Multivariate binary time series models for absence/presence data in ecology

Lionel Truquet (CREST-ENSAI, Rennes),

Motivation

- In ecology, the study of **absence-presence of species** in an ecosystem is an important problem widely considered in the literature.

- Such studies require to **explain or to forecast some binary vectors with coordinates** 0 **or** 1, depending if a given species is present or absent in a specific area.

- How to model the presence/absence data across the time and to identify possible patterns (attraction hypothesis between the species, influence of the environment, time dependencies...)?

- **Time series analysis of binary vectors is far from being well documented** if we target complex modeling (study of autoregressive processes, modeling the influence of exogenous regressors, spatio-temporal analysis if data are sampled at different sites).

Motivation through an example

- In Sebastián-González et al. [Proc. R. Soc. B, 2010], waterbird surveys are considered in a set of irrigation ponds.

- At each pond, the absence/presence data of 7 waterbirds were recorded during several years.

- Many covariates are available:
  - Fixed environmental and spatial covariates (pond area, presence or absence of shore/submerged/reed vegetation...).
  - Absence/presence of the same species at time $t-1$.
  - Absence/presence of other species at time t.

- The various covariates (time-varying and non time-varying) seem to have an impact on the dynamic.

Main questions

- How to develop an autoregressive time-series model for binary data in which various type of covariates can be included ?
- How to get statistical guarantees for inference when a longitudinal analysis is necessary ?
- The model used in a aforementioned reference is a **multivariate logistic model**. At a given pond, let $Y_t \in \{0,1\}^k$ the absence/presence vector of k species at time t.

$$Y_{it} = \mathbb{1}_{\lambda_{it}+logit(\Phi(\varepsilon_{it}))>0}, \quad \lambda_t = X_t\beta.$$

  - Φ is the Gaussian cdf.
  - ε_t is a Gaussian vector with mean 0 and correlation matrix R.
  - X_t available covariates at time t.
- This multivariate extension of the logistic model is a standard choice in the static case. See O'Brien and Dunson [Biometrics, 2004].
- An alternative (with $logit \leftrightarrow \Phi^{-1}$) is the multivariate probit model widely used in econometrics (Chib and Greenberg [Biometrica, 1998]).

A dynamic version of multivariate probit models

- We present a time series analogue of the multivariate probit model.
- We investigate a **frequentist** approach for parameters inference.
- We derive **stationarity properties** for such models (useful at least for deriving short-term interactions).
- We adapt the single path framework to a **longitudinal type approach**, taking in account of the information available at different observation sites.
- We focus on the multivariate probit case but the multivariate logistic model can be studied in the same way.

Outline

Outline

Dynamic multivariate probit model

- The models writes as

$$Y_{it} = \mathbb{1}_{\lambda_{it}+\varepsilon_{it}>0}, \quad \lambda_t = d + \sum_{j=1}^{p} A_j Y_{t-j} + B X_{t-1}.$$

- $X_t \in \mathbb{R}^d$ denotes the (random) covariates available at time t.
- $d \in \mathbb{R}^k$, $A_1, \ldots, A_p$ are $k \times k$ matrices and B is a matrix of size $k \times d$
- The noise components ε_t are i.i.d. $\mathcal{N}_k(0, R)$.
- The process $(X_t, \varepsilon_t)_{t \in \mathbb{Z}}$ is assumed to be stationary and ε_t is independent of $(\varepsilon_s, X_s)_{s \leq t-1}$.
- The process (X_t) is not required to be ergodic (i.e. partial sums will not necessarily converge to a non-random limit). For instance, $X_t = (Z, W_t)$, Z being the **non time-varying random covariates** and W_t the **time-varying random covariates**.

Existence of a stationary solution

- Without covariates, the model is an irreducible finite-state Markov chain. There then exists a unique invariant probability measure, without any other condition.

- With covariates, $(Y_t)_{t\in\mathbb{Z}}$ is no more a Markov chain and the stationarity conditions are less clear.

- Intuitively, the result should remain the same: ε_t has a full support and from any set of past binary vectors, the probability of reaching any arbitrary subsequent binary vector is positive.

