Dynamic modeling of abundance data in ecology

Guillaume Franchi

ENSAI, Bruz

ECODEP Conference
12 February 2024

Outlines

Outlines

I. Introduction

II. Modeling relative abundance

III. Modeling Absence/Presence of species

Ecological definitions

Consider an ecosystem with a number d of species.

Ecological definitions

Consider an ecosystem with a number d of species.

Definition

Relative abundance: the vector of proportions of each species in the whole ecosystem.

Ecological definitions

Consider an ecosystem with a number d of species.

Definition

Relative abundance: the vector of proportions of each species in the whole ecosystem.

It is an element of the **simplex**

$$\mathcal{S}_{d-1} = \left\{ (y_1, \ldots, y_d) \in]0, +\infty[^d \;/\; \sum_{i=1}^{d} y_i = 1 \right\}.$$

Ecological definitions

Consider an ecosystem with a number d of species.

Definition

Relative abundance: the vector of proportions of each species in the whole ecosystem.

It is an element of the **simplex**

$$\mathcal{S}_{d-1} = \left\{ (y_1, \ldots, y_d) \in]0, +\infty[^d \ / \ \sum_{i=1}^{d} y_i = 1 \right\}.$$

One can also consider the **Shannon entropy**

$$\forall y = (y_1, \ldots, y_d) \in \mathcal{S}_{d-1}, \ I_S(y) = -\sum_{i=1}^{d} y_i \log(y_i).$$

Objectives

- Make predictions about the relative abundance of an ecosystem over time.

Objectives

- Make predictions about the relative abundance of an ecosystem over time.
- Understand the dynamics of this ecosystem:

Objectives

- Make predictions about the relative abundance of an ecosystem over time.
- Understand the dynamics of this ecosystem:
  - ➜ The dynamic of each species.

Objectives

- Make predictions about the relative abundance of an ecosystem over time.
- Understand the dynamics of this ecosystem:
  - The dynamic of each species.
  - The interaction between species.

Objectives

- Make predictions about the relative abundance of an ecosystem over time.
- Understand the dynamics of this ecosystem:
  - ➜ The dynamic of each species.
  - ➜ The interaction between species.
  - ➜ The impact of exogenous variables.

An example

We consider a population of insects studied in a sugar cane field in La Réunion, during the years 2022 and 2023.

An example

We consider a population of insects studied in a sugar cane field in La Réunion, during the years 2022 and 2023.

We focus on three groups of species:

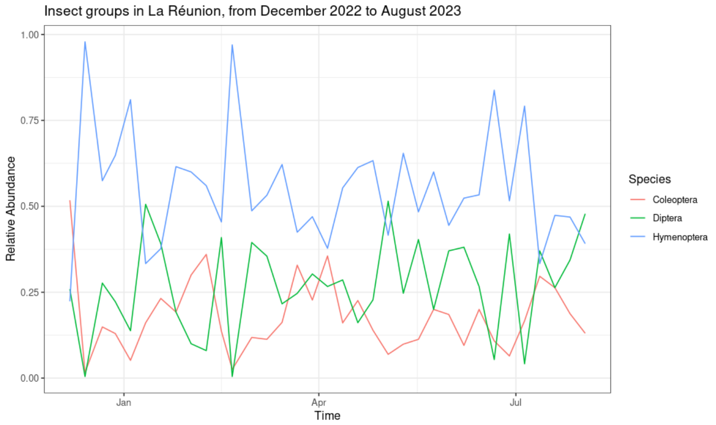
(a) Coleoptera

(b) Hymenoptera

(c) Diptera

An example

We consider a population of insects studied in a sugar cane field in La Réunion, during the years 2022 and 2023.



Outlines

Chain with complete connections (1/2)

We propose to model our abundance along time by a time series $(Y_t)_{t\in\mathbb{Z}}$ valued in the simplex $\mathcal{S}_{d-1}$.

Chain with complete connections (1/2)

We propose to model our abundance along time by a time series $(Y_t)_{t\in\mathbb{Z}}$ valued in the simplex $\mathcal{S}_{d-1}$.

The idea is to define $(Y_t)_{t\in\mathbb{Z}}$ as a **chain with complete connections**

$$\mathbb{P}\left(Y_{t+1} \in A \mid Y_t^- = y_t^-\right) = P\left(A \mid y_t^-\right)$$

Chain with complete connections (1/2)

We propose to model our abundance along time by a time series $(Y_t)_{t\in\mathbb{Z}}$ valued in the simplex $\mathcal{S}_{d-1}$.

The idea is to define $(Y_t)_{t\in\mathbb{Z}}$ as a **chain with complete connections**

$$\mathbb{P}\left(Y_{t+1} \in A \ \middle|\ Y_t^- = y_t^-\right) = P\left(A \ \middle|\ y_t^-\right)$$

where:

- P is a transition kernel from source $\mathcal{S}_{d-1}^{\mathbb{N}}$ and target $\mathcal{S}_{d-1}$,

Chain with complete connections (1/2)

We propose to model our abundance along time by a time series $(Y_t)_{t\in\mathbb{Z}}$ valued in the simplex $\mathcal{S}_{d-1}$.

The idea is to define $(Y_t)_{t\in\mathbb{Z}}$ as a **chain with complete connections**

$$\mathbb{P}\left(Y_{t+1} \in A \;\middle|\; Y_t^- = y_t^-\right) = P\left(A \;\middle|\; y_t^-\right)$$

where:

- P is a transition kernel from source $\mathcal{S}_{d-1}^{\mathbb{N}}$ and target $\mathcal{S}_{d-1}$,
- Y_t^- denotes the entire past of the time series at time t:

$$Y_t^- = \left(Y_t, Y_{t-1}, \ldots\right).$$

Chain with complete connections (2/2)

Remark

Chain with complete connections (2/2)

Remark

- The process $(Y_t)_{t \in \mathbb{Z}}$ has possibly an infinite memory.

