Le modèle Poisson log-normal pour l'analyse de distributions jointes d'abondance

Julien Chiquet, Marie-Josée Cros, Mahendra Mariadassou, Nathalie Peyrard, Stéphane Robin

8

Le modèle Poisson log-normal pour l'analyse de distributions jointes d'abondance

Julien CHIQUET[1], Marie-Josée CROS[2], Mahendra MARIADASSOU[3], Nathalie PEYRARD[2] et Stéphane ROBIN[1]

[1] *MIA Paris, INRAE, AgroParisTech, Université Paris-Saclay, Paris, France*

[2] *MIAT, INRA, Université de Toulouse, Castanet-Tolosan, France*

[3] *MaIAGE, INRAE, Université Paris-Saclay, Jouy-en-Josas, France*

8.1. Introduction

Le fonctionnement d'un écosystème dépend essentiellement des interactions des espèces qui le composent avec leur environnement (interactions abiotiques) et des interactions que ces espèces entretiennent entre elles (interactions biotiques). Savoir caractériser la diversité d'une telle communauté permet de suivre son évolution dans le temps, ou de comprendre comment des patterns d'une communauté peuvent varier d'un environnement à un autre. Dans une optique de conservation (ou de contrôle) cela permet également d'évaluer les effets de mesures de protection, voire de les cibler (Tylianakis *et al*. 2010 ; Xiao *et al*. 2018).

La diversité d'un écosystème s'étudie aussi bien en microbiologie comme pour le microbiome intestinal (Layeghifard *et al*. 2017), qu'à une échelle plus macroscopique comme les communautés d'arbres dans les forêts (Clark *et al*. 2014). Dans les deux

Approches statistiques pour les variables cachées en écologie,
coordonné par Nathalie PEYRARD et Olivier GIMENEZ.

cas, l'information de base est un tableau de présence-absence ou de comptages des individus de chaque espèce dans différents échantillons, caractérisés par des propriétés environnementales différentes. En microbiologie, les données de métabarcoding vont fournir un comptage de « reads » (ou séquences), alors qu'en écologie forestière ou marine par exemple, on compte directement les individus.

Pour obtenir une description fine de ces écosystèmes, il ne suffit pas de regarder la richesse spécifique de la communauté, ni même son indice de Shannon. Ces résumés des comptages perdent l'information sur la manière dont les espèces répondent ensemble à l'environnement et sur les associations (positives ou négatives) entre les espèces.

Pour répondre à ces limites, des modèles statistiques ont été proposés pour l'étude des associations d'espèces et de leur réaction conjointe à l'environnement. Ils modélisent conjointement les présences-absences (Harris 2015 ; Ovaskainen *et al*. 2017) ou les abondances (Popovic *et al*. 2018) de l'ensemble de ces espèces. Il s'agit des modèles de distributions jointes d'espèces (notés JSDM pour *Joint Species Distribution Models*), par opposition notamment aux modèles d'abondance d'espèce (SDM (Elith et Leathwick 2009)) qui visent à étudier l'influence de l'environnement sur l'abondance d'une seule espèce. L'application des JSDM au cas des données de présence-absence est abordé dans le chapitre 7 de cet ouvrage. Ici, nous abordons le cas des données de comptage.

D'un point de vue statistique, la modélisation conjointe de données d'abondance présente plusieurs difficultés liées à la nature même de ces données et notamment le fait (i) qu'il s'agisse de comptages et (ii) que leur dispersion observée est souvent beaucoup plus grande que celle attendue sous la distribution de référence pour les comptages, à savoir la loi de Poisson. Concernant le point (i), contrairement au cas de données continues, pour lequel la loi normale multivariée est la référence, il n'existe pas de loi multivariée naturelle pour les données de comptage. Il existe bien sûr des modèles multivariés de comptage mais ceux-ci imposent souvent des contraintes fortes, notamment de signe, sur les dépendances existant entre les espèces (Inouye *et al*. 2017).

Les modélisateurs intéressés par les JSDM se sont ainsi naturellement orientés vers des modèles à variables latentes (Warton *et al*. 2015) qui offrent une plus grande flexibilité en termes de modélisation de la dépendance. Ils permettent notamment de se ramener à des données continues au niveau des variables latentes. Plusieurs de ces modèles reposent ainsi sur une modélisation gaussienne de la couche latente (Ovaskainen *et al*. 2017 ; Popovic *et al*. 2018) : la dépendance entre les espèces y est décrite par la matrice de covariance du vecteur latent associé à chaque échantillon. Le modèle Poisson log-normal (PLN (Aitchison et Ho 1989)) sur lequel porte ce chapitre entre exactement dans ce cadre. L'avantage des modèles à variables latentes est

qu'ils induisent par construction une surdispersion, du simple fait qu'ils font intervenir un aléa supplémentaire dans la distribution des observations. Enfin, ces modèles permettent de prendre assez facilement en compte l'effet de l'environnement, donc des interactions abiotiques, sur l'abondance des différentes espèces, *via* une régression.

Nous illustrons l'intérêt du modèle PLN sur des données de comptage d'espèces marines issues du programme de recherche PISCO (PISCO Research Consortium 2019b). Les écosystèmes côtiers marins sont actuellement soumis à de fortes perturbations (surpêche, destruction d'habitats, pollution) ainsi qu'au changement climatique. Cela a des conséquences (températures extrêmes, acidification, invasion d'espèces) tant écologiques que socio-économiques (Pan *et al*. 2013). Mieux connaître ces écosystèmes est un enjeu pour mieux les protéger et gérer les activités humaines impactantes. Depuis 1999, c'est l'objectif du programme de recherche PISCO (PISCO Research Consortium 2019b) qui étudie les communautés marines le long de la côte ouest de l'Amérique du Nord. L'objectif est d'accroître la compréhension des causes et conséquences des changements de l'écosystème. Pour ce faire, un programme à long terme d'échantillonnage des espèces est mené dont un des buts est de mieux connaître la distribution des espèces et leurs interactions.

Un des écosystèmes étudiés est celui des forêts de laminaires géantes (*kelp forest*, (PISCO Research Consortium 2019a)) qui peut être observé autour des îles Channel Islands, situées au large de Santa Barbara (Basse-Californie). La mise en place de zones protégées autour des îles en fait des refuges pour la vie marine sauvage (Caselle 2013). Pour cette étude, nous nous focalisons sur l'île Anacapa (composée de 3 îlots, que nous appellerons sites) dont les abords sont assez bien protégés et dont les côtes marines ont été observées dès 1999. Sur la base d'un jeu de données, que nous appellerons MariNet, extrait des données PISCO, nous allons illustrer comment utiliser le modèle PLN pour 3 types de questions :

– évaluer l'influence de covariables comme le site ou l'année, sur l'abondance des espèces des covariables ;

– identifier des espèces réagissant de la même manière à ces covariables ;

– identifier des interactions directes entre espèces.

8.2. Le modèle Poisson log-normal

8.2.1. *Le modèle*

Nous présentons tout d'abord le modèle Poisson log-normal général et précisons comment il prend en compte à la fois les effets abiotiques et les interactions biotiques, de même que l'effort d'échantillonnage.

Le modèle Poisson log-normal multivarié (dans la suite PLN) introduit par (Aitchison et Ho 1989) est un modèle à variable latente permettant de décrire un vecteur de comptage S-dimensionnel sur-dispersé.

On considère un échantillon de taille N issu de tirages indépendants d'un tel vecteur. Dans le cas de l'étude des données MariNet, S représente le nombre d'espèces et un échantillon est défini par 4 éléments dont le site et l'année (section 8.3). Le modèle PLN relie chaque observation $\mathbf{Y}_n = (Y_{n1}, \ldots, Y_{nS}) \in \mathbb{N}^S$ $(1 \leq n \leq N)$ du vecteur de comptage à un vecteur gaussien latent $\mathbf{Z}_n \in \mathbb{R}^S$, de telle sorte que les coordonnées de $\mathbf{Y}_n$ soient tirées indépendamment conditionnellement à $\mathbf{Z}_n$ selon une distribution de Poisson :

$$\begin{array}{l} \text{espace latent } \mathbf{Z}_n \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \\ \text{espace des observations } Y_{ns} \mid Z_{ns} \quad \text{indep.} \mathbf{Y}_n \mid \mathbf{Z}_n \sim \mathcal{P}\left(\exp\{\mathbf{Z}_n\}\right) \end{array} \qquad [8.1]$$

Le vecteur de moyennes $\boldsymbol{\mu} \in \mathbb{R}^S$ correspond aux effets principaux, tandis que la matrice de variance-covariance $\boldsymbol{\Sigma}$ décrit la structure de dépendance entre les S coordonnées du vecteur $\mathbf{Z}_n$.

Dans le cadre des données d'abondance, il s'agit de la structure de dépendance entre les espèces au sein des échantillons collectés. La structure de dépendance du modèle PLN est représentée graphiquement sur la figure 8.1. La figure 8.2 propose une vue géométrique du modèle PLN lorsque deux espèces sont en jeu.