- We use a random mapping approach. For instance if $p = 1$, $Y_t = F_{X_{t-1},\varepsilon_t}(Y_{t-1})$ and a meaningful approach for deriving a stationary solution is to study the backward iterations of the random maps:

$$Y_t := \lim_{s\to\infty} F_{X_{t-1},\varepsilon_t} \circ \cdots \circ F_{X_{t-s-1},\varepsilon_{t-s}}(y).$$

- One can show that such almost sure limit always exists and does not depend on the initial binary vector y.

A proof with a picture ($p = 1$, $k = 2$ in the ergodic case)

Set $C_t = \cap_{i=1}^k \left\{\varepsilon_{i,t} + \sum_{\ell=1}^d B(i,\ell)X_{\ell,t} + h > 0\right\}$ with

$$h = \min_{1 \le i \le k} \min_{y' \in \{0,1\}^k} \left\{ d_i + \sum_{\ell=1}^k A_1(i,\ell) y'_\ell \right\}.$$

Then $\mathbb{P}(C_t) = \mathbb{P}(C_0) > 0$ and $T(\omega) = \inf\{h \ge 1 : \omega \in C_{t-h}\} < \infty$ a.s.

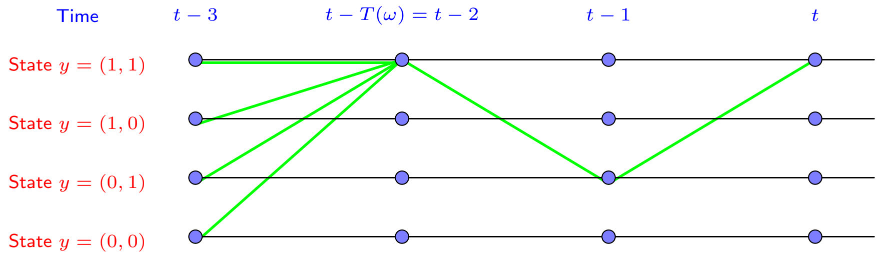

Figure: Coalescence of the paths for backward iterations

Formal result

$$Y_{i,t} = \mathbb{1}_{\lambda_{i,t}+\varepsilon_{i,t}>0}, \quad \lambda_t = d + \sum_{j=1}^{p} A_j Y_{t-j} + B X_t.$$

Let

$$\mathcal{F}_t = \sigma\left((X_{s-1}, \varepsilon_s) : s \leq t\right).$$

The previous convergence can also be obtained (with more tedious arguments) under the non-ergodic scenario.

Theorem

There exists a unique stationary and $(\mathcal{F}_t)_t$-adapted process $(Y_t)_{t\in\mathbb{Z}}$ solutions of the previous recursions.

1. *There exists a representation $Y_t = H\left(\varepsilon_t, X_{t-1}, \varepsilon_{t-1}, X_{t-2}, \ldots\right)$ where $H = \left(\mathbb{R}^k \times \mathbb{R}^d\right)^{\mathbb{Z}} \to \{0,1\}^k$ is a measurable function.*
2. *If the process $(X_t, \varepsilon_t)_{t\in\mathbb{Z}}$ is ergodic, so is the process $(Y_t)_{t\in\mathbb{Z}}$.*

Outline

Drawbacks of (conditional) likelihood inference

- Setting $I_1 = (0, \infty)$ and $I_0 = (-\infty, 0]$,

$$\begin{aligned} & \mathbb{P}\left(\cap_{i=1}^{k}\{Y_{i,t} = s_i\}|\mathcal{F}_{t-1}\right) \\ = \; & \mathbb{P}\left(\cap_{i=1}^{k}\left\{\lambda_{i,t} + \varepsilon_{i,t} \in I_{s_i}\right\}|\mathcal{F}_{t-1}\right) \\ = \; & \int_{I_{s_1} - \lambda_{1,t}} \cdots \int_{I_{s_k} - \lambda_{k,t}} \phi_R(x_1, \ldots, x_k) dx_1 \cdots dx_k, \end{aligned}$$

  where ϕ_R denotes the Gaussian density in $\mathbb{R}^k$ with mean 0 and correlation matrix R.

