non-significant, which means that the Wald test accepts the hypothesis that CADCAM does not Granger cause FMS ($Y^1 \nrightarrow Y^2$). In fact, the economic intuition suggests that CADCAM is a powerful design tool even without an FMS. The correlation between the error terms of the latent regressions relative to CADCAM and FMS (ρ) is positive and highly significant: the strong simultaneous dependence suggests that the decision of joint simultaneous adoption occurs more often than what would be expected if the two decisions were taken independently. The χ^2_3 distributed likelihood ratio test for the joint exclusion of β_{11}, β_{21} and ρ takes on value 26.043, strongly rejecting the hypothesis. Similar results are obtained when testing is performed in model S_1. It is worth noticing that the results of non-causality tests depend on the information set: however, the results of the non-causality analysis are substantially unchanged even when the variables conditioned upon are all those included in Colombo and Mosconi (1995).

9. Conclusions

In this paper we make a step towards rendering an important tool of applied macroeconometric analysis, such as Granger non-causality, available and operational for those situations in which the processes involved in the analysis are binary, as often happens in microeconometric analysis. The paper is grounded on a rigorous mathematical definition of non-causality, which is shown to be easily fitted into a dynamic version of the bivariate probit model. Particular attention is placed in including covariates into the analysis, and in specializing the definitions for longitudinal data sets. Panel type data for heterogeneous individuals are in fact typical in microeconometric applications.

The paper implicitly suggests so many extensions to fill up a research agenda. To make some examples: a more parsimonious representation of the dynamics; allowing for unobserved heterogeneity; generalizing to a multivariate setting; considering multinomial instead of binary variables; mixing binary and continuous variables; analyzing the impact of time aggregation.

Acknowledgements

This paper has benefited at different stages from discussions with Per Kragh Andersen, Clelia Di Serio, Clive Granger, Philip Hougaard, Søren Johansen and Niels Keiding. We wish to thank Peter Boswijk and two anonymous referees for helpful comments and suggestions. The authors are particularly indebted to Fabio Sartori, who significantly contributed to the development of the main ideas presented in the paper. We would like to thank Riccardo "Jack" Lucchetti for his help with Gauss and Christine Choirat for her support with Ox and her TEXpertise. The usual disclaimer applies.

References

Aalen, O.O., 1987. Dynamic modelling and causality. Scandinavian Actuarial Journal 177–190.

Aalen, O.O., Borgan, Ø., Keiding, N., Thormann, J., 1980. Interaction between life history events, nonparametric analysis for prospective and retrospective data in the presence of censoring. Scandinavian Journal of Statistics 7, 161–171.

Amemiya, T., 1973. Regression analysis when the dependent variable is truncated normal. Econometrica 41, 997–1016.

Amemiya, T., 1981. Qualitative response models: a survey. Journal of Economic Literature XIX, 1483–1536.

Andersen, P.K., Borgan, Ø., Gill, R.D., Keiding, N., 1993. Statistical Models Based on Counting Processes. Springer, Berlin.

Anderson, T.W., Goodman, L.A., 1957. Statistical inference about Markov chains. Annals of Mathematical Statistics 28, 89–110.

Ashford, J.R., Sowden, R.R., 1970. Multi-variate probit analysis. Biometrics 26, 535–546.

Basmann, R.L., 1988. Causality tests and observationally equivalent representations of econometric models. Journal of Econometrics 39, 69–104.

Billingsley, P., 1961. Statistical methods in Markov chains. Annals of Mathematical Statistics 32, 12–40 correction ibidem 1343.

Billingsley, P., 1995. Probability and Measure, third ed. Wiley, New York.

Bouissou, M.B., Laffont, J.J., Vuong, Q.H., 1986. Tests of non-causality under Markov assumptions for qualitative panel data. Econometrica 54, 395–414.

Chamberlain, G., 1982. The general equivalence of Granger and Sims causality. Econometrica 50, 569–581.

Chamberlain, G., 1984. Panel data. In: Griliches, Z., Intriligator, M.D. (Eds.), Handbook of Econometrics, vol. 2. North-Holland, Amsterdam (Chapter 22).

Colombo, M., Mosconi, R., 1995. Complementarity and cumulative learning effects in the early diffusion of multiple technologies. Journal of Industrial Economics XLIII, 13–48.

Cox, D.R., 1975. Partial likelihood. Biometrika 62, 269–276.

Davidson, J., 2000. Econometric Theory. Blackwell Publishers, London.

Doornik, J.A., 1999. Object-Oriented Matrix Programming Using Ox, third ed. Timberlake Consultants Press, London.

Dueker, M.J., 2001. Forecasting qualitative variables with vector autoregressions: a qualitative VAR model of U.S. recessions, The Federal Reserve Bank of St. Louis Working Paper 2001-012A.

Dueker, M.J., Wesche, K., 2001. Forecasting output with information from business cycle turning points: a qualitative variable VAR. The Federal Reserve Bank of St. Louis Working Paper 2001-019A.

Dufour, J.M., Renault, E., 1998. Short run and long run causality in time series: theory. Econometrica 66, 1099–1126.

Engle, R.F., Hendry, D.F., Richard, J.F., 1983. Exogeneity. Econometrica 51, 277–304.

Fisher, R.A., 1921. On the ‘probable error’ of a coefficient of correlation deduced from a small sample. Metron 1, 1–32.

Florens, J.-P., Fougère, D., 1996. Non-causality in continuous time. Econometrica 64, 1195–1212.