Chain with complete connections (2/2)

Remark

- The process $(Y_t)_{t\in\mathbb{Z}}$ has possibly an infinite memory.
- If $P(A \mid Y_t^-)$ depends only on the $p+1$ first values of Y_t^-

$$P(A \mid Y_t^-) = \tilde{P}(A \mid Y_t, Y_{t-1}, \ldots, Y_{t-p}),$$

we obtain a Markov chain.

Chain with complete connections (2/2)

Remark

- The process $(Y_t)_{t\in\mathbb{Z}}$ has possibly an infinite memory.
- If $P(A \mid Y_t^-)$ depends only on the $p+1$ first values of Y_t^-

$$P(A \mid Y_t^-) = \tilde{P}(A \mid Y_t, Y_{t-1}, \ldots, Y_{t-p}),$$

  we obtain a Markov chain.
- It is possible to add a process of exogenous variables $(X_t)_{t\in\mathbb{Z}}$ to the dynamic of the process $(Y_t)_{t\in\mathbb{Z}}$

$$\mathbb{P}\left(Y_{t+1} \in A \mid Y_t^- = y_t^-, X_t^- = x_t^-\right) = P\left(A \mid y_t^-, x_t^-\right)$$

Existence of the process

Existence of the process

Theorem 1

Under assumptions **A1** and **A2** below, there exists a time series $(Y_t)_{t\in\mathbb{Z}}$ which is strictly stationary such that

$$\forall t \in \mathbb{Z},\ \mathbb{P}(Y_{t+1} \in A \mid Y_t^- = y_t^-) = P(A \mid y_t^-).$$

Existence of the process

Theorem 1

Under assumptions **A1** and **A2** below, there exists a time series $(Y_t)_{t\in\mathbb{Z}}$ which is strictly stationary such that

$$\forall t \in \mathbb{Z},\ \mathbb{P}(Y_{t+1} \in A \mid Y_t^- = y_t^-) = P(A \mid y_t^-).$$

Furthermore, its distribution is unique and it is ergodic.

Existence of the process

Theorem 1

Under assumptions **A1** and **A2** below, there exists a time series $(Y_t)_{t\in\mathbb{Z}}$ which is strictly stationary such that

$$\forall t \in \mathbb{Z},\ \mathbb{P}(Y_{t+1} \in A \mid Y_t^- = y_t^-) = P(A \mid y_t^-).$$

Furthermore, its distribution is unique and it is ergodic.

Assumptions

A1 $b_0 = \sup\left\{ d_{TV}\left(P(\cdot \mid y), P(\cdot \mid z)\right) \ / \ y, z \in \mathcal{S}_{d-1}^{\mathbb{N}} \right\} < 1.$

Existence of the process

Theorem 1

Under assumptions **A1** and **A2** below, there exists a time series $(Y_t)_{t\in\mathbb{Z}}$ which is strictly stationary such that

$$\forall t \in \mathbb{Z},\ \mathbb{P}(Y_{t+1} \in A \mid Y_t^- = y_t^-) = P(A \mid y_t^-).$$

Furthermore, its distribution is unique and it is ergodic.

Assumptions

A1 $b_0 = \sup\left\{d_{TV}\left(P(\cdot \mid y), P(\cdot \mid z)\right)\ /\ y, z \in \mathcal{S}_{d-1}^{\mathbb{N}}\right\} < 1.$

A2 For $m \geqslant 1$, we denote

$$b_m = \sup\left\{d_{TV}\left(P(\cdot \mid y), P(\cdot \mid z)\right)\ /\ y, z \in \mathcal{S}_{d-1}^{\mathbb{N}}, y \overset{m}{=} z\right\}.$$

Existence of the process

Theorem 1

Under assumptions **A1** and **A2** below, there exists a time series $(Y_t)_{t\in\mathbb{Z}}$ which is strictly stationary such that

$$\forall t \in \mathbb{Z},\ \mathbb{P}(Y_{t+1} \in A \mid Y_t^- = y_t^-) = P(A \mid y_t^-).$$

Furthermore, its distribution is unique and it is ergodic.

Assumptions

A1 $b_0 = \sup\left\{d_{TV}\left(P(\cdot \mid y), P(\cdot \mid z)\right)\ /\ y, z \in \mathcal{S}_{d-1}^{\mathbb{N}}\right\} < 1.$

A2 For $m \geqslant 1$, we denote

$$b_m = \sup\left\{d_{TV}\left(P(\cdot \mid y), P(\cdot \mid z)\right)\ /\ y, z \in \mathcal{S}_{d-1}^{\mathbb{N}}, y \overset{m}{=} z\right\}.$$

We have $\sum_{m\in\mathbb{N}} b_m < \infty$.

Dirichlet model (1/3)

A natural proposal for $P(\cdot \mid Y_t^-)$ is a **Dirichlet distribution**

$$P(\cdot \mid Y_t^-) = \mathrm{Dir}(\lambda_t, \varphi_t).$$

Dirichlet model (1/3)

A natural proposal for $P(\cdot \mid Y_t^-)$ is a **Dirichlet distribution**

$$P(\cdot \mid Y_t^-) = \mathrm{Dir}(\lambda_t, \varphi_t).$$

In the context of Ecology, it has been suggested in Marquet et al. (2017) that the relative abundance of a given species in large ecosystems is often compatible with a Beta distribution.

Dirichlet model (1/3)

A natural proposal for $P(\cdot \mid Y_t^-)$ is a **Dirichlet distribution**

$$P(\cdot \mid Y_t^-) = \mathrm{Dir}(\lambda_t, \varphi_t).$$

In the context of Ecology, it has been suggested in Marquet et al. (2017) that the relative abundance of a given species in large ecosystems is often compatible with a Beta distribution.

