Modeling abundance time series through a GLM model

Guillaume Franchi

ENSAI, Bruz

Ecodep Conference,
22 June 2022

Outlines

Outlines

Definition of abundance

Definition

Definition of abundance

Definition

Standard abundance: counts of each species in the ecosystem. It describes the commonness and rarity of species.

Definition of abundance

Definition

Standard abundance: counts of each species in the ecosystem. It describes the commonness and rarity of species.

Relative abundance: proportions of each species in the whole ecosystem. It describes the biodiversity of the ecosystem.

Notations

A relative abundance $y = (y_1, \ldots, y_d)$ is an element of the **simplex**:

$$\mathcal{S}_{d-1} = \left\{ (x_1, \ldots, x_d) \in]0; +\infty[^d \mid \sum_{i=1}^{d} x_i = 1 \right\}.$$

Notations

A relative abundance $y = (y_1, \dots, y_d)$ is an element of the **simplex**:

$$\mathcal{S}_{d-1} = \left\{ (x_1, \dots, x_d) \in]0; +\infty[^d \mid \sum_{i=1}^{d} x_i = 1 \right\}.$$

For an abundance $y \in \mathcal{S}_{d-1}$, we define the **Shannon index**:

$$I_S(y) = -\sum_{i=1}^{d} y_i \log(y_i).$$

Notations

A relative abundance $y = (y_1, \dots, y_d)$ is an element of the **simplex**:

$$\mathcal{S}_{d-1} = \left\{ (x_1, \dots, x_d) \in]0; +\infty[^d \;\middle|\; \sum_{i=1}^{d} x_i = 1 \right\}.$$

For an abundance $y \in \mathcal{S}_{d-1}$, we define the **Shannon index**:

$$I_S(y) = -\sum_{i=1}^{d} y_i \log(y_i).$$

Furthermore, we denote $\overline{y} = (y_1, \dots, y_{d-1})$.

Constraints related to the relative abundance

When studying time series of relative abundance $(Y_t)_{t\in\mathbb{Z}}$ where $Y_t = (Y_{t,1},\ldots,Y_{t,d})$, difficulties arise.

Constraints related to the relative abundance

When studying time series of relative abundance $(Y_t)_{t\in\mathbb{Z}}$ where $Y_t = (Y_{t,1}, \ldots, Y_{t,d})$, difficulties arise.

- ✘ From the positivity constraint: $\forall i \in \{1, \ldots, d\}\,,\; Y_{t,i} > 0$.

Constraints related to the relative abundance

When studying time series of relative abundance $(Y_t)_{t\in\mathbb{Z}}$ where $Y_t = (Y_{t,1}, \ldots, Y_{t,d})$, difficulties arise.

- ✘ From the positivity constraint: $\forall i \in \{1, \ldots, d\}\,,\; Y_{t,i} > 0$.
- ✘ From the sum constraint: $\sum_{1\leqslant i\leqslant d} Y_{t,i} = 1$.

Constraints related to the relative abundance

When studying time series of relative abundance $(Y_t)_{t\in\mathbb{Z}}$ where $Y_t = (Y_{t,1}, \ldots, Y_{t,d})$, difficulties arise.

- ✘ From the positivity constraint: $\forall i \in \{1, \ldots, d\}\,,\ Y_{t,i} > 0$.
- ✘ From the sum constraint: $\sum_{1\leqslant i\leqslant d} Y_{t,i} = 1$.

➜ It is impossible to apply our favourite time series models...

Outlines

Traditional approach (1/2)

The idea proposed by Aitchison (1982) is simple:

Traditional approach (1/2)

The idea proposed by Aitchison (1982) is simple:

1. Transform the time series:

$$Z_t = f(Y_t)$$

where $f : \mathcal{S}_{d-1} \longrightarrow \mathbb{R}^n$ is a one to one mapping.

Traditional approach (1/2)

The idea proposed by Aitchison (1982) is simple:

1. Transform the time series:

$$Z_t = f(Y_t)$$

   where $f : \mathcal{S}_{d-1} \longrightarrow \mathbb{R}^n$ is a one to one mapping.