Florens, J.-P., Mouchart, M., 1982. A note on non-causality. Econometrica 50, 583–591.

Fokianos, K., Kedem, B., 1998. Prediction and classification of nonstationary categorical time series. Journal of Multivariate Analysis 67, 277–296.

Geweke, J., 1984. Inference and causality in economic time series models. in: Griliches, Z., Intriligator, M.D. (Eds.), Handbook of Econometrics, vol. 2. North-Holland, Amsterdam (Chapter 19).

Gouriéroux, C., 2000. Econometrics of Qualitative Dependent Variables. Cambridge University Press, Cambridge.

Gouriéroux, C., Monfort, A., Renault, E., 1987. Kullback causality measures. Annales d'Économie et de Statistique 67, 369–410.

Granger, C.W.J., 1969. Investigating causal relations by econometric models and cross-spectral methods. Econometrica 37, 424–438.

Granger, C.W.J., 1988. Recent developments in a concept of causality. Journal of Econometrics 39, 199–211.

Gray, R.M., Kieffer, J.C., 1980. Asymptotically mean stationary measures. Annals of Probability 8, 962–973.

Grether, D.M., Maddala, G.S., 1982. A time series model with qualitative variables. In: Deistler, M., Furst, E., Schwodiauer, G. (Eds.), Games, Economic Dynamics, and Time Series Analysis. Physica-Verlag, Wien-Wurzburg.

Gumbel, E.J., 1961. Bivariate logistic distributions. Journal of the American Statistical Association 56, 335–349.

Heckman, J.J., 1978. Simple statistical models for discrete panel data developed and applied to test the hypothesis of true state dependence against the hypothesis of spurious state dependence. Annales de l'INSEE 30–31, 227–269.

Heckman, J.J., 1981. Statistical models for discrete panel data. In: Manski, C.F., McFadden, D.L. (Eds.), Structural Analysis of Discrete Data with Econometric Applications. MIT Press, Cambridge.

Honoré, B.E., 2002, Non-linear models with panel data. Cemmap working paper CWP13/02.

Hotelling, H., 1953. New light on the correlation coefficient and its transforms. Journal of the Royal Statistical Society. Series B 15, 193–232.

Hougaard, P., 1987. Modelling multivariate survival. Scandinavian Journal of Statistics 14, 291–304.

Johnson, N.L., Kotz, S., 1972. Continuous Multivariate Distributions. Wiley, New York.

Kalbfleisch, J.D., Prentice, R.L., 1980. The Statistical Analysis of Failure Time Data. Wiley, New York.

Kaufmann, H., 1987. Regression models for nonstationary categorical time series: asymptotic estimation theory. Annals of Statistics 15, 79–98.

Koo, H.P., Janowitz, B.K., 1983. Interrelationships between fertility and marital disolution: results of a simultaneous logit model. Demography 20, 129–145.

Lazarsfeld, P.F., 1948. The use of panels in social research. Proceedings of the American Philosophical Society 92, 406–410.

Lillard, L.A., 1993. Simultaneous equations for hazards. Marriage duration and fertility timing. Journal of Econometrics 56, 189–217.

Lillard, L.A., Waite, L.J., 1993. A joint model of marital childbearing and marital disruption. Demography 30, 653–681.

Maddala, G.S., 1987. Limited dependent variable models using panel data. Journal of Human Resources 22, 307–338.

McCullough, B.C., 1978. Effects of variables using panel data: a review of techniques. Public Opinion Quarterly 42, 199–200.

Morimune, K., 1979. Comparisons of normal and logistic models in the bivariate dichotomous analysis. Econometrica 47, 957–976.

Murphy, A., 1994. Testing normality in bivariate probit models: a simple artificial regression based LM test, WP94/27, Department of Economics at University College Dublin.

Petersen, T., 1995. Models for interdependent event-history data: specification and estimation. Sociological Methodology 25, 317–375.

Schwarz, G., 1978. Estimating the dimension of a model. Annals of Statistics 6, 461–464.

Schweder, T., 1970. Composable Markov processes. Journal of Applied Probability 7, 400–410.

Slud, E., Kedem, B., 1994. Partial likelihood analysis of logistic regression and autoregression. Statistica Sinica 4, 89–106.

Stolnitz, G.J., 1983. Three to five main challenges to demographic research. Demography 20, 415–432.

Swamy, P.A.V.B., von zur Muehlen, P., 1988. Further thoughts on testing for causality with econometric models. Journal of Econometrics 39, 105–147.

Tuma, N.B., 1980. When can interdependence in a dynamic system of qualitative variables be ignored? Sociological Methodology 11, 358–391.

Waite, L.J., Lillard, L.A., 1991. Children and marital disruption. American Journal of Sociology 96, 930–953.

Winship, C., 1986. Heterogeneity and interdependence: a test using survival models. Sociological Methodology 16, 250–282.

Wong, W.H., 1986. Theory of partial likelihood. Annals of Statistics 14, 88–123.

Wooldridge, J.M., 2000. A framework for estimating dynamic, unobserved effects panel data models with possible feedback to future explanatory variables. Economics Letters 68, 245–250.

Wooldridge, J.M., 2002. Simple solutions to the initial conditions problem for dynamic, nonlinear panel data models with unobserved heterogeneity, Working Paper (http://www.msu.edu/~ec/faculty/wooldridge/wooldridge.html)

Yamaguchi, K., 1990. Logit and multinomial logit models for discrete-time event-history analysis: a causal analysis of interdependent discrete-state processes. Quality & Quantity 24, 323–341.