Remark

The Dirichlet distribution is actually the generalization of the Beta distribution.

Dirichlet model (1/3)

A natural proposal for $P(\cdot \mid Y_t^-)$ is a **Dirichlet distribution**

$$P(\cdot \mid Y_t^-) = \mathrm{Dir}(\lambda_t, \varphi_t).$$

In the context of Ecology, it has been suggested in Marquet et al. (2017) that the relative abundance of a given species in large ecosystems is often compatible with a Beta distribution.

Remark

The Dirichlet distribution is actually the generalization of the Beta distribution.

The Dirichlet distribution $\mathrm{Dir}(\lambda, \varphi)$, supported by $\mathcal{S}_{d-1}$ is characterized by

Dirichlet model (1/3)

A natural proposal for $P(\cdot \mid Y_t^-)$ is a **Dirichlet distribution**

$$P(\cdot \mid Y_t^-) = \mathrm{Dir}(\lambda_t, \varphi_t).$$

In the context of Ecology, it has been suggested in Marquet et al. (2017) that the relative abundance of a given species in large ecosystems is often compatible with a Beta distribution.

Remark

The Dirichlet distribution is actually the generalization of the Beta distribution.

The Dirichlet distribution $\mathrm{Dir}(\lambda, \varphi)$, supported by $\mathcal{S}_{d-1}$ is characterized by

- its mean vector $\lambda = (\lambda_1, \ldots, \lambda_d)$;

Dirichlet model (1/3)

A natural proposal for $P(\cdot \mid Y_t^-)$ is a **Dirichlet distribution**

$$P(\cdot \mid Y_t^-) = \mathrm{Dir}(\lambda_t, \varphi_t).$$

In the context of Ecology, it has been suggested in Marquet et al. (2017) that the relative abundance of a given species in large ecosystems is often compatible with a Beta distribution.

Remark

The Dirichlet distribution is actually the generalization of the Beta distribution.

The Dirichlet distribution $\mathrm{Dir}(\lambda, \varphi)$, supported by $\mathcal{S}_{d-1}$ is characterized by

- its mean vector $\lambda = (\lambda_1, \ldots, \lambda_d)$;
- a dispersion parameter $\varphi > 0$.

Dirichlet model (2/3)

➜ In the spirit of the logistic regression, we propose that for all $t \in \mathbb{Z}$

$$\mathrm{alr}(\lambda_t) = \eta_0 + \sum_{k \geqslant 1} \eta_k \overline{Y}_{t-k} + \sum_{k \geqslant 1} \zeta_k X_{t-k},$$

where alr is the mapping

$$\begin{aligned} \mathrm{alr} : \quad & \mathcal{S}_{d-1} \longrightarrow \mathbb{R}^{d-1} \\ & y = (y_1, \dots, y_d) \longmapsto \left(\log\left(\frac{y_1}{y_d}\right), \dots, \log\left(\frac{y_{d-1}}{y_d}\right) \right), \end{aligned}$$

$\overline{Y} = (Y_1, \dots, Y_{d-1})$, and the η's and ζ's are matrices.

Dirichlet model (2/3)

➜ In the spirit of the logistic regression, we propose that for all $t \in \mathbb{Z}$

$$\mathrm{alr}(\lambda_t) = \eta_0 + \sum_{k \geqslant 1} \eta_k \overline{Y}_{t-k} + \sum_{k \geqslant 1} \zeta_k X_{t-k},$$

where alr is the mapping

$$\begin{aligned} \mathrm{alr} : \quad & \mathcal{S}_{d-1} \longrightarrow \mathbb{R}^{d-1} \\ & y = (y_1, \dots, y_d) \longmapsto \left(\log\left(\frac{y_1}{y_d}\right), \dots, \log\left(\frac{y_{d-1}}{y_d}\right) \right), \end{aligned}$$

$\overline{Y} = (Y_1, \dots, Y_{d-1})$, and the η's and ζ's are matrices.

➜ We also propose that for all $t \in \mathbb{Z}$

$$\varphi_t = \exp\left(\theta_0 + \sum_{k \geqslant 1} \theta_k I_S(Y_{t-k+1}) \right),$$

where the θ_k's are real numbers.

Dirichlet model (3/3)

The matrices η's give us precise information about the interactions between species.

Dirichlet model (3/3)

- The matrices η's give us precise information about the interactions between species.

- The matrices ζ's give us precise information about the impact of exogenous variables on the abundance of species.

Dirichlet model (3/3)

- The matrices η's give us precise information about the interactions between species.

- The matrices ζ's give us precise information about the impact of exogenous variables on the abundance of species.

- The θ's give us information about the volatility of the abundance. The idea is to connect the biodiversity of the ecosystem and its variability.

Return to our example (1/2)

We fit the following Dirichlet model to the population of insects in La Réunion

$$\mathrm{alr}(\lambda_t) = \eta_0 + \eta_1 Y_t + \zeta_1 X_t$$

Return to our example (1/2)

We fit the following Dirichlet model to the population of insects in La Réunion

$$\mathrm{alr}(\lambda_t) = \eta_0 + \eta_1 Y_t + \zeta_1 X_t$$

and

$$\varphi_t = \exp\left(\theta_0 + \theta_1 I_S(Y_t)\right).$$

Return to our example (1/2)

We fit the following Dirichlet model to the population of insects in La Réunion

$$\mathrm{alr}(\lambda_t) = \eta_0 + \eta_1 Y_t + \zeta_1 X_t$$

and

$$\varphi_t = \exp\left(\theta_0 + \theta_1 I_S(Y_t)\right).$$

The vector of covariates X_t is composed by climatic variables such as the total rainfall amount, the temperature, the ground radiation and evapotranspiration.