2. Apply the desired model on the time series $(Z_t)_{t \in \mathbb{Z}}$.

Traditional approach (1/2)

The idea proposed by Aitchison (1982) is simple:

1. Transform the time series:

$$Z_t = f(Y_t)$$

   where $f : \mathcal{S}_{d-1} \longrightarrow \mathbb{R}^n$ is a one to one mapping.
2. Apply the desired model on the time series $(Z_t)_{t\in\mathbb{Z}}$.
3. Eventually transform back the fitted values $\widehat{Z_t}$:

$$\widehat{Y_t} = f^{-1}(\widehat{Z_t}).$$

Traditional approach (1/2)

The idea proposed by Aitchison (1982) is simple:

1. Transform the time series:

$$Z_t = f(Y_t)$$

   where $f : \mathcal{S}_{d-1} \longrightarrow \mathbb{R}^n$ is a one to one mapping.
2. Apply the desired model on the time series $(Z_t)_{t \in \mathbb{Z}}$.
3. Eventually transform back the fitted values $\widehat{Z_t}$:

$$\widehat{Y}_t = f^{-1}(\widehat{Z_t}).$$

Remark

Traditional approach (1/2)

The idea proposed by Aitchison (1982) is simple:

1. Transform the time series:

$$Z_t = f(Y_t)$$

   where $f : \mathcal{S}_{d-1} \longrightarrow \mathbb{R}^n$ is a one to one mapping.
2. Apply the desired model on the time series $(Z_t)_{t\in\mathbb{Z}}$.
3. Eventually transform back the fitted values $\widehat{Z_t}$:

$$\widehat{Y_t} = f^{\;-1}(\widehat{Z_t}).$$

Remark

A very popular choice for f is the **additive log-ratio**:

$$\begin{aligned} f :\ \mathcal{S}_{d-1} &\longrightarrow \mathbb{R}^{d-1} \\ (y_1, \ldots, y_d) &\longmapsto \left(\log\left(\frac{y_1}{y_d}\right), \ldots, \log\left(\frac{y_{d-1}}{y_d}\right)\right) \end{aligned}$$

Traditional approach (2/2)

This method has shown good results in terms of fitted values or previsions.

Traditional approach (2/2)

This method has shown good results in terms of fitted values or previsions.

BUT:

The interpretation of the model's parameters *(applied to* $(Z_t)_{t\in\mathbb{Z}}$*)* is very difficult, if not impossible.

Traditional approach (2/2)

This method has shown good results in terms of fitted values or previsions.

BUT:

The interpretation of the model's parameters *(applied to* $(Z_t)_{t\in\mathbb{Z}}$*)* is very difficult, if not impossible.

OUR GOAL:

To propose a model for the time series $(Y_t)_{t\in\mathbb{Z}}$ which is easily interpretable.

Outlines

Staying in the simplex

The traditional approach lacks of interpretation due to the transformation of the initial time series.

The idea is thus to propose a model for time series which stays in the simplex:

$$\mathbb{P}(Y_{t+1} \in A \mid Y_t = y_t) = P(A/y_t)$$

where P is a transition kernel with source $\mathcal{S}_{d-1}$ and target $\mathcal{S}_{d-1}$.

Remark

Staying in the simplex

The traditional approach lacks of interpretation due to the transformation of the initial time series.

The idea is thus to propose a model for time series which stays in the simplex:

$$\mathbb{P}(Y_{t+1} \in A \ \mid Y_t = y_t) = P(A/y_t)$$

where P is a transition kernel with source $\mathcal{S}_{d-1}$ and target $\mathcal{S}_{d-1}$.

Remark

It is obviously possible to consider several lag-values *(even an infinity)*.

How do we choose P ?

We propose that $P(\cdot/y_t)$ follows a **Dirichlet distribution**.

How do we choose P ?

We propose that $P(\cdot/y_t)$ follows a **Dirichlet distribution**.

Why ?

How do we choose P ?

We propose that $P(\cdot/y_t)$ follows a **Dirichlet distribution**.

Why ?

Because this distribution allows us to approach almost any kind of distribution on the simplex *(it is the generalization of the Beta distribution)*.

Reminders about the Dirichlet distribution

The Dirichlet distribution can be characterized by:

Reminders about the Dirichlet distribution

The Dirichlet distribution can be characterized by:

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

Reminders about the Dirichlet distribution

The Dirichlet distribution can be characterized by:

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

Reminders about the Dirichlet distribution

The Dirichlet distribution can be characterized by:

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

We denote the Dirichlet distribution: $\mathrm{Dir}(\lambda, \varphi)$.

Reminders about the Dirichlet distribution

The Dirichlet distribution can be characterized by:

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

We denote the Dirichlet distribution: $\mathrm{Dir}(\lambda, \varphi)$.

Remark

Reminders about the Dirichlet distribution

The Dirichlet distribution can be characterized by:

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

We denote the Dirichlet distribution: $\mathrm{Dir}(\lambda, \varphi)$.

