Classification de données multivariées multitypes basée sur des modèles de mélange : application à l'étude d'assemblages d'espèces en écologie

Vera Georgescu

ACADÉMIE D'AIX-MARSEILLE
UNIVERSITÉ D'AVIGNON ET DES PAYS DE VAUCLUSE

THÈSE

présentée pour obtenir le grade de
Docteur en Sciences de l'Université d'Avignon et des Pays de Vaucluse

Spécialité : Biostatistique

Classification de données multivariées multitypes basée sur des modèles de mélange

Application à l'étude d'assemblages d'espèces en écologie

par

Vera GEORGESCU

soutenue publiquement le 17 décembre 2010 devant un jury composé de :

M. Christian Lavergne	Professeur, Université Montpellier III	Président du jury
M. Jean-Jacques Boreux	Professeur, Université de Liège	Rapporteur
M. Gilles Celeux	Directeur de recherche, INRIA Rocquencourt	Rapporteur
M. Frédéric Mortier	Chargé de recherche, CIRAD Montpellier	Examinateur
M. Rachid Senoussi	Directeur de recherche, INRA Avignon	Directeur de thèse
M. André Kretzschmar	Directeur de recherche, INRA Avignon	Directeur de thèse
M. Samuel Soubeyrand	Chargé de recherche, INRA Avignon	Directeur de thèse

École doctorale ED 166 Information, Structures, Systèmes
Laboratoire BioSP INRA Avignon

Remerciements

Cette thèse est le fruit d'un heureux hasard. Je ne voulais pas faire une thèse, je ne l'ai pas cherchée et je ne m'attendais pas non plus à une telle proposition, lorsqu'en juin 2007 je suis allée aux "Premières rencontres de statistiques spatiales de Niolon" sur l'invitation de Pascal Monestiez. C'est André Kretzschmar, par la suite mon directeur de thèse, qui m'a proposé ce projet, un soir, autour d'une bière, dans l'unique bar de Niolon.

Une des raisons pour lesquelles j'ai choisi de faire cette thèse a été ma curiosité de découvrir enfin le laboratoire BioSP d'Avignon, dont plusieurs sources m'avaient vanté l'ambiance exceptionnelle. Ce fut un privilège de travailler dans un cadre aussi chaleureux et propice au travail en équipe que l'a été le laboratoire BioSP de l'INRA d'Avignon. Je remercie tous les membres de cette unité pour leur accueil et je tiens à remercier certains de mes collègues tout particulièrement pour leur aide précieuse au cours de mon séjour.

J'ai eu la chance d'avoir, non pas un, mais trois directeurs de thèse, qui se sont tous impliqués et m'ont apporté des points de vue différents, des aides et connaissances complémentaires. Je les remercie pour le temps et l'attention qu'ils m'ont consacré durant ces trois années.

André sans qui il n'y aurait rien eu : ni la thèse, ni le projet, ni le financement, et puis surtout sans qui je n'aurais probablement jamais emprunté ce chemin. Merci d'avoir veillé à ce que tout se passe bien, de m'avoir poussée à me fixer des objectifs et à m'y tenir et de m'avoir encouragée et soutenue dans les moments difficiles. Je garde d'excellents souvenirs de nos deux voyages en Corse à la recherche de pucerons, même si les quelques rares colonies que nous avons déniché n'ont pas suffi pour alimenter le modèle spatio-temporel ambitieux vers lequel on tendait.

Merci à Rachid Senoussi pour sa patience face à mes lacunes théoriques, qu'il a réussi à combler en partie, ses vérifications pointilleuses de mes calculs et ses relectures attentives. Il a été mon appui et référent théorique face aux calculs et démonstrations.

Je remercie Samuel Soubeyrand, qui, bien que sans rapport direct avec ma thèse au départ, est rapidement devenu mon troisième directeur de thèse. Son implication a été telle qu'il a été à l'origine des grandes lignes directrices de ma thèse. Son imagination, sa compétence et son pragmatisme m'ont été d'une aide inestimable.

Je tiens à leur dire à quel point j'ai apprécié leur gentillesse et leur sens de l'humour, que j'ai éprouvé plus d'une fois.[1]

[1] notamment en leur envoyant par mail, un mois avant la date prévue pour la soumission de mon manuscrit de thèse, ma soi-disant décision d'arrêter la science pour me consacrer à des activités artistiques. L'effet escompté s'est produit et j'ai ainsi réussi à fixer une réunion en un temps record.

Un grand merci à tous mes autres collaborateurs et en particulier à Nicolas Desassis, qui a rendu possible toute la deuxième partie de ma thèse grâce à ses compétences algorithmiques et son enthousiasme à toute épreuve, et Anna-Liisa Laine, qui m'a donné la possibilité de me pencher sur d'autres jeux de données que les données de pucerons Corses. L'application biologique qui en a découlé a sans aucun doute joué un rôle décisif dans la publication de mon premier article.

Je remercie également les membres de mon comité de pilotage pour leur suivi et leur contribution au bon déroulement de ma thèse : Radu Stoica (Université Lille 1), Jean-Noël Bacro (Université Montpellier 2), et tout spécialement Dominique Agostini (Présidente du Centre INRA de Corse) pour son accueil en Corse et son aide dans la mise en place de notre travail de collecte de données.

Je remercie mon jury de thèse qui m'a fait l'honneur de venir (malgré quelques émotions) un vendredi 17 décembre avant les vacances de Noël.

Cette thèse a été financée par l'INRA et la région PACA, mais ce financement a été rendu possible grâce à la participation du GRCETA de Basse Durance. Je remercie en particulier Pascal Borioli pour sa bienveillance et sa compréhension lorsque les circonstances (certaines lacunes dans les données pucerons que nous n'avons pas réussi à combler grâce à de nouvelles données) ont éloigné ma thèse de la problématique initiale des pucerons pour lui faire prendre une tournure plus méthodologique et générale.

La thèse a aussi été l'occasion de faire des rencontres formidables, je pense notamment au groupe de Rochebrune (sensiblement le même que celui de Niolon), ainsi qu'au groupe de Model-Based Clustering. Bien plus que les congrès, ces groupes de travail d'une semaine m'ont permis de connaître des chercheurs français et étrangers dans une bonne ambiance et un esprit de collaboration.[2]

J'ai également une pensée pour Gilles Caraux, mon professeur de statistiques en deuxième année à l'Agro Montpellier, qui a éveillé mon intérêt pour les statistiques.

Je suis très reconnaissante à tous mes collègues et amis du labo BioSP d'avoir été là :

Flo, mon amie de longue date, toujours là pour discuter et aider, dont les conseils avisés m'ont permis d'avancer et de voir plus loin dans ma thèse,

Julien, dont l'insouciance contagieuse a eu raison de mes doutes et inquiétudes du début,

Emily, qui m'a encouragée et aidée tout au long de la rédaction finale malgré mon humeur morose,

Jimmy, de bonne humeur en toutes circonstances,

Lionel, toujours cynique et toujours prêt à rire de tout,

Etienne, pour sa gentillesse, ses relectures et ses conseils,

Pascal, dont j'ai toujours apprécié les discussions malgré son point de vue rarement optimiste,

Joel, grand[3] esprit critique, qui a contribué à mettre ma thèse sur les rails,

Denis, pour ses coups de pouce qui m'ont ouvert des nouvelles opportunités,

Sylvie, qui a toujours veillé à ce que tout se passe bien dans le labo,

Hervé, mon soutien informatique et Linux (c'est les meilleurs !),

Franck, pour son soutien Matlab et sa bonne humeur,

Marie-Odile, qui, sans pouvoir être là à mon pot de thèse, a pensé à préparer quelque chose.

Merci encore à Emily et André pour mes pneus neufs en période de crise.

[2] ainsi que d'améliorer mon niveau de ski

[3] je dirais même *excellent*

Ces trois années ont été marquées par des moments inoubliables, concours de pétanque, nombreuses soirées entre amis chez André et Marianne, Flo et Gwen, Julien et Manon, Samuel, Etienne, organisations secrètes avec Julie et Marianne ...

Et enfin je remercie mes amis et ma famille pour leur soutien. Je pense à Virginie et Pascal de Montpellier ; à Cécile et Kawthar, mes amies depuis mon arrivée en France, qui ont toujours été là pour les événements importants comme le jour de ma soutenance, et à Mélanie, ma première amie d'Avignon qui n'était pas à l'INRA.

Merci à mes parents et ma soeur pour leur soutien, mon père pour m'avoir sensibilisé aux mathématiques et ma mère qui m'a toujours encouragée dans la voie de la recherche, et qui m'en a toujours cru capable.

Et enfin mon Benjamin, qui a réussi à transformer les derniers mois de ma thèse en une période heureuse, malgré la fatigue et le travail.

Avant-propos

Dans cette thèse, nous présentons des contributions méthodologiques à :

1. l'étude de données multivariées spatiales en écologie,
2. la classification automatique de données multivariées basée sur les modèles de mélange de lois.

Nous nous intéressons à la description statistique de la manière dont plusieurs espèces s'assemblent en densité sur un support spatial. Pour cela nous disposons de mesures spatiales ponctuelles d'abondances de plusieurs espèces observées aux mêmes sites. La disposition et l'espacement des sites ne sont pas contraints et des espaces non euclidiens peuvent être considérés (structure arborescente hiérarchique par exemple).

La première étape de cette thèse consiste à définir une notion d'assemblage d'espèces. Nous proposons de définir le terme "assemblage d'espèces" comme un ensemble de sites pour lesquels les combinaison d'abondances d'espèces sont similaires, et d'identifier les assemblages d'espèces dans un jeu de données grâce à une méthode de classification automatique basée sur des modèles de mélange de lois multivariées. Dans cette démarche, un assemblage d'espèces est assimilé à une composante du mélange de lois, donc à une distribution de probabilité multivariée (partie 2 de la thèse).

La diversité des types de données en écologie (e.g. discret, continu, binaire, ordinal) et l'existence de variables de types différents dans un même jeu de données, (liée au protocole expérimental et aux limites des moyens de mesure), rend l'utilisation de distributions multivariées classiques difficile dans certains cas. Ceci nous amène à considérer des classes de distributions qui permettent de s'adapter facilement aux types des données et de prendre en compte des données corrélées de types différents. Nous définissons ces classes de distributions par des modèles hiérarchiques (partie 3).

Après avoir défini ces distributions multivariées génériques pour décrire les assemblages d'espèces, nous proposons d'étendre les méthodes de classification basées sur des modèles de mélange existantes à des données de types différents, en utilisant des mélanges des distributions génériques que nous avons défini (partie 4).

Cette thèse est organisée en 5 parties et divisée en 9 chapitres.

La première partie est une introduction générale et inclut les deux premiers chapitres qui introduisent le contexte écologique et statistique dans lequel cette thèse s'encadre. Le premier chapitre présente quelques problématiques spécifiques à l'écologie spatiale et les méthodes statistiques qui ont été proposées pour les résoudre, le deuxième chapitre introduit la classification basée sur les modèles de mélange.

La deuxième partie propose une méthode d'analyse statistique de données écologiques multivariées spatiales basée sur une classification multivariée à l'aide de mélanges de lois gaussiennes suivie d'une étape d'analyse spatiale des classes créées.

Les troisième et quatrième parties définissent une nouvelle classe de distributions qui permettent d'étendre le champ d'application de la méthode proposée en partie II à des données de types différents.

Enfin, la cinquième et dernière partie est une discussion générale qui résume les idées et contributions scientifiques de cette thèse et aborde plusieurs perspectives ouvertes par ce travail.

Les chapitres 3 et 6 sont des chapitres introductifs des parties II et III respectivement. Ils complètent l'introduction générale en présentant de manière plus approfondie et technique les problèmes abordés dans chacune de ces parties ainsi que les perspectives et limites des méthodes proposées.

Les chapitres 4 et 7 se présentent sous forme d'articles en anglais, dont le premier a déjà été publié et le deuxième est soumis à publication :

Georgescu, V., Soubeyrand, S., Kretzschmar, A. & Laine, A.-L. (2009). Exploring spatial and multitype species assemblages. *Biometrical Journal* **51** (6) 979–995.

Georgescu, V., Desassis, N., Soubeyrand, S., Kretzschmar, A. & Senoussi, R. A hierarchical model for multivariate data of different types and maximum likelihood estimation. (soumis à *Statistics and Computing*).

Table des matières abrégée

partie I Contexte écologique et statistique

1 Contexte écologique, questions et approche 3

2 Introduction à la classification basée sur les modèles de mélange 19

partie II Explorer des données spatiales multivariées d'abondances d'espèces

3 Etudier les assemblages d'abondances d'espèces pour explorer les interactions interspécifiques 31

4 Exploring Spatial and Multitype Assemblages of Species Abundances . . 37

5 Une extension de CASA aux données de comptages : Utilisation de mélanges de lois de Poisson multivariées 65

partie III Un modèle hiérarchique pour données multivariées de types différents

6 Problématique : un cadre de modélisation et d'estimation général pour des types de données variés 71

7 A Hierarchical Model for Multivariate Data of Different Types and Maximum Likelihood Estimation 81

partie IV Utiliser des lois à structure hiérarchique pour classer des données multivariées de types différents

8 Classer des données de types différents en utilisant des modèles hiérarchiques multivariés 105

partie V Conclusion et perspectives

9 Les assemblages : un outil d'étude des interactions entre espèces 133

Références 143

Table des matières

partie I Contexte écologique et statistique

1 Contexte écologique, questions et approche . . . 3
1.1 Introduction à l'écologie spatiale . . . 3
1.2 Méthodes statistiques pour l'analyse de données spatiales en écologie . . . 6
1.2.1 Les approches exploratoires . . . 7
1.2.2 Les approches basées sur des modèles mécanistes spatiaux . . . 9
1.3 Objectif et démarche de la thèse . . . 10
1.4 Jeux de données . . . 12
1.4.1 Répartition de trois espèces de pucerons sur des rameaux de clémentinier . . . 12
1.4.2 Distribution spatiale d'une plante-hôte, le plantain lancéolé, et de son pathogène, l'oïdium . . . 14
1.4.3 Assemblages de couleurs d'une image aérienne . . . 16
1.4.4 Suivi de l'activité de colonies d'abeilles domestiques (*Apis mellifera*) . . . 17

2 Introduction à la classification basée sur les modèles de mélange . . . 19
2.1 Classification basée sur des modèles de mélange de lois . . . 19
2.1.1 Définition des mélanges finis de lois . . . 20
2.1.2 Identifiabilité des mélanges finis de lois . . . 22
2.2 Estimation des modèles de mélange . . . 23
2.2.1 Approche par maximum de vraisemblance : L'algorithme EM . . . 23
2.2.2 Approche bayésienne . . . 26

partie II Explorer des données spatiales multivariées d'abondances d'espèces

3 Etudier les assemblages d'abondances d'espèces pour explorer les interactions interspécifiques . . . 31
3.1 Classification des abondances et structure spatiale des assemblages d'espèces . . . 32
3.2 Limites et perspectives . . . 33
3.2.1 Limites de l'étape de classification en assemblages d'espèces . . . 33

3.2.2 Limites des outils d'analyse spatiale présentés 34

4 Exploring Spatial and Multitype Assemblages of Species Abundances . . 37
4.1 Introduction 37
4.2 The CASA Method 41
4.2.1 Stage 1 : Classification into types of species assemblages 41
4.2.2 Stage 2 : Spatial analysis of the types of assemblages 43
4.3 Simulation Study 45
4.3.1 Three modes of migration 45
4.4 Case-study 50
4.4.1 The dataset 50
4.4.2 Results 51
4.5 Discussion 54
4.6 Supplementary Material 57
4.6.1 Simulation Study : A Model with a Hidden Interacting Species 57
4.6.2 Extension 61
4.6.3 Clustering Models in CASA 63

5 Une extension de CASA aux données de comptages : Utilisation de mélanges de lois de Poisson multivariées 65
5.1 Modèle de mélange de lois de Poisson bivariées 65
5.2 Estimation par l'algorithme EM 66
5.3 Application à l'étude d'assemblages d'espèces de pucerons sur des feuilles de clémentinier 66

partie III Un modèle hiérarchique pour données multivariées de types différents

6 Problématique : un cadre de modélisation et d'estimation général pour des types de données variés 71
6.1 Une classe de modèles qui s'adaptent aux types des données 71
6.2 Estimation par maximum de vraisemblance et discussion 72
6.3 Limites et extensions 75
6.3.1 Tests d'adéquation 75
6.3.2 Décrire la dépendance entre variables de types différents 75
6.3.3 Extensions du modèle non explorées dans cette thèse 76

7 A Hierarchical Model for Multivariate Data of Different Types and Maximum Likelihood Estimation 81
7.1 Introduction 81
7.2 Multivariate hierarchical model 83
7.2.1 Definition of the general model 83

7.2.2 Exponential families . . . 84
7.2.3 Submodel examples . . . 85
7.3 Maximum likelihood estimation via the MCEM algorithm . . . 87
7.3.1 Expectation step : . . . 88
7.3.2 Maximization step . . . 90
7.3.3 Stopping rule . . . 90
7.4 Simulation studies . . . 91
7.4.1 BPLN model . . . 91
7.4.2 Binomial-Poisson model . . . 93
7.5 An application to beehive data . . . 94
7.5.1 Beehive dataset . . . 94
7.5.2 Results . . . 95
7.6 Discussion . . . 97
7.7 Appendices . . . 98
7.7.1 Unconditional moments of the bivariate Binomial-Poisson model . . . 98
7.7.2 PQL estimators of the conditional moments for different distributions and link functions . . . 100
7.7.3 Maximum likelihood estimators of the multivariate normal parameters in the M-step of the MCEM . . . 101

partie IV Utiliser des lois à structure hiérarchique pour classer des données multivariées de types différents