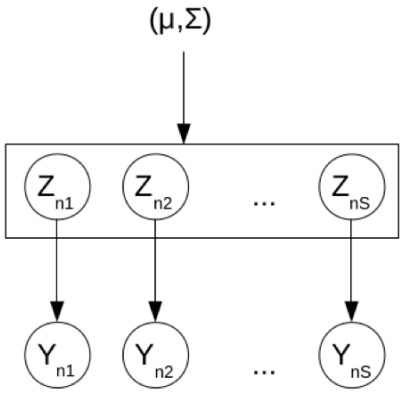


Figure 8.1. *Représentation graphique des dépendances dans le modèle PLN. Les variables aléatoires sont encerclées, les paramètres non*

La distribution PLN est naturellement sur-dispersée par rapport à la distribution de Poisson, comme attendu dans ce type de contexte applicatif. En effet, si $\mathbf{\Sigma} = [\sigma_{sr}]_{1 \leq j,k \leq S}$, alors $\mathbb{E}(Y_{ns}) = e^{\mu_s + \sigma_{ss}/2}$ et $\mathbb{V}(Y_{ns}) = \mathbb{E}(Y_{ns}) + (e^{\sigma_{ss}} - 1)\mathbb{E}(Y_{ns})^2 \geq \mathbb{E}(Y_{ns})$.

De plus, la covariance entre les comptages observés de 2 espèces peut prendre des signes arbitraires : $\mathbb{C}\text{ov}(Y_{ns}, Y_{nr}) = (e^{\sigma_{sr}} - 1)\mathbb{E}(Y_{ns})\mathbb{E}(Y_{nr})$, et ainsi $\mathbb{C}\text{ov}(Y_{ns}, Y_{nr})$ a le même signe que $\mathbb{C}\text{ov}(Z_{ns}, Z_{nr}) = \sigma_{sr}$.

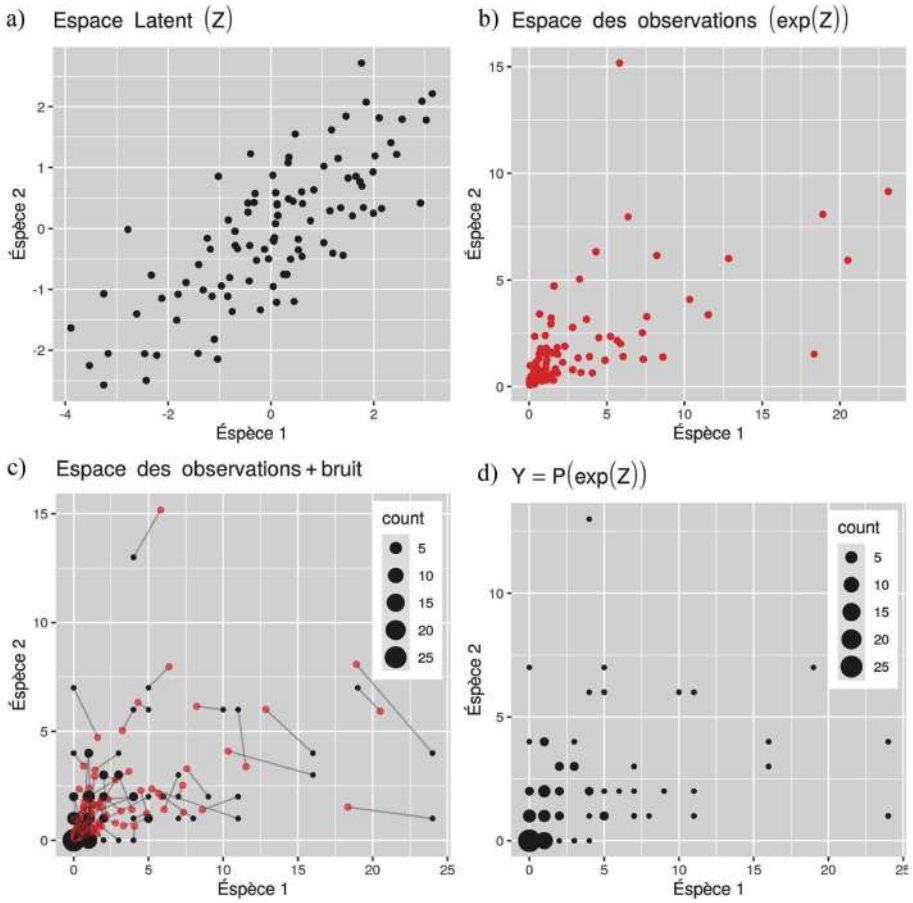


Figure 8.2. *PLN : vue géométrique du modèle pour 2 espèces. a) Positions dans l'espace latent : log-abondances moyennes. b) Comptages moyens. c) Comptages moyens (rouge) et comptages observés (noir). d) Comptages observés*

8.2.1.1. *Covariables et offsets*

On généralise naturellement le modèle [8.1] à une formulation proche du modèle linéaire général (c'est-à-dire à multiples réponses), où les effets principaux correspondent à une combinaison linéaire de D covariables fixes que l'on note $\mathbf{x}_n$[1]. Dans notre contexte, il est également naturel d'intégrer une matrice d'offsets au modèle, c'est-à-dire un décalage fixe dans la régression (connu du modélisateur), qui dépendent de l'échantillon, et éventuellement de l'espèce. Ceci permet en particulier de modéliser la notion d'intensité d'observation, comme nous le verrons plus tard. On note $\mathbf{o}_n \in \mathbb{R}^S$ le vecteur d'offsets de l'échantillon n. Ainsi, le modèle [8.1] se généralise naturellement à :

$$\mathbf{Y}_n \mid \mathbf{Z}_n \sim \mathcal{P}\left(\exp\{\mathbf{Z}_n\}\right), \qquad \mathbf{Z}_n \sim \mathcal{N}(\mathbf{o}_n + t(\mathbf{x}_n)\mathbf{B}, \mathbf{\Sigma}) \tag{8.2}$$

où $\mathbf{B}$ est la matrice $D \times S$ des coefficients de régression.

8.2.1.2. *Modélisation de la matrice de variance-covariance*

La paramétrisation utilisée pour décrire la matrice de variance-covariance peut être précisée selon les besoins, afin notamment de réduire le nombre total de paramètres du modèle. Dans l'hypothèse la plus générale, la matrice $\mathbf{\Sigma}$ possède $S(S+1)/2$ paramètres (S paramètres de variances et $S(S-1)/2$ termes de covariance). Cependant, le modélisateur peut choisir de décrire uniquement les variances des espèces, à l'aide d'une matrice $\mathbf{\Sigma}$ diagonale avec seulement S paramètres. Le modèle est alors équivalent à S SDM indépendants. Dans une situation extrême[2], on peut être amené à utiliser un seul paramètre de variance pour la matrice tout entière, de sorte que $\mathbf{\Sigma} = \sigma \mathbf{I}_p$. Dans les sections 8.2.3 et 8.2.4, nous présenterons d'autres types de modélisation de la matrice $\mathbf{\Sigma}$, adaptés à la réduction de dimension et à l'inférence de réseau.

8.2.1.3. *Notation additionnelle*

Par la suite, la totalité des données disponibles pour les N échantillons sera représentée sous la forme de trois matrices, où la n^{e} ligne est associée au n^{e} échantillon, avec $\mathbf{Y}$ la matrice $N \times S$ des comptages, $\mathbf{X}$ la matrice $N \times D$ des covariables et $\mathbf{O}$ la matrice $N \times S$ des offsets.

8.2.2. *Méthode d'inférence*

Nous décrivons ici brièvement comment les paramètres du modèle PLN peuvent être estimés et en quoi cette estimation pose des difficultés typiques des modèles à variables latentes.

1. Dans la suite, le vecteur des covariables $\mathbf{x}_n$ inclut également la constante.
2. Par exemple, lorsqu'une série de modèles PLN sont issus des composantes d'un mélange.

La question de l'inférence concerne l'estimation des paramètres de régression $\mathbf{B}$ et de la matrice de variance-covariance $\mathbf{\Sigma}$. On notera $\boldsymbol{\theta} = \{\mathbf{B}, \mathbf{\Sigma}\}$ l'ensemble des paramètres du modèle.

8.2.2.1. *Inférence dans les modèles à variables latentes*

Le modèle PLN est un modèle à variable latente (appelé aussi à données incomplètes) dans le cadre duquel la méthode d'estimation par maximum de vraisemblance n'est pas applicable. En effet, l'évaluation de la log-vraisemblance des données observées, c'est-à-dire :

$$\log p_{\boldsymbol{\theta}}(\mathbf{Y}) = \log \int_{\mathcal{Z}} p_{\boldsymbol{\theta}}(\mathbf{Y}, \mathbf{Z}) \mathrm{d}\mathbf{Z}$$

est impossible, du fait de l'intégration sur l'ensemble $\mathcal{Z} = \mathbb{R}^S$ décrivant l'espace des valeurs possibles de la variable latente.