Remark

Actually, for $Y \sim \mathrm{Dir}(\lambda, \varphi)$, then $\mathrm{Cov}(Y_i, Y_j) = -\dfrac{\lambda_i \lambda_j}{\varphi + 1}$ if $i \neq j$ and $\mathrm{Var}(Y_i) = \dfrac{\lambda_i(1-\lambda_i)}{\varphi + 1}$.

Back to the model

We thus propose that:

$$P(\cdot/y_t) = \mathrm{Dir}\left(\lambda(\eta, y_t), \varphi(\theta, y_t)\right)$$

where θ and η are the parameters of the model.

Back to the model

We thus propose that:

$$P(\cdot/y_t) = \mathrm{Dir}\left(\lambda(\eta, y_t), \varphi(\theta, y_t)\right)$$

where θ and η are the parameters of the model.

Following the GLM framework, we also propose that

$$\eta = (A, B) \in \mathbb{R}^{(d-1)\times(d-1)} \times \mathbb{R}^{d-1} \quad \text{and} \quad \theta = (\theta_1, \theta_2) \in \mathbb{R}^2$$

with the link functions:

$$\mathrm{alr}\left(\lambda(\eta, y_t)\right) = A \cdot \overline{y_t} + B \tag{1}$$

and

$$\varphi(\theta, y_t) = \exp\left(\theta_1 + \theta_2 \cdot I_S(y_t)\right). \tag{2}$$

Existence of our process

Theorem 1

Existence of our process

Theorem 1

Assume the following assumptions hold true:

Existence of our process

Theorem 1

Assume the following assumptions hold true:

A1: θ belongs to a compact set Θ of $\mathbb{R}^2$;

Existence of our process

Theorem 1

Assume the following assumptions hold true:

A1: θ belongs to a compact set Θ of $\mathbb{R}^2$;

A2: η belongs to a compact set H of $\mathbb{R}^{(d-1)\times(d-1)} \times \mathbb{R}^{d-1}$.

Existence of our process

Theorem 1

Assume the following assumptions hold true:

A1: θ belongs to a compact set Θ of $\mathbb{R}^2$;

A2: η belongs to a compact set H of $\mathbb{R}^{(d-1)\times(d-1)} \times \mathbb{R}^{d-1}$.

Then there exists a unique time series $(Y_t)_{t\in\mathbb{Z}}$ which is strictly stationary and ergodic such that for all $t \in \mathbb{Z}$:

$$\mathcal{L}(Y_{t+1} \mid Y_t = y_t) = \mathrm{Dir}\left(\lambda(\eta, y_t), \varphi(\theta, y_t)\right)$$

with λ and φ satisfying equations (1) and (2).

Existence of our process

Theorem 1

Assume the following assumptions hold true:

A1: θ belongs to a compact set Θ of $\mathbb{R}^2$;

A2: η belongs to a compact set H of $\mathbb{R}^{(d-1)\times(d-1)} \times \mathbb{R}^{d-1}$.

Then there exists a unique time series $(Y_t)_{t\in\mathbb{Z}}$ which is strictly stationary and ergodic such that for all $t \in \mathbb{Z}$:

$$\mathcal{L}(Y_{t+1} \mid Y_t = y_t) = \mathrm{Dir}\left(\lambda(\eta, y_t), \varphi(\theta, y_t)\right)$$

with λ and φ satisfying equations (1) and (2).

Remark

Existence of our process

Theorem 1

Assume the following assumptions hold true:

A1: θ belongs to a compact set Θ of $\mathbb{R}^2$;

A2: η belongs to a compact set H of $\mathbb{R}^{(d-1)\times(d-1)} \times \mathbb{R}^{d-1}$.

Then there exists a unique time series $(Y_t)_{t\in\mathbb{Z}}$ which is strictly stationary and ergodic such that for all $t \in \mathbb{Z}$:

$$\mathcal{L}(Y_{t+1} \mid Y_t = y_t) = \mathrm{Dir}\left(\lambda(\eta, y_t), \varphi(\theta, y_t)\right)$$

with λ and φ satisfying equations (1) and (2).

Remark

Under assumptions **A1** and **A2**, the kernel P actually satisfies the **Doeblin's condition**.

Outlines

Interpretation of parameter θ

Recall that the dispersion parameter of our model satisfies:

$$\varphi(\theta, y_t) = \exp(\theta_1 + \theta_2 \cdot I_S(y_t))$$

where the Shannon index $I_S(y_t) > 0$ is a measure of biodiversity in the ecosystem.

Interpretation of parameter θ

Recall that the dispersion parameter of our model satisfies:

$$\varphi(\theta, y_t) = \exp(\theta_1 + \theta_2 \cdot I_S(y_t))$$

where the Shannon index $I_S(y_t) > 0$ is a measure of biodiversity in the ecosystem.