Return to our example (1/2)

We fit the following Dirichlet model to the population of insects in La Réunion

$$\operatorname{alr}(\lambda_t) = \eta_0 + \eta_1 Y_t + \zeta_1 X_t$$

and

$$\varphi_t = \exp\left(\theta_0 + \theta_1 I_S(Y_t)\right).$$

The vector of covariates X_t is composed by climatic variables such as the total rainfall amount, the temperature, the ground radiation and evapotranspiration.

An optimization of the conditional likelihood is performed to obtain an estimation of the parameters.

Return to our example (2/2)

We kept the 12 last weeks of our data apart from our estimation sample, they are indeed used to compare our predictions with the values observed in the reality.

Return to our example (2/2)

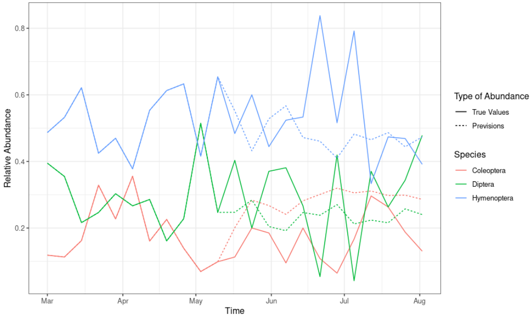
We kept the 12 last weeks of our data apart from our estimation sample, they are indeed used to compare our predictions with the values observed in the reality.

The predictions are made by simulating a thousand trajectories of the abundance for the last 12 weeks, and taking the mean of these trajectories gives us our predicted values.

Return to our example (2/2)

We kept the 12 last weeks of our data apart from our estimation sample, they are indeed used to compare our predictions with the values observed in the reality.

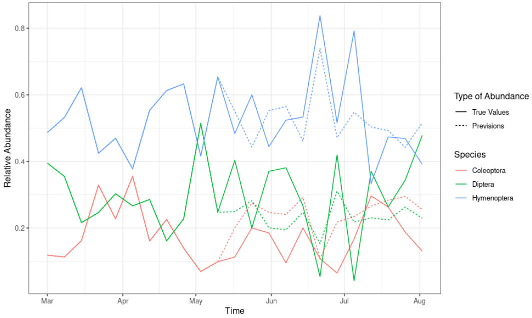
The predictions are made by simulating a thousand trajectories of the abundance for the last 12 weeks, and taking the mean of these trajectories gives us our predicted values.



Return to our example (2/2)

We kept the 12 last weeks of our data apart from our estimation sample, they are indeed used to compare our predictions with the values observed in the reality.

The predictions are made by simulating a thousand trajectories of the abundance for the last 12 weeks, and taking the mean of these trajectories gives us our predicted values.



Limits of the model

Limits of the model

- ✘ In practice, ecological data do not have a lot of observations along time.

Limits of the model

- ✗ In practice, ecological data do not have a lot of observations along time.
- ✗ In practice, there are a lot of zero values in the abundances observed, which will lead to an error when computing our estimators.

Limits of the model

- ✗ In practice, ecological data do not have a lot of observations along time.
- ✗ In practice, there are a lot of zero values in the abundances observed, which will lead to an error when computing our estimators.

- Use panel data.

Limits of the model

- ✗ In practice, ecological data do not have a lot of observations along time.
- ✗ In practice, there are a lot of zero values in the abundances observed, which will lead to an error when computing our estimators.

- Use panel data.
- Model the absence/presence of species in the ecosystem.

Outlines

I. Introduction

II. Modeling relative abundance

III. Modeling Absence/Presence of species

Dynamic probit regression

We model here the absence/presence of d species in an ecosystem at time t by

$$Y_t = (Y_{1,t}, \ldots, Y_{d,t}) \in \{0,1\}^d .$$

Dynamic probit regression

We model here the absence/presence of d species in an ecosystem at time t by

$$Y_t = (Y_{1,t}, \ldots, Y_{d,t}) \in \{0,1\}^d .$$

We assume that for all $i \in \{1, \ldots, d\}$

$$Y_{i,t} = \mathbf{1}_{]0,+\infty[}\,(\lambda_{i,t} + \varepsilon_{i,t})$$

Dynamic probit regression

We model here the absence/presence of d species in an ecosystem at time t by

$$Y_t = (Y_{1,t}, \ldots, Y_{d,t}) \in \{0,1\}^d .$$

We assume that for all $i \in \{1, \ldots, d\}$

$$Y_{i,t} = \mathbf{1}_{]0,+\infty[}\left(\lambda_{i,t} + \varepsilon_{i,t}\right)$$

where

- $\lambda_t = \sum_{l=1}^{p} A_l \cdot Y_{t-l} + B \cdot X_{t-1}$;

Dynamic probit regression

We model here the absence/presence of d species in an ecosystem at time t by

$$Y_t = (Y_{1,t}, \ldots, Y_{d,t}) \in \{0,1\}^d .$$

We assume that for all $i \in \{1, \ldots, d\}$

$$Y_{i,t} = \mathbf{1}_{]0,+\infty[}\left(\lambda_{i,t} + \varepsilon_{i,t}\right)$$

where

- $\lambda_t = \sum_{l=1}^{p} A_l \cdot Y_{t-l} + B \cdot X_{t-1}$;
- $(X_t)_{t\in\mathbb{Z}}$ is a process of covariates;

Dynamic probit regression

We model here the absence/presence of d species in an ecosystem at time t by

$$Y_t = (Y_{1,t}, \dots, Y_{d,t}) \in \{0,1\}^d .$$

We assume that for all $i \in \{1, \dots, d\}$

$$Y_{i,t} = \mathbf{1}_{]0,+\infty[}\left(\lambda_{i,t} + \varepsilon_{i,t}\right)$$

where

- $\lambda_t = \sum_{l=1}^{p} A_l \cdot Y_{t-l} + B \cdot X_{t-1}$;
- $(X_t)_{t\in\mathbb{Z}}$ is a process of covariates;
- $(\varepsilon_t)_{t\in\mathbb{Z}}$ is a sequence of i.i.d random variables with distribution $\mathcal{N}_{\mathbb{R}^d}(0, R)$.