Une méthode populaire pour contourner cette difficulté consiste à utiliser l'algorithme *Expectation-Maximization* (EM (Dempster *et al*. 1977)) pour maximiser localement la log-vraisemblance en s'appuyant sur l'espérance conditionnelle de la log-vraisemblance des données complétées, c'est-à-dire :

$$\mathbb{E}_{\boldsymbol{\theta}} \left[\log p_{\boldsymbol{\theta}}(\mathbf{Y}, \mathbf{Z}) \mid \mathbf{Y} \right] \qquad [8.3]$$

Cette approche permet de traiter l'estimation d'un grand nombre de modèles à variables latentes. Malheureusement, le modèle PLN ne se prête pas directement à son utilisation lorsque le nombre d'espèces S augmente : l'évaluation de [8.3] nécessite d'être capable d'intégrer selon la distribution de chaque vecteur latent $\mathbf{Z}_n$, conditionnellement au vecteur de comptage $\mathbf{Y}_n$. Or, cette distribution n'a pas de forme close dans le cadre du modèle PLN et il faudrait avoir recours à des schémas d'intégration numérique ou des approches de type Monte-Carlo qui passent difficilement à l'échelle de données de plus de quelques dizaines d'espèces.

8.2.2.2. *Approximation variationnelle*

Pour contourner ce problème, nous nous appuyons sur une méthode d'approximation variationnelle, qui consiste à trouver une distribution approchée pour $p_{\boldsymbol{\theta}}(\mathbf{Z}_n \mid \mathbf{Y}_n)$ qui simplifie l'intégration. L'approche variationnelle telle qu'introduite par (Wainwright et Jordan 2008) consiste à minimiser la mesure de divergence de Kullback-Leibler KL entre la vraie loi conditionnelle et la loi approchée, cette dernière étant choisie dans une classe de distribution prédéfinie et simplifiant le calcul de [8.3]. Dans le cas des modèles PLN, nous proposons d'approcher $p_{\boldsymbol{\theta}}(\mathbf{Z}_n \mid \mathbf{Y}_n)$

par une distribution gaussienne multivariée notée q_n, de vecteur de moyenne $\mathbf{m}_n$ et de matrice de variance-covariance diagonale $\mathbf{S}_n = \text{diag}(\mathbf{s}_n^2)$. L'ensemble des paramètres variationnels en jeu sont collectés dans le vecteur $\boldsymbol{\psi} = (\mathbf{M}, \mathbf{S})$, où $\mathbf{M} = t([t(\mathbf{m}_1) \dots t(\mathbf{m}_n)])$, $\mathbf{S} = t([(t(\mathbf{s}_1^2)) \dots t((\mathbf{s}_n^2)])$.

L'utilisation de la divergence de Kullback-Leibler pour mesurer la qualité de l'approximation mène à une version approchée de l'algorithme EM (ou EM variationnel), qui maximise une borne inférieure de la log-vraisemblance des observations, définie par :

$$\begin{aligned} J(\mathbf{Y}; \boldsymbol{\psi}, \theta) &\triangleq \log p_{\boldsymbol{\theta}}(\mathbf{Y}) - KL\left[q_{\boldsymbol{\psi}}(\mathbf{Z}) || p_{\boldsymbol{\theta}}(\mathbf{Z} \mid \mathbf{Y})\right] \\ &= \mathbb{E}_{q_{\boldsymbol{\psi}}}\left[\log p_{\boldsymbol{\theta}}(\mathbf{Y}, \mathbf{Z})\right] - \mathbb{E}_{q_{\boldsymbol{\psi}}}\left[\log q_{\boldsymbol{\psi}}(\mathbf{Z})\right] \end{aligned} \quad [8.4]$$

où $\mathbb{E}_{q_\psi}$ est l'espérance conditionnelle approchée au sens de la distribution $q_{\boldsymbol{\psi}}$. Une telle approche peut être généralisée à l'ensemble des variantes du modèle PLN que nous utilisons dans ce chapitre. La borne inférieure de la vraisemblance est alors optimisée en $\boldsymbol{\psi}$ et $\boldsymbol{\theta}$ à l'aide d'une approche type montée de gradient (plus précisément, nous utilisons l'algorithme CCSA proposé par (Svanberg 2002) et implémenté dans la librairie C++ `nlopt` (Johnson 2011)).

8.2.3. *Réduction de dimension*

Nous présentons ici une première variante du modèle PLN général, spécialement adaptée pour les études portant sur un grand nombre d'espèces. Cette variante est notamment utile pour la visualisation de grands jeux de données.

8.2.3.1. *Présentation du modèle*

Dans le modèle PLN général, on suppose que la variable latente appartient à un espace latent de même dimension S que l'espace des observations. Cette propriété résulte de l'absence de contrainte sur la matrice de covariance $\boldsymbol{\Sigma}$ de la loi des vecteurs latents $\mathbf{Z}_n$. Cette hypothèse peut être coûteuse dans le cas où le nombre d'espèces est grand ; il est alors tentant de supposer que les Z_{ns} appartiennent à un espace de dimension intrinsèque $K \ll S$. On peut pour cela définir, dans le cadre du modèle PLN, le strict analogue du modèle d'analyse en composantes principales (ACP) probabilistes proposé par (Tipping et Bishop 1999) dans le cadre gaussien : le modèle PLNPCA (Chiquet *et al.* 2018).

Une première formulation du modèle PLNPCA consiste à supposer que la matrice $\boldsymbol{\Sigma}$ est de rang $K \ll S$. Cette hypothèse impose qu'il existe une matrice $\mathbf{T}$ de dimension $S \times K$ telle que :

$$\boldsymbol{\Sigma} = \mathbf{T}t(\mathbf{T}) \tag{8.5}$$

Cette formulation en suggère une autre, plus intuitive qui repose sur la définition d'un vecteur de facteurs latents $\mathbf{W}_n$ associé à chaque échantillon :

$$\begin{aligned} &\text{espace des facteurs latents } \mathbf{W}_n \sim \mathcal{N}(\mathbf{0}_K, \boldsymbol{I}_{K\times K}) \\ &\text{espace latent } \mathbf{Z}_n = \boldsymbol{\mu} + \mathbf{T}\mathbf{W}_n \\ &\Rightarrow \mathbf{Z}_n \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{T}t(\mathbf{T})) \end{aligned} \tag{8.6}$$

On peut remarquer que le vecteur latent $\mathbf{Z}_n$ est complètement déterminé par le vecteur $\mathbf{W}_n$: le modèle PLNPCA n'induit donc pas à proprement parler de couche latente supplémentaire par rapport au modèle PLN défini à l'équation [8.1]. Comme PLN, PLNPCA s'accommode aisément de covariables : il suffit pour cela de reprendre la distribution des variables latentes définie à l'équation [8.2].

L'ensemble des paramètres du modèle est désormais $\boldsymbol{\theta} = \{\mathbf{B}, \mathbf{T}\}$ mais il faut noter que le modèle est inchangé si on remplace $\mathbf{T}$ par $\mathbf{TO}$, où $\mathbf{O}$ est une matrice de rotation de $\mathbb{R}^K$. Autrement dit, $\mathbf{T}$ n'est identifiable qu'à une matrice de rotation près, au travers de la matrice $\mathbf{T}t(\mathbf{T})$, exactement comme dans l'ACP probabiliste[3].

8.2.3.2. *Inférence du modèle*

L'inférence du modèle PLNPCA s'appuie sur la même mécanique que celle du modèle PLN avec deux différences majeures. En premier lieu, l'approximation variationnelle porte désormais sur $p_{\boldsymbol{\theta}}(\mathbf{W}_n \mid \mathbf{Y}_n)$. Les vecteurs $\mathbf{m}_n$ et $\mathbf{s}_n$ sont donc de taille K, et non S. En second lieu, le choix du nombre K de facteurs latents, n'est généralement pas connu et cela nécessite donc d'utiliser un critère de sélection.

La méthode d'inférence adoptée en pratique est la suivante : (i) sélectionner un nombre de facteurs latents maximum (noté $K_{\max}$), (ii) estimer un modèle PLNPCA de taille K pour toutes les valeurs de K entre 1 et $K_{\max}$ et (iii) sélectionner la valeur $\hat{K}$ qui maximise le critère de vraisemblance pénalisée suivant :

$$BIC(K) = J_K(\mathbf{Y}; \boldsymbol{\psi}, \theta) - \frac{S(D+K)}{2}$$

où J_K est la vraisemblance variationnelle [8.4] calculée pour le modèle PLNPCA à K facteurs. Ce critère est adapté du critère BIC de (Schwarz 1978) en remplaçant la vraisemblance par son approximation variationnelle.

3. Pour laquelle les contraintes d'orthogonalité des axes et de tri des axes par valeur propre décroissante permettent néanmoins de lever le problème de non-identifiabilité.