Thus, if $\theta_2 > 0$, the more diversity there is at time t, the less variability we will have at time $t + 1$.

Interpretation of parameter θ

Recall that the dispersion parameter of our model satisfies:

$$\varphi(\theta, y_t) = \exp(\theta_1 + \theta_2 \cdot I_S(y_t))$$

where the Shannon index $I_S(y_t) > 0$ is a measure of biodiversity in the ecosystem.

Thus, if $\theta_2 > 0$, the more diversity there is at time t, the less variability we will have at time $t + 1$.
A negative value for θ_2 is interpreted the opposite way.

Interpretation of parameter θ

Recall that the dispersion parameter of our model satisfies:

$$\varphi(\theta, y_t) = \exp(\theta_1 + \theta_2 \cdot I_S(y_t))$$

where the Shannon index $I_S(y_t) > 0$ is a measure of biodiversity in the ecosystem.

Thus, if $\theta_2 > 0$, the more diversity there is at time t, the less variability we will have at time $t + 1$.
A negative value for θ_2 is interpreted the opposite way.

θ_1 corresponds to a variability inherent to the process: the lower it is, the more volatile the time series is.

Interpretation of parameter $\eta = (A, B)$ (1/3)

It is a more tricky explanation, and the mean parameter of our model satisfies:

$$\lambda(\eta, y_t) \\ = \mathrm{alr}^{-1}(A \cdot \overline{y_t} + B)$$

Interpretation of parameter $\eta = (A, B)$ (1/3)

It is a more tricky explanation, and the mean parameter of our model satisfies:

$$\begin{aligned} &\lambda(\eta, y_t) \\ &= \mathrm{alr}^{-1}(A \cdot \overline{y_t} + B) \\ &= \left(\frac{\exp(A_{1*} \cdot \overline{y_t} + B_1)}{1 + \sum_{j=1}^{d-1} \exp(A_{j*} \cdot \overline{y_t} + B_j)}, \dots, \frac{1}{1 + \sum_{j=1}^{d-1} \exp(A_{j*} \cdot \overline{y_t} + B_j)} \right) \end{aligned}$$

where A_{j*} denotes the j^{th} line of matrix A.

Interpretation of parameter $\eta = (A, B)$ (2/3)

The coefficients of B can be interpreted as an inherent dynamic for the abundance of each species: the higher B_i is, the higher the expected value of species i will be.

Interpretation of parameter $\eta = (A, B)$ (2/3)

The coefficients of B can be interpreted as an inherent dynamic for the abundance of each species: the higher B_i is, the higher the expected value of species i will be.

In order to interpret A, one can consider the means ratio between species i and j at time $t+1$:

$$MR(i, j, y_t) = \exp\left((A_{i*} - A_{j*}) \cdot \overline{y_t}\right).$$

Interpretation of parameter $\eta = (A, B)$ (2/3)

The coefficients of B can be interpreted as an inherent dynamic for the abundance of each species: the higher B_i is, the higher the expected value of species i will be.

In order to interpret A, one can consider the means ratio between species i and j at time $t+1$:

$$MR(i, j, y_t) = \exp\left((A_{i*} - A_{j*}) \cdot \overline{y_t}\right).$$

Assume for example that the abundance is modified at time t: species of reference d increases by p, at the expense of species i and j, resulting in a new abundance:

$$z_t = y_t + \left(0, \dots, 0, -\underset{\substack{\uparrow \\ i}}{\alpha} \cdot p, 0, \dots, 0, (\alpha - \underset{\substack{\uparrow \\ j}}{1}) \cdot p, 0, \dots, 0, \underset{\substack{\uparrow \\ d}}{p}\right)$$

Interpretation of parameter $\eta = (A, B)$ (3/3)

Considering the impact the means ratio, we get:

$$\frac{MR(i,j,z_t)}{MR(i,j,y_t)} = \exp\left(p\left(\alpha(A_{ij} + A_{ji} - A_{ii} - A_{jj}) + A_{jj} - A_{ij}\right)\right).$$

Interpretation of parameter $\eta = (A, B)$ (3/3)

Considering the impact the means ratio, we get:

$$\frac{MR(i,j,z_t)}{MR(i,j,y_t)} = \exp\left(p\left(\alpha(A_{ij} + A_{ji} - A_{ii} - A_{jj}\right) + A_{jj} - A_{ij}\right)).$$

Thus, the coefficients in A give us a precise information about how the changes in an abundance would affect the means ratio between two species.