8 Classer des données de types différents en utilisant des modèles hiérarchiques multivariés . . . 105
8.1 Introduction au modèle de mélange gaussien latent . . . 105
8.2 Modèle général . . . 107
8.3 Exemples de sous-modèles . . . 107
8.3.1 Modèle pour données de comptage : Mélange de MPLN . . . 107
8.3.2 Modèles pour données mixtes continues et discrètes : mélange Normal-Poisson, mélange Normal-Binomial . . . 108
8.4 Estimation par maximum de vraisemblance avec l'algorithme MCEM . . . 109
8.4.1 Etape E . . . 110
8.4.2 Etape M . . . 111
8.4.3 Règle d'arrêt . . . 113
8.4.4 Initialisation . . . 114
8.4.5 Algorithme de classification automatique . . . 114
8.5 Etude de trois sous-modèles par simulation . . . 115
8.5.1 Estimation à nombre de classes fixé pour un modèle de mélange Normal-Poisson à deux classes . . . 116

8.5.2 Estimation et choix du nombre de classes : modèle de mélange Normal-Poisson . . . 117
8.5.3 Estimation à nombre de classes fixé sur des composantes fortement non gaussiennes : modèle de mélange Normal-Binomial . . . 120
8.5.4 Lorsque les deux variables gaussiennes sont latentes : modèles de mélange BPLN . . . 122
8.6 Discussion - Perspectives - Extensions . . . 124
8.6.1 Perspectives . . . 125
8.6.2 Extensions . . . 127

partie V Conclusion et perspectives

9 Les assemblages : un outil d'étude des interactions entre espèces . . . 133
9.1 CASA : une approche de classification pour analyser la structure spatiale des assemblages d'espèces . . . 133
9.2 Extensions de CASA en vue de perspectives d'application . . . 135
9.2.1 Extension de l'étape de classification de CASA . . . 136
9.2.2 Extensions de l'étape d'analyse spatiale de CASA . . . 138
9.3 Vers une dynamique des assemblages . . . 139
9.3.1 CASA temporel : suivi de données temporelles . . . 139
9.3.2 Simulation de la dynamique d'espèces en interaction : comprendre la formation des assemblages par une approche mécaniste . . . 140

Références . . . 143

Première partie

Contexte écologique et statistique

1

Contexte écologique, questions et approche

Dans ce chapitre nous commençons par une courte introduction à quelques problèmatiques générales en écologie spatiale et aux méthodes statistiques qui ont été proposées pour les traiter. Ceci permet de placer le travail méthodologique effectué dans cette thèse dans un cadre d'application qui a guidé notre démarche. Nous présentons ensuite notre approche d'étude de la coexistence d'espèces dans un cadre spatial, en définissant le concept d'assemblage d'espèces. Enfin nous présentons les jeux de données qui ont guidé et illustré ce travail.

1.1 Introduction à l'écologie spatiale

L'importance de la prise en compte de l'espace sur les résultats et les théories écologiques est discutée dans Tilman & Kareiva, eds. (1997).

L'espace joue un rôle prépondérant dans la plupart des problèmes traditionnellement étudiés en écologie tels que la compétition pour une ressource (Lehman & Tilman, 1997), la dynamique proie-prédateur et la dynamique hôte-parasite, ainsi que la coexistence de nombreuses espèces utilisant un nombre réduit de ressources communes.

En effet, les premières théories écologiques, s'appuyant sur des modèles de dynamique des populations qui ne tenaient pas compte de l'espace (et supposaient ainsi implicitement que les populations étaient bien mélangées dans un milieu homogène), ne permettaient pas de comprendre comment autant d'espèces peuvent coexister dans la nature. Ainsi, le principe d'exclusion compétitive et les modèles de Lotka et Volterra pour la prédation, sont confrontés au paradoxe de la diversité de Hutchinson (1961), qui a remarqué que dans les eaux des lacs et des océans une centaine ou plus d'espèces de phytoplancton peuvent coexister, toutes en compétition pour les mêmes ressources limitées (Lehman & Tilman, 1997). Or la théorie prédisait que le nombre d'espèces en compétition ne peut dépasser le nombre de ressources. Introduire l'espace explicitement permet de résoudre en partie ce paradoxe, car les interactions entre voisins et la dispersion locale induisent une agrégation des espèces et une ségrégation spatiale. L'intensité de cet effet dépend de l'intensité de la compétition et des échelles des phénomènes de dispersion et de compétition entre voisins.

Donc certaines coexistences d'espèces sont possibles uniquement dans un contexte spatial où l'habitat est subdivisé. Ce résultat a été suggéré en premier par Gause (1935), qui a constaté que la mise en présence en laboratoire de l'espèce *Paramecium* et de son prédateur *Didinium* conduisait invariablement à l'extinction d'une ou des deux espèces, et a émis l'hypothèse qu'une recolonisation périodique par l'une ou l'autre espèce pourrait expliquer leur coexistence dans la nature.

Ces travaux témoignent de l'importance de la prise en compte de l'hétérogénéité spatiale pour comprendre les processus écologiques naturels.

Distribution spatiale d'une espèce et associations spatiales entre espèces

En écologie, on étudie la distribution spatiale d'une espèce et le lien entre les distributions spatiales de deux ou plusieurs espèces (appelée parfois "association spatiale" entre espèces) afin d'inférer les processus écologiques sousjacents. Les processus écologiques désignent les mécanismes par lesquels un organisme, une espèce ou une communauté écologique interagit avec son environnement biotique et abiotique.

La distribution spatiale des organismes résulte d'une combinaison de processus écologiques endogènes, qui sont liés aux interactions intraspécifiques et à la dynamique de la population, et exogènes, qui sont liés aux interactions interspécifiques et à l'influence de l'environnement (Tilman & Kareiva, eds., 1997). Les processus endogènes qui structurent les espèces sont par exemple la compétition intraspécifique et la densité-dépendance. Les processus exogènes peuvent être abiotiques (climat) ou biotiques (espèces compétitrices, prédation).

L'étude des distributions spatiales et des associations d'espèces se font généralement grâce à des méthodes statistiques qui prennent en compte explicitement la nature corrélée des données spatiales (voir 1.2).

La dépendance des processus à l'échelle spatiale

Comprendre comment la distribution d'une espèce (occupation et autocorrélation spatiale) et les associations de plusieurs espèces (co-distribution de plusieurs espèces) changent lorsqu'on change d'échelle est fondamental pour comprendre la formation des patterns en écologie (Hui, 2009).

Lorsqu'on parle d'échelle spatiale dans un cadre écologique il faut distinguer plusieurs notions : l'échelle du processus écologique d'intérêt, l'échelle d'observation et l'échelle de l'analyse spatiale des données (Bellier, 2007).

Chaque processus structurant la distribution des organismes s'exprime à une échelle particulière et changer d'échelle implique changer de processus. Ainsi, Hui (2009) montre dans une étude théorique que l'intensité de la corrélation spatiale et l'intensité de l'association spatiale entre espèces diminuent lorsqu'on augmente l'échelle d'analyse (ou la taille de l'unité d'échantillonnage considérée).

D'autres auteurs ont constaté la dépendance de la structure spatiale (ou *pattern*) à l'échelle d'étude dans des systèmes naturels. En effet, les études de systèmes proies-prédateurs dans le milieu marin de Rose & Legget (1981) et Fauchald et al. (2000) montrent

que le pattern des corrélations spatiales entre proies et prédateurs varie selon l'échelle spatiale considérée : ces corrélations sont négatives à petite échelle, positives à plus large échelle, et presque nulles à très grande échelle.

L'échantillonnage dans un cadre spatial

L'acquisition de données pour l'étude de la distribution spatiale d'une ou plusieurs espèces passe par le choix d'un plan d'échantillonnage, qui doit être adapté au processus écologique étudié (et notamment à l'échelle de ce processus). Un plan d'échantillonnage peut être caractérisé par trois parmètres (Perry et al., 2002) :

- l'étendue de la zone d'étude,
- l'unité d'échantillonnage (ou résolution),
- la distance entre unités d'échantillonnage.

La modification de ces trois paramètres peut influencer les inférences de l'analyse spatiale, donc leur choix est essentiel pour les écologues. En effet si la zone d'étude est trop petite le processus sera difficile à identifier, et si elle est trop grande, alors elle pourra inclure plusieurs processus écologiques agissant différemment dans des sous-régions, ce qui rendra plus difficile le traitement des données par des méthodes de statistiques spatiales classiques. En effet, la plupart des méthodes supposent l'homogénéité en terme de processus écologique au sein de la zone d'étude, qui se traduit par la notion de stationnarité dans les modèles (Bellier, 2007, p. 21).

On peut à partir de ces trois caractéristiques distinguer plusieurs types d'échantillonnages selon le découpage de la zone d'étude :

- L'unité d'échantillonnage est une sous-région de la zone d'étude et l'ensemble des unités d'échantillonnage forment une partition complète de la zone d'étude. Autrement dit, la zone d'étude est subdivisée en cellules contiguës, par exemple une grille régulière, l'unité d'échantillonnage est la cellule et la distance entre unités d'échantillonnage est nulle. Dans ce cas on parle d'échantillonnage exhaustif de la zone d'étude. On dit parfois que l'espace est discrétisé.
- L'unité d'échantillonnage est le quadrat, c'est à dire des sous-régions qui ne forment pas une partition de la zone d'étude. Les données sont récoltées par quadrats lorsque l'échantillonnage exhaustif n'est pas possible. Souvent les quadrats sont rectangulaires, de même surface, et placés soit au hasard, soit pour former un arrangement prédéfini. Donc la distance entre unités d'échantillonnage peut varier.
- L'unité d'échantillonnage est un point dans l'espace (échantillonnage ponctuel).

Enfin on peut distinguer plusieurs types de données spatiales selon le type d'information disponible. Perry et al. (2002) proposent de distinguer les données ponctuelles (fréquentes en écologie des plantes), les données régionales associées à une partition de la zone d'étude en sous-régions (fréquentes en écologie du paysage et en géographie), et les données spatialement implicites (telles la distance au plus proche voisin pour chaque unité d'échantillonnage,

Diggle, 2003). Ils distinguent les types de données ponctuelles selon la quantité d'information disponible :

- $(\boldsymbol{x})$: échantillonnage exhaustif de la position des individus le long d'un transect (1 coordonnée),
- $(\boldsymbol{x}, \boldsymbol{y})$: échantillonnage exhaustif de la position des individus dans un espace à deux dimensions (2 coordonnées),
- $(\boldsymbol{x}, \boldsymbol{z})$ ou $(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z})$: la position de l'individu (à 1 ou 2 coordonnées) est récoltée ainsi qu'un attribut $\boldsymbol{z}$ (variable qualitative ou quantitative).

Les données récoltées par quadrats peuvent être transformées en données ponctuelles de type $(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z})$, où les coordonnées $(\boldsymbol{x}, \boldsymbol{y})$ sont souvent définies comme le barycentre des quadrats, et les valeurs $\boldsymbol{z}$ représentent par exemple le nombre d'individus dans le quadrat.

Dans cette thèse nous nous sommes intéressés à des données avec plusieurs attributs, c'est-à-dire qu'à chaque site (ou unité d'échantillonnage) on observe un vecteur multivarié $\boldsymbol{z}$, qui contient dans notre cas les abondances de plusieurs espèces.

Nous allons présenter maintenant quelques méthodes statistiques pour traiter les données spatiales en écologie, qui sont souvent spécifiques à un type d'échantillonnage.

1.2 Méthodes statistiques pour l'analyse de données spatiales en écologie

De nombreuses méthodes statistiques ont été développées pour analyser la distribution spatiale des données écologiques ou pour tenter d'inférer les processus écologiques sousjacents. On peut distinguer les méthodes selon le degré d'information qu'elles fournissent. Certaines méthodes sont uniquement descriptives, d'autres permettent de tester des hypothèses nulles (parfois plus complexes que l'hypothèse de distribution aléatoire ou *complete spatial randomness* - CSR), et enfin certaines permettent d'ajuster des modèles spatiaux. Beaucoup de méthodes sont empiriques et ne s'appuient pas sur un modèle, qui ne peut être créé avant d'avoir récolté des données suffisantes sur les espèces étudiées, leur cycle de vie, leur dynamique, leur démographie et leurs interactions. Ces méthodes dépendent du type d'échantillonnage des données et la plupart sont utilisées pour des études exploratoires.

On peut distinguer les méthodes globales et les méthodes locales. Les méthodes globales résument l'information spatiale sur toute la zone d'étude (par exemple sous forme d'un indice d'agrégation), alors que les méthodes locales proposent de quantifier la variation locale (par exemple sous formes d'indices locaux). Les méthodes locales peuvent être très utiles car elles peuvent mettre en évidence une hétérogénéité spatiale de la zone d'étude (le processus écologique sousjacent varie dans différentes sous-régions de la zone d'étude). Or beaucoup de méthodes globales sont basées sur une hypothèse de stationnarité du processus, donc d'homogénéité spatiale.

Enfin un des objectifs principaux des statistiques spatiales est de mesurer la dépendance spatiale. Il y a plusieurs manières de tenir compte de la dépendance spatiale selon le type d'échantillonnage et de découpage de l'espace. Par exemple sur une grille régulière on peut décrire la dépendance par des matrices de voisinage à partir de schémas de connexion

entre voisins. Par exemple on peut définir un voisinage à 1 niveau par les (4 ou 8) cellules adjacentes. Lorsque les unités d'échantillonnage sont espacées de manière irrégulière, on peut utiliser des schémas de connection tels que la triangulation de Delaunay. On peut définir des schémas de connexion de niveau supérieur et on peut également utiliser des pondérations différentes pour différents niveaux de voisins. Dans le cas de données ponctuelles on utilise généralement les distances euclidiennes entre points.

Nous présentons brièvement quelques approches exploratoires puis les approches basées sur des modèles, selon que l'on s'intéresse à la distribution spatiale d'une espèce ou au lien spatial entre plusieurs espèces.

1.2.1 Les approches exploratoires

Étudier la distribution spatiale d'une espèce

Souvent la première question qui se pose lors de l'étude de la distribution spatiale d'une espèce (ou pattern) est de savoir s'il y a structuration spatiale par rapport à l'hypothèse nulle de distribution spatiale aléatoire (CSR). On caractérise ensuite le pattern spatial non aléatoire suivant qu'il est agrégé (interactions positives ou d'attraction entre individus) ou régulier (intéractions négatives ou de répulsion entre individus).

Données ponctuelles $(\boldsymbol{x}, \boldsymbol{y})$

Pour ce type de données la position de chaque individu est connue et les méthodes existantes s'appuient sur la théorie des processus ponctuels (Diggle, 2003). Dans ce cas on teste si la distribution est significativement différente d'un processus ponctuel de Poisson (hypothèse nulle de CSR). S'il est significativement différent on dit que le pattern est structuré spatialement et on le définit comme agrégé ou régulier. Les indices $K(d)$ et $L(d)$ de Ripley sont basés sur le calcul du nombre moyen d'individus qui se trouvent à une distance d d'un individu choisi au hasard. On calcule ensuite l'évolution de $L(d)$ en fonction de la distance d. La comparaison à des graphiques calculés pour des pattern aléatoires générés par simulation de Monte Carlo permet de tester la structure spatiale du pattern.

Données ponctuelles $(\boldsymbol{x}, \boldsymbol{z})$, $(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z})$

Lorsque l'attribut $\boldsymbol{z}$ correspond à des comptages d'individus dans des quadrats, on peut utiliser des méthodes de ratio variance/moyenne, basées sur le fait que des données provenant d'un processus de Poisson (hypothèse CSR) devraient avoir un ratio de 1 (Dale et al, 2002). D'autres méthodes sont prévues pour des échantillonnages exhaustifs de l'espace qui est divisé en cellules contiguës, telles les méthodes de *quadrat variance* en une dimension (le long d'un transect) et *block quadrat variance* en deux dimensions. Dans ces méthodes, on calcule la variance en se basant sur des couples d'unités d'échantillonnage adjacents au lieu de la totalité comme dans les méthodes de ratio variance/moyenne.

Enfin beaucoup de méthodes s'intéressent à quantifier l'autocorrélation spatiale des données en fonction de la distance. On parle de données autocorrélées ou de dépendance

spatiale lorsque des unités d'échantillonnage proches ont des valeurs plus similaires que des unités éloignées. Le principe de ces méthodes, apparues dans le domaine des géostatistiques (Chilès & Delfiner, 1999), est proche des méthodes de *quadrat variance* en remplaçant les blocs de quadrats par les unités d'échantillonnage (sites). La technique la plus utilisée est basée sur l'étude du variogramme expérimental (de l'échantillon), obtenu en calculant la variance moyenne des attributs $\boldsymbol{z}$ pour tous les couples de sites distants de d, pour une gamme de distances d. Le plus souvent on découpe la distance maximale entre sites en intervalles et à chaque intervalle de distance on associe une variance moyenne.

L'idée générale de ces méthodes est que l'autocorrelation diminue (et donc la variance augmente) avec la distance jusqu'à un maximum est atteint au bout d'une distance d qu'on appelle en géostatistiques la portée. La portée peut être vue comme l'estimation de la taille moyenne des aggrégats et des "trous". D'autres mesures d'autocorrélation utilisées en écologie sont l'indice I de Moran et l'indice c de Geary qui permettent de tester l'hypothèse de CSR. Ces méthodes sont des méthodes globales qui supposent l'homogénéité spatiale du processus.