8.2.3.3. *Exploitation des résultats d'estimation*

À l'issue de l'estimation des paramètres du modèle, on dispose d'un estimateur $\widehat{\boldsymbol{\theta}}$ de $\boldsymbol{\theta}$ et $\widetilde{\boldsymbol{\psi}}$ de $\boldsymbol{\psi}$. Ces derniers permettent (i) de calculer la déviance expliquée par le modèle, (ii) d'estimer la position des échantillons dans l'espace latent et (iii) d'explorer la structure résiduelle, c'est-à-dire non expliquée par les covariables.

La déviance se calcule à l'aide de la formule :

$$D_{\hat{K}} = \frac{J_{\hat{K}}(\mathbf{Y}, \widetilde{\boldsymbol{\psi}}, \widehat{\boldsymbol{\theta}}) - J_{\min}}{J_{\max} - J_{\min}}$$

où $J_{\max}$ est la vraisemblance du modèle saturé, obtenue en imposant $\mathbf{Z}_n = \log(\mathbf{Y}_{ns})$ dans [8.4], et $J_{\min}$ est la vraisemblance du modèle à 0 facteurs, obtenue en imposant $\mathbf{Z}_n = \mathbf{o}_n + t(\mathbf{x}_n)\widehat{\mathbf{B}}$ dans la même équation. La position des échantillons dans l'espace latent est donnée par $\widetilde{\mathbf{Z}_n} = \mathbf{o}_n + t(\mathbf{x}_n)\widehat{\mathbf{B}} + \widehat{\mathbf{T}}\widetilde{\mathbf{m}}_n$. Pour faciliter la visualisation, on s'intéresse uniquement au terme $\widehat{\mathbf{T}}\widetilde{\mathbf{m}}_n$. Ce dernier correspond à la structure *résiduelle*, celle qui reste après avoir corrigé l'hétérogénéité des efforts d'observations ($\mathbf{o}_n$) et l'effet des covariables ($(\mathbf{x}_n)\widehat{\mathbf{B}}$). Nous en donnons une illustration dans la section 8.3.3.

8.2.4. *Inférence de réseaux d'interaction*

Nous introduisons maintenant une seconde variante du modèle PLN qui vise à reconstruire le réseau des interactions écologiques, c'est-à-dire notamment à distinguer entre les associations statistiques (corrélations) et les interactions directes entre chaque paire d'espèces.

8.2.4.1. *PLN comme un modèle graphique gaussien*

Le modèle PLN peut également être utilisé pour mieux comprendre les interactions entre les espèces composant un écosystème. La matrice de covariance $\boldsymbol{\Sigma} = [\sigma_{sr}]_{1\leq j,k\leq S}$ fournit une première indication sur ces relations, au travers des corrélations $\rho_{sr} = \sigma_{sr}/\sqrt{\sigma_{ss}\sigma_{rr}}$. On sait cependant que la simple analyse de ces corrélations ne permet pas de distinguer les interactions directes entre espèces des associations qui sont le fruit de liens indirects. Ainsi, les abondances de deux proies d'un même prédateur peuvent être corrélées du seul fait des fluctuations de l'abondance du prédateur, même si ces deux proies n'entretiennent aucune interaction entre elles.

Les modèles graphiques (Lauritzen 1996) fournissent un cadre probabiliste général pour opérer cette distinction. Sans entrer dans une présentation générale des modèles graphiques, on peut retenir que, dans le cas des modèles graphiques gaussiens (GGM), la matrice de *précision* $\boldsymbol{\Omega} = [\omega_{sr}]_{1\leq s,r\leq S} := \boldsymbol{\Sigma}^{-1}$ est associée aux corrélations *partielles* entre les espèces $\widetilde{\rho}_{sr} = -\omega_{sr}/\sqrt{\omega_{ss}\omega_{rr}}$. Dans le cadre gaussien, la corrélation

partielle est en fait une corrélation conditionnelle, ce qui signifie que $\widetilde{\rho}_{sr} = 0$ si et seulement si les variables latentes Z_{ns} et Z_{nr} sont indépendantes conditionnellement à toutes les autres $\{Z_{nq}\}_{q \neq s,r}$. L'interprétation écologique de cette propriété est que la nullité de $\widetilde{\rho}_{sr}$ indique que les espèces j et k n'interagissent pas directement.

8.2.4.2. *Inférence*

L'inférence des réseaux écologiques vise ainsi à distinguer les associations indirectes des associations directes, qu'on suppose généralement peu nombreuses. Cette dernière hypothèse revient à supposer que la matrice $\mathbf{\Omega}$ est creuse, c'est-à-dire contient une majorité de termes nuls. (Chiquet *et al*. 2019) proposent ainsi une version du modèle PLN dédiée à l'inférence de réseau qui reprend le modèle décrit dans les équations [8.1] et [8.2] en ajoutant lors de son inférence un terme de pénalité visant à rendre la matrice $\mathbf{\Omega}$ creuse. Plus précisément, les paramètres du modèle sont estimés en maximisant la fonction :

$$J_\lambda(\mathbf{Y}; \boldsymbol{\psi}, \theta) := J(\mathbf{Y}; \boldsymbol{\psi}, \theta) - \lambda \sum_{s<r} |\omega_{sr}| \qquad [8.7]$$

où $J(\mathbf{Y}; \boldsymbol{\psi}, \theta)$ est la borne inférieure définie à l'équation [8.4]. Le paramètre de régularisation λ contrôle la parcimonie de la matrice $\mathbf{\Omega}$: plus λ est élevé, moins les interactions inférées ($\widehat{\omega}_{sr} \neq 0$) seront nombreuses. Cette fonction objectif s'avère convexe à la fois en $\boldsymbol{\psi}$ et en θ, ce qui permet d'utiliser un algorithme de descente de gradient efficace. Le choix de λ est évidemment critique : il peut se faire soit au moyen de critère de vraisemblance pénalisée de type BIC ou eBIC (Foygel et Drton 2010), soit par rééchantillonnage (voir par exemple (Liu *et al*. 2010)).

8.3. Analyse des données d'espèces marines

8.3.1. *Description des données*

Les données considérées ici sont les abondances d'espèces marines (poissons, invertébrés, algues) observées sur l'île Anapaca au large des côtes de Basse Californie. L'île est constituée de trois îlots. Nous considérons les données relatives aux deux îlots Est (site AEI) et Milieu (site AMI). Seules les observations réalisées les années où les côtes étaient protégées sont considérées, soit 1999-2014 (16 années) pour AEI et 2003-2014 (12 années) pour AMI. Sur chaque îlot (que l'on appellera aussi site), des régions d'observation sont définies à l'est (côté E), au centre (côté CEN) et à l'ouest (côté W). Enfin, pour chaque côté, des zones plus ou moins éloignées de la côte sont définies (zones INNER, MID et OUTER). Quatre protocoles d'observation ont été définis, adaptés aux espèces observées et basés sur des transects (sorte de couloirs virtuels sous l'eau) à différentes profondeurs pour observer les

poissons ou positionnés sur le fond marin et intégrant des quadrats (surfaces carrées) de différentes dimensions pour observer les algues, les invertébrés et aussi certains poissons. Ainsi, les quatre protocoles n'ont pas été systématiquement utilisés dans les transects et certaines espèces peuvent être observées par plusieurs protocoles.

Afin de se ramener à une table de comptage, les données brutes ont été agrégées par échantillon, défini comme une combinaison unique année $\times$ site $\times$ côté $\times$ zone. Dans chaque échantillon, l'abondance de chaque espèce est définie comme le nombre total d'occurrences de cette dernière sur l'ensemble des transects. L'intensité d'observation (utilisé comme offset dans les modèles PLN) est de même définie comme le nombre de transects dans lequel un protocole permettant observer l'espèce a été mis en œuvre.

La table de comptages a ensuite été filtrée pour ne conserver que (i) les échantillons avec une intensité d'observation strictement positive pour au moins 80 % des espèces et (ii) les espèces avec une abondance moyenne supérieure à 1 et une intensité d'observation strictement positive dans au moins 80 % des échantillons restants. Ces filtrages conservent 66 espèces (sur 195) et 142 échantillons (sur 169). Ils permettent de filtrer toutes les intensités d'observation nulles, que l'on ne sait pas traiter numériquement, et de réduire fortement la proportion de 0s dans la table de comptages (de 76 % à 44 %).

On appelle données MariNet, les données issues de cette phase de prétraitement des données brutes de comptage d'espèces. Les données MariNet sont constituées par (i) une définition d'échantillons, (ii) l'abondance de chaque espèce et (iii) l'intensité d'observation pour chaque échantillon, ce qui permet de construire les 3 matrices $\mathbf{Y}$, $\mathbf{X}$ et $\mathbf{O}$. Enfin, les espèces sont identifiées par un code dans le texte et le tableau 8.1 donne le dictionnaire entre ce code et le nom scientifique et le nom anglais courant de l'espèce.