Outlines

I. Abundance and compositional data

II. Dirichlet GLM model for time series

III. Estimation
- Maximum of conditional likelihood
- Maximum of conditional pseudo-likelihood

IV. Application of the model

V. Conclusion

Settings

We consider now that we have a sample $(y_t)_{0 \leqslant t \leqslant n}$ of the abundance of an ecosystem along time.

Settings

We consider now that we have a sample $(y_t)_{0 \leqslant t \leqslant n}$ of the abundance of an ecosystem along time.

We assume that this observed abundance a realization derived from the ergodic process $(Y_t)_{t \in \mathbb{Z}}$ mentioned in theorem 1.

Settings

We consider now that we have a sample $(y_t)_{0 \leqslant t \leqslant n}$ of the abundance of an ecosystem along time.

We assume that this observed abundance a realization derived from the ergodic process $(Y_t)_{t \in \mathbb{Z}}$ mentioned in theorem 1.

Furthermore, we consider the couple of parameters (θ, η) under its vectorized form.

Outlines

I. Abundance and compositional data

II. Dirichlet GLM model for time series

III. Estimation

- Maximum of conditional likelihood
- Maximum of conditional pseudo-likelihood

IV. Application of the model

V. Conclusion

Maximum of conditional likelihood: definition

It is the most natural estimator, and it is defined by

$$\left(\hat{\theta}_n, \hat{\eta}_n\right) = \underset{(\theta,\eta)\in\Theta\times H}{\operatorname{argmin}} \; -\sum_{t=1}^{n} \log\left(p_{(\theta,\eta)}(y_t, y_{t-1})\right)$$

where $p_{(\theta,\eta)}(\cdot, y_{t-1})$ is the density of the kernel $P(\cdot/y_{t-1})$:

Maximum of conditional likelihood: definition

It is the most natural estimator, and it is defined by

$$\left(\hat{\theta}_n, \hat{\eta}_n\right) = \underset{(\theta,\eta)\in\Theta\times H}{\operatorname{argmin}} \; -\sum_{t=1}^{n} \log\left(p_{(\theta,\eta)}(y_t, y_{t-1})\right)$$

where $p_{(\theta,\eta)}(\cdot, y_{t-1})$ is the density of the kernel $P(\cdot/y_{t-1})$:

$$\begin{aligned} p_{(\theta,\eta)}(y_t, y_{t-1}) = \frac{\Gamma(\varphi(\theta, y_{t-1}))}{\prod_{i=1}^{d} \Gamma(\alpha_i(\theta, \eta, y_{t-1}))} \times \prod_{i=1}^{d-1} y_{t,i}^{\alpha_i(\theta,\eta,y_{t-1})-1} \\ \times \left(1 - \sum_{i=1}^{d-1} y_{t,i}\right)^{\alpha_d(\theta,\eta,y_{t-1})-1} \end{aligned}.$$

Maximum of conditional likelihood: properties

Proposition 1

Maximum of conditional likelihood: properties

Proposition 1

Under assumptions **A1** and **A2**, the estimator $\left(\hat{\theta}_n, \hat{\eta}_n\right)$ is **strongly consistent** and **asymptotically normal**.

Maximum of conditional likelihood: properties

Proposition 1

Under assumptions **A1** and **A2**, the estimator $\left(\hat{\theta}_n, \hat{\eta}_n\right)$ is **strongly consistent** and **asymptotically normal**.

Remark

Maximum of conditional likelihood: properties

Proposition 1

Under assumptions **A1** and **A2**, the estimator $\left(\hat{\theta}_n, \hat{\eta}_n\right)$ is **strongly consistent** and **asymptotically normal**.

Remark

This estimator is yet difficult to compute.

Maximum of conditional likelihood: properties

Proposition 1

Under assumptions **A1** and **A2**, the estimator $\left(\hat{\theta}_n, \hat{\eta}_n\right)$ is **strongly consistent** and **asymptotically normal**.

Remark

This estimator is yet difficult to compute. In practice, the optimization of the application:

$$(\theta, \eta) \longmapsto -\sum_{t=1}^{n} \log\left(p_{(\theta,\eta)}(y_t, y_{t-1})\right)$$

strongly depends on the initial value chosen for (θ, η), and can numerically fail.

Outlines

I. Abundance and compositional data

II. Dirichlet GLM model for time series

III. Estimation

- Maximum of conditional likelihood
- Maximum of conditional pseudo-likelihood

IV. Application of the model

V. Conclusion

Maximum of conditional pseudo-likelihood: definition

We focus here on the parameter $\eta = (A, B)$.

Maximum of conditional pseudo-likelihood: definition

We focus here on the parameter $\eta = (A, B)$.