Des méthodes locales ont également été proposées. Anselin (1995) a adapté les indices de Moran et de Geary pour obtenir un indice local qui permet d'analyser comment l'autocorrélation spatiale varie dans la zone d'étude. Enfin, Perry et al (1999) proposent la méthode SADIE (Spatial Analysis by Distance InDicEs) qui permet de détecter des agrégats en attribuant à chaque site un indice d'agrégation et en proposant un test de l'hypothèse nulle de CSR.

Étudier les distributions de plusieurs espèces et leurs relations

L'étude de la dépendance entre deux ou plusieurs jeux de données spatiaux récoltés sur la même zone d'étude (par exemple les distributions spatiales de plusieurs espèces) est souvent appelée étude de l'association spatiale entre les jeux de données (ou entre espèces).

Remarquons qu'on a plusieurs manières d'obtenir des distributions de plusieurs espèces à partir de la définition des données ponctuelles de type $(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z})$. Soit l'attribut $\boldsymbol{z}$ est catégoriel (cas 1), par exemple à 2 catégories qui désignent respectivement l'espèce 1 et l'espèce 2, et alors on est ramenées à des méthodes de processus ponctuels marqués (Diggle, 2003). Soit l'attribut $\boldsymbol{z}$ est un vecteur multivarié contenant l'abondance de plusieurs espèces mesurées aux mêmes sites (cas 2).

Cas 1 : Processus ponctuels marqués

Bar-Hen & Picard (2006) ont fait une synthèse sur les principaux indices globaux d'association pour définir la relation entre deux nuages de points (patterns). Ils proposent de distinguer 3 grandes approches :

– l'approche géométrique, qui quantifie la proportion de mélange entre deux nuages (mélange homogène ou pas, mélange partiel-total),

– l'approche de *random labelling*, qui quantifie la dissimilarité entre les répartitions observées des deux nuages de points et la répartition qui serait obtenue si on permutait les labels des processus au hasard sans changer les positions des points, (c'est-à-dire qui consiste à regarder s'il y a un lien entre la répartition des points et le label),
– le test d'indépendance entre distributions marginales des deux espèces.

Une approche locale pour les processus ponctuels marqués a été proposée par Allard, Brix & Chadœuf (2001), qui proposent un test local d'indépendance entre deux nuages de points.

Cas 2 : Données ponctuelles $(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z})$ *lorsque* $\boldsymbol{z}$ *est un vecteur multivarié d'attributs*

Il y a très peu d'indices d'association dans la littérature pour des données d'abondances de plusieurs espèces observées aux mêmes sites. Dans le cadre de données de comptages de deux espèces observées aux mêmes sites, Perry & Dixon (2002) ont proposé un indice global d'association spatiale, qui peut également être calculé localement pour chaque site. Leur méthode est basée sur une comparaison des indices d'agrégation obtenus pour chaque espèce par la méthode SADIE (Perry et al, 1999).

Une approche différente a été proposée par Karunanayake & Laverty (2006) dans le cadre de l'étude des distributions spatiales de données de comptages de trois espèces. Ils proposent d'identifier des zones d'association entre données de comptages homogènes dans l'espace par des méthodes de classification spatiale. Plus précisément ils proposent d'effectuer une classification basée sur un modèle de mélange de lois de Poisson multivariées.

Dans cette thèse, nous nous sommes intéressés aux approches exploratoires locales pour décrire et analyser l'association spatiale de données écologiques qui rentrent dans le cas 2 présenté ici.

1.2.2 Les approches basées sur des modèles mécanistes spatiaux

Les modèles mécanistes constituent un autre type d'approche souvent utilisé par les écologues pour tenter de comprendre la dynamique d'une espèce, les interactions intraspécifiques et interspécifiques. Comme nous l'avons déjà mentionné, les modèles nécessitent (et incorporent) des connaissances approfondies sur la biologie des espèces (par exemple sur leur durée de vie, leur déplacement, leur dynamique, leurs interactions).

Les approches les plus courantes sont déterministes, mais des approches stochastiques, telles que les processus de branchements (Haccou, Jagers & Vatutin, 2007; Révész, 1994) peuvent également être utilisées. On peut distinguer différents types d'outils de modélisation :

– les équations de réaction-diffusion prenant en compte l'espace de manière continue (Durett & Levin, 1994),
– les automates cellulaires pour lesquels l'espace est divisé en cases discrètes contiguës (Durett & Levin, 1994), ou de manière plus générale les systèmes de particules en interaction (*interacting particle systems*, voir par exemple Révész, 1994),
– les modèles en métapopulations considérant que l'espace est divisé en *patchs* plus ou moins structurés spatialement (Hanski, 1999),

– des modèles spatialement explicites pour lesquels chaque individu est placé dans un espace continu (Pacala, 1996).

Nous n'avons pas utilisé ce type d'approche dans cette thèse, mais nous discuterons dans le chapitre 9 des perspectives offertes par l'utilisation de modèles mécanistes basées sur les processus de branchements.

1.3 Objectif et démarche de la thèse

Dans cette thèse nous nous sommes intéressés à la manière dont plusieurs espèces s'assemblent en densité sur un support spatial. En particulier, notre objectif a été de caractériser des assemblages d'abondances d'espèces par des méthodes exploratoires, dans le cas de données ponctuelles d'abondances multivariées de type $(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z})$, selon la typologie que nous avons exposée précédemment.

Notion d'assemblage d'abondances d'espèces

La notion d'assemblage d'espèces est utilisée habituellement en écologie des communautés pour désigner la diversité d'espèces observées sur un site donné, c'est-à-dire la richesse spécifique de l'écosystème sur ce site (Cristofoli & Mahy, 2010). La question posée dans ce cas est surtout : quelles espèces coexistent à un endroit ? Dans ces études, les jeux de données contiennent des mesures sur des centaines d'espèces, qui ne sont pas toujours présentes simultanément aux mêmes sites et dont certaines sont même très rares.

Dans cette thèse le terme "assemblage d'espèces" n'est pas utilisé dans ce sens. En effet, nous allons nous intéresser aux assemblages d'abondances d'espèces pour un nombre d'espèces restreint et qui coexistent sur la plupart des sites. Comme nous l'avons vu précédemment, les relations entre les distributions de deux ou plusieurs espèces sont souvent appelées associations spatiales entre espèces dans la littérature (Perry & Dixon, 2002; Hui, 2009), où l'on cherche souvent à quantifier un degré d'association ou de dissociation entre espèces.

Dans cette thèse nous ne nous limitons pas à cette description qui quantifie le degré d'association ou de dissociation, mais nous cherchons à identifier tous les types d'assemblages d'abondances d'espèces d'un jeu de données. La raison écologique principale qui justifie cette démarche est la possibilité d'avoir de l'hétérogénéité spatiale dans la zone d'étude. En effet, nous avons vu que la plupart des méthodes existantes sont basées sur le principe de stationnarité du processus écologique sur la zone d'étude, qui se traduit par une homogénéité spatiale des données.

Dans un cadre de données écologiques ponctuelles $(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z})$, où à chaque site les abondances de plusieurs espèces ont été mesurées, nous définissons **un assemblage d'espèces** (ou un type d'assemblage d'espèces) comme **un ensemble de sites d'observations pour lesquels les combinaisons d'abondances d'espèces observées sont similaires** (ou proches).

La question posée ici est : comment les espèces coexistent à un endroit ? Peut-on définir une typologie des combinaisons d'abondances de ces espèces (que nous appelons assemblages) afin de pouvoir, à terme, émettre des hypothèses sur leur origine (suggérer des facteurs environnementaux et des types d'interaction entre espèces à tester ou à prendre en compte dans des approches de modélisation) ? Notre but est donc de caractériser les associations spatiales entre les espèces, qui sont le reflet des interactions de ces espèces entre elles et avec leur environnement, en passant par cette notion d'assemblage d'espèces que nous venons de définir.

Dans la définition que nous donnons des assemblages d'espèces, il y a l'idée implicite d'une classification des sites basée sur les mesures multivariées d'abondances. Nous proposons dans le chapitre 4 de définir les types d'assemblages par une méthode de classification basée sur des mélanges de lois gaussiennes multivariées.

Nous avons choisi de définir les assemblages d'espèces sans introduire l'autocorrélation spatiale des abondances entre sites voisins et d'étudier la dépendance spatiale des assemblages dans une deuxième étape. Les deux principales raisons écologiques qui justifient ce choix sont l'échelle du phénomène étudié et l'hétérogénéité spatiale du processus.

Premièrement, pour la généricité de la méthode, il est préférable de ne pas introduire d'hypothèses spécifiques à un certain type de données spatiales. En effet nous avons vu que selon le type de données, on utilise soit des matrices de voisinage, soit des matrices de distances. Karunanayake & Laverty (2006) proposent d'introduire un certain type de dépendance spatiale dans la classification de données de comptages de plusieurs espèces observées dans des quadrats disposés selon un plan d'échantillonnage régulier dans l'espace. Ils utilisent pour cela une chaine de Markov de niveau 1, en supposant que deux unités proches ont des valeurs proches. Or, si les sites sont suffisamment éloignés par rapport à l'échelle du processus écologique, il se pourrait qu'il n'y ait plus de corrélation spatiale à cette distance. Inversement, si les sites sont relativement proches par rapport à l'échelle du processus, un modèle à voisinage d'ordre 2 ou plus serait mieux adapté. Et enfin pour des sites espacés de manière irrégulière il faudrait introduire autrement l'information de dépendance dans la classification.

Enfin, le fait de ne pas introduire d'hypothèses spatiales permet de s'affranchir de l'hypothèse de stationnarité du processus, et de mettre en évidence des zones hétérogènes par rapport aux assemblages d'espèces, qui pourront être reliées par exemple à l'hétérogénéité d'un facteur environnemental ou à des états d'équilibres différents des espèces. En effet, selon la théorie écologique qui postule l'existence d'états d'équilibres multiples d'un écosystème, les communautés d'espèces peuvent se trouver à des états d'équilibres différents dans des régions différentes, qui se manifestent par des assemblages d'abondances différents dans ces régions. Cette théorie est étayée par de nombreuses études écologiques, telles celles de Walker et al. (1981); Peterson (1984); Dublin et al. (1990); Messier (1994); Estes & Duggins (1995); Petraitis & Latham (1999); Seabloom & Richards (2003); le Roux & McGeoch (2008).

1.4 Jeux de données

Dans cette section nous présentons brièvement les quatre jeux de données qui ont été utilisés dans cette thèse pour illustrer la notion d'assemblage d'espèces que nous venons de définir et notre démarche méthodologique :

- des comptages de trois espèces de pucerons spatialisés sur un support arborescent,
- des données de suivi temporel annuel de la distribution spatiale à large échelle d'une plante hôte et de son pathogène (dans les îles Åland en Finlande sur les années 2001-2008),
- une image aérienne d'un paysage agricole situé dans l'État de New York aux États-Unis,
- un suivi temporel de l'activité de ruches d'abeilles au cours d'une saison de production (de la dynamique de sa population, d'un parasite infectant les abeilles et de sa production de miel).

1.4.1 Répartition de trois espèces de pucerons sur des rameaux de clémentinier

Ce jeu de données, acquis entre 1996 et 1998 au cours d'un projet de l'INRA de Corse et de Sophia-Antipolis, a constitué le point de départ de cette thèse. Des comptages exhaustifs des espèces de pucerons présentes en fin de saison de végétation sur les feuilles d'environ 300 fragments de rameaux de clémentinier ont été réalisés. Un fragment de rameau est constitué de 10 pousses feuillées consécutives (à nombre de feuilles variables). Il s'agit de données non temporelles, sur un espace à structure hiérarchique (rameau - pousse - feuille) dont on connait la disposition relative et l'ordre des éléments. Les trois espèces présentes dans ces données sont : le puceron du melon *Aphis gossypii*, le puceron vert du citrus *Aphis spiraecola* et le puceron brun du citrus *Toxoptera citricida*. Les trois espèces de pucerons observées cohabitent sous forme d'assemblages qui sont structurés spatialement (voir figure 1.1).

Ce jeu de données sera analysé dans le chapitre 5. Nous nous sommes intéressés principalement aux deux espèces les plus abondantes *Aphis gossypii* (espèce 1) et *Aphis spiraecola* (espèce 2) et nous présenterons des analyses seulement sur les 230 fragments de rameaux occupés par les deux espèces.

Les particularités de ce jeu de données ont contribué à notre choix de définir les assemblages d'espèces par une classification non spatiale à cause de la difficulté d'étudier l'autocorrelation des données dans ce contexte (Kretzschmar, Soubeyrand & Desassis, 2010). En effet, nous disposons de très peu de données spatiales sur un rameau, qui est la zone d'étude la plus grande sur laquelle on pourrait définir une structure de voisinage, et il est difficile de tenir compte des autres fragments de rameaux dans un tel cadre.

Nous allons donc définir des assemblages sur ce jeu de données en supposant que les données de comptages observées sur chaque feuille sont indépendantes. Nous proposons de classer les feuilles selon les vecteurs d'abondances, donc d'effectuer une classification sur les données représentées dans le graphique 1.2, pour définir les assemblages d'espèces. Nous

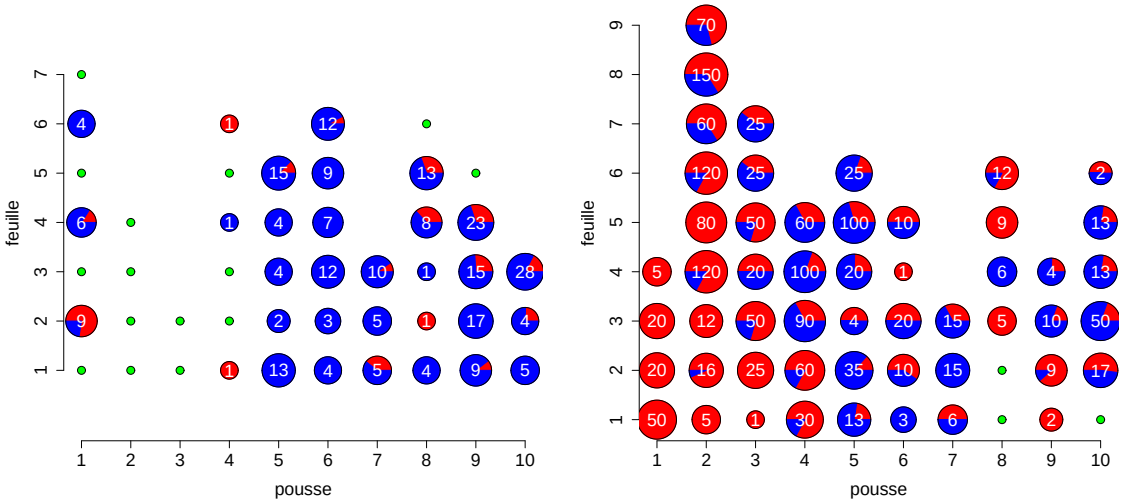


Fig. 1.1. Exemple de deux fragments de rameaux de 10 pousses occupés par les 2 espèces de pucerons les plus fréquentes (*Aphis gossypii* en rouge et *Aphis spiraecola* en bleu, les feuilles sans pucerons en vert). La surface des cercles est proportionnelle au comptage total (marqué en blanc), le comptage respectif de chaque espèce est représenté par la surface des couleurs.

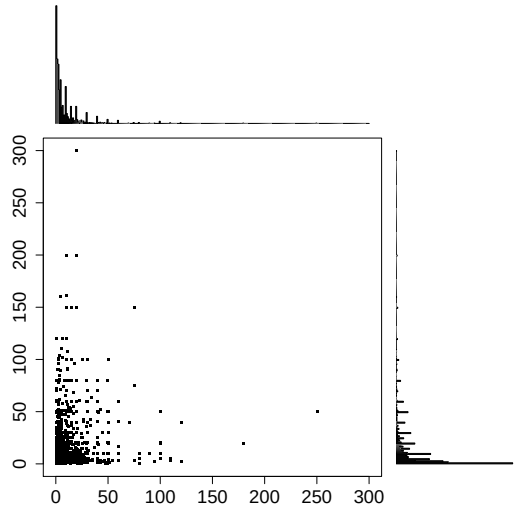


Fig. 1.2. Données de comptages des deux espèces de pucerons *Aphis gossypii* (en abscisse) et *Aphis spiraecola* (en ordonnée) sur 1731 feuilles où les espèces sont en coprésence. Un point du graphique correspond à une feuille. Les histogrammes des comptages marginaux sont représentées sur les côtés du graphique.

pouvons observer sur ce graphique une des caractéristiques principales de ce jeu de données, à savoir la forte surdispersion des abondances marginales des espèces (pour l'espèce 1, en abscisse, une médiane à 4 et un maximum à 250, pour l'espèce 2, en ordonnée, une médiane à 6 et un maximum à 300). De plus, un test du χ^2 d'indépendance (test d'adéquation à une densité jointe simulée à partir des marginales sous l'hypothèse d'indépendance), rejette l'hypothèse d'indépendance des comptages des deux espèces (p-valeur$= 0.01 < 0.05$).

Précisons que l'étude des assemblages des espèces de pucerons de ce jeu de données présente de plus un enjeu agronomique et économique, car les pucerons sont des vecteurs potentiels de virus, le plus important étant le Citrus Tristeza Virus (CTV). Gottwald et al. (1999) ont montré que la composition des assemblages d'espèces de pucerons affecte la distribution spatiale du CTV. Le puceron brun (*Toxoptera citricida*), est le meilleur vecteur de la Tristeza. La transmission du virus par le pucerons brun est jusqu'à 25 fois plus efficace qu'avec le 2$^{\text{ème}}$ vecteur le plus efficace, le puceron du melon *Aphis gossypii* (Yokomi et al., 1994). L'étude de la structuration spatiale des assemblages d'espèces de pucerons peut s'avérer utile pour comprendre et limiter la dispersion du virus.