Dynamic probit regression

We model here the absence/presence of d species in an ecosystem at time t by

$$Y_t = (Y_{1,t}, \dots, Y_{d,t}) \in \{0,1\}^d .$$

We assume that for all $i \in \{1, \dots, d\}$

$$Y_{i,t} = \mathbf{1}_{]0,+\infty[}\left(\lambda_{i,t} + \varepsilon_{i,t}\right)$$

where

- $\lambda_t = \sum_{l=1}^{p} A_l \cdot Y_{t-l} + B \cdot X_{t-1}$;
- $(X_t)_{t\in\mathbb{Z}}$ is a process of covariates;
- $(\varepsilon_t)_{t\in\mathbb{Z}}$ is a sequence of i.i.d random variables with distribution $\mathcal{N}_{\mathbb{R}^d}(0, R)$.

It is actually a dynamic version of a multivariate probit regression.

Existence of the process

Existence of the process

Theorem 2

Assume that the process $(\zeta_t)_{t\in\mathbb{Z}}$ defined by

$$\zeta_t = (X_{t-1}, \varepsilon_t)$$

is strongly stationary.

Existence of the process

Theorem 2

Assume that the process $(\zeta_t)_{t\in\mathbb{Z}}$ defined by

$$\zeta_t = (X_{t-1}, \varepsilon_t)$$

is strongly stationary.
There exists a strongly stationary process $(Y_t)_{t\in\mathbb{Z}}$ satisfying

$$\forall i \in \{1,\dots,d\}\,,\ Y_{i,t} = \mathbf{1}_{]0,+\infty[}\left(\lambda_{i,t} + \varepsilon_{i,t}\right).$$

Existence of the process

Theorem 2

Assume that the process $(\zeta_t)_{t\in\mathbb{Z}}$ defined by

$$\zeta_t = (X_{t-1}, \varepsilon_t)$$

is strongly stationary.
There exists a strongly stationary process $(Y_t)_{t\in\mathbb{Z}}$ satisfying

$$\forall i \in \{1, \dots, d\}\,,\ Y_{i,t} = \mathbf{1}_{]0,+\infty[}\left(\lambda_{i,t} + \varepsilon_{i,t}\right).$$

In addition, its distribution is unique.

Remark

Furthermore, if $(\zeta_t)_{t\in\mathbb{Z}}$ is ergodic, $(Y_t)_{t\in\mathbb{Z}}$ is also ergodic.

Estimation results (1/3)

We consider first a single trajectory of an absence/presence process $(Y_t)_{1\leqslant t\leqslant T}$, and we are interested in the estimation of

$$\theta = (A_1, \dots, A_p, B, R)\,.$$

Estimation results (1/3)

We consider first a single trajectory of an absence/presence process $(Y_t)_{1\leqslant t\leqslant T}$, and we are interested in the estimation of

$$\theta = (A_1, \ldots, A_p, B, R)\,.$$

➜ Optimizing the **pseudo conditional log-likelihood**

$$\hat{\theta} = \mathrm{argmax}\ \sum_{t=p+1}^{T} \log\left(\int_{\mathbb{R}^k} \prod_{i=1}^{k} \mathbb{1}_{I_{Y_{i,t}}}\left(\lambda_{i,t} + x_i\right) \varphi_R(x)\mathrm{d}x\right)$$

where φ_R is the density of the distribution $\mathcal{N}(0, R)$ and

$$I_{Y_{i,t}} = \begin{cases}]0, +\infty[& \text{if } Y_{i,t} = 1 \\ ]-\infty, 0] & \text{if } Y_{i,t} = 0 \end{cases}\,.$$

Estimation results (1/3)

We consider first a single trajectory of an absence/presence process $(Y_t)_{1\leqslant t\leqslant T}$, and we are interested in the estimation of

$$\theta = (A_1, \dots, A_p, B, R)\,.$$

➜ Optimizing the **pseudo conditional log-likelihood**

$$\hat{\theta} = \operatorname{argmax} \sum_{t=p+1}^{T} \log\left(\int_{\mathbb{R}^k} \prod_{i=1}^{k} \mathbb{1}_{I_{Y_{i,t}}}\left(\lambda_{i,t} + x_i\right) \varphi_R(x)\mathrm{d}x\right)$$

where φ_R is the density of the distribution $\mathcal{N}(0, R)$ and

$$I_{Y_{i,t}} = \begin{cases}]0, +\infty[& \text{if } Y_{i,t} = 1 \\ ]-\infty, 0] & \text{if } Y_{i,t} = 0 \end{cases}.$$

✘ Difficult function to optimize...

Estimation results (2/3)

We thus propose a two-step method.

Estimation results (2/3)

We thus propose a two-step method.

- We first optimize with respect to $\gamma = (A_1, \dots, A_p, B)$

$$\hat{\gamma} = \operatorname{argmax} \sum_{t=p+1}^{T} \sum_{i=1}^{k} Y_{i,t} \log(\Phi(\lambda_{i,t})) + (1 - Y_{i,t}) \log(\Phi(-\lambda_{i,t}))$$

where Φ denotes the cdf of the gaussian distribution.

Estimation results (2/3)

We thus propose a two-step method.