This estimator is defined by:

$$\hat{w}_n = \underset{\eta \in H}{\operatorname{argmin}} \; -\sum_{t=1}^{n} \sum_{i=1}^{d} y_{t,i} \log\left(\lambda_i(\eta, y_{t-1})\right).$$

Maximum of conditional pseudo-likelihood: definition

We focus here on the parameter $\eta = (A, B)$.

This estimator is defined by:

$$\hat{w}_n = \underset{\eta \in H}{\operatorname{argmin}} \; - \sum_{t=1}^{n} \sum_{i=1}^{d} y_{t,i} \log\left(\lambda_i(\eta, y_{t-1})\right).$$

Proposition 2

Maximum of conditional pseudo-likelihood: definition

We focus here on the parameter $\eta = (A, B)$.

This estimator is defined by:

$$\hat{w}_n = \underset{\eta \in H}{\operatorname{argmin}} \; -\sum_{t=1}^{n} \sum_{i=1}^{d} y_{t,i} \log\left(\lambda_i(\eta, y_{t-1})\right).$$

Proposition 2

The application

$$\eta \longmapsto -\sum_{t=1}^{n} \sum_{i=1}^{d} y_{t,i} \log\left(\lambda_i(\eta, y_{t-1})\right)$$

is convex.

Maximum of conditional pseudo-likelihood: properties

Proposition 3

Maximum of conditional pseudo-likelihood: properties

Proposition 3

Under assumptions **A1** and **A2**, the estimator $\hat{w}_n$ is **strongly consistent** and **asymptotically normal**.

Remark

Maximum of conditional pseudo-likelihood: properties

Proposition 3

Under assumptions **A1** and **A2**, the estimator $\hat{w}_n$ is **strongly consistent** and **asymptotically normal**.

Remark

This time, $\hat{w}_n$ is numerically easy to compute, due to convexity.

Maximum of conditional pseudo-likelihood: properties

Proposition 3

Under assumptions **A1** and **A2**, the estimator $\hat{w}_n$ is **strongly consistent** and **asymptotically normal**.

Remark

This time, $\hat{w}_n$ is numerically easy to compute, due to convexity.

It can be a good strategy to initialize the value of η with $\hat{w}_n$ when looking for the maximum of conditional likelihood.

Outlines

Simulation settings

We simulate a thousand time series of abundance $(y_t)_{1\leqslant t\leqslant 1000}$ with $d = 3$ species, according to our model, with the following parameters:

Simulation settings

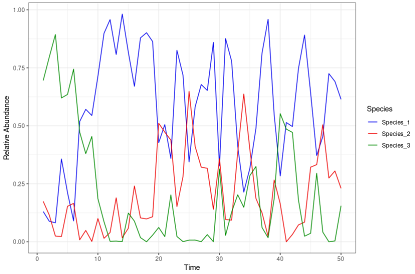
We simulate a thousand time series of abundance $(y_t)_{1 \leqslant t \leqslant 1000}$ with $d = 3$ species, according to our model, with the following parameters:

$$A = \begin{pmatrix} 4.5 & 3 \\ 5 & 6.5 \end{pmatrix}, \quad B = \begin{pmatrix} -2 \\ -4 \end{pmatrix} \quad \text{and} \quad \theta = \begin{pmatrix} 2 \\ 0.5 \end{pmatrix}.$$

Simulation settings

We simulate a thousand time series of abundance $(y_t)_{1\leqslant t\leqslant 1000}$ with $d = 3$ species, according to our model, with the following parameters:

$$A = \begin{pmatrix} 4.5 & 3 \\ 5 & 6.5 \end{pmatrix}, \quad B = \begin{pmatrix} -2 \\ -4 \end{pmatrix} \quad \text{and} \quad \theta = \begin{pmatrix} 2 \\ 0.5 \end{pmatrix}.$$



Simulation of an abundance of three species

Estimated results

With the maximum of conditional pseudo-likelihood, the mean of the estimates obtained is given by:

$$\hat{A}_1 = \begin{pmatrix} 4.48 & 2.99 \\ 5.00 & 6.49 \end{pmatrix} \quad \text{and} \quad \hat{B}_1 = \begin{pmatrix} -1.98 \\ -3.99 \end{pmatrix}.$$

Estimated results

With the maximum of conditional pseudo-likelihood, the mean of the estimates obtained is given by:

$$\hat{A}_1 = \begin{pmatrix} 4.48 & 2.99 \\ 5.00 & 6.49 \end{pmatrix} \quad \text{and} \quad \hat{B}_1 = \begin{pmatrix} -1.98 \\ -3.99 \end{pmatrix}.$$