1.4.2 Distribution spatiale d'une plante-hôte, le plantain lancéolé, et de son pathogène, l'oïdium

Ce jeu de données à grande échelle s'intéresse à la répartition d'un champignon, l'oïdium (*Podosphaera plantaginis*), infectant sa plante hôte sauvage, le plantain lancéolé (*Plantago lanceolata*). La zone d'étude s'étend sur une surface de 50 $\times$ 70 km dans les îles Åland, en Finlande. Des mesures ont été effectuées pendant 7 années consécutives (2001-2008) sur environ 3000 prairies (en général $<$ 1 ha), en fin de saison d'infection (septembre). Pour chaque prairie la surface occupée par le plantain et la présence-absence d'oïdium a été notée (voir figure 1.3).

Des parties de ce jeu de données ont été étudiées à plusieurs reprises par Laine (2004); Laine & Hanski (2006); Soubeyrand et al. (2009), Ces auteurs ont proposé une étude de la dynamique spatio-temporelle du pathogène, formée d'extinctions et de récolonisations fréquentes, soit par une étude de type GLMM (Laine & Hanski, 2006), soit par des modèles mécanistes (Soubeyrand et al., 2009). L'étude de Laine & Hanski (2006) montre que les colonisations sont plus fréquentes à l'est de Åland, ce qui pourrait être relié à la dispersion des spores d'oïdium dans la direction dominante du vent, et que l'occurrence et la persistence du pathogène est plus probable à proximité des côtes, ce qui suggère des facteurs physiques de l'environnement (humidité, température, précipitations, vents).

Nous proposons d'étudier ces données en termes d'assemblages des deux espèces.

L'avantage de la notion d'assemblage d'espèces telle que nous l'avons définie (i.e. sans autocorrélation spatiale) est de pouvoir être directement applicable à ce jeu de données à grande échelle, ainsi qu'au premier jeu de données de comptages d'espèces de pucerons à une échelle bien plus fine (les feuilles). Il serait difficile d'introduire un type de dépendance spatiale entre données dans la méthode de classification basée sur les modèles de mélange, qui nous permet de définir les assemblages, tel qu'il puisse s'adapter aux deux cas.

Une analyse sur une partie de ce jeu de données (années 2001–2006), basée sur une définition des assemblages d'espèces par une classification à l'aide de modèles de mélange de lois gaussiennes bivariées, est présentée dans le chaptire 4. Pour obtenir des données d'abondances d'espèces pseudo-continues qui soient plus adaptées à l'hypothèse gaussienne, nous proposons d'agréger le jeu de données dans l'espace (on utilise une grille régulière avec

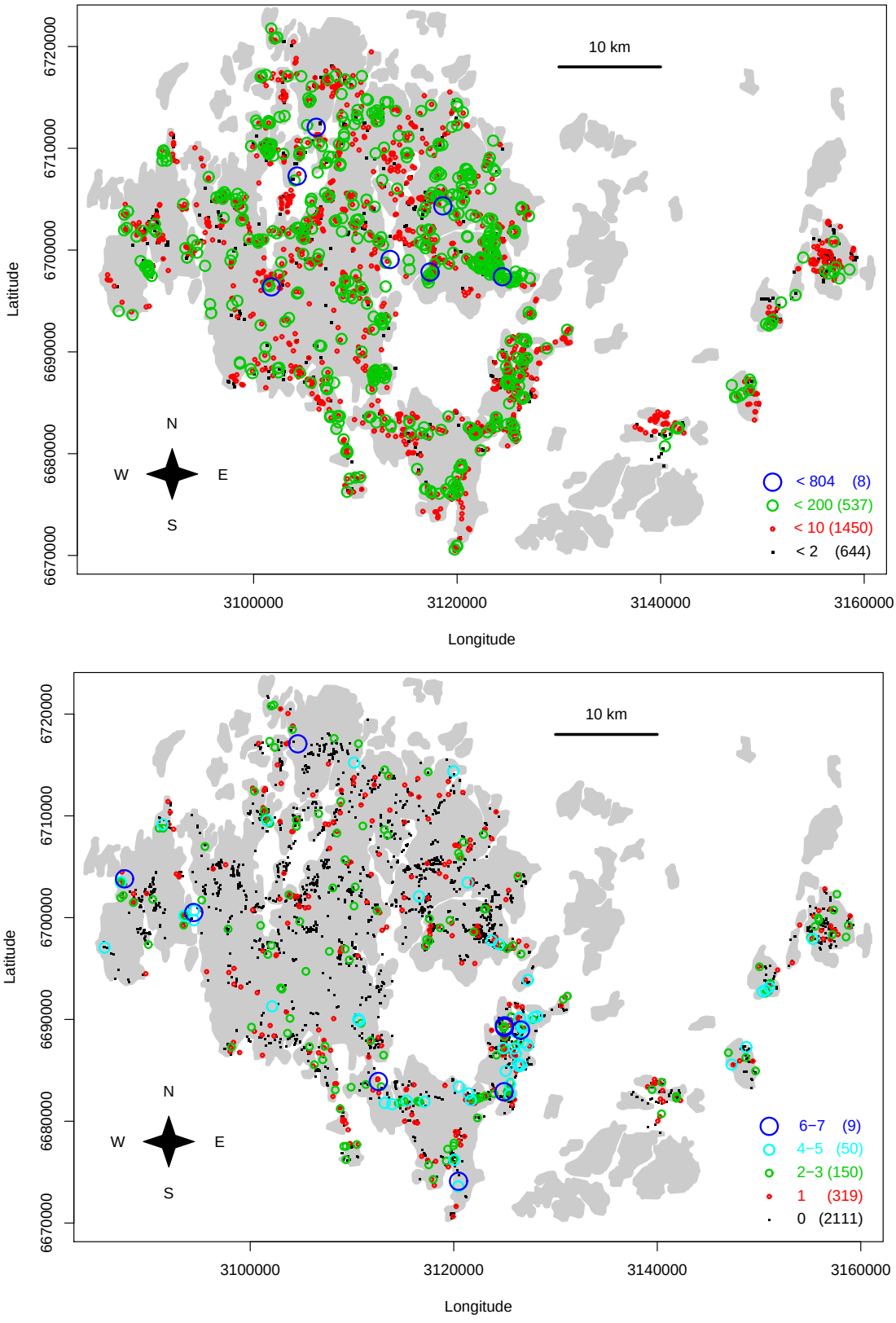


Fig. 1.3. Données d'abondances d'un système hôte-pathogène sur les îles Åland en Finlande agrégées sur les années 2001-2008. Haut : Surface de plantain moyennée sur les 7 années. Bas : Occurrences de l'oïdium cumulées sur les 7 années. La taille des ronds est proportionnelle à l'abondance et pour chaque intensité le nombre de sites concernés est marqué entre parenthèses dans la légende. (2639 sites en tout)

des cellules de côté 1.5 km) et dans le temps (sur les années 2001–2006). Nous obtenons 600 sites occupés par le plantain, dont 340 sont infectés par l'oïdium.

Ce jeu de données nous permet également de voir les limites des méthodes de classification basées sur les modèles de mélange de gaussiennes bivariées, qui ne sont pas adaptées pour tous les types de données. Ce constat a motivé la recherche de classes de distributions plus génériques que les lois gaussiennes multivariées, qui puissent s'adapter à plusieurs types de données, dans les parties III et IV. Dans le chapitre 9 nous discutons de la possibilité de prise en compte des données brutes, en utilisant des lois à structure hiérarchique définies dans la partie III de cette thèse.

1.4.3 Assemblages de couleurs d'une image aérienne

Dans le chapitre 4 nous considérons l'étude d'une photographie aérienne d'un paysage agricole, provenant de l'État de New York aux États-Unis, pour illustrer le fait que la notion d'assemblage que nous avons défini n'est pas limité à des données d'abondances d'espèces. Ici nous considérons des assemblages de trois couleurs par pixels. Pour cela nous avons estimé les proportions moyennes de rouge, vert et bleu sur 3481 cellules (d'environ 200×200 m^2) obtenues à partir de la figure 1.4. Nous avons utilisé le logiciel ImageJ pour diviser l'image en composantes RGB.

Fig. 1.4. Image aérienne d'un paysage agricole dans l'État de New York, États-Unis.

Cette image permet également de vérifier la cohérence spatiale des résultats obtenues par la méthode de classification non spatiale que nous utilisons pour définir des assemblages. En

effet notre méthode choisit 4 classes qui identifient les principaux éléments vus par l'œuil nu (voir figure 4.9 et l'explication correspondate en 4.6). L'intérêt de ne pas avoir introduit l'autocorrélation spatiale explicitement est de bien détecter les barrières du paysage (telles les haies) ; le désavantage est de rajouter du bruit au sein des prairies homogènes (qui sera plus ou moins grand selon la taille des pixels).

1.4.4 Suivi de l'activité de colonies d'abeilles domestiques (*Apis mellifera*)

Ce jeu de données provient d'un projet INRA - ADAPI (Association pour le développement de l'apiculture provençale), financé par le Fonds européen agricole de garantie (FEAGA).

L'activité de 300 colonies d'abeilles situées dans 20 ruchers (15 ruches par rucher) sur un large observatoire au sud de la France a été suivie au cours d'une saison de production.

Chaque ruche a été pesée tous les deux jours pendant 24 jours au cours du mois de juin 2009. Un modèle logistique décrit la variation temporelle du poids d'une ruche et permet d'estimer le gain de poids maximal sur cette période pour chaque ruche. Ce gain de poids correspond principalement à la production de miel pendant ces 24 jours.

Le nombre de cellules operculées (voir figure 1.5) a été compté 3 fois à 12 jours d'intervalle dans chaque ruche (jours 0, 12 et 24). En effet, le développement à partir d'une cellule operculée jusqu'à l'émergence d'une abeille ouvrière est de 12 jours. Donc en comptant à intervalle de 12 jours on peut espérer avoir une estimation fiable du recrutement de nouvelles abeilles, qui est un indicateur de l'activité de la ruche.

Enfin le nombre d'acariens *Varroa jacobsoni*, un parasite des abeilles, a été mesuré pour chaque ruche sur une échantillon de 20 g d'abeilles adultes (environ 150 abeilles) à 0 et 24 jours. Ce parasite contribue à l'affaiblissement des colonies d'abeilles, comme le montre une étude récente de Guzmán et al. (2010), et a un impact économique sur l'apiculture.

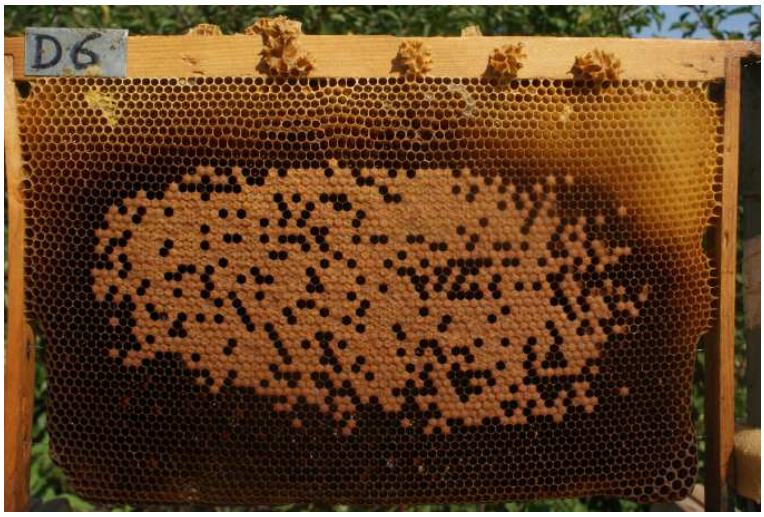


Fig. 1.5. Photographie du couvain d'un cadre de ruche. Le couvain correspond aux cellules operculées, qui contiennent chacune une larve d'abeille et qui se distinguent des autres cellules par une couleur beige clair.

Ce jeu de données non spatial est traité dans la partie III pour illustrer le modèle hiérarchique spatial, que nous présentons dans le chapitre 7, pour prendre en compte des variables de types différents. Les trois variables : gain de poids, nombre de cellules operculées et nombre de parasites ont été considérées. Il permet d'illustrer la notion d'assemblage sur des variables qui ne sont pas forcément des abondances d'espèces d'une part (telles que le gain de poids), et la modélisation d'assemblages formés de variables de types différents d'autre part.

2

Introduction à la classification basée sur les modèles de mélange

Dans le chapitre précédent, nous avons défini la notion d'assemblage d'espèces comme un ensemble de sites pour lesquels les observations multivariées d'abondances d'espèces sont similaires. Regrouper des objets similaires est l'objet des méthodes de classification, donc construire les assemblages d'espèces à partir d'un jeu de données revient à classer les données multivariées d'abondances d'espèces observées. Cette classification se fait dans l'espace des abondances d'espèces, sans prendre en compte la position spatiale des sites d'observations. Nous avons choisi d'utiliser une méthode de classification basée sur les modèles de mélange de lois. Dans ce cadre, à chaque classe (qui représente ici un assemblage d'espèces) on associe une distribution de probabilité multivariée.

Nous présentons maintenant quelques considérations générales sur les méthodes de classification à l'aide de modèles de mélange.

2.1 Classification basée sur des modèles de mélange de lois

Nous nous sommes intéressés exclusivement aux classifications basées sur des modèles de mélanges finis de lois paramétriques. L'approche de classification basée sur les modèles de mélange est apparue il y a plus de cent ans avec l'étude remarquée de Pearson (1894), mais n'a réellement pris son essor qu'à partir d'une trentaine d'années, grâce à l'apparition de méthodes efficaces d'estimation par maximum de vraisemblance telles que l'algorithme EM de Dempster, Laird & Rubin (1977). A présent plusieurs ouvrages de référence existent sur la question, le plus récent étant celui de McLachlan & Peel (2000). Notons cependant que pendant longtemps les méthodes de classification étaient heuristiques et basées pour la plupart sur des critères de distances classiques. Parmi les méthodes heuristiques on distingue souvent deux catégories principales (Jain, Murty & Flynn, 1999) :

- des méthodes de partitionnement itératif (à nombre de classes fixés les observations sont ré-attribuées à chaque étape aux classes en optimisant un critère) telles que $k - means$, basé sur un critère de somme des moindres carrés (MacQueen, 1967),
- des méthodes hiérarchiques ascendantes (par fusion de classes) ou descendantes (par division de classes). On parle de méthode ascendante hiérarchique (*hierarchical agglomerative clustering*) lorsqu'on part de classes contenant une observation et qu'à chaque

étape on fusionne deux classes pour optimiser un critère jusqu'à arriver à une seule classe finale. Le critère pour décider quels groupes fusionner est basé sur une mesure de similarité entre groupes, telle que la distance minimale entre groupes (*single-linkage clustering*).

Les méthodes heuristiques de classification proposent de regrouper les données en classes en utilisant un critère basé sur une mesure de similarité telle que la distance euclidienne. L'approche de classification à l'aide de modèles de mélange se base sur la densité des données, en supposant que chaque groupe correspond à une distribution de probabilité. A titre d'exemple, la classification à l'aide de mélanges de lois normales suppose que les différents groupes proviennent de lois normales de paramètres différents. L'ensemble des données proviennent donc de plusieurs distributions de probabilité distinctes qui constituent un mélange de lois. Ce mélange est entièrement connu lorsqu'on connait les paramètres des différentes lois qui le composent (appelées les composantes du mélange) et dans quelles proportions les données proviennent de chacune de ces lois (proportions de mélange). A nombre de classes fixés, il s'agit d'estimer les paramètres de ces différentes lois normales et les proportions de mélange qui permettent d'ajuster le mieux les données. Dans un contexte de classification, le but est surtout d'obtenir pour chaque observation la composante dont elle provient, ce qui définira sa classe.

De nombreux auteurs discutent des avantages de l'approche basée sur des modèles de mélange par rapport aux approches heuristiques, un des principaux étant qu'elle fournit un cadre théorique à la classification, et permet de résoudre de manière élégante des problèmes tels que le choix du nombre de classes, qui est placé alors dans un cadre statistique de sélection de modèles (Fraley & Raftery, 2002; McLachlan & Peel, 2000).

Remarquons également que, outre leur utilisation en classification automatique ou analyse discriminante, les modèles de mélange peuvent être utilisés pour estimer une densité et notamment pour ajuster des données surdispersées par rapport aux distribution classiques. Dans cette problématique on peut utiliser soit les mélanges discrets à nombre de composantes finis, qui sont utilisés également en classification, soit les mélanges continus de lois. Nous verrons des exemples de mélanges continus dans la partie III.

Notations

Par la suite, les variables aléatoires sont notées avec des majuscules quand cela est possible (sauf dans le cas de θ), leurs réalisations sont notées avec la lettre minuscule correspondante et les vecteurs et matrices sont notés en gras. La notation générique f est utilisée pour la fonction de densité d'une variable aléatoire et f_X désigne la densité de la variable aléatoire X. Dans le cadre des modèles de mélange f_k désigne la densité de la $k^{\text{ème}}$ composante. La densité conditionnelle d'une variable X sachant Z sera notée $f_{X|Z}$.

2.1.1 Définition des mélanges finis de lois

Les modèles de mélange utilisés en classification automatique sont des mélanges discrets à nombre de composantes fini (mais souvent inconnu). On se place dans le cadre multivarié.