- We first optimize with respect to $\gamma = (A_1, \ldots, A_p, B)$

$$\hat{\gamma} = \operatorname{argmax} \sum_{t=p+1}^{T} \sum_{i=1}^{k} Y_{i,t} \log(\Phi(\lambda_{i,t})) + (1 - Y_{i,t}) \log(\Phi(-\lambda_{i,t}))$$

where Φ denotes the cdf of the gaussian distribution.

- We then maximize all pairwise conditional likelihoods

$$\widehat{R}(i,j) = \underset{r\ \in]-1,1[}{\operatorname{argmax}} \sum_{t=p+1}^{T} \log \left\{ \int_{I_{Y_{i,t}} - \hat{\lambda}_{i,t}} \Phi\left((2Y_{j,t} - 1) \frac{\hat{\lambda}_{j,t} + r x_i}{\sqrt{1 - r^2}} \right) \varphi(x_i) \mathrm{d}x_i \right\}$$

Estimation results (3/3)

Proposition 1

Assume the process ζ_t is ergodic. Under some reasonable assumptions on the covariates:

1) All estimators $\hat{\theta}$, $\hat{\gamma}$ and $\hat{R}$ are strongly consistent.

2) Moreover, we have the asymptotic normality of

$$\sqrt{T-p}\left(\hat{\theta}-\theta_0\right) \quad \text{and} \quad \sqrt{T-p}\left(\hat{\gamma}-\gamma_0, \widehat{R}-R_0\right).$$

The case of panel data

We now consider a number of n trajectories of an absence/presence process $(Y_{j,t})_{1\leqslant j\leqslant n, 1\leqslant t\leqslant T}$, and are still interested in the estimation of θ.

The case of panel data

We now consider a number of n trajectories of an absence/presence process $(Y_{j,t})_{1\leqslant j\leqslant n, 1\leqslant t\leqslant T}$, and are still interested in the estimation of θ.

The aim is to improve the speed of convergence of our estimators with the number of sites.

The case of panel data

We now consider a number of n trajectories of an absence/presence process $(Y_{j,t})_{1\leqslant j\leqslant n, 1\leqslant t\leqslant T}$, and are still interested in the estimation of θ.

The aim is to improve the speed of convergence of our estimators with the number of sites.

Obtain a general version of Birkhoff's ergodic theorem (Giap & Van Quang, 2016).

The case of panel data

We now consider a number of n trajectories of an absence/presence process $(Y_{j,t})_{1\leqslant j\leqslant n, 1\leqslant t\leqslant T}$, and are still interested in the estimation of θ.

The aim is to improve the speed of convergence of our estimators with the number of sites.

Obtain a general version of Birkhoff's ergodic theorem (Giap & Van Quang, 2016).

Generalize the results about consistency and central limit theorems for M-estimators.

Results about M-estimators (1/2)

Results about M-estimators (1/2)

Usually, if we consider the estimator

$$\hat{\theta} = \text{argmax} \sum_{t=1}^{T} m_\theta(Z_t)$$

where m_θ is a measurable mapping and $(Z_t)_{t \in \mathbb{Z}}$ an ergodic process,

Results about M-estimators (1/2)

Usually, if we consider the estimator

$$\hat{\theta} = \text{argmax} \sum_{t=1}^{T} m_\theta(Z_t)$$

where m_θ is a measurable mapping and $(Z_t)_{t\in\mathbb{Z}}$ an ergodic process, the consistency of $\hat{\theta}$ relies in particular on

$$\mathbb{E}\left(\sup_\theta |m_\theta(Z_0)|\right) < +\infty,$$

Results about M-estimators (1/2)

Usually, if we consider the estimator

$$\hat{\theta} = \operatorname{argmax} \sum_{t=1}^{T} m_\theta(Z_t)$$

where m_θ is a measurable mapping and $(Z_t)_{t\in\mathbb{Z}}$ an ergodic process, the consistency of $\hat{\theta}$ relies in particular on

$$\mathbb{E}\left(\sup_\theta |m_\theta(Z_0)|\right) < +\infty,$$

and its asymptotic normality on

$$\mathbb{E}\left(\|\dot{m}_{\theta_0}(Z_0)\|^2\right) < +\infty.$$

Results about M-estimators (2/2)

In the case of panel data, the same results can be obtained with

$$\hat{\theta} = \operatorname{argmax} \sum_{j=1}^{n} \sum_{t=1}^{T} m_\theta(Z_{j,t}),$$

Results about M-estimators (2/2)

In the case of panel data, the same results can be obtained with

$$\hat{\theta} = \operatorname{argmax} \sum_{j=1}^{n} \sum_{t=1}^{T} m_\theta(Z_{j,t}),$$

➜ by assuming that all processes $(Z_{1,t})_{t\in\mathbb{Z}}, \ldots, (Z_{n,t})_{t\in\mathbb{Z}}$ are mutually independent, and their distributions are identical;

Results about M-estimators (2/2)

In the case of panel data, the same results can be obtained with

$$\hat{\theta} = \operatorname{argmax} \sum_{j=1}^{n} \sum_{t=1}^{T} m_\theta(Z_{j,t}),$$

- ➜ by assuming that all processes $(Z_{1,t})_{t\in\mathbb{Z}}, \ldots, (Z_{n,t})_{t\in\mathbb{Z}}$ are mutually independent, and their distributions are identical;
- ➜ by modifying the "order conditions"

$$\mathbb{E}\left(\sup_\theta \, |m_\theta(Z_{0,0})|^{1+\delta}\right) < +\infty \quad \text{and} \quad \mathbb{E}\left(\|\dot{m}_{\theta_0}(Z_{0,0})\|^{2(1+\delta)}\right) < +\infty;$$

Results about M-estimators (2/2)

In the case of panel data, the same results can be obtained with

$$\hat{\theta} = \operatorname{argmax} \sum_{j=1}^{n} \sum_{t=1}^{T} m_\theta(Z_{j,t}),$$

- by assuming that all processes $(Z_{1,t})_{t\in\mathbb{Z}}, \ldots, (Z_{n,t})_{t\in\mathbb{Z}}$ are mutually independent, and their distributions are identical;
- by modifying the "order conditions"

$$\mathbb{E}\left(\sup_\theta \ |m_\theta(Z_{0,0})|^{1+\delta}\right) < +\infty \quad \text{and} \quad \mathbb{E}\left(\|\dot{m}_{\theta_0}(Z_{0,0})\|^{2(1+\delta)}\right) < +\infty;$$

- and by adding the following "order condition"

$$\mathbb{E}\left(\|\ddot{m}_{\theta_0}(Z_{0,0})\|^{1+\delta}\right) < +\infty$$

for some $\delta > 0$.