With the maximum of conditional likelihood, we have:

$$\hat{A}_2 = \begin{pmatrix} 4.53 & 3.16 \\ 5.10 & 6.63 \end{pmatrix}, \quad \hat{B}_2 = \begin{pmatrix} -2.03 \\ -4.05 \end{pmatrix} \quad \text{and} \quad \hat{\theta} = \begin{pmatrix} 1.53 \\ 0.02 \end{pmatrix}.$$

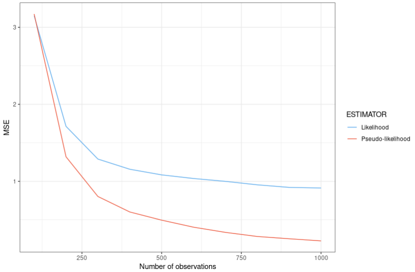
Comparison with the true values

How many observations are necessary to obtain a "good" estimation ?

Comparison with the true values

How many observations are necessary to obtain a "good" estimation ?

One can find below the MSE of our estimates, depending on the number of observations used.



MSE of both estimators

Outlines

The dataset

We consider here a population of alpine birds during 38 years, from 1964 to 2001 *(see (Svensson, 2006) for details)*.

The dataset

We consider here a population of alpine birds during 38 years, from 1964 to 2001 *(see (Svensson, 2006) for details)*.

For simplification purpose, we will focus on three particular species: Anthus pratensis, Calcarius lapponicus and Oenanthe oenanthe.

(a) Anthus pratensis

(b) Calcarius lapponicus

(c) Oenanthe oenanthe

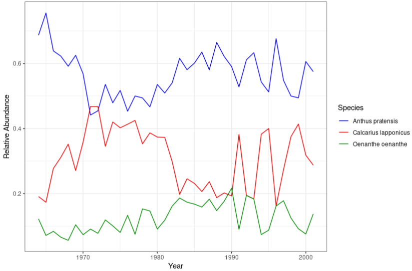
Scandinavian Birds

The dataset

We consider here a population of alpine birds during 38 years, from 1964 to 2001 *(see (Svensson, 2006) for details)*.

For simplification purpose, we will focus on three particular species: Anthus pratensis, Calcarius lapponicus and Oenanthe oenanthe.

One can find below the graphics of relative abundance for these species.



Abundance of Scandinavian Birds

Estimation results and previsions (1/2)

We assume our time series of birds satisfies our model, and we use the 30 first observations to estimate the parameters with the maximum of conditional likelihood:

Estimation results and previsions (1/2)

We assume our time series of birds satisfies our model, and we use the 30 first observations to estimate the parameters with the maximum of conditional likelihood:

$$\hat{A} = \begin{pmatrix} 4.68 & 3.43 \\ 3.83 & 5.13 \end{pmatrix}, \quad \hat{B} = \begin{pmatrix} -2.21 \\ -2.87 \end{pmatrix} \quad \text{and} \quad \hat{\theta} = \begin{pmatrix} 1.34 \\ 0.65 \end{pmatrix}.$$

Estimation results and previsions (2/2)

We can try to make previsions on the last eight years with the obtained estimates.

Estimation results and previsions (2/2)

We can try to make previsions on the last eight years with the obtained estimates.

The previsions $\hat{y}_t$ are made "step by step" using the estimates obtained previously, and the conditional mean:

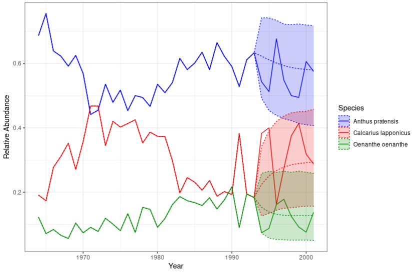
$$\widehat{y_{t+1}} = \lambda(\hat{\eta}, \hat{y}_t).$$

Estimation results and previsions (2/2)

We can try to make previsions on the last eight years with the obtained estimates.

The previsions $\hat{y}_t$ are made “step by step” using the estimates obtained previously, and the conditional mean:

$$\widehat{y_{t+1}} = \lambda(\hat{\eta}, \hat{y}_t).$$



Prevision of the abundance for Scandinavian Birds

Interpretation of the results (1/2)

➜ Coefficients in $\widehat{B}$ are both negative.

Interpretation of the results (1/2)

➜ Coefficients in $\widehat{B}$ are both negative.
The abundance of the two first species are intrinsically more likely to decline than the last species.

Interpretation of the results (1/2)

➜ Coefficients in $\widehat{B}$ are both negative.
The abundance of the two first species are intrinsically more likely to decline than the last species.

➜ Second coefficient in $\widehat{\theta}$ is positive.

