Dynamical modeling of abundance data in ecology

written by Guillaume FRANCHI , Supervisors : Lionel TRUQUET – ENSAI and Paul DOUKHAN – Université Cergy-Pontoise

Introduction

The **relative abundance** of a species in an ecosystem is the proportion of individuals of the species among all individuals from all species.
Abundance data of $d \geqslant 2$ given species are therefore **compositional data**, i.e. elements of the simplex :

$$\mathcal{S}_{d-1} = \left\{ (y_1, \dots, y_d) \in [0;1]^d \;\middle|\; \sum_{1 \leqslant i \leqslant d} y_i = 1 \right\}.$$

Even if the study of abundance data at a fixed time has been very developped in the litterature, it is less common to find models considering the abundance as a time series.

Model

We denote $(Y_t)_{t\in\mathbb{Z}}$ the time series of abundance : $Y_t = \left(Y_{t,1}, \dots Y_{t,d}\right)$, where $Y_{t,i}$ is the relative abundance of the i^{th} species at time t. We also denote $(X_t)_{t\in\mathbb{Z}}$ the time series of exogenous variables that we might consider.
We want our model to take into account three phenomena :

- ▷ the dynamic of the abundance ;
- ▷ the influence of species between them ;
- ▷ the impact of exogenous variables.

In order to model compositional data, Douma and Weedon (2019) recommand to building a model on the original data rather than applying some transformations on them. Thus, we assume that $(Y_t)_{t\in\mathbb{Z}}$ is some kind of **Markov process** in the sense that :

$$\mathbb{P}(Y_t \in A \mid Y_{t-1} = y_{t-1}, X = x) = K_\theta^{x_{t-1}}(y_{t-1}, A)$$

where for a parameter $\theta = (\phi, \eta, \beta, \gamma)$ and $\overline{y} = (y_1, ..., y_{d-1})$, the markovian kernel $K_\theta^{x_{t-1}}(y_{t-1}, \cdot)$ follows a **Dirichlet distribution** $\mu^{(d)}_{\alpha(\theta, y_{t-1}, x_{t-1})}$ defined by

$$\forall i \in \{1, \dots, d-1\},\ \alpha_i(\theta, y, x_0) = \phi \cdot \frac{\exp(\eta_i' \cdot \overline{y} + \beta_i' \cdot x_0 + \gamma_i)}{1 + \sum_{j=1}^{d-1} \exp(\eta_j' \cdot \overline{y} + \beta_j' \cdot x_0 + \gamma_j)}$$

$$\text{and} \qquad \alpha_d(\theta, y, x_0) = \frac{\phi}{1 + \sum_{j=1}^{d-1} \exp(\eta_j' \cdot \overline{y} + \beta_j' \cdot x_0 + \gamma_j)}.$$

The conditional means $\lambda_1, ..., \lambda_d$ of this distribution are given for $\psi = (\eta, \beta, \gamma)$ by

$$\forall i \in \{1, \dots, d-1\},\ \lambda_i(\psi, y, x_0) = \frac{\exp(\eta_i' \cdot \overline{y} + \beta_i' \cdot x_0 + \gamma_i)}{1 + \sum_{j=1}^{d-1} \exp(\eta_j' \cdot \overline{y} + \beta_j' \cdot x_0 + \gamma_j)}$$

$$\text{and} \qquad \lambda_d(\psi, y, x_0) = \frac{1}{1 + \sum_{j=1}^{d-1} \exp(\eta_j' \cdot \overline{y} + \beta_j' \cdot x_0 + \gamma_j)}.$$

Note that for identifiability reasons, species d is here a species of reference, so we do not study its influence on the other species. Let us recall that a Dirichlet distribution $\mu_\alpha^{(d)}$ is defined by

$$\int f \, \mathrm{d}\mu_\alpha^{(d)} = \int f\left(x_1, ..., x_{d-1}, 1 - \sum_{i=1}^{d-1} x_i\right) \frac{\Gamma\left(\sum_{i=1}^d \alpha_i\right)}{\Gamma(\alpha_1) \dots \Gamma(\alpha_d)} \prod_{i=1}^{d-1} x_i^{\alpha_i - 1} \left(1 - \sum_{i=1}^{d-1} x_i\right)^{\alpha_d - 1} \mathrm{d}x_i.$$

Estimators

We propose here two methods to estimate the values of parameter θ. Assume we get a sample $(X_t, Y_t)_{0 \leqslant t \leqslant n}$ of both abundance and exogenous variables.
The first one is based on the **conditional likelihood** of our model. If $p_\theta(y_t, y_{t-1}, x_{t-1})$ is the density of kernel $K_\theta^{x_{t-1}}(y_{t-1}, \cdot)$, then we estimate θ by

$$\widehat{\theta}_n = \underset{\theta}{\operatorname{argmin}} \left(-\sum_{t=1}^{n} \log(p_\theta(y_t, y_{t1}, x_{t-1}) \right).$$

Estimator $\widehat{\theta}_n$ is both strongly consistent and asymptotically normal.
The second estimator is based on a **convex optimization problem**. We do not consider parameter ϕ anymore and we focus on $\psi = (\eta, \beta, \gamma)$. Actually, ϕ is not necessary for prediction. We estimate ψ by

$$\widehat{\psi}_n = \underset{\psi}{\operatorname{argmin}} \sum_{t=1}^{n} \left(-\sum_{i=1}^{d} y_{t,i} \log(\lambda_i(\psi, y_{t-1}, x_{t-1})) \right).$$

Once again, estimator $\widehat{\psi}_n$ is strongly consistent and asymptotically normal.