Estimation Results for panel data (1/2)

In the case of panel data, we can consider similar estimators as the ones mentioned previously

$$\hat{\theta} = \operatorname{argmax} \sum_{j=1}^{n} \sum_{t=p+1}^{T} \log \left(\int_{\mathbb{R}^k} \prod_{i=1}^{k} \mathbb{1}_{I_{Y_{i,j,t}}} (\lambda_{i,j,t} + x_i) \varphi_R(x) \mathrm{d}x \right),$$

$$\hat{\gamma} = \operatorname{argmax} \sum_{j=1}^{n} \sum_{t=p+1}^{T} \sum_{i=1}^{k} Y_{i,j,t} \log(\Phi(\lambda_{i,j,t})) + (1 - Y_{i,j,t}) \log(\Phi(-\lambda_{i,j,t}))$$

Estimation Results for panel data (1/2)

In the case of panel data, we can consider similar estimators as the ones mentioned previously

$$\hat{\theta} = \operatorname{argmax} \sum_{j=1}^{n} \sum_{t=p+1}^{T} \log \left(\int_{\mathbb{R}^k} \prod_{i=1}^{k} \mathbb{1}_{I_{Y_{i,j,t}}}(\lambda_{i,j,t} + x_i) \varphi_R(x) \mathrm{d}x \right),$$

$$\hat{\gamma} = \operatorname{argmax} \sum_{j=1}^{n} \sum_{t=p+1}^{T} \sum_{i=1}^{k} Y_{i,j,t} \log(\Phi(\lambda_{i,j,t})) + (1 - Y_{i,j,t}) \log(\Phi(-\lambda_{i,j,t}))$$

and

$$\widehat{R}(i_1, i_2) = \operatorname{argmax} \sum_{j=1}^{n} \sum_{t=p+1}^{T} \log \int_{I_{Y_{i_1,j,t}} - \hat{\lambda}_{i_1,j,t}} \Phi\left((2Y_{i_2,j,t} - 1) \frac{\hat{\lambda}_{i_2,j,t} + r x_{i_1}}{\sqrt{1 - r^2}} \right) \varphi(x_{i_1}) \mathrm{d}x_{i_1}.$$

Estimation Results for panel data (2/2)

Proposition 2

Under some reasonable assumptions on the processes $(\zeta_{j,t})_{t\in\mathbb{Z}}$'s:

1) All estimators $\hat{\theta}$, $\hat{\gamma}$ and $\widehat{R}$ are strongly consistent.

2) Moreover, we have the asymptotic normality of

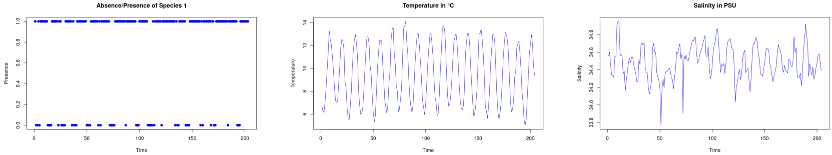
$$\sqrt{n(T-p)}\left(\hat{\theta}-\theta_0\right) \quad \text{and} \quad \sqrt{n(T-p)}\left(\hat{\gamma}-\gamma_0, \widehat{R}-R_0\right).$$

Simulations (1/3)

We simulated the absence/presence of 3 fish species, depending on the temperature and salinity of the water, over 5 sites.

Simulations (1/3)

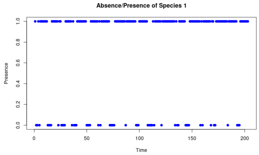
We simulated the absence/presence of 3 fish species, depending on the temperature and salinity of the water, over 5 sites.

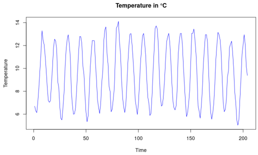


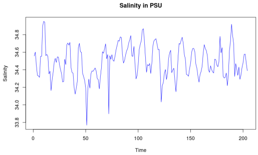
Then, four of these sites are used for estimation, the last one for testing.

Simulations (1/3)

We simulated the absence/presence of 3 fish species, depending on the temperature and salinity of the water, over 5 sites.







Then, four of these sites are used for estimation, the last one for testing. Here, we have

$$\lambda_t = A \cdot Y_{t-1} + B \cdot X_{t-1},$$

where $(X_t)_{t \in \mathbb{Z}}$ is the process composed by the temperature and salinity.