Interpretation of the results (1/2)

➜ Coefficients in $\widehat{B}$ are both negative.
The abundance of the two first species are intrinsically more likely to decline than the last species.

➜ Second coefficient in $\widehat{\theta}$ is positive.
The more diversity there is, the more volatile the ecosystem will be.

Interpretation of the results (1/2)

➜ Coefficients in $\widehat{B}$ are both negative.
The abundance of the two first species are intrinsically more likely to decline than the last species.

➜ Second coefficient in $\widehat{\theta}$ is positive.
The more diversity there is, the more volatile the ecosystem will be.

➜ For the interpretation $\widehat{A}$, we can compute some different impacts on means ratios.

Interpretation of the results (1/2)

➜ Coefficients in $\widehat{B}$ are both negative.
The abundance of the two first species are intrinsically more likely to decline than the last species.

➜ Second coefficient in $\widehat{\theta}$ is positive.
The more diversity there is, the more volatile the ecosystem will be.

➜ For the interpretation $\widehat{A}$, we can compute some different impacts on means ratios.

For example, if species 3 increases its abundance by p, at the expense of species 1 and 2:

$$z_t = y_t + (-\alpha \cdot p, (\alpha - 1) \cdot p, p).$$

Interpretation of the results (2/2)

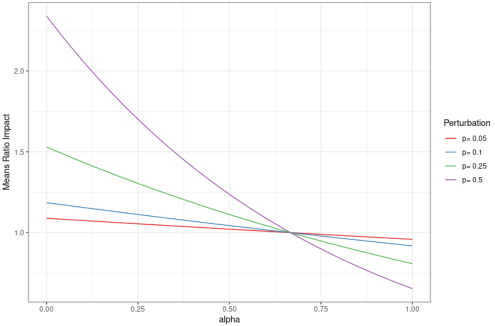
We compute:

$$\frac{MR(1,2,z_t)}{MR(1,2,y_t)} = \exp\left(p\left(\alpha\left(\hat{A}_{12} + \hat{A}_{21} - \hat{A}_{11} - \hat{A}_{22}\right)\right) + \hat{A}_{22} - \hat{A}_{12}\right)$$
$$= \exp(p(-2.55\alpha + 1.7))$$

Interpretation of the results (2/2)

We compute:

$$\frac{MR(1,2,z_t)}{MR(1,2,y_t)} = \exp\left(p\left(\alpha\left(\hat{A}_{12}+\hat{A}_{21}-\hat{A}_{11}-\hat{A}_{22}\right)\right)+\hat{A}_{22}-\hat{A}_{12}\right)$$
$$= \exp(p(-2.55\alpha+1.7))$$



Impacts on the means ratio

Outlines

Conclusion

Conclusion

➜ Construction of an ergodic time series

Conclusion

- Construction of an ergodic time series
- Interpretation of the model's parameters

Conclusion

- Construction of an ergodic time series
- Interpretation of the model's parameters
- Theoretical estimation results

Conclusion

- ➜ Construction of an ergodic time series
- ➜ Interpretation of the model's parameters
- ➜ Theoretical estimation results
- ✗ In ecology, time series do not possess many observations

Conclusion

- ➜ Construction of an ergodic time series
- ➜ Interpretation of the model's parameters
- ➜ Theoretical estimation results
- ✗ In ecology, time series do not possess many observations
    - Use of panel data ?

Conclusion

- ➜ Construction of an ergodic time series
- ➜ Interpretation of the model's parameters
- ➜ Theoretical estimation results
- ✗ In ecology, time series do not possess many observations
  - Use of panel data ?
- ✗ Data sets often possess "zero values"

Conclusion

- ➜ Construction of an ergodic time series
- ➜ Interpretation of the model's parameters
- ➜ Theoretical estimation results
- ✗ In ecology, time series do not possess many observations
  - Use of panel data ?
- ✗ Data sets often possess "zero values"
  - Infer missing values ?

Conclusion

- ➜ Construction of an ergodic time series
- ➜ Interpretation of the model's parameters
- ➜ Theoretical estimation results
- ✗ In ecology, time series do not possess many observations
  - Use of panel data ?
- ✗ Data sets often possess "zero values"
  - Infer missing values ?
  - Construction of a model on the simplex, but with a dimension which can vary ?

Thank you !

References

Aitchison, J. (1982). The statistical analysis of compositional data. *Journal of the Royal Statistical Society: Series B (Methodological)*, *44*(2), 139–160.

Svensson, S. (2006). Species composition and population fluctuations of alpine bird communities during 38 years in the scandinavian mountain range. *Ornis Svecica*, *16*(4), 183–210.