Soit $\boldsymbol{Y} = (\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n)$ un échantillon aléatoire de taille n, où $\boldsymbol{Y}_j$ est un vecteur aléatoire de dimension d. On dit que $\boldsymbol{Y}$ suit un mélange fini de lois à K composantes si sa densité peut s'écrire :

$$f_{\boldsymbol{Y}}(\boldsymbol{y}; \boldsymbol{\Phi}) = \sum_{k=1}^{K} \pi_k f_k(\boldsymbol{y}; \boldsymbol{\phi}_k)$$

où $\boldsymbol{\Phi}$ est le vecteur de l'ensemble des paramètres $\boldsymbol{\Phi} = (\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_K, \pi_1, \ldots, \pi_K)$, f_k désigne la densité de la composante k de paramètre $\boldsymbol{\phi}_k$ et π_k est la probabilité qu'une observation appartient à la composante k, avec les conditions $\pi_k > 0$ et $\sum_{k=1}^{K} \pi_k = 1$. La quantité π_k sera appelée par la suite la proportion de mélange de la composante k pour éviter toute confusion avec les probabilités d'appartenance a posteriori des observations p_{ik} définies en (2.2). Dans un contexte de classification, les composantes de la loi mélange sont supposées correspondre aux classes et le but est de reconstituer les appartenances des données à ces différentes classes.

Un modèle de mélange fini de lois peut être vu comme un modèle à structure hiérarchique et représenté sous la forme d'un graphique orienté acyclique (usuellement abrégé par l'acronyme anglais DAG pour *Directed Acyclic Graph*). Ce type de graphique illustre les dépendances conditionnelles entre grandeurs (paramètres et variables) sous forme d'arbres (voir le schéma général figure 2.1). Les noeuds initiaux sont des grandeurs conditionnantes qui ne dépendent d'aucune autre variable aléatoire (il s'agit des paramètres du modèle), les noeuds intermédiaires sont des grandeurs conditionnées et conditionnantes (ce sont les variables latentes), et les noeuds terminaux (feuilles de l'arbre) sont les grandeurs conditionnées (les variables observées). Nous représentons ces derniers par des carrés.

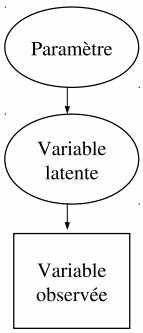


Fig. 2.1. Organisation d'un graphe orienté acyclique (DAG) dans le cadre de modèles hiérarchiques

Dans le cas du mélange gaussien (cf figure 2.2) les noeuds initiaux sont les paramètres π_k, à savoir les proportions de mélange, qui conditionnent, pour chaque observation i, l'appartenance aux classes (variable latente non observée). A chaque classe sont associés des paramètres de moyenne et de variance d'une loi gaussienne multivariée, qui conditionnent à leur tour les variables observées. Le graphe est souvent lu en sens inverse, en remontant l'arbre, de la manière suivante : chaque observation multivariée i correspond à la réalisation de d variables aléatoires $Y_{i1}, \ldots, Y_{id}$ (variables observées). Ces réalisations vont dépendre

de la classe de laquelle provient l'observation i donc des paramètres de la loi normale qui définit cette classe. A leur tour, les appartenances aux classes des différentes observations i dépendent des proportions de mélange $\pi_1, \ldots, \pi_K$, qui doivent être respectées.

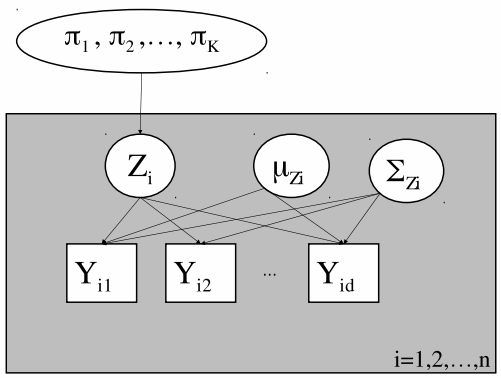


Fig. 2.2. Graphe orienté acyclique (*Directed Acyclic Graph* ou DAG) du modèle de mélange gaussien

2.1.2 Identifiabilité des mélanges finis de lois

L'identifiabilité des modèles de mélange est définie de la manière suivante : Soient deux membres d'une famille paramétrique de densités mélange :

$$f_{\boldsymbol{Y}}(\boldsymbol{y};\boldsymbol{\Phi}) = \sum_{k=1}^{K} \pi_k f_k(\boldsymbol{y};\boldsymbol{\phi}_k) \quad \text{et} \quad f_{\boldsymbol{Y}}(\boldsymbol{y};\boldsymbol{\Phi}^*) = \sum_{k=1}^{K^*} \pi_k^* f_k(\boldsymbol{y};\boldsymbol{\phi}_k^*).$$

Cette famille de densités mélanges est identifiable (pour les paramètres $\boldsymbol{\Phi}$) si l'égalité pour presque tous les $\boldsymbol{y}_i$ (par rapport à la mesure dominante définissant les densités) :

$$f_{\boldsymbol{Y}}(\boldsymbol{y};\boldsymbol{\Phi}) = f_{\boldsymbol{Y}}(\boldsymbol{y};\boldsymbol{\Phi}^*)$$

est obtenue si et seulement si $K = K^*$ et qu'il existe une permutation des indices des composantes telle que :

$$\pi_k = \pi_{k^*} \quad \text{et} \quad f_k(\boldsymbol{y}_i;\boldsymbol{\phi}) = f_k(\boldsymbol{y}_i;\boldsymbol{\phi}^*).$$

Ceci est à l'origine du *label-switching* qui pose problème dans l'approche d'estimation bayésienne discutée en 2.2.2. L'identifiabilité des modèles de mélange a été étudiée pour plusieurs lois. McLachlan & Peel (2000, p.26) discutent ce problème et fournissent d'autres références. La conclusion est que la plupart des mélanges finis de distributions continues sont identifiables. On peut citer en contre-exemple les mélanges de distributions uniformes. L'identifiabilité des mélanges finis de Poisson et de distributions binomiales négatives a également été prouvée, alors que les mélanges de binomiales ne sont pas identifiables si $n_b < 2K-1$, où n_b est le paramètre de la binomiale correspondant au nombre de répétitions de l'épreuve de Bernoulli et K est le nombre de classes (ce qui implique que les mélanges de Bernoulli ne sont pas identifiables).

2.2 Estimation des modèles de mélange

La méthode d'estimation la plus couramment employée pour les modèles de mélange de lois est l'algorithme d'espérance-maximisation (EM) (McLachlan & Peel, 2000). Cet algorithme sera souvent utilisé dans cette thèse sous différentes variantes, de ce fait, nous lui consacrons un paragraphe ici. L'approche bayésienne est brièvement commentée dans le deuxième paragraphe.

2.2.1 Approche par maximum de vraisemblance : L'algorithme EM

L'algorithme d'espérance maximisation (*Expectation Maximization*, EM) a été introduit par Dempster, Laird & Rubin (1977) pour calculer les estimateurs du maximum de vraisemblance des paramètres d'un modèle lorsque celui-ci comporte des données manquantes ou des variables latentes (Tanner, 1991). De ce fait, il est particulièrement adapté à l'estimation des modèles de mélange de lois, car il prend en compte la structure latente inhérente au problème de classification en complétant ou en augmentant les données observées avec des données non observées qui indiquent les appartenance inconnues aux classes. En effet, le problème de l'estimation des paramètres d'un mélange fini dans un contexte de classification peut être formulé ainsi :

Supposons que les observations $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ sont indépendantes et issues de K composantes d'un mélange de lois de paramètres inconnus $\boldsymbol{\phi}_k$ avec $k \in \{1, \ldots, K\}$. Introduisons la variable latente $\boldsymbol{Z}$ correspondant aux appartenances aux classes définie ainsi :

$$Z_{ik} = \begin{cases} 1 \text{ si l'observation } i \text{ provient de la } k^{\text{ème}} \text{ composante} \\ 0 \text{ sinon.} \end{cases}$$

L'idée de l'EM consiste à calculer de manière itérative les paramètres $\boldsymbol{\phi}_k, \boldsymbol{\pi}_k$ qui maximisent la vraisemblance $l_c(\boldsymbol{\Phi}; \boldsymbol{Y}, \boldsymbol{Z})$ des données complètes $(\boldsymbol{Y}, \boldsymbol{Z})$ sachant les données observées $\boldsymbol{y}$. L'algorithme EM procède de la manière suivante : après initialisation de $\boldsymbol{\Phi}$ deux étapes se succèdent. A l'itération t+1 :

L'étape E (espérance) consiste à calculer l'espérance conditionnelle de la vraisemblance des données complètes sachant les données observées et les valeurs des estimateurs à l'étape t (appelée fonction Q) :

$$Q(\boldsymbol{\Phi}, \boldsymbol{\Phi}^{(t)}) = \mathbb{E}_{\boldsymbol{\Phi}^{(t)}}[l_c(\boldsymbol{\Phi}; \boldsymbol{Y}, \boldsymbol{Z}) | \boldsymbol{Y} = \boldsymbol{y}] = \sum_{i=1}^{n} \sum_{k=1}^{K} p_{ik}^{(t)} \log[\pi_k f_k(\boldsymbol{y}_i; \boldsymbol{\phi}_k)]. \tag{2.1}$$

Dans le cadre des modèles de mélange cela consiste à calculer les probabilités a posteriori d'appartenance aux classes :

$$p_{ik}^{(t)} = \mathbb{P}_{\boldsymbol{\Phi}^{(t)}}(Z_{ik} = 1 | \boldsymbol{Y}_i = \boldsymbol{y}_i) = \frac{\pi_k^{(t)} \ f_k(\boldsymbol{Y}_i | \boldsymbol{\phi}_k^{(t)})}{\sum_{j=1}^{K} \pi_j^{(t)} \ f_j(\boldsymbol{Y}_i | \boldsymbol{\phi}_j^{(t)})}. \tag{2.2}$$

L'étape M (maximisation) consiste à actualiser les estimateurs des paramètres par :

$$\boldsymbol{\Phi}^{(t+1)} = \arg\max_{\boldsymbol{\Phi}} Q(\boldsymbol{\Phi}, \boldsymbol{\Phi}^{(t)}).$$

Ces étapes peuvent être directes comme dans le cas de mélanges gaussiens multivariés (McLachlan & Peel, 2000). Dans d'autres cas plus compliqués, l'étape E ou l'étape M ne sont pas explicites, et il faut alors recourir à des étapes supplémentaires, soit de Monte Carlo pour approcher l'espérance de l'étape E, soit d'optimisation numérique dans l'étape M. Par la suite, dans les parties III et IV, nous verrons des cas où l'étape E n'est pas explicite, car l'expression de 2.1 fait intervenir des intégrales en grande dimension, ce qui nous amènera à parler d'une version modifiée de l'EM : le Monte Carlo EM, introduite par Wei & Tanner, 1990.

Limites de l'approche fréquentiste

Les limites de l'approche fréquentiste d'estimation du maximum de vraisemblance des modèles de mélange est lié à la forme de la fonction de vraisemblance.

En effet, dans le cas des modèles de mélange, la surface de la vraisemblance comporte de nombreux maxima locaux. Ceci rend l'algorithme EM très dépendant de l'initialisation. En effet une mauvaise initialisation va d'une part allonger la durée nécessaire à la convergence et aura tendance, d'autre part, à détecter les maxima locaux à la place du maximum absolu.

Un autre problème lié à la fonction de vraisemblance est que, pour beaucoup de modèles de mélange, et notamment les mélanges multivariés gaussiens avec une matrice de covariance non contrainte, la vraisemblance n'est pas bornée. Donc pour certaines séquences de paramètres elle peut tendre vers l'infini, ce qui cause la divergence de l'EM (on parle aussi de dégénèrescences, ou de solutions dégénérées de vraisemblance infinie). Par exemple lorsque dans un mélange gaussien la moyenne tend vers l'observation et la variance vers 0 (loi de Dirac), c'est-à-dire lorsqu'une composante est réduite à une observation, la vraisemblance va tendre vers l'infini. En pratique, selon Fraley & Raftery (2007), ceci est causé par des singularités dans l'estimation de la matrice de covariance et arrive le plus souvent lorsque la covariance est autorisée à varier entre composantes et lorsqu'il y a beaucoup de composantes.

Initialisation de l'EM

Comme nous venons de le mentionner, l'initialisation de l'EM est un problème assez délicat, qui peut avoir des conséquences sur les estimations obtenues. Ce problème a été discuté par de nombreux auteurs (Biernacki, Celeux & Govaert, 2003; McLachlan & Peel, 2000, p. 54) qui ont proposé plusieurs solutions pour faciliter la convergence de l'algorithme et éviter de tomber sur des maxima locaux.

Sans information a priori sur les paramètres, plusieurs stratégies d'initialisation existent :

– Effectuer plusieurs initialisations des paramètres au hasard et comparer les vraisemblances obtenues après convergence pour éliminer les maxima locaux. Une solution plus économe en temps consiste à initialiser l'EM au hasard et arrêter l'algorithme au

bout d'un nombre d'itérations fixé à l'avance pour choisir l'initialisation (une ou plusieurs) qui a permis la plus grande augmentation de vraisemblance (Biernacki, Celeux & Govaert, 2003). Ceci permet de tester beaucoup d'initialisations (ce qui augmente les chances de tomber sur une bonne initialisation, à savoir une initialisation proche des vraies valeurs des paramètres) pour n'en garder que les meilleures qui seront poursuivies jusqu'à la convergence de l'EM (ce qui permet d'économiser du temps de calcul).

– Partir d'une partition initiale obtenue avec un autre type de classification (ou définies au hasard), initialiser les probabilités a posteriori d'appartenance aux classes p_{ik} à 0 ou 1 selon la partition initiale et démarrer l'EM par l'étape M à la première itération (ce qui permettra d'initialiser les paramètres).
– Tirer au hasard les paramètres en utilisant les moments de l'échantillon. Par exemple pour initialiser un mélange de lois normales on peut calculer la moyenne $\bar{\boldsymbol{y}}$ et la variance $\boldsymbol{V}$ empirique de l'échantillon, puis tirer les paramètres d'espérance $\boldsymbol{\mu}_k$ de chaque classe k, pour $k \in \{1, \ldots, K\}$, dans une distribution normale $\mathcal{N}(\bar{\boldsymbol{y}}, \boldsymbol{V})$. Les matrices de covariance et les proportions de mélange peuvent être initialisés par $\boldsymbol{\Sigma}_k = \boldsymbol{V}$ et $\pi_k = 1/K$.

Des méthodes plus élaborées peuvent être utilisées, par exemple, dans un cadre d'estimation de mélanges gaussiens multivariés, Fraley & Raftery (2002) proposent d'effectuer une pré-classification hiérarchique ascendante basée sur un modèle de mélange de gaussiennes qui leur fournit directement des valeurs initiales pour les paramètres de chaque modèle.

D'autres approches ont été proposées pour diminuer l'impact de l'initialisation sur les estimations finales des paramètres. Elles consistent à modifier l'EM déterministe afin de lui permettre une plus grande exploration de l'espace des paramètres et d'éviter qu'il reste bloqué sur une mauvaise initialisation. Ainsi, par exemple Ueda & Nakano (1998) ont proposé une approche EM avec recuit déterministe (*Deterministic Annealing EM algorithm* ou DAEM, où la "température du recuit" est mise sur les probabilités a posteriori d'appartenance aux classes p_{ik}), et Celeux & Diebolt (1985) ont proposé un algorithme EM avec une étape stochastique (*Stochastic EM* ou SEM). Biernacki, Celeux & Govaert (2003) et McLachlan & Peel (2000) (p. 61) recommandent d'utiliser l'algorithme SEM plutôt dans une phase d'initialisation puis de continuer jusqu'à la convergence avec un EM classique.

Sélection du meilleur modèle dans le cadre d'une classification automatique

Les approches de classification automatique basées sur des modèles de mélanges consistent à considérer plusieurs modèles, correspondant chacun à un nombre de classes et éventuellement à des contraintes sur les paramètres des densités des composantes (par exemple imposer la même matrice de covariance pour toutes les composantes), et à choisir le meilleur modèle parmi ceux-ci par un critère de sélection (Fraley & Raftery, 2002).

La démarche de classification la plus courante peut être résumée ainsi :

– Choix d'un nombre maximal de classes à considérer $K_{\max}$,
– Initialisation puis estimation des paramètres par l'algorithme EM pour chaque modèle (nombre de classes et éventuellement pour chaque modèle contraint).

– Calcul d'un critère de sélection de modèle basé sur la vraisemblance du mélange calculé avec les estimations finales des paramètres. Dans cette thèse les critères BIC (*Bayesian Information Criterion*), approximation du facteur de Bayes proposée par Schwarz (1978), et ICL (*Integrated Classification Likelihood*) proposé par Biernacki, Celeux & Govaert (2000) ont été utilisés.
– Sélection du modèle qui maximize le critère choisi.

2.2.2 Approche bayésienne

L'estimation bayésienne des modèles de mélange est possible depuis peu grâce à des méthodes de Monte Carlo par chaînes de Markov (*Markov Chain Monte Carlo* ou MCMC) telles que l'échantillonneur de Gibbs.

Simulations MCMC

Le principe des méthodes de Monte Carlo par chaînes de Markov est de construire une chaîne ergodique de Markov dont la distribution stationnaire est la distribution a posteriori du paramètre d'intérêt, qui est dans ce cas le vecteur $\boldsymbol{\Phi}$ des paramètres du modèle de mélange. Plusieurs méthodes peuvent être utilisées pour échantillonner dans cette distribution a posteriori. L'échantillonneur de Gibbs et l'algorithme de Metropolis-Hastings en font partie. L'échantillonneur de Gibbs utilise les distributions conditionelles d'un paramètre (ou d'un sous-vecteur de $\boldsymbol{\Phi}$) sachant tous les autres paramètres et les données observées (appellées conditionelles complètes).