8.3.2. *Effets du site et de la date*

Dans un premier temps, nous avons utilisé le modèle Poisson log-normal en incluant l'effet de plusieurs covariables, une par une, afin de déterminer celles de plus forte influence. Ces covariables sont : le site, le côté d'observation, la zone, et enfin la période. La période est un groupement d'années successives. En effet, une première analyse n'a pas révélé d'effet année fort, en revanche, si l'on regroupe les années en deux périodes (de 1999 à 2001, et après 2001), un effet apparaît. La coupure à l'année 2001 a été obtenue de manière automatique, à partir du résultat d'une hiérarchique ascendante (avec critère Ward) sur le jeu de données. Les covariables de plus fort effet sont déterminées grâce au critère de sélection de modèles BIC (ICL est également

disponible, mais plus adapté aux problèmes de , donc non retenu). Le modèle incluant un effet site puis celui incluant un effet période sont les deux modèles maximisant le BIC. Les autres modèles ont un BIC inférieur à celui du modèle sans covariable.

Code	Nom scientifique	Nom commun anglais/description
BFRE	*Brachyistius frenatus*	Kelp surfperch
EMOR	*Engraulis mordax*	Northern anchovy
KGB	*Sebastes* (*atrovirens, carnatus, Chrysomelas, caurinus*)	Rockfish
SJAP	*Scomber japonicus*	Greenback mackerel
TSYM	*Trachurus symmetricus*	Jack mackerel
ANTSOL	*Anthopleura sola*	Green anemone
APLCAL	*Aplysia californica*	California brown
BARNAC		Barnacle
CYPSPA	*Cypraea spadicea*	Chestnut cowrie
DICTYOTALES	*Dictyota spp.* et *Dictyopteris undulata*	
LOPCHI	*Lophogorgia chilensis*	Red gorgonian
LYTANAAD	*Lytechinus anamesus*	White urchin, adult > 2,5 cm
MEGSPP	*Megastrea spp.*	Turban snail
MEGUND	*Megastrea undosum*	Wavy turban snail
PANINT	*Panulirus interruptus*	Spiny lobster
STRFRAAD	*Strongylocentrotus franciscanus*	Red urchin, adult > 2,5 cm
STRPURAD	*Strongylocentrotus purpuratus*	Purple urchin, adult > 2,5 cm
BROWN	*Colpomenia spp.*	Brown algae
BUSHY	*Gelidium, Pterocladia, Gastroclonium, Gracilaria, Condracanthus canaliculatus*	Red algae with cylindrical branches
CYSOSMAD	*Cystoseira osmundacea*	Bladder chain, adult diameter > 6 cm
ENCRED		Encrusting non-coralline red algae
LAMSPP	*Laminaria spp.*	
MACPYR_HF	*Macrosystis* holdfast	
MACPYRAD	*Laminaria spp.*	Giant kelp, adult height
PTECALAD	*Pterygophora californica*	

Tableau 8.1. *Dictionnaire entre le code d'une espèce et le nom scientifique et le nom commun anglais, pour les espèces pour lesquelles le modèle infère une interaction avec l'une des deux espèces d'oursin STRFRAAD ou STRPURAD. Les noms des poissons sont écrits en bleu, ceux des invertébrés en gris et ceux des algues en rouge.*

L'effet site semble s'expliquer par la présence de quelques espèces spécifiques d'un îlot, comme le montre la valeur des coefficients de chacun des deux sites dans le modèle de régression (figure 8.3). Ces espèces sont LAMSPP (une algue), PTECALAD (une algue), MEGSPP (un escargot de mer), SJAP (un maquereau) pour l'îlot AEI et TSYM (un maquereau) pour l'îlot AMI.

Figure 8.3. *Modèle Poisson log-normal avec covariable 'site'. Représentation des coefficients de chacun des deux sites dans le modèle de régression*

L'effet période observé pourrait s'expliquer par le fait pour la première période (de 1999 à 2001), les données proviennent uniquement du site AEI, lequel n'est un site protégé que depuis 1999. Cette première période pourrait être une phase de transition dans la composition et la structure de la communauté d'espèces présentes sur l'îlot, alors que durant la seconde période, la communauté protégée s'est stabilisée. L'estimation du modèle Poisson log-normal avec covariable 'période' sur les échantillons uniquement issus de l'îlot AEI, montre que la différence entre les deux périodes est due à quelques espèces plus fortement représentées dans la période 2 (LAMPPS, MEGSPP, TSYM, figure 8.4).

Figure 8.4. *Modèle Poisson log-normal avec covariable « période ». Représentation des coefficients de chacune des deux périodes dans le modèle de régression, dans le cas du site AEI*

8.3.3. *Réduction de dimension*

Nous avons ensuite regardé comment se comporte le modèle avec réduction de dimension, selon que l'on n'intègre aucune covariable, la covariable site seule (la plus influente d'après notre première analyse) ou toutes les covariables. Pour chacun de ces trois modèles, la dimension est choisie en utilisant le critère BIC. Le critère BIC sélectionne systématiquement 19 dimensions (contre 66 pour un modèle plein), même si la variance reste concentrée sur les 5 à 10 premiers axes de l'espace latent (figure 8.5, première ligne).

L'ajout du site, puis des autres covariables, permet de diminuer leur impact sur la structuration des communautés dans l'espace latent (figure 8.5, deuxième ligne) pour se concentrer sur l'espace résiduel. Formellement, l'effet du site est toujours présent

dans les positions latentes $\widetilde{\mathbf{Z}}_n$. Dans le premier modèle (sans covariable), il est intégré à la structure résiduelle $\widehat{\mathbf{T}}\widetilde{\mathbf{m}}_n$ (et donc visible dans les graphiques) tandis que dans les deux autres, il disparaît de la structure résiduelle (et donc des graphiques) pour intégrer le terme correctif $\mathbf{x}_n^\top \widehat{\mathbf{B}}$ associé aux covariables. Cela a pour conséquence de mieux répartir les espèces sur le cercle des corrélations (figure 8.5, troisième ligne), et notamment d'en avoir moins qui sont fortement associées à l'axe 1.

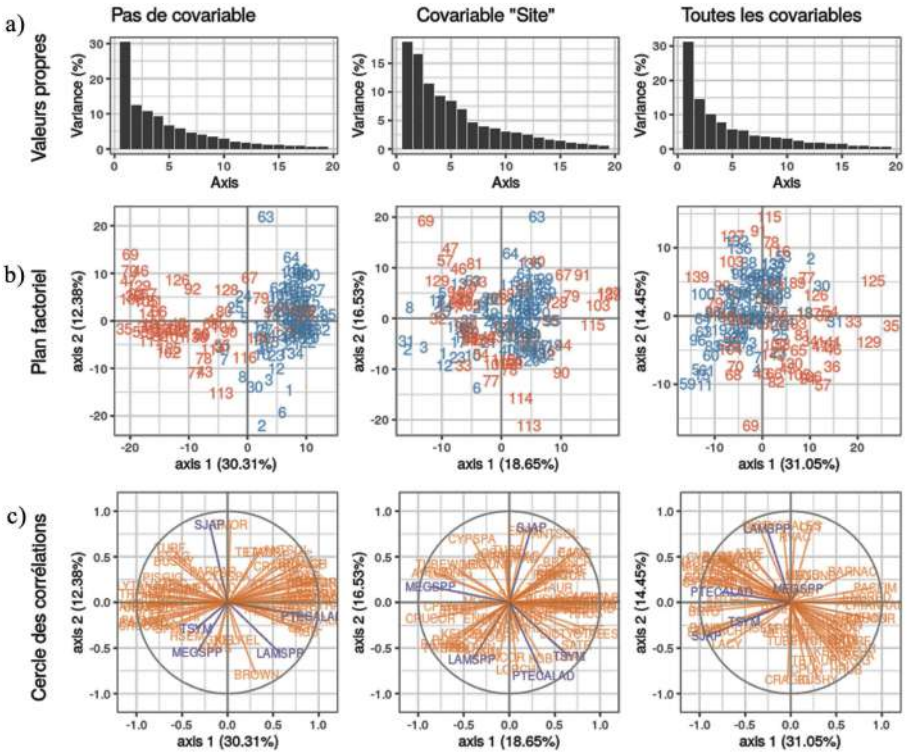


Figure 8.5. *Réduction de dimension pour un modèle sans covariable, avec introduction de la covariable site et avec introduction de toutes les covariables (de gauche à droite)*

COMMENTAIRE SUR LA FIGURE 8.5.– *a) Tracé des valeurs propres pour chacune des dimensions ; b) représentation des échantillons dans le premier plan principal (en bleu le site AEI, en rouge le site AMI) ; c) représentation des espèces sur le cercle des corrélations des deux premières dimensions (les espèces en bleu sont les espèces discriminantes des sites, telles qu'obtenues dans l'analyse précédente).*

Par ailleurs, si l'on applique la réduction de dimension uniquement aux échantillons provenant de l'îlot AEI, dans un modèle avec effet site, on observe à nouveau l'effet période identifié en première analyse : les échantillons de la période 1 se détachent nettement de ceux de la période 2 sur le premier plan principal (figure 8.6).