- The log-likelihood function for (θ, R), $\theta = (d, A_1, \ldots, A_p, B)$, is defined by

$$\mathcal{L}_n(\theta, R) = \sum_{t=p+1}^{T} \log\left[\int_{I_{s_1} - \lambda_{1,t}(\theta)} \cdots \int_{I_{s_k} - \lambda_{k,t}(\theta)} \phi_R(x) dx_1 \cdots dx_k\right]. \quad (1)$$

- **For multivariate probit models, numerical evaluation of the likelihood is difficult.**

Alternative: Pseudo-likelihood inference for θ (step 1)

$$Y_{i,t} = \mathbb{1}_{\lambda_{i,t}+\varepsilon_{i,t}>0}, \quad \lambda_t = d + \sum_{j=1}^{p} A_j Y_{t-j} + B X_{t-1}.$$

- Set

$$\overline{\mathcal{L}}(\theta) = \sum_{t=p+1}^{n} \sum_{i=1}^{k} \left[Y_{i,t} \log \Phi\left(\lambda_{i,t}(\theta)\right) + (1 - Y_{i,t}) \log \Phi\left(-\lambda_{i,t}(\theta)\right) \right]$$

  and $\hat{\theta} = \arg\max_{\theta \in \Theta} \overline{\mathcal{L}}(\theta)$.

- Estimation is done as if $\varepsilon_{1,t}, \ldots, \varepsilon_{k,t}$ were independent: **Pseudo-likelihood approach**.

- Maximization can be obtained "equation by equation" since for $1 \leq i \leq k$ and $t \in \mathbb{Z}$,

$$\lambda_{i,t}(\theta) = \sum_{h=1}^{p} \sum_{\ell=1}^{k} A_h(i,\ell) Y_{j,t-h} + \sum_{\ell=1}^{d} B(i,\ell) X_{\ell,t-1}$$

Pairwise composite likelihood estimation for R (step 2)

- For $1 \leq i < i' \leq k$, set $R_{i,i'} = \begin{pmatrix} 1 & r_{i,i'} \\ r_{i,i'} & 1 \end{pmatrix}$. $\hat{\theta}$ pseudo-likelihood estimator.
- Set

$$\hat{r}_{i,i'} = \text{argmax} \sum_{t=p+1}^{n} \log \int_{I_{Y_{i,t}} - \lambda_{i,t}(\hat{\theta})} \int_{I_{Y_{i',t}} - \lambda_{i',t}(\hat{\theta})} \phi_{r_{i,i'}}(x_i, x_{i'}) dx_i dx_{i'}$$

$$\text{argmax} \sum_{t=p+1}^{n} \log \left\{ \int_{I_{Y_{i,t}} - \lambda_{i,t}(\hat{\theta})} \Phi \left((2Y_{i',t} - 1) \frac{\lambda_{i',t}(\hat{\theta}) - r_{i,i'} x_i}{\sqrt{1 - r_{i,i'}^2}} \right) \phi(x_i) dx_i \right\}.$$

- For $s_i, s_j \in \{0, 1\}$,

$$\int_{I_{s_i} - \lambda_{i,t}(\theta_0)} \Phi \left((2s_{i'} - 1) \frac{\lambda_{i',t}(\theta) - r_{0,i,i'} x_i}{\sqrt{1 - r_{i,i'}^2}} \right) \phi(x_i) dx_i$$

  is simply equal to $\mathbb{P}_{\theta,R}(Y_{i,t} = s_i, Y_{i',t} = s_{i'} | \mathcal{F}_{t-1})$, which explains the terminology pairwise (conditional) likelihood.
- See Varin et al. [Stat. Sinica, 2011] for an overview of composite likelihood methods.

Asymptotic results for ergodic paths

Theorem

Assume that (θ, R) *are in a compact set and* $\mathbb{E}|X_1|^2 < \infty$. *Then, up to an identifiability constraint on the covariates* X_t*:*

1. $(\hat{\theta}, \hat{R})$ *is strongly consistent and* $\sqrt{T}\left(\hat{\theta} - \theta\right)$ *converges in distribution towards a Gaussian distribution with mean* 0.
2. *If additionally,* $\mathbb{E}\left[\exp\left(\kappa|X_1|^2\right)\right] < \infty$ *for some* $\kappa > 0$, $\sqrt{T}\left(\hat{R} - R\right)$ *is also asymptotically Gaussian with mean* 0.