Comme dans l'approche fréquentiste par EM, on introduit les variables latentes $\boldsymbol{Z}$ indicatrices des classes. L'algorithme produit alternativement des échantillons de la variable latente $\boldsymbol{Z}$ et du vecteur de paramètres $\boldsymbol{\Phi}$, c'est-à-dire une chaîne de Markov pour les données manquantes et une chaîne pour les paramètres (voir McLachlan & Peel, 2000, p. 120 pour plus de détails).

Intérêt de l'approche bayésienne

L'approche bayésienne est intéressante lorsqu'on possède de l'information a priori sur les composantes. Elle peut également résoudre certains problèmes qui apparaissent dans l'estimation par maximum de vraisemblance telles que les dégénérescences et les singularités. L'approche bayésienne permet de corriger ce problème en lissant la vraisemblance (Fraley & Raftery, 2002, p. 627). Fraley & Raftery (2007) proposent de combiner l'approche EM et l'estimateur du maximum a posteriori bayésien pour éviter les problèmes de dégénérescence et de singularité. Ils utilisent un prior sur les paramètres qui permet d'éliminer les problèmes de convergence de l'EM dûs aux singularités. Ceci ne permet pas de résoudre les problèmes de l'approche bayésienne abordés dans le paragraphe suivant, mais a l'avantage d'être plus économe en temps de calcul que les approches MCMC. Notons également que Richardson & Green (1997) ont utilisé une approche MCMC par sauts réversibles (*Reversible Jump MCMC*) afin de pouvoir se déplacer entre différents modèles et valeurs de paramètres, ce

qui permet d'obtenir les facteurs de Bayes et d'estimer le nombre de classes directement. Cependant Fraley & Raftery (2002) ont constaté que l'implémentation et l'utilisation de cette méthode dans le cas multivarié gaussien est difficile.

Limites de l'approche bayésienne

Deux problèmes majeurs apparaissent dans l'estimation des modèles de mélange par des méthodes bayésiennes.

Le premier est le problème des priors impropres. On rappelle qu'une loi $f(.)$ est dite impropre si $I = \int f(\phi)d\phi$ est infini. Les estimateurs bayésiens pour des modèles de mélange sont bien définis si les distributions a priori (priors) sont propres. En effet l'utilisation de priors impropres va aboutir à des distributions a posteriori impropres. Lorsqu'il n'y a pas d'information a priori sur les composantes, utiliser des priors non informatifs classiques n'est pas possible dans le cadre des mélanges. En effet il y a toujours la possibilité qu'aucune observation ne soit allouée à une ou plusieurs composantes et les données seront donc non informatives sur celles-ci. Une des solutions qui a été proposée est l'utilisation de priors partiellement propres (McLachlan & Peel, 2000, p. 125).

Le deuxième problème est le *label-switching* qu'on peut appeler échange d'indices ou permutation d'indices. Ce problème apparaît lorsque le prior ne contient pas d'information a priori suffisante pour discriminer entre les composantes d'un modèle de mélange appartenant à la même famille paramétrique. En effet l'échange des indices entre composantes n'a pas d'effet sur la logvraisemblance, et comme il y a $K!$ indexations possibles, il y a $K!$ composantes de la distribution a posteriori qui sont identiques à l'exception de l'indexation si le prior est symétrique par rapport aux indexations. Pour un algorithme itératif, l'effet des échanges d'indices est très pénalisant, car d'une itération à l'autre les indices peuvent être permutés entre composantes. Ceci ne pose pas de problème dans le calcul des estimateurs du maximum de vraisemblance via l'EM, mais constitue un sérieux handicap dans la simulation de réalisations du vecteur de paramètres $\boldsymbol{\Phi}$ à partir de sa distribution a posteriori. Ceci a pour conséquence néfaste par exemple que les estimations a posteriori des moyennes des composantes du mélange seront toutes égales. Celeux, Hurn & Robert (2000) vont jusqu'à dire, en rapport avec le problème d'échange d'indices dans le contexte bayésien, que presque tous les échantillonneurs MCMC implémentés pour les modèles de mélange n'ont pas réussi à converger.

Pour une bibliographie approfondie sur le sujet, ainsi que des détails concernant les méthodes utilisées et les problèmes que nous venons de mentionner, voir par exemple McLachlan & Peel (2000), chapitre 4.

Deuxième partie

Explorer des données spatiales multivariées d'abondances d'espèces

3

Etudier les assemblages d'abondances d'espèces pour explorer les interactions interspécifiques

Dans cette partie une nouvelle approche de l'analyse de l'association spatiale locale entre données d'abondances de plusieurs espèces est présentée. Le point de vue pris ici est différent de celui de Perry & Dixon (2002), qui proposent un indice d'association locale (basé sur des mesures de distance) pour mesurer l'intensité de l'association spatiale entre deux ensembles de données de comptages. Ces auteurs mesurent l'association par un indice continu, et s'intéressent uniquement aux extrêmes : association fortement positive (attraction) et fortement négative (répulsion). De plus leur méthode est limitée à deux espèces. L'approche présentée ici s'apparente plus à celle de Karunanayake & Laverty (2006), qui s'intéressent à des données spatiales de comptages de trois espèces d'adventices (mauvaises herbes). Ils proposent de définir des zones homogènes (que nous pouvons interpréter comme des types d'associations spatiales) en utilisant une méthode de classification trivariée. Leur méthode de classification est basée sur des modèles de mélange de lois de Poisson multivariées (Karlis & Meligkotsidou, 2006) et tient compte de la corrélation spatiale des données en utilisant un champ de Markov caché.

Nous proposons également de classer les données pour obtenir des assemblages d'espèces, qui sont la signature des interactions des espèces entre elles et avec leur environnement. Mais notre approche est différente de celle de Karunanayake & Laverty (2006) car nous définissons les assemblages d'espèces en utilisant seulement les vecteurs de comptages observés à chaque site, c'est-à-dire en faisant abstraction de leur lien spatial. Ce dernier est exploré dans une deuxième étape.

La méthode que nous présentons ne comporte donc que peu d'hypothèses (aucune hypothèse a priori sur l'autocorrélation spatiale), ce qui la rend plus flexible. En effet, comme nous l'avons argumenté en 1.3, cela permet de s'affranchir, dans un premier temps, de deux problèmes majeurs en écologie spatiale, à savoir du problème d'échelle et de l'hypothèse d'homogénéité spatiale (ou hypothèse de stationnarité du processus écologique). Notre méthode peut s'avérer utile dans un stade préliminaire d'une étude écologique pour générer des hypothèses à tester et aider à construire des modèles d'inférence.

3.1 Classification des abondances et structure spatiale des assemblages d'espèces

Dans le chapitre 4 nous proposons une méthode exploratoire permettant de détecter des assemblages d'espèces à partir de données spatiales d'abondances de plusieurs espèces, de les cartographier et d'analyser leur structure spatiale (Georgescu et al., 2009). Cette méthode, intitulée Classification And Spatial Analysis (CASA), effectue d'abord une classification non spatiale des vecteurs d'abondances d'espèces observées, puis une analyse spatiale des classes ainsi créées. Nous appelons ces classes des assemblages d'espèces.

1. Classification : Des assemblages d'espèces sont obtenus en classant les vecteurs d'abondances d'espèces par une méthode probabiliste basée sur des mélanges de lois gaussiennes multivariées (méthode et algorithme proposés par Fraley & Raftery, 2002, 2006). Cette classification ne tient pas compte de la corrélation spatiale des données et permet d'obtenir pour chaque vecteur d'abondances sa probabilité d'appartenance à chacun des types d'assemblage. A l'issue de cette étape, une carte des classes est obtenue en utilisant la règle du maximum a posteriori (MAP), qui consiste à affecter les données à la classe la plus probable.
2. Analyse spatiale : Les assemblages obtenus sont analysés par des méthodes classiques d'analyse spatiale qui sont à adapter au cas étudié. Ces méthodes utilisent les probabilités d'appartenance des observations aux classes afin de prendre en compte l'incertitude de la première étape (seulement l'incertitude due à l'affectation des observations aux classes est prise en compte).

La séparation de l'étape d'identification des assemblages d'espèces et de l'étape d'analyse spatiale, rend la méthode CASA plus flexible, car elle n'introduit pas d'hypothèses sur la structure de corrélation entre sites. Par conséquent, la première étape de classification de CASA peut être utilisée en l'état pour des types d'échantillonnage différents, des échelles différentes, des cas où la dépendance spatiale n'est pas stationnaire, et des espaces à structure quelconque (par exemple arborescente). Dans ces cas, seule la deuxième étape de CASA est à adapter aux spécificités de l'espace et de l'échantillonnage en choisissant des outils d'analyse spatiale adéquats.

Cette méthode est illustrée sur le jeu de données présenté en 1.4.2 concernant la répartition de l'oïdium sur les prairies de plantain dans les îles Åland. Un pré-traitement est nécessaire pour transformer les données de présence/absence de l'oïdium en données "pseudo-continues". Pour cela, les données sont agrégées dans le temps et l'espace sur une grille de 1.5×1.5 km^2. Une classification en 5 assemblages est choisie par l'algorithme de classification grâce à un critère de sélection de modèle (BIC). L'analyse spatiale met en évidence un gradient d'infection du plantain par l'oïdium. L'abondance du plantain n'est donc pas le seul facteur qui explique la distribution de l'oïdium. Celle-ci peut être due soit à une structuration génétique de la plante hôte, qui formerait des zones où les plantes sont résistantes au pathogène et des zones plus sensibles, soit à l'effet de facteurs environnementaux, tels que l'humidité et la température, qui pourraient influencer l'installation

du pathogène, ou le vent, qui pourrait influencer sa dispersion. Pour aller plus loin, l'effet de facteurs environnementaux sur la distribution des assemblages pourrait être testé par régression. Les résultats de cette étude exploratoire génèrent donc un ensemble d'hypothèses à tester pour une étude plus approfondie ou pour alimenter une approche de modélisation du processus.

Cette approche n'est pas limitée aux données écologiques mais peut présenter un intérêt dans d'autres domaines où l'on dispose de données multivariées spatiales, et où la corrélation des variables est susceptible d'intervenir sur la structuration spatiale des données (tels que la sociologie, l'épidémiologie spatiale, la génétique des populations).

3.2 Limites et perspectives

3.2.1 Limites de l'étape de classification en assemblages d'espèces

La méthode de classification utilisée dans la première partie de CASA suppose que les données d'abondances des espèces sont issues d'un mélange de lois. Donc un assemblage d'espèces est modélisé par une loi de probabilité multivariée dans l'espace des abondances. Dans cette partie de la thèse, la méthode CASA utilise uniquement les lois normales multivariées, donc les données d'abondance considérées doivent être continues (e.g. poids et surface) ou pseudo-continues (c'est-à-dire pouvoir être quantifiées finement par une transformation).

Or, la transformation des données entraîne en général une perte d'information. Par exemple, dans l'illustration sur le jeu de données présenté en 1.4.2, l'agrégation des données entraîne une perte d'information sur la taille des prairies de l'hôte, qui est susceptible d'influencer la distribution du pathogène.

De plus, lorsque l'hypothèse de normalité n'est pas plausible, l'algorithme de classification automatique de Fraley & Raftery (2006) n'arrive pas, en général, à estimer le nombre de classes et tend à le surestimer.

Or les données d'abondances en écologie sont souvent :

- récoltées sous formes diverses, qui dépendent des espèces étudiées, des moyens mis en oeuvre et de la précision du matériel : discrètes (comptages), binaires (présence-absence), ordinales,
- loin de l'hypothèse de normalité : surdispersées, sousdispersées, 0-inflatées.

Pour pouvoir utiliser CASA sur des données brutes qui ne vérifient pas l'hypothèse de normalité (et éviter toute perte d'information liée à leur transformation en variables pseudo-continues), il faudrait adapter les lois dans l'étape de classification aux types des données.

Dans la suite de cette thèse l'objectif a été de généraliser la méthode de classification utilisée dans la première partie de CASA à d'autres types de données. Dans le chapitre 5, nous envisageons une extension de CASA aux données de comptages dans le cas bivarié, basée sur l'utilisation de lois de Poisson bivariées à la place des gaussiennes dans la classification, que l'on applique sur les données de comptages de pucerons présentées en 1.4.1.

Dans les parties III et IV nous définissons un modèle hiérarchique capable de prendre en compte des données de plusieurs types.

Enfin, nous avons argumenté en 1.3 et au début de ce chapitre, de l'intérêt d'une approche de classification qui n'introduit aucune hypothèse sur les autocorrélations spatiales des abondances d'espèces entre sites. Notons que, dans certains cas, il peut être utile d'introduire des connaissances a priori sur l'autocorrélation spatiale. Ceci a surtout été montré dans le cadre de la segmentation d'images (Besag, 1986; Ambroise, Dang & Govaert, 1996). La classification à l'aide de modèles de mélange peut s'étendre à des données dépendantes $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$, en supposant un modèle markovien stationnaire pour la distribution des vecteurs $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n$, indicateurs des composantes d'origine des données (McLachlan & Peel, 2000, p. 33 et chapitre 13). En une dimension ce modèle markovien est une chaîne de Markov (utilisée par Karunanayake & Laverty, 2006 dans leur classification d'assemblages d'adventices) ; en deux dimensions ou plus c'est un champ markovien (*Markov random field* ou MRF). L'estimation de ce modèle peut se faire en modifiant l'algorithme EM, par exemple en introduisant une pénalisation basée sur le voisinage, comme dans l'algorithme NEM (*Neighbourhood EM*) proposé par Ambroise, Dang & Govaert (1996), ou en introduisant un échantillonneur de Gibbs pour simuler à partir de la distribution a posteriori de $\boldsymbol{Z}$ sachant $\boldsymbol{Y}$ dans l'étape E de l'EM, comme proposé par Chalmond (1989).

3.2.2 Limites des outils d'analyse spatiale présentés

Les analyses spatiales présentées dans le chapitre 4 se limitent à des outils basiques qui sont utilisées dans des espaces euclidiens. Nous utilisons :

- les variogrammes pour tester l'existence d'une structuration spatiale des classes,
- des distances entre classes pour étudier les relations entre classes,
- les barycentres des classes pour voir des tendances lorsque les classes sont relativement bien séparées spatialement,
- une méthode de lissage des probabilités d'appartenance aux classes qui utilise les probabilités des plus proches voisins.

Les probabilités d'appartenance des sites aux différentes classes sont utilisées pour obtenir des intervalles de confiance des variogrammes, barycentres et distances, afin de tenir compte de l'erreur commise à la première étape de notre procédure. Nous n'avons pas présenté d'outils utilisables dans des espaces non euclidiens. L'adaptation d'outils d'analyse spatiale existants pour la deuxième étape de CASA est assez facile : il suffit de prendre en compte l'erreur de classification en utilisant les probabilités d'appartenance aux classes. Pour cela nous proposons de procéder en simulant un grand nombre de fois des nouvelles cartes d'appartenance des classes à partir des probabilités d'appartenance obtenues à la première étape et de calculer la statistique choisie pour chaque simulation. On pourrait également chercher à inclure explicitement/analytiquement l'erreur de classification dans l'expression de l'outil choisi (distance, variogramme) au lieu de procéder par simulation.

Notons enfin que dans cette étape nous tenons compte uniquement de l'incertitude de l'affectation aux classes par le MAP, et non de l'incertitude d'estimation des probabilités a posteriori d'appartenances aux classes ($\hat{p}_{ik}$). Or les $\hat{p}_{ik}$ sont calculés conditionnellement aux estimations du maximum de vraisemblance des paramètres du modèle. Mais comme ceci ne tient pas compte de l'incertitude dans l'estimation des paramètres, il est probable que l'incertitude globale soit sousestimée, ce qui biaise les estimations des $\hat{p}_{ik}$ vers une plus grande certitude (vers 0 ou 1). Cet effet diminue lorsque la taille de l'échantillon augmente (Fraley & Raftery, 2002, p. 627).

4

Exploring Spatial and Multitype Assemblages of Species Abundances

By Vera Georgescu[1,*], Samuel Soubeyrand[1], André Kretzschmar[1], Anna-Liisa Laine[2]

Biometrical Journal 51 (2009), 979–995

[1] INRA, UR546 Biostatistique et Processus Spatiaux, Agroparc, F-84914 Avignon, France

[2] Department of Applied Biology, PO Box 27, 00014 University of Helsinki, Finland

[*] Corresponding author : vera.georgescu@avignon.inra.fr – FAX : +33(0)432722182

The ecological theory of the existence of multiple stable states between species, or the spatial heterogeneity of some unobserved environmental factor, support the idea of multitype interactions between species. These multitype interactions can lead to different assemblages of species abundances. An exploratory tool for the detection of these species assemblages and for their spatial analysis is presented in this article. A two-stage analysis is proposed. First, a classification into types of species assemblages using only the species abundances at each site, regardless of their spatial location, is performed. The clustering procedure is based on multivariate normal mixtures and provides a measure of the classification uncertainty. Second, some tools for the study of the spatial structure of these types of assemblages are presented. We transfer the classification uncertainty to the spatial analysis of the classes in order to draw more accurate conclusions. This Classification And Spatial Analysis (CASA) method is used to point out a spatial gradient of infection in a host-pathogen system in the Åland Islands in Finland. It can be a useful preliminary tool for ecological studies involving the spatial distributions of several species.