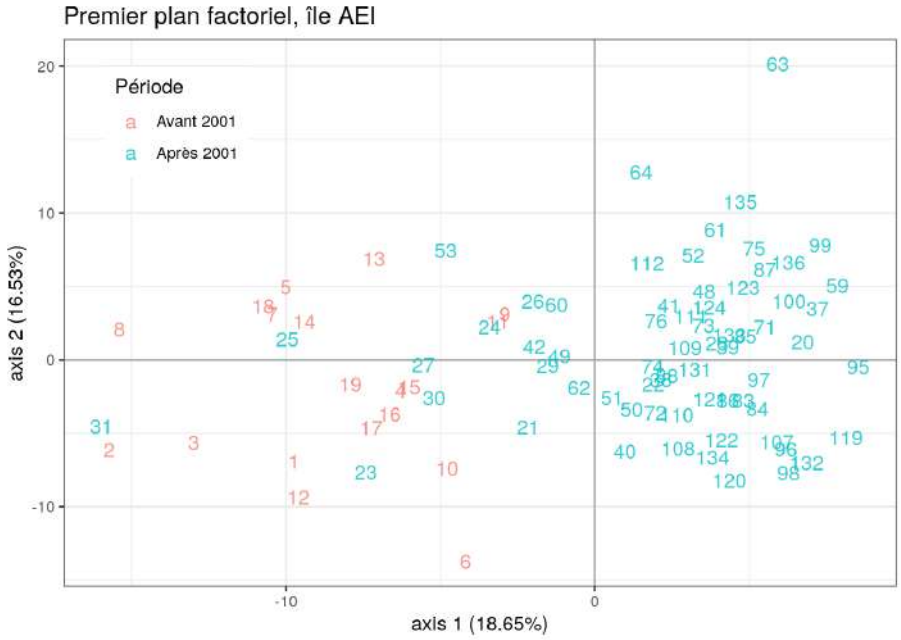


Figure 8.6. *Représentation des échantillons sur l'îlot AEI dans le premier plan principal. Les échantillons de la période 1 (rouge) se détachent de ceux de la période 2 (bleu)*

8.3.4. *Inférence d'interactions écologiques*

Nous tentons maintenant d'identifier des interactions directes entre espèces au moyen de l'approche décrite à la section 8.2.4. Pour mémoire, le réseau est obtenu en forçant la matrice de précision à contenir un grand nombre de zéros, la proportion de zéros (donc d'arêtes dans le réseau) étant contrôlée par le paramètre λ.

8.3.4.1. *Effet de la pénalité*

La figure 8.7 montre l'effet du paramètre λ sur la densité (c'est-à-dire la proportion d'arêtes présentes) et l'ajustement des 16 modèles possibles obtenus en combinant les 4 covariables année, site, côté, et zone. Comme attendu, la densité du réseau

(figure 8.7a) et l'ajustement du modèle (mesurée par J_λ définie en [8.7], figure 8.7b) augmentent systématiquement quand le paramètre de régularisation λ diminue. L'utilisation d'un critère de sélection de modèle (ici BIC) permet de corriger cet effet et de déterminer une valeur optimale pour λ.

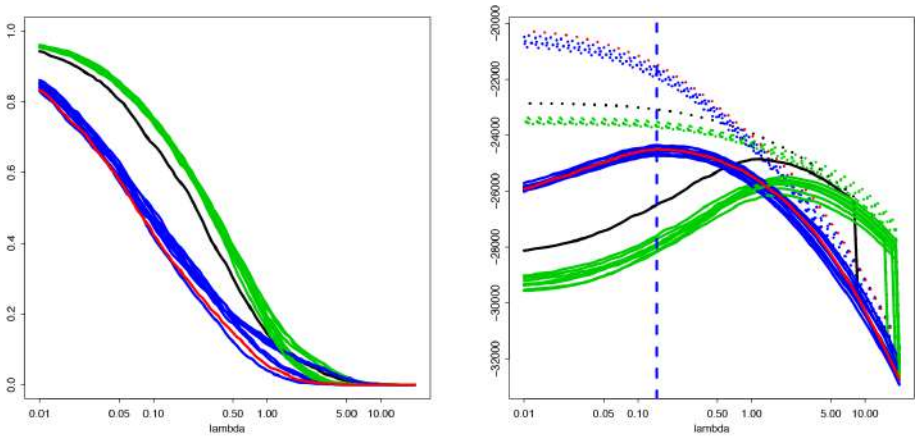


Figure 8.7. *Effet du paramètre λ sur la densité (a) et l'ajustement (b) des modèles. Légende : noir = modèle sans covariable, rouge = modèle complet, vert = modèles n'incluant pas l'effet année, bleu = modèles incluant l'effet année. À droite : pointillés = borne inférieure J_λ, traits pleins = critère BIC. Pointillé vertical : valeur optimale de λ pour le modèle « année + site ».*

La figure 8.7 montre deux comportements très différents, selon que le modèle inclut ou non l'effet de l'année. Les modèles incluant l'année (bleu et rouge) sont à la fois mieux ajustés et plus parcimonieux en termes d'arêtes. Cette observation confirme l'effet majeur de l'année sur l'abondance des différentes espèces. Au total, le modèle qui donne le meilleur critère BIC ($-24348,52$) est le modèle prenant en compte l'année et le site ; cet optimum est atteint pour $\lambda = 0,147$.

8.3.4.2. *Robustesse des arêtes*

Le choix du paramètre de régularisation est évidemment critique et a une grande influence sur le réseau finalement obtenu, et notamment sur sa densité (figure 8.7a). La robustesse des résultats peut être améliorée en utilisant une approche par rééchantillonnage telle que celle proposée par (Liu *et al.* 2010) : elle consiste à ajuster le modèle sur un grand nombre de sous-échantillons et à associer à chaque arête une fréquence de sélection. La figure 8.8 montre les résultats de cette procédure pour le modèle incluant l'année et le site.

La figure figure 8.8a montre une séparation assez nette entre des arêtes qui sont presque systématiquement sélectionnées et des arêtes qui ne le sont presque jamais. La figure 8.8b montre que, dans le cas présent, la distribution de ces fréquences est très cohérente avec la liste des arêtes fournie directement par le critère BIC. La procédure de rééchantillonnage indique donc une bonne robustesse de la sélection d'arêtes opérée par BIC.

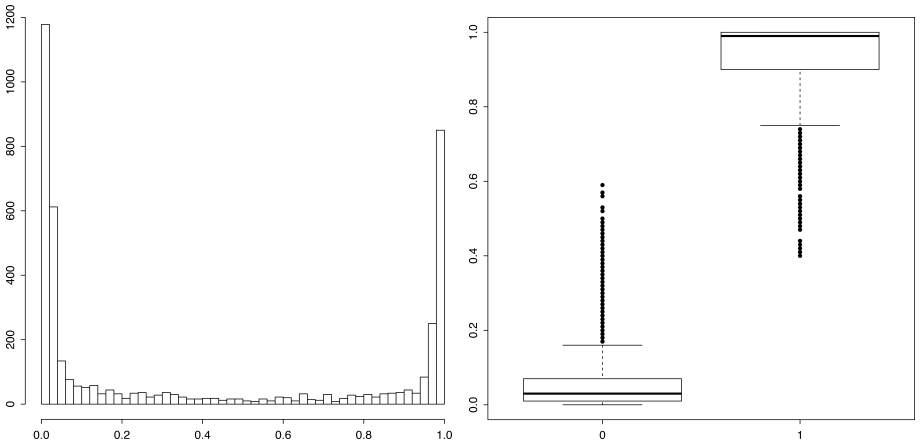


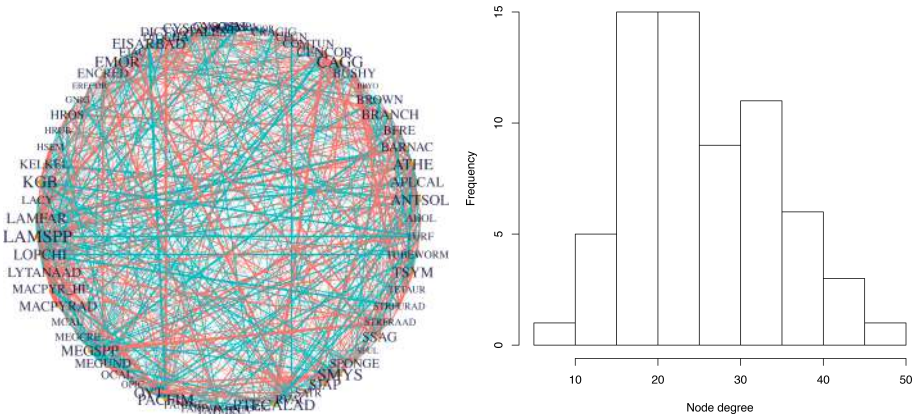
Figure 8.8. *Stabilité des arêtes sélectionnées pour le modèle incluant les effets des covariables année et site. a) Histogramme des fréquences de sélection des arêtes dans les sous-échantillons. b) Distribution de ces fréquences pour les arêtes non sélectionnées avec le critère BIC ('0') et les arêtes sélectionnées par ce même critère ('1').*

8.3.4.3. *Réseau inféré*

Le réseau inféré (figure 8.9) est un réseau assez dense (densité de 0,4), certaines espèces étant liées à de nombreuses autres, ce qui est compatible avec un écosystème complexe. Du fait de la complexité des relations écologiques (prédation, parasitisme, symbiose, etc.), la non-direction des relations trouvées, ainsi que la non-prise en compte de toutes les espèces et le faible nombre d'observation, il convient d'analyser le réseau avec précaution (Blanchet *et al*. 2020), non comme un réseau d'interactions directes réelles, mais comme un support de réflexion en le confrontant à la connaissance actuelle.