Simulations (2/3)

We obtain the following estimations results

Simulations (2/3)

We obtain the following estimations results

$$A = \begin{pmatrix} 0.2 & 0.1 & -0.2 \\ 0.5 & 0.1 & -0.2 \\ -0.5 & 0.3 & 0.2 \end{pmatrix} \quad \text{and} \quad \hat{A} = \begin{pmatrix} 0.296 & -0.499 & -0.590 \\ 0.444 & 0.320 & -0.138 \\ -0.183 & 0.385 & 0.198 \end{pmatrix},$$

Simulations (2/3)

We obtain the following estimations results

$$A = \begin{pmatrix} 0.2 & 0.1 & -0.2 \\ 0.5 & 0.1 & -0.2 \\ -0.5 & 0.3 & 0.2 \end{pmatrix} \quad \text{and} \quad \hat{A} = \begin{pmatrix} 0.296 & -0.499 & -0.590 \\ 0.444 & 0.320 & -0.138 \\ -0.183 & 0.385 & 0.198 \end{pmatrix},$$

$$B = \begin{pmatrix} 0.5 & -0.1 \\ 0.2 & -0.1 \\ -0.3 & 0.1 \end{pmatrix} \quad \text{and} \quad \hat{B} = \begin{pmatrix} 0.582 & -0.118 \\ 0.232 & -0.110 \\ -0.317 & 0.096 \end{pmatrix},$$

Simulations (2/3)

We obtain the following estimations results

$$A = \begin{pmatrix} 0.2 & 0.1 & -0.2 \\ 0.5 & 0.1 & -0.2 \\ -0.5 & 0.3 & 0.2 \end{pmatrix} \quad \text{and} \quad \hat{A} = \begin{pmatrix} 0.296 & -0.499 & -0.590 \\ 0.444 & 0.320 & -0.138 \\ -0.183 & 0.385 & 0.198 \end{pmatrix},$$

$$B = \begin{pmatrix} 0.5 & -0.1 \\ 0.2 & -0.1 \\ -0.3 & 0.1 \end{pmatrix} \quad \text{and} \quad \hat{B} = \begin{pmatrix} 0.582 & -0.118 \\ 0.232 & -0.110 \\ -0.317 & 0.096 \end{pmatrix},$$

and

$$R = \begin{pmatrix} 1 & 0.2 & -0.5 \\ 0.2 & 1 & -0.3 \\ -0.5 & -0.3 & 1 \end{pmatrix} \quad \text{and} \quad \hat{R} = \begin{pmatrix} 1 & 0.204 & -0.436 \\ 0.204 & 1 & -0.303 \\ -0.436 & 0.204 & 1 \end{pmatrix}.$$

Simulations (3/3)

We then make previsions at horizon 1 for the testing site, and obtain the following accuracy.

Simulations (3/3)

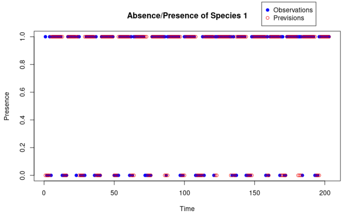
We then make previsions at horizon 1 for the testing site, and obtain the following accuracy.

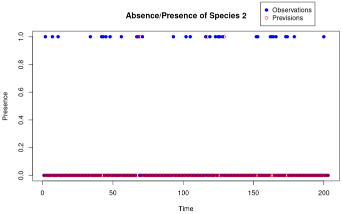
	Species 1	Species 2	Species 3
Accuracy	79.3%	83.7%	78.3%
Mean Presence	73.5%	15.7%	57.8%

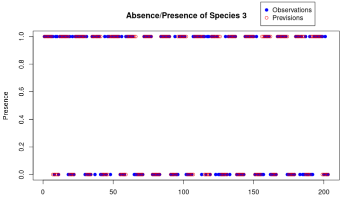
Simulations (3/3)

We then make previsions at horizon 1 for the testing site, and obtain the following accuracy.

	Species 1	Species 2	Species 3
Accuracy	79.3%	83.7%	78.3%
Mean Presence	73.5%	15.7%	57.8%







Real data (1/2)

The previous simulation is based upon a real dataset collected by the government of Scotland: https://data.marine.gov.scot/.

Real data (1/2)

The previous simulation is based upon a real dataset collected by the government of Scotland: https://data.marine.gov.scot/.

We study here the absence/presence of two aquatic micro-organisms: Alexandrium and Dinophysis.

Real data (1/2)

The previous simulation is based upon a real dataset collected by the government of Scotland: https://data.marine.gov.scot/.

We study here the absence/presence of two aquatic micro-organisms: Alexandrium and Dinophysis.

The data were collected monthly from 1997 to 2013 on 5 different locations in Scotland, and we have access to the covariates: Temperature, Salinity and Oxidised Nitrogen.

Real data (2/2)

Once again, 4 sites were used for estimation, and we use the last site to perform previsions at horizon 1.

Real data (2/2)

Once again, 4 sites were used for estimation, and we use the last site to perform previsions at horizon 1.

We obtain the following accuracy.

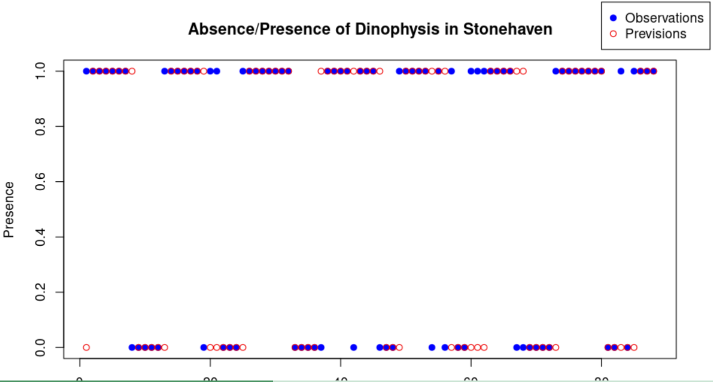
	Alexandrium	Dinophysis
Accuracy	72.7%	75.0%
Mean Presence	65.2%	64.0%

Real data (2/2)

Once again, 4 sites were used for estimation, and we use the last site to perform previsions at horizon 1.

We obtain the following accuracy.

	Alexandrium	Dinophysis
Accuracy	72.7%	75.0%
Mean Presence	65.2%	64.0%



Thank you !