Results

In order to apply the previous results, we propose to consider a dataset of abundance of three bird species (Anthus pratensis, Calcarius lapponicus and Oenanthe oenanthe) in Scandinavia, from 1964 to 2001 (See Svensson (2006) for details). No exogenous variable is considered here, and one can find in Figure 1 a representation of these abundances. We choose Oenanthe oenanthe as species of reference.
The estimators introduced previously give us the results presented in Table 1. Even if the number of observations is quite low, we obtain quite similar results with both methods.

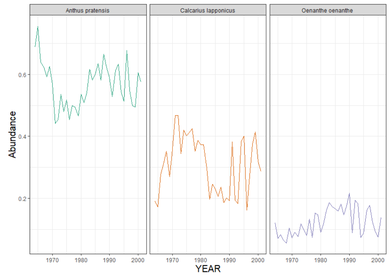

Figure 1: Abundance of scandinavian birds

Parameter	Estimates with $\widehat{\theta}_n$	Estimates with $\widehat{\psi}_n$
ϕ	49.55	NA
$\eta_{1,1}$	3.87	4.08
$\eta_{1,2}$	2.78	2.84
$\eta_{2,1}$	2.94	3.21
$\eta_{2,2}$	4.52	4.71
γ_1	-1.60	-1.79
γ_2	-2.09	-2.15

Table 1: Estimation results with both methods

Note that both species Anthus and Calcarius are benefic to each other, since the entries of η are all positive. We also observe $\eta_{2,2}$ is slightly higher than $\eta_{1,1}$, which would indicate a slightly higher birth rate for the species Calcarius. It is indeed the case, as species Calcarius can have one more egg per nest.
We also remark that both species have a negative coefficient γ_i : independently to the abundance or exogenous variables at time $t-1$, their populations have a natural inclination to decrease, to the profit of the species Oenanthe. Actually, species Calcarius has a lower life expectancy than species Oenanthe. On the contrary, species Anthus has a higher life expectancy, but for some other reasons, its species are more endangered than the other ones. We can guess that an unindentified factor lowers its natural life expectancy, thus modifying its abundance trend.

Model assumptions

The main necessary assumption is the **stationarity** of process $(Y_t)_{t\in\mathbb{Z}}$ and the **ergodicity** of process $(X_t)_{t\in\mathbb{Z}}$: it ensures the ergodicity of $(X_t, Y_t)_{t\in\mathbb{Z}}$, that is necessary to obtain good properties for estimation.
Another quite reasonable assumption on the process $(X_t)_{t\in\mathbb{Z}}$ is that it takes its **values in compact set** and its variables are not **linearly correlated almost surely** :

$$\forall Q \neq 0,\ \forall R,\ \mathbb{P}(Q \cdot X_0 + R = 0) < 1.$$

If it was not the case, we could actually replace one variable by a linear combination of the others, so this variable would not be necessary in our model.
Finally, it is assumed that parameter θ can only take its **values in a compact set**.

Model interpretation

The dynamic of our model supposes that abundance at time t depends on abundance at time $t-1$ and some covariables at time $t-1$, but how exactly ?
It actually depends on the parameters in θ. The matrix η contains information on the influence of species between them. A positive entry $\eta_{i,j}$ means that species j has a positive influence on the abundance of species i. A negative entry indicates a negative influence.
The same interpretation holds for the coefficients of matrix β : a positive entry $\beta_{i,k}$ indicates that a high value of the exogenous variable k stimulates the abundance of species i.
Parameter ϕ controls the variability of the abundances : the higher it is, the more stable abundances will be.
Finally, the entries γ_i contain information about the trend of species i, independently from the abundance or exogenous variables at a given time t: the higher γ_i is, the more likely species i will have a high abundance at step $t+1$. We thus expect γ_i to gather informations about the natural life expectancy of species i.

Conclusion

The model presented here suffers from two major inconvenients.
The first one is that we only obtain asymptotical results for our estimators, and it is obviously difficult to conduct ecological surveys during hundreds of years...
The second one is that our model does not allow zero values for abundance, although in real life, it is possible not to observe a species at a given time t, but to observe it later.
Thus, the model we presented is only a beginning, and needs more sophistication. Nevertheless, it presents a strong advantage in terms of interpretability. Indeed, all the parameters characterizing the model are easily identifiable and can give good insights on the dynamic of abundance.

References

Douma, J. C., & Weedon, J. T. (2019). Analysing continuous proportions in ecology and evolution: A practical introduction to beta and dirichlet regression. *Methods in Ecology and Evolution*, *10*(9), 1412–1430.
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.