Afin d'analyser le réseau, une visualisation de la matrice des corrélations partielles (figure 8.10 avec regroupement des espèces en poissons, invertébrés, algues) est intéressante. Cette matrice est symétrique, les relations n'étant pas dirigées. On peut noter que la région qui correspond aux interactions entre les poissons d'un côté et les invertébrés ou les algues de l'autre est moins riche en relations de poids supérieurs à 0,1

que l'ensemble de la matrice. La fréquence est de 0,12 pour les interactions entre poissons et 0,04 pour les interactions entre poissons x (invertébrés, algues). Les poissons se nourrissant d'invertébrés et d'algues (même si certains poissons mangent d'autres poissons), on aurait pu s'attendre à voir apparaître des relations trophiques fortes entre les poissons et les invertébrés ou les algues. On pourrait penser que ces interactions sont plus difficiles à détecter lorsque la relation est plus complexe (alimentation diversifiée, habitat, etc.). On peut aussi noter une relation entre MACPYRAD (*giant kelp*, adulte) et MACPYR_HF (*giant kelp*, grappin) dont le poids (corrélation partielle de 0,37) est très supérieur aux autres (le poids maximal suivant en valeur absolue est de 0,28). Ceci s'explique par le fait que ce soit la même espèce considérée sous des formes différentes.



a) b)

Figure 8.9. *Réseau d'interaction sélectionné : visualisation du réseau avec le package R* `PLNmodels` *(a) et histogramme des degrés des nœuds du réseau (nombre d'interactions de chaque espèce, b)*

Pour une étude, il peut être intéressant d'analyser les interactions de plus fort poids, ou les espèces les plus fortement connectées, ou bien encore des espèces d'intérêt. C'est ce dernier point que nous allons poursuivre en nous concentrant sur une espèce invasive d'oursin violet (STRPURAD) qui colonise les fonds marins au détriment notamment d'une espèce d'oursin rouge (STRFRAAD) comestible, mais aussi de tout l'écosystème (Woody 2020). La figure 8.11 visualise les interactions trouvées pour ces deux espèces d'oursins. On trouve effectivement une interaction forte entre les deux espèces. Il semble difficile d'interpréter d'un point de vue écologique le signe positif de la corrélation partielle associé à l'interaction. STRPURAD

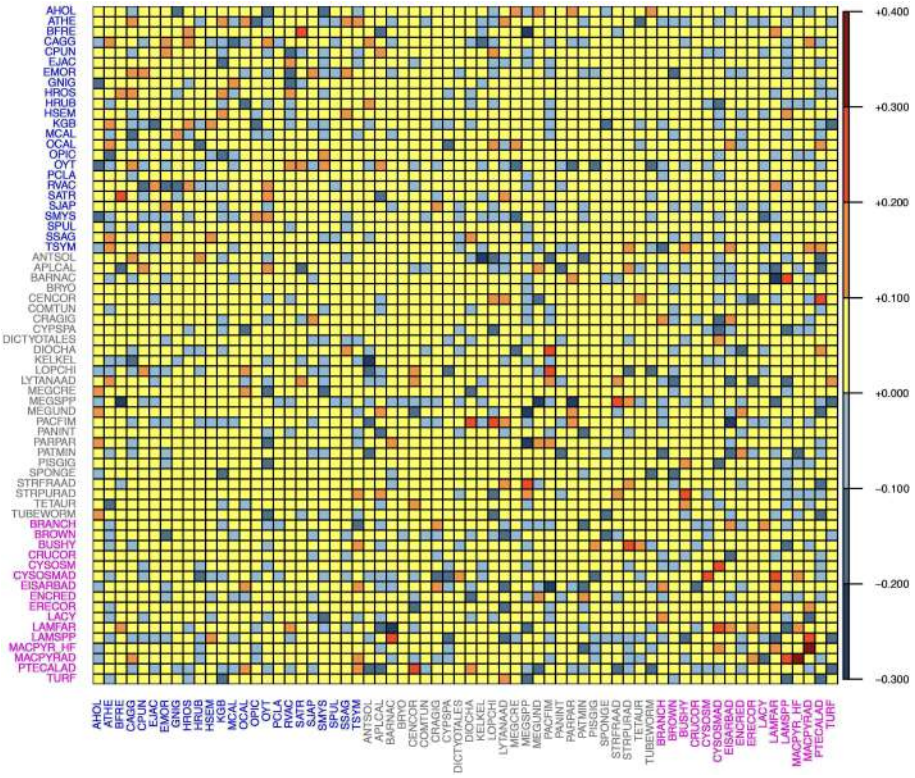
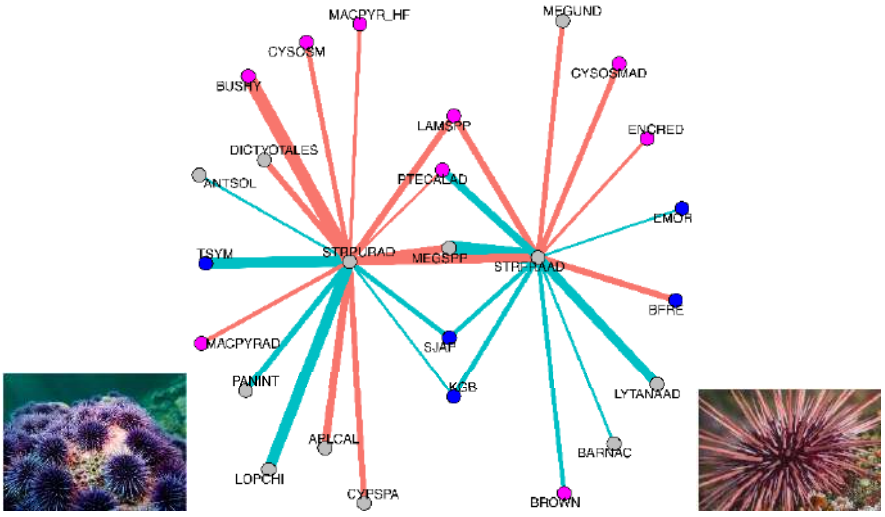


Figure 8.10. *Réseau d'interaction sélectionné : visualisation de la matrice des corrélations partielles en groupant les espèces poissons (bleu), invertébrés (gris) et algues (rouge)*

étant une espèce invasive, l'augmentation de sa population ne favorise *a priori* pas l'augmentation de population de STRFRAAD. En revanche, des actions peuvent avoir été menées pour protéger STRFRAAD, compte tenu de l'augmentation de la population de STRPURAD. Chaque espèce d'oursin est en interaction avec un nombre comparable d'espèces qui se répartissent de manière assez équitable entre algues, invertébrés et poissons. Ceci est compatible avec un point de vue trophique, les oursins sont à la fois des prédateurs d'algues et des proies pour certains poissons. Seules cinq espèces interagissent avec les deux espèces d'oursin considérées. Cela pourrait être le reflet de réseaux d'interactions écologiques assez différents pour chaque espèce. Quatre de ces cinq espèces sont les espèces fortement liées à l'îlot AEI identifiées dans l'étude de l'effet des covariables. Elles pourraient être structurantes de la communauté sur ce site. Enfin, on trouve plusieurs espèces d'intérêt direct pour l'homme : homard (PANINT), maquereau (TSYM, SJAP), anchois (EMOR), perche

de mer (BFRE) et sébastes (KGB), ce qui est en accord avec l'intérêt porté par les biologistes aux oursins dans l'écosystème.



a) b)

Figure 8.11. *Interactions trouvées pour deux espèces d'oursin*

COMMENTAIRE SUR LA FIGURE 8.11.– *a) Oursin rouge (STRFRAAD, Strongylocentrotus franciscanus). b) Oursin violet (STRPURAD, Strongylocentrotus purpuratus) (photos de piscoweb.org). Les nœuds bleus correspondent à des poissons, les gris à des invertébrés et les rouges à des algues. La couleur des arêtes est rouge pour une relation positive et cyan pour une relation négative. L'épaisseur des arêtes est proportionnelle à l'intensité de la relation.*

8.4. Discussion

La compréhension du fonctionnement d'un écosystème passe par celle des interactions existantes entre les espèces qui le composent. De ce fait, les analyses statistiques visant à aider à cette compréhension doivent nécessairement reposer sur une modélisation jointe de l'abondance de l'ensemble des espèces. Le modèle PLN offre un cadre flexible et aisément interprétable pour comprendre à la fois les effets environnementaux (interactions abiotiques) et la structure de dépendance entre les espèces (interactions biotiques). Comme la plupart des modèles joints de distribution d'espèces (JSDM), le modèle PLN utilise une couche latente pour modéliser la dépendance entre les espèces. Contrairement à d'autres modèles présentés dans cet ouvrage, ces

variables latentes n'ont pas de réalité biologique, mais servent seulement d'auxiliaire de modélisation.