Key words : CASA method ; Coexistence ; Error propagation ; Model-based clustering ; Spatial data.

4.1 Introduction

The possibility that different assemblages of species abundances may represent stable alternative community states has been considered in several ecological studies and is referred to in the literature as the theory of multiple stable states. These alternative states can occur at various spatial and temporal scales.

Many examples of multitype interactions between species in natural ecosystems can be found in ecological literature, where we find variation among locations in the abundances of the coexisting species (le Roux & McGeoch, 2008). Possible explanations for having different interaction types between species are extrinsic environmental variations (not induced by the studied species) and the existence of alternative stable states of an ecological system. However, as pointed out by Peterson (1984), these explanations are sometimes difficult to distinguish. These multitype interactions can lead to different assemblages of species abundances.

Several studies show different assemblages of species abundances located side by side in apparently similar environment. Systems involving plant and animal communities, which show spatial structure due to multiple stable states have been studied by several authors. These include the studies of plant communities in semi-arid savannas by Walker et al. (1981) and Dublin et al. (1990), where the two alternative states are grass-dominated and woody-dominated territories. Seabloom & Richards (2003) studied the grassland communities structured by herbivores, the two stable states being annual- or perennial-dominated territories. The case of animal species was considered by Messier (1994) in a predator-prey model, namely the moose-wolf interactions over several observation sites in Canada, and a three-trophic-level system in Alaska (sea otters - invertebrate herbivores - macroalgae) with two stable points was studied by Estes & Duggins (1995). Peterson (1984) and Petraitis & Latham (1999) give further examples of alternative assemblages of species evolving side-by-side.

The authors of these ecological studies proposed models for the dynamics of the species which lead to alternative stable equilibria (Walker et al., 1981; Messier, 1994), or suggested ways of experimentally testing their existence (Seabloom & Richards, 2003; Estes & Duggins, 1995; Petraitis & Latham, 1999). Our approach is different in the sense that we present a statistical method that explores the existence of spatial assemblages of species abundances in a given dataset. Our method may help to identify and map significantly constrasted types of species assemblages. These assemblages might correspond to stable states but could also be transitory assemblages. Classes that are frequent in space are more likely to be stable states than the other classes. However, dynamical data would be required to determine if the classes are persistent.

Most of the existing methods for handling spatial data are based on the distance methodology reviewed by Diggle (2003) and Illian et al. (2008) and concern spatial point patterns, where individual positions are mapped for the studied species. Many of these analyse only spatial point patterns representing a single species, fewer are interested in describing the relationship between two point patterns. Bar-Hen & Picard (2006) reviewed the leading indices of association used in ecology. Few authors have tried to map the local interactions between species. Allard, Brix & Chadœuf (2001) addressed this problem by testing local independence between two point processes for the case in which the positions of individuals are known for each species. Perry & Dixon (2002) proposed a method for measuring globally and locally the spatial association for ecological count data measured at known spatial loca-

tions. Their method maps a continuous measure of the degree of association or dissociation at each observation site. Only two types of interaction are considered : positive or negative.

In this article we propose a new approach to explore the local interactions between species for spatially referenced abundance data of two species or more : the Classification And Spatial Analysis (CASA) method. It identifies types of species assemblages based on their abundances and provides a spatial analysis of these assemblages. Since we have no prior knowledge of the spatial structures of the types of assemblages, we separated the identification of the types of assemblages (which we obtain by a non spatial clustering procedure) from the spatial analysis of the assemblages. This two-stage analysis has to take into account the estimation error at each stage in order to provide accurate conclusions. Therefore, to obtain the estimation error due to the classification in the first stage we used a probabilistic clustering procedure that provides not only a partition of the sites into classes but also the probabilities of the sites to belong to each of these classes. These probabilities can thus be used in the second stage of CASA.

CASA method proceeds as follows :

In the first stage the sites are classified into types of species assemblages, using a non spatial clustering procedure based on two assumptions : the observations (i.e. the vectors of species abundances) are assumed to be independent, and to follow a multivariate normal mixture distribution. Each component of the mixture determines a cluster. The number of Gaussian components and their parameters are estimated using a maximum likelihood approach combined with a model selection criterion (Fraley & Raftery, 2002).

In the second stage of CASA, a spatial analysis of the types of species assemblages is performed. The tools we propose for this purpose are variograms of the variables defining the classes, with a permutation procedure to test whether the classes are spatially structured or not, and class to class distances, to study the pairwise relationship between classes. These tools take account of the uncertainty of the classification estimated in the first stage of CASA.

To illustrate CASA, we use a large-scale ecological dataset of the spatial distribution of powdery mildew infecting its wild host plant, *Plantago lanceolata*, in the Åland Islands in south-west Finland over the years 2001 to 2006 (Laine, 2004; Laine & Hanski, 2006). Figure 4.1 shows the abundances of the host plant and of the pathogen over the area, which was previously divided into a grid of contiguous cells of 1.5 km^2. The abundances were averaged over the 6 years of survey. We can see that the relationship between the abundances of the two species is not straightforward. We use CASA to define types of species assemblages and study their spatial distribution.

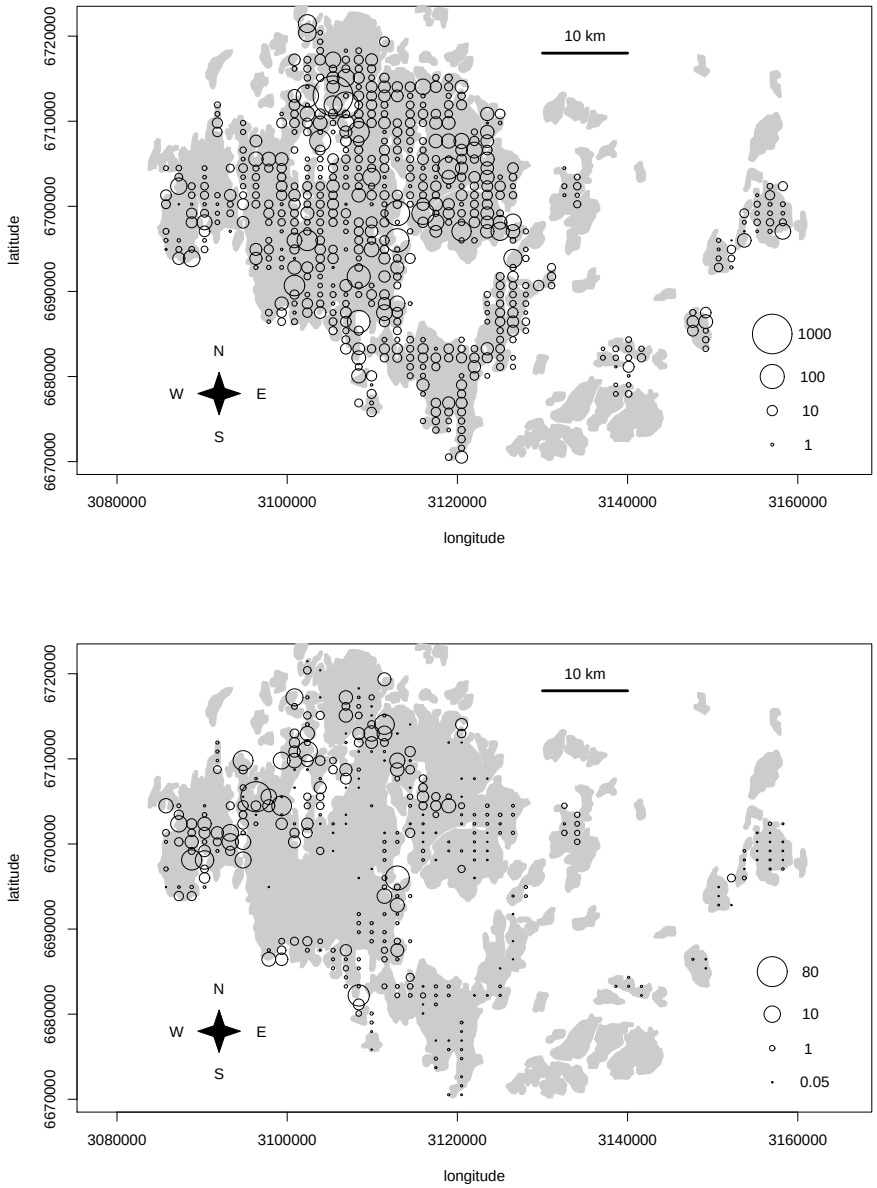


Fig. 4.1. Maps of the abundances of two species : a host plant and its pathogen. Top : Mean cover of *Plantago lanceolata* (host) over the survey years 2001-2006. Bottom : Mean incidence of powdery mildew (pathogen) over the 6 survey years. The size of the circles are proportional to the logarithm of the abundances.

This article is organized as follows. The CASA method is presented in Section 2. Its performance is evaluated on simulated data in Section 3. In Section 4 we use CASA to explore the host-pathogen interactions in the ecological dataset in the Åland Islands. Finally, a discussion is proposed in Section 5. The simulated datasets and the computer code is available in the Supplementary Material sectio on the journal's website.

4.2 The CASA Method

The Classification And Spatial Analysis (CASA) method provides a map of the various types of abundance assemblages, given the maps of the abundances of each species over the same observation sites, and a spatial analysis of these assemblages, i.e. how each assemblage is organized and how a given assemblage is organized with respect to the others.

The data we consider are species abundances measured at sites with specified location. The grid of the sites may be regular or not. The abundance variables are assumed to be continuous, e.g. measurements of species biomass, species coverage (i.e. proportions of an area occupied by the species).

Let X_{is} be the abundance of species $s \in \{1, \ldots, S\}$, at observation site $i \in \{1, \ldots, n\}$, where S and n are, respectively, the numbers of studied species and observation sites. We present below the classification of the abundance vectors $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{iS})$ into types of assemblages (Section 2.1), and the tools for the spatial analysis of the classes (Section 2.2).

4.2.1 Stage 1 : Classification into types of species assemblages

We define the types of species assemblages without any information of proximity or distance between sites, in order to avoid strong assumptions on the spatial structure of the assemblages.

We described the dependence structure between species abundances by assuming that the abundance vector $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{iS})$ is drawn independently for each site from a multivariate normal mixture distribution, up to a known transformation t :

$$t(\boldsymbol{X}_i) \sim \sum_{k=1}^{K} \tau_k \, \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

where $k \in \{1, \ldots, K\}$ labels the mixture components, the τ_k are the mixture proportions (with $\sum_{k=1}^{K} \tau_k = 1$), $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ are, respectively, the mean vector and the covariance matrix of the k^{th} component of the mixture. Each component corresponds to a type of species assemblage.

The problem of determining the types of species assemblages becomes a problem of model-based clustering. To model the covariance matrices, estimate the parameters of the mixture model, the posterior probability for each observation to belong to each component distribution and the number of components, we used the methodology presented by Fraley & Raftery (2002) based on multivariate normal mixture models.

Identifiability of finite mixture models has been proved for the Gaussian distribution family (see Teicher, 1963). McLachlan & Peel (2000) discuss the identifiability issues for mixtures of multivariate normal distributions. The mixture model is identifiable if :

$$\tau_k \neq 0 \quad (k = 1, \ldots, K)$$

and

$$(\boldsymbol{\mu}_h, \boldsymbol{\Sigma}_h) \neq (\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \quad (h \neq k = 1, \ldots, K).$$

A K-component normal mixture model with unrestricted component-covariance is a highly parameterized model. Various parameterizations for $\boldsymbol{\Sigma}_k$ are considered by Fraley & Raftery (2002). These parameterizations are based on eigenvalue decomposition, which can be interpreted in terms of shape, volume and orientation of the component distributions :

$$\boldsymbol{\Sigma}_k = \underbrace{\lambda_k}_{\text{volume}} \overbrace{\boldsymbol{D}_k}^{\text{orientation}} \underbrace{\boldsymbol{A}_k}_{\text{shape}} \boldsymbol{D}_k^T, \tag{4.1}$$

where $\lambda_k \boldsymbol{A}_k$ is the diagonal matrix of eigenvalues, and $\boldsymbol{D}_k$ is the matrix of eigenvectors for class k. A reduction in the number of parameters can be achieved by imposing various constraints on the geometric characteristics of the component distributions. They can be constrained to be the same for all clusters or may vary between clusters (see Supplementary Material section 4.6.3 for further details about the covariance models used).

The mixture parameters are estimated by an Expectation-Maximization (EM) algorithm initialized by a model-based hierarchical agglomeration algorithm (Dempster, Laird & Rubin, 1977; Tanner, 1996).

In general, the EM algorithm is used to estimate parameters in probabilistic models when the model includes unobserved latent variables. In our case, we have $(\boldsymbol{X}_i, \boldsymbol{Z}_i)$, where the $\boldsymbol{X}_i$ are the multivariate observations of species abundances at each site, and $\boldsymbol{Z}_i = (Z_{i1}, \ldots, Z_{iK})$ are the latent variables, with

$$Z_{ik} = \begin{cases} 1 \text{ if observation site } i \text{ belongs to the } k^{th} \text{ class} \\ 0 \text{ otherwise,} \end{cases}$$

and $P(Z_{ik} = 1) = \tau_k$. Let $\boldsymbol{C}_k = \{i = 1, \ldots, n : Z_{ik} = 1\}$ denote the k^{th} class.

The EM algorithm provides estimates $\hat{\boldsymbol{\theta}}_k = (\hat{\tau}_k, \hat{\boldsymbol{\mu}}_k, \hat{\boldsymbol{\Sigma}}_k)$ for the parameters $\boldsymbol{\theta}_k = (\tau_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, as well as the estimate $\hat{p}_{ik}$ of the conditional probability $p_{ik} = \mathrm{P}(Z_{ik} = 1 | \boldsymbol{X}, \boldsymbol{\theta})$ that i belongs to class $\boldsymbol{C}_k$ given the data and the parameters :

$$\hat{p}_{ik} = \frac{\hat{\tau}_k \; f_k(\boldsymbol{X}_i | \hat{\boldsymbol{\mu}}_k, \hat{\boldsymbol{\Sigma}}_k)}{\sum_{j=1}^{K} \hat{\tau}_j \; f_j(\boldsymbol{X}_i | \hat{\boldsymbol{\mu}}_j, \hat{\boldsymbol{\Sigma}}_j)}.$$

To select the parameterization of the model and the number of clusters, the Bayesian Information Criterion (BIC) is used :

$$BIC_m = 2\mathcal{L}_m^\star(\boldsymbol{X}, \boldsymbol{Z}) - \nu_m \log(n), \tag{4.2}$$

where $\mathcal{L}_m^\star(\boldsymbol{X}, \boldsymbol{Z})$ is the maximized loglikelihood of the model m and the data, ν_m is the number of independent parameters estimated for the model m, and n the number of observation sites.

A classification of the observations into clusters (denoted by the binary variables $\hat{Z}_{ik}$) is obtained by the maximum *a posteriori* (MAP) method, which simply affects each observation site i to the cluster that maximizes the probability $\hat{p}_{ik}$ for all $k \in \{1, \ldots, K\}$:

$$\hat{Z}_{ik} = \begin{cases} 1 \text{ if } \hat{p}_{ik} = \max\limits_{j \in \{1,\ldots,K\}} \hat{p}_{ij} \\ 0 \text{ otherwise.} \end{cases}$$

The `R package MCLUST` proposed by Fraley & Raftery (2006) was used to run the clustering procedure.

4.2.2 Stage 2 : Spatial analysis of the types of assemblages

We present in this section some of the specific tools used to study the spatial structure of the classes of assemblages identified, as well as the spatial relationships between classes. Variograms were used for the study of single classes, and class to class distances for the pairwise relationships between classes. The first stage of the CASA method provides a classification $\{\hat{Z}_{ik} : i = 1, \ldots, n; k = 1, \ldots, K\}$ of the sites in terms of types of assemblages and the probabilities $\{\hat{p}_{ik} : i = 1, \ldots, n; k = 1, \ldots, K\}$ for the sites to belong to each of the K classes found. The probabilities $\hat{p}_{ik}$ are used to transfer the classification uncertainty to the spatial analysis.

The CASA method is intended for exploratory studies, so the tools for the spatial analysis of the abundance assemblages can vary and should be adapted to the hypotheses one wants to test. Emphasis is on the propagation of the classification error to the spatial analysis.

Spatial structure of single classes : Variograms

To describe the spatial pattern of the estimated class $\hat{\boldsymbol{C}}_k$, defined by $\hat{\boldsymbol{Z}}_k = \{\hat{Z}_{ik} : i = 1, \ldots, n\}$, the sample variogram is used (Chilès & Delfiner, 1999) :

$$\hat{\gamma}_k(d) = \frac{1}{2n_d} \sum_{\{i,j : d(i,j) \backsimeq d\}} (\hat{Z}_{ik} - \hat{Z}_{jk})^2$$

where d denotes the distance and n_d the number of pairs of observation sites (i, j) placed at an approximate distance d. The experimental variogram is implemented with a certain tolerance around d, which defines classes of distances.