Outline

Two scenarios

- The model is now fitted using data coming from different observations sites $j = 1, \ldots, n$.

$$Y_{i,j,t} = \mathbb{1}_{\lambda_{i,j,t}+\varepsilon_{i,j,t}>0}, \quad \lambda_{j,t} = d + \sum_{\ell=1}^{p} A_\ell Y_{j,t-h} + BX_{j,t}, \quad 1 \leq t \leq T.$$

- The model is simplistic. No heterogeneity or individual effects for the different sites (e.g. d does not depend on i) as in the classical framework of panel data.
- We want to get an asymptotic for parameters inference when both n and T grows to infinity (not necessarily $T = T_n$ and $n \to \infty$).
- **Scenario** 1: $X_{j,t} = (Z_j, W_{j,t})$. In this case, $(X_{j,t}, \varepsilon_{j,t})_{t\in\mathbb{Z}}$ are i.i.d. across the index j. Z_j, $(W_{j,t})_t$ and $(\varepsilon_{j,t})_t$ are mutually independent. Moreover, $(W_{j,t})_t$ is an ergodic process.
- **Scenario** 2: we assume existence of common factors $X_{j,t} = (Z_j, W_t)$. In this case $(Z_j)_j$, $(W_t)_t$ and $(\varepsilon_{j,t})_{j,t}$ are assumed to be independent, the Z_j's are i.i.d. and $(W_t)_{t\in\mathbb{Z}}$ is an ergodic process.

Law of large numbers over two indices

In each scenario, we have the following law of large numbers,

$$\frac{1}{nT}\sum_{j=1}^{n}\sum_{t=1}^{T} H_{j,t} \to \mathbb{E}\left(H_{1,1}\right)$$

if

$$H_{j,t} = H\left(\varepsilon_{j,t}, X_{j,t-1}, \varepsilon_{j,t-1}, X_{j,t-2}, \ldots\right)$$

satisfies $\mathbb{E}|H_{1,1}|\log^+ H_{1,1} < \infty$ and $\min(n,T) \to \infty$.

The random field $(H_{j,t})_{j,t}$ is stationary and the problem is related to ergodic properties for the two $\mathbb{Z}^2-$actions, $\theta_1 H_{j,t} = H_{j+1,t}$ and $\theta_2 H_{j,t} = H_{j,t+1}$

1. In the first scenario, θ_1 is ergodic (i.i.d assumption).
2. In the second scenario, none of the transformation θ_1, θ_2 are ergodic. However, the intersection of their respective invariant sigma-fields is trivial.

This is sufficient to get parameter consistency in the longitudinal case.

Martingale central limite theorem for two indices

- The second problem for asymptotic normality of parameter estimates concerns limit theorems for sum of square integrable martingale differences $S_{nT} := \frac{1}{\sqrt{nT}} \sum_{j=1}^{n} \sum_{t=1}^{T} H_{j,t}$,

$$H_{j,t} = H\left(\varepsilon_{j,t}, W_{j,t-1}, \varepsilon_{j,t-1}, W_{j,t-2}, \ldots\right).$$

- Here, $\mathcal{F}_{j,t} = \sigma\left((W_{i,s}, \varepsilon_{i,s}) : i \leq j, s \leq t\right)$ and

$$\mathbb{E}\left[H_{j,t} \middle| \vee_{1 \leq i \leq n} \mathcal{F}_{i,t-1}\right] = \mathbb{E}\left[H_{j,t} \middle| \vee_{1 \leq t \leq T} \mathcal{F}_{j-1,t}\right] = 0.$$

- In the **first scenario**, $S_{n,T}$ has a **Gaussian limit**.
- In the **second scenario** with common factors across the sites, $S_{n,T}$ is asymptotically distributed as a **mixture of Gaussian distributions**.
- A general result of Volný [SPA, 2019] applies to the second scenario.

To take away...

- It is possible to develop a theory for time series analogues of multivariate binary models (probit, logistic,...) that take in account both endogenous and exogenous regressors.
- Some numerically tractable inference procedures are possible.
- One can also fit the model to panel type data (at least under some stringent assumptions for the asymptotic guarantees).
- Finite-sample accuracy of the proposed inference procedure remains to evaluate in the time series context (in progress).
- It could be interesting to get a more realistic modeling for longitudinal analysis (heterogeneous intercepts $d = d_j$, spatial correlation of the errors $(\varepsilon_{j,t})_j$).