Nous avons décrit dans ce chapitre plusieurs variantes utiles en écologie, comme la réduction de dimension, ou la comparaison de sites ou d'échantillons ou encore l'inférence de réseaux. Du fait de son interprétabilité, il est facile de confronter *a posteriori* les résultats du modèle PLN avec des données exogènes : on peut ainsi comparer la réponse de chaque espèce à l'environnement (décrite par les coefficients de régression) avec des traits ou des groupes fonctionnels d'espèces pour mieux comprendre les règles gouvernant cette réponse. L'ensemble de ces variantes sont implémentées dans le package R `PLNmodels` disponible sur `cran.r-project.org`. La syntaxe de ce package est semblable à celle de la plupart des modèles sous R. D'autre généralisations sont en cours de développement et seront bientôt disponibles. On peut notamment introduire un modèle de mélange gaussien (McLahan et Peel 2000) dans la partie latente du modèle [8.1] afin de permettre une structuration des sites en groupes homogènes.

On peut noter que la réduction de dimension telle que nous l'avons présentée à la section 8.2.3 est spécialement intéressante quand le nombre d'espèces S est grand. Elle est donc complémentaire de la méthode SCGLR proposée dans le chapitre 9 qui traite du cas où le nombre de covariables D est grand : SGCLR vise à exhiber automatiquement quelques composantes explicatives, définies comme des combinaisons des covariables originales.

Comme indiqué à la section 8.2.2, la méthode d'inférence utilisée repose sur une approximation variationnelle qui apporte une grande efficacité computationnelle mais qui ne permet pas d'obtenir des mesures d'incertitude (écart-type, intervalle de confiance) pour les estimations des paramètres ou de définir des tests de significativité. Plusieurs séries de travaux sont poursuivis en ce sens à l'heure actuelle.

Le modèle PLN permet de prendre en compte le fait que l'effort d'échantillonnage peut varier selon les sites, les échantillons ou les espèces. Cette prise en compte est évidemment essentielle pour éviter des biais qui feraient perdre toute pertinence aux résultats. Dans bon nombre d'expériences, aucune mesure directe de cet effort n'est disponible et seules des estimations (comme le nombre total d'individus observés) peuvent être utilisées. La qualité de ces estimations influe évidemment sur la qualité des résultats de l'analyse.

8.5. Remerciements

Nous remercions Jennifer Caselle pour avoir mis à notre disposition les données issues du projet PISCO, et Jennifer Caselle et Laura Dee pour les discussions sur la compréhension de ces données.

8.6. Bibliographie

Aitchison, J. and Ho, C. (1989). The multivariate Poisson-log normal distribution. *Biometrika*, 76(4), 643–653.

Blanchet, F.G., Cazelles, K., Gravel, D. (2020). Co-occurrence is not evidence of ecological interactions. *Ecology Letters*, 23(7), 1050–1063 [Online]. Available at : https://onlinelibrary.wiley.com/doi/abs/10.1111/ele.13525.

Caselle, J. (2013). A decade of protection, 10 years of change at the channel islands [Online]. Available at : http ://www.piscoweb.org/sites/default/files/portfolios /CI_10-Yr_Brochure_web.pdf.

Chiquet, J., Mariadassou, M., Robin, S. (2018). Variational inference for probabilistic Poisson PCA. *The Annals of Applied Statistics*, 12(4), 2674–2698.

Chiquet, J., Mariadassou, M., Robin, S. (2019). A variational Bayesian framework for graphical models. *International Conference on Machine Learning*, Long Beach, CA, USA.

Clark, J.S., Gelfand, A.E., Woodall, C.W., Zhu, K. (2014). More than the sum of the parts : Forest climate response from joint species distribution models. *Ecological Applications*, 24(5), 990–999.

Dempster, A.P., Laird, N.M., Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39, 1–38.

Elith, J. and Leathwick, J.R. (2009). Species distribution models : Ecological explanation and prediction across space and time. *Annual Review of Ecology, Evolution, and Systematics*, 40, 677–697.

Foygel, R. and Drton, M. (2010). Extended Bayesian information criteria for Gaussian graphical models. *Advances in Neural Information Processing Systems*, 604–612.

Harris, D.J. (2015). Generating realistic assemblages with a joint species distribution model. *Methods in Ecology and Evolution*, 6(4), 465–473.

Inouye, D.I., Yang, E., Allen, G.I., Ravikumar, P. (2017). A review of multivariate distributions for count data derived from the Poisson distribution. *Wiley Interdisciplinary Reviews: Computational Statistics*, 9(3), e1398.

Johnson, S.G. (2011). The NLopt nonlinear-optimization package.

Lauritzen, S.L. (1996). *Graphical Models*, volume 17. Clarendon Press, Oxford.

Layeghifard, M., Hwang, D.M., Guttman, D.S. (2017). Disentangling interactions in the microbiome : A network perspective. *Trends in Microbiology*, 25(3), 217–228 [Online]. Available at : http://www.sciencedirect.com/science/article/pii/S0966842X16301858.

Cette bibliographie est identique à celle de l'ouvrage correspondant en anglais publié par ISTE.

Liu, H., Roeder, K., Wasserman, L. (2010). Stability approach to regularization selection (stars) for high dimensional graphical models. *NIPS'10, Curran Associates Inc.*, 1432–1440.

McLahan, G. and Peel, D. (2000). *Finite Mixture Models*. John Wiley & Sons, Inc, Brisbane.

Ovaskainen, O., Tikhonov, G., Norberg, A., Blanchet, F.G., Duan, L., Dunson, D., Roslin, T., Abrego, N. (2017). How to make more out of community data ? A conceptual framework and its implementation as models and software. *Ecology Letters*, 20(5), 561–576.

Pan, J., Marcoval, M., Bazzini, S., Vallina, M., De Marco, S. (2013). Coastal marine biodiversity : Challenges and threats. In *Marine Ecology in a Changing World*, Arias, A.h., Menendez A.C. (eds.). CRC Press, Boca Raton.

PISCO Research Consortium (2019a). Kelp forest sampling protocols [Online]. Available at : http://www.piscoweb.org/kelp-forest-sampling-protocols.

PISCO Research Consortium (2019b). Partnership for interdisciplinary studies of coastal oceans [Online]. Available at : http://piscoweb.org.

Popovic, G.C., Hui, F.K., Warton, D.I. (2018). A general algorithm for covariance modeling of discrete data. *Journal of Multivariate Analysis*, 165, 86–100.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6(2), 461–464 [Online]. Available at : http://dx.doi.org/10.1214/aos/1176344136.

Svanberg, K. (2002). A class of globally convergent optimization methods based on conservative convex separable approximations. *SIAM Journal on Optimization*, 12(2), 555–573.

Tipping, M.E. and Bishop, C.M. (1999). Probabilistic principal component analysis. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 61(3), 611–622 [Online]. Available at : http://dx.doi.org/10.1111/1467-9868.00196.

Tylianakis, J.M., Laliberté, E., Nielsen, A., Bascompte, J. (2010), Conservation of species interaction networks. *Biological Conservation*, 143(10), 2270–2279 [Online]. Available at : http://www.sciencedirect.com/science/article/pii/S0006320709005126.

Wainwright, M.J. and Jordan, M.I. (2008). Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1–2), 1–305.

Warton, D.I., Blanchet, F.G., O'Hara, R.B., Ovaskainen, O., Taskinen, S., Walker, S.C., Hui, F.K. (2015). So many variables: Joint modeling in community ecology. *Trends in Ecology & Evolution*, 30(12), 766–779.

Woody, T. (2020). California's critical kelp forests are disappearing in a warming world. Can they be saved ? [Online]. Available at : https://www.nationalgeographic.com/science/2020/04/california-critical-kelp-forests-disappearing-warming-world-can-they-be-saved/.

Xiao, H., Dee, L.E., Chadès, I., Peyrard, N., Sabbadin, R., Stringer, M., McDonald-Madden, E. (2018). Win-wins for biodiversity and ecosystem service conservation depend on the trophic levels of the species providing services. *Journal of Applied Ecology*, 55(5), 2160–2170 [Online]. Available at : https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/1365-2664.13192.