In order to take into account possible misassignments to the class $\boldsymbol{C}_k$, a confidence interval for the variogram of $\hat{\boldsymbol{Z}}_k$ is calculated by drawing $q = 1000$ sets $\boldsymbol{Z}_k^r = \{Z_{ik}^r : i = 1, \ldots, n\}$, with $r \in \{1, \ldots, q\}$, from a Bernoulli distribution :

$$Z_{ik}^r \sim \mathscr{B}ernoulli(\hat{p}_{ik}).$$

The sample variograms of the simulated $\boldsymbol{Z}_k^r$ are used to obtain a 95% confidence interval for $\hat{\gamma}_k(d)$. In a second step, random permutation tests (Manly, 1997) are used to assess whether the spatial pattern of the class $\hat{\boldsymbol{C}}_k$ is structured or not. Permutation of the $\{\hat{Z}_{ik} : i = 1, \ldots, n\}$ among all the sites i would not take into account the classification error (due to the variability of the MAP decision rule). Therefore, we repeat 1000 times the following procedure for a given class $\hat{\boldsymbol{C}}_k$: First, we permute randomly the probabilities

$\{\hat{p}_{ik} : i = 1, \ldots, n\}$ among all the sites i, secondly, we draw the randomized classes $\boldsymbol{Z}_k^{r\star} = \{\hat{Z}_{ik}^{r\star} : i = 1, \ldots, n\}$ from a Bernoulli distribution :

$$\hat{Z}_{ik}^{r\star} \sim \mathscr{B}ernoulli(\hat{p}_{ik}^{r}),$$

where $\hat{p}_{ik}^{r}$ are the randomly permuted $\hat{p}_{ik}$, and thirdly we calculate the variograms of the permuted classes $\boldsymbol{Z}_k^{r\star}$. A 95% confidence band is then calculated for the variograms of $\boldsymbol{Z}_k^{r\star}$, corresponding to the randomized patterns of the class $\hat{\boldsymbol{C}}_k$.

The confidence band of the variogram of the k^{th} class (built using the $\boldsymbol{Z}_k^r$) is then compared to the confidence band of the variogram of the randomized pattern of $\hat{\boldsymbol{C}}_k$ (built on the basis of the $\boldsymbol{Z}_k^{r\star}$). The classes for which the confidence band of the variograms of the $\boldsymbol{Z}_k^r$ goes outside of the confidence band of the $\boldsymbol{Z}_k^{r\star}$ are considered spatially structured.

Spatial organization between classes : Point-to-set distances

The spatial organization between classes deserves a detailed study, because it gives information on the underlying interactions between species. Abrupt changes in the abundances of the two species (a class of low abundances close to a class of high abundances) might indicate a relevant environmental variable, which influences the interactions between species. Conversely, a gradual spatial decrease of the abundances would rather indicate migration (i.e. populations that spread around the sources), or a gradual variation of an environmental factor.

For this analysis we used a point-to-set distance : the mean distance from the points of the class $\hat{\boldsymbol{C}}_k$ to the set $\hat{\boldsymbol{C}}_l$, defined by :

$$\hat{D}_l(k) = \frac{1}{\#\hat{\boldsymbol{C}}_k} \sum_{i \in \hat{\boldsymbol{C}}_k} \min_{j \in \hat{\boldsymbol{C}}_l} d(i,j) \tag{4.3}$$

where $\#\hat{\boldsymbol{C}}_k$ is the number of sites belonging to class $\hat{\boldsymbol{C}}_k$ and $d(i,j)$ is the euclidean distance between sites i and j.

To take account of the variability of the estimations $\hat{\boldsymbol{C}}_k$ and $\hat{\boldsymbol{C}}_l$, the class of a site i is simulated according to a multinomial distribution with probabilities $\hat{p}_{i1}, \ldots, \hat{p}_{iK}$. More precisely, for $r \in \{1, \ldots, q\}$, with $q = 1000$, $\boldsymbol{Z}_i^r = (Z_{i1}^r, \ldots, Z_{iK}^r)$ is drawn from a multinomial distribution with parameters 1 and $\hat{\boldsymbol{p}}_i = (\hat{p}_{i1}, \ldots, \hat{p}_{iK})$:

$$\begin{cases} \boldsymbol{Z}_i^r \sim \mathscr{M}ultinomial(1, \hat{\boldsymbol{p}}_i) \\ C_k^r = \{i | Z_{ik}^r = 1\} \\ C_l^r = \{i | Z_{il}^r = 1\} \end{cases}$$

Then, for each iteration r, we calculate $D_l^r(k)$, according to formula (4.3), to obtain an empirical distribution $\{D_l^r(k) : r = 1, \ldots, q\}$ for the point-to-set distance from class $\boldsymbol{C}_k$ to class $\boldsymbol{C}_l$. This distribution can be used to describe the spatial co-organization between classes.

Remark : Other distance statistics can be constructed, for example the point-to-point distance :

$$\frac{1}{\#\hat{\boldsymbol{C}}_k}\frac{1}{\#\hat{\boldsymbol{C}}_l}\sum_{i\in\hat{\boldsymbol{C}}_k}\sum_{j\in\hat{\boldsymbol{C}}_l} d(i,j)$$

and the set-to-set distance :

$$\min_{i\in\hat{\boldsymbol{C}}_k, j\in\hat{\boldsymbol{C}}_l} d(i,j).$$

Here, we proposed the asymmetric point-to-set distance $D_l(k)$ because it allows us to show interesting features of the studied data sets.

4.3 Simulation Study

To illustrate the behaviour of CASA, we simulated the spatial distribution of two species with multitype interactions on a 25×25 grid. This example shows that our method is capable of capturing different types of abundance assemblages from the study of the joint density of abundances if the hypothesis of normal mixture distribution is satisfied. The maps of these classes can be helpful to infer underlying mecanisms (species dynamics, environment) by studying the spatial structure of species assemblages and the relationships between them. Another example is given in Supplementary Material section 4.6.1.

4.3.1 Three modes of migration

Three types of abundance assemblages are modeled in this example : Low abundances for both species, corresponding to areas weakly colonized (low-low abundance assemblage), high abundances for both species, corresponding to areas where both species are well established (high-high), high abundances for species 1 associated with low abundances for species 2 (high-low).

Areas with low abundances for both species can be interpreted as areas where the environment is poor, and species 1 can be seen as a pioneer species which colonizes the poor areas, modifies the environment locally and leads the way for species 2. So species 2 depends on species 1. The assemblage where only species 1 is abundant corresponds then to the areas newly colonized by species 1, but in which species 2 has not developed yet, or the environment has not yet become favorable for species 2. The assemblage where both species are abundant corresponds to areas with favorable environment, from which species 1 can migrate (or be dispersed in the case of plants).

We simulated the migration of species 1 in three different ways : Migration with a dispersal kernel (Case 1), uniform dispersal (Case 2), and uniform and aggregated migration (Case 3).

An ecological interpretation in terms of seed dispersal strategies can be given for the considered cases. The three cases could correspond respectively to dispersal by gravity (barochory), wind (anemochory) and animals (zoochory) ; see for example Jordano et al. (2007) and Muller-Landau et al. (2008).

Model

Let $\boldsymbol{X}_i = (X_{i1}, X_{i2})$ be the abundances of two species observed at site i ($i \in \{1, \ldots, n\}$, $n = 625$). Conditional on the states of two hidden variables $\{E_i : i = 1, \ldots, n\}$ and $\{H_i : i = 1, \ldots, n\}$, the abundances $\{\boldsymbol{X}_i : i = 1, \ldots, n\}$ are mutually independent and drawn from the lognormal distributions :

$$\boldsymbol{X}_i | E_i, H_i \sim \mathscr{LN}\left(\boldsymbol{\mu}(E_i, H_i), \sigma^2 \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right).$$

E_i and H_i are both binary variables and have the following meaning : H_i indicates the sites in which both species are abundant, and E_i indicates the sites where only species 1 is abundant.

Here, $\sigma^2 = 0.3$ and the mean vector, $\boldsymbol{\mu}(e, h)$, of the conditional distribution of the $\boldsymbol{X}_i$ is defined over $\{0, 1\} \times \{0, 1\}$ and satisfies :

$$\boldsymbol{\mu}(e, h) = \begin{cases} (1, 1) & \text{if } h = 1 \\ (0, 0) & \text{if } h = 0 \text{ and } e = 0 \\ (1, 0) & \text{if } h = 0 \text{ and } e = 1 \end{cases}$$

where e and h denote the realized values of the random variables E and H.

H_i is obtained by truncating a spatial Gaussian process :

$$H_i = \mathbb{I}\{H_0(i) > 0.7\}$$

where $\mathbb{I}$ is the indicator function, H_0 is a Gaussian field with a mean vector of zeros of length n, $\boldsymbol{\mu_0}$, a covariance matrix $\boldsymbol{\Sigma_0} = 0.5\ e^{-(25\ \boldsymbol{D})^2/5}$ and $\boldsymbol{D}$ is the matrix of distances among sites [1] : $H_0 \sim \mathscr{N}(\boldsymbol{\mu_0}, \boldsymbol{\Sigma_0})$.

The variable E_i is simulated in three different ways over the considered area, corresponding to three different migration strategies for species 1.

Case 1.

Migration with a dispersal kernel

In this case the migration of species 1 depends on the distance from the sources (high-high abundance assemblage). For example, in the case of plants, this could correspond to dispersion of the seeds by gravity (barochory).

The variable E_i is obtained by : $E_i = \mathbb{I}\{E_0(i) \geq 1\}$, where E_0 is simulated conditionally on H :

$$E_0(i) \sim \mathscr{P}oisson(\lambda_i)$$

[1] The distances between the centres of the sites are considered. Their coordinate values are within the range [0,1] on the two coordinate axes, thus the minimum distance between 2 different sites equals 0.04.

$$\lambda_i = \sum_j H_j \Phi(d(i,j))$$

with $d(i,j)$ the distance between the sites i and j, and Φ the Gaussian dispersal kernel (Soubeyrand et al., 2008) defined by :

$$\Phi(d(i,j)) = \frac{3}{2\pi} e^{-\frac{d(i,j)^2}{2}}$$

E_0 can be interpreted as the number of colonization events. The sites where $H = 1$ (high-high abundance assemblage) are considered to be sources of dispersion of the species, so E_0 will take higher values when it is near to a source. $E = 1$ indicates the presence of colonists of species 1 at the sites where $H = 0$ (it corresponds to the high-low abundance assemblage).

Case 2.

Uniform migration

Here, the migration of species 1 is independent of the distance from the sources. If we consider once more the example of plants, this could correspond to dispersion of the seeds by wind (anemochory).

This is achieved by randomly permuting the cells i where $E_i = 1$ and $H_i = 0$.

Case 3.

Uniform and aggregated migration

Here, species 1 migrates in the form of aggregates, whose positions are independent of the distance from the sources. This could correspond to dispersion of plant seeds by animals (zoochory).

To achieve this, another Gaussian field H' is simulated for the cells where $H = 0$:

$$H'_i | H_i = 0 \sim \mathcal{N}(\boldsymbol{\mu_0}, \boldsymbol{\Sigma_0})$$

E is replaced by E' :

$$E' = \mathbb{I}\{H' \geq H'_\alpha\}$$

where H'_α is the r^{th} greatest value of H' such that $r = \sum_{i=1}^n E'_i = \sum_{i=1}^n E_i$. In other words the number of cells occupied by the new settlers of species 1 (high-low abundance assemblage) is kept equal to the value in the first situation.

In the three cases the joint distribution of the abundances of the two species $(\boldsymbol{X}_1, \boldsymbol{X}_2)$ remains the same. Only the spatial repartition of the areas newly colonized by species 1 (high-low abundance assemblage) changes.

Results

The results of a simulation of this model are represented in Figure 4.2 and Figure 4.3.

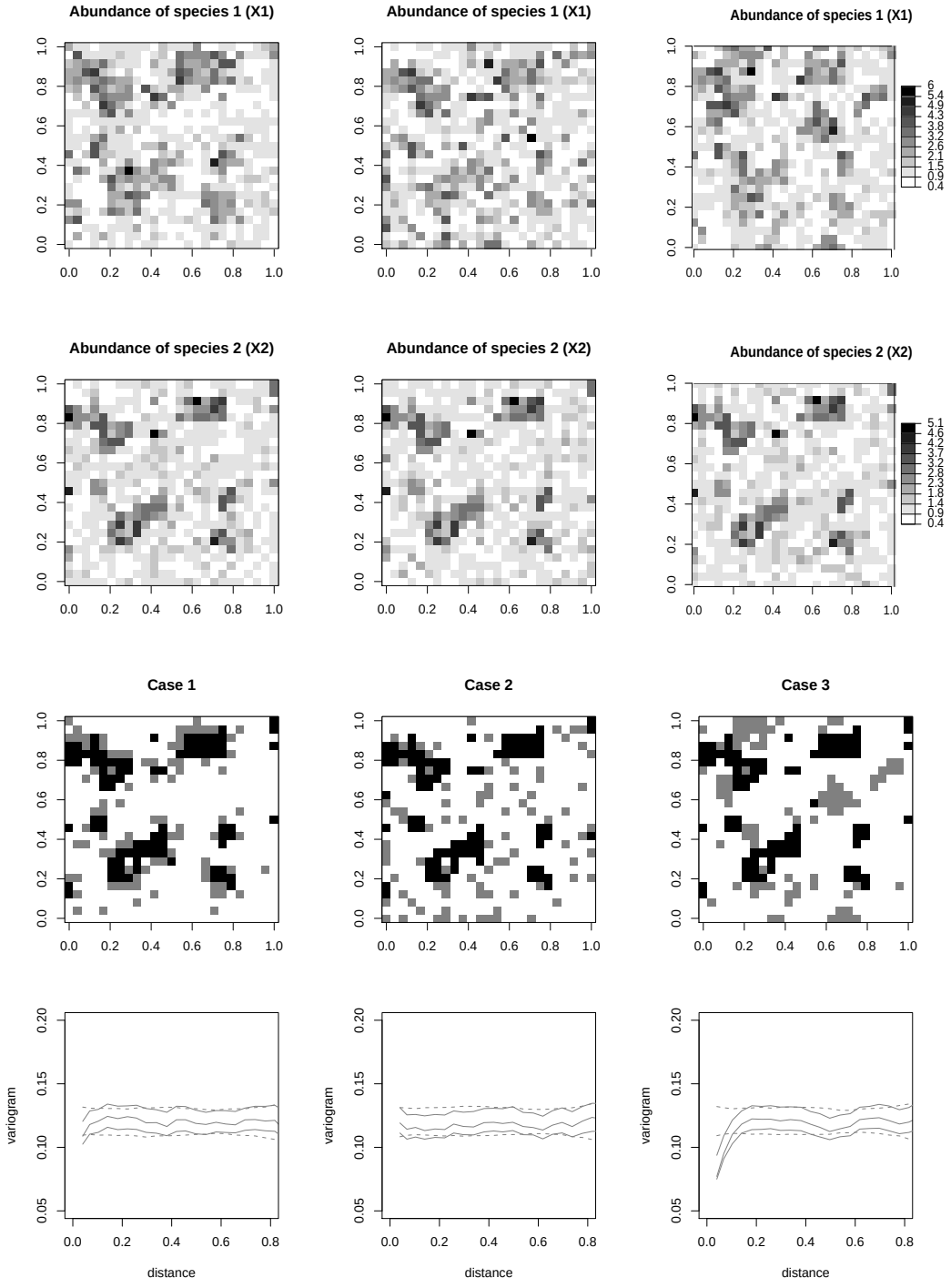


Fig. 4.2. Results of CASA on the simulated dataset. Each case of migration of species 1 is represented on a different column. From top to bottom : Maps of the abundances for each species, map of the 3 types of assemblages, variograms of the grey class, corresponding to the spread of species 1. The variograms show that the grey class is structured only in case 3.

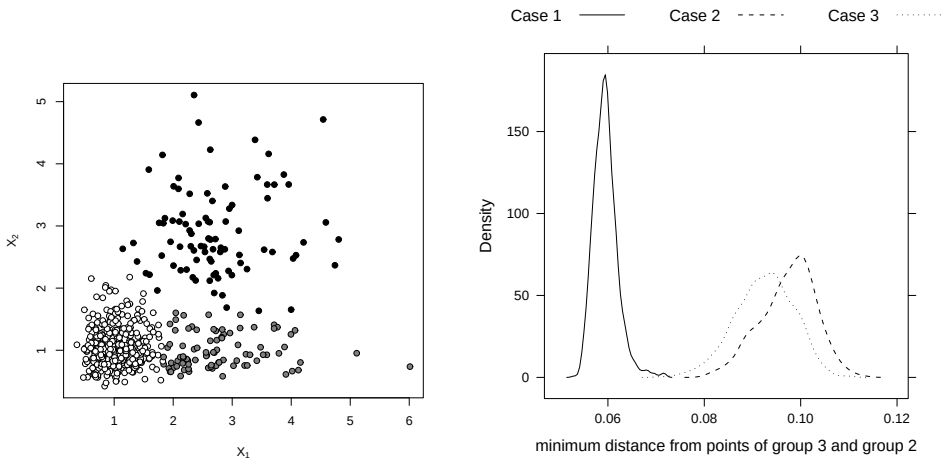


Fig. 4.3. Results of CASA on the simulated dataset. Left : Classification on the scatterplot of $(\boldsymbol{X}_1, \boldsymbol{X}_2)$. Right : Comparison of the minimum distance between the grey and the black class in the 3 cases.

Classification

The clustering procedure is performed here on the logarithm of the abundances, which validate the hypotheses of the clustering procedure (by construction), namely that the observations are independent and follow a normal mixture distribution. A 3-cluster model with an equal-volume, spherical variance $\boldsymbol{\Sigma}_k = \lambda \boldsymbol{I}$, where $\boldsymbol{I}$ is the identity matrix and λ a scalar estimated by the EM algorithm, is chosen by the BIC. The 3 types of assemblages are correctly identified : Low-low abundance assemblage in white, high-high abundance assemblage in black, and high-low abundance assemblage in grey.

The uncertainty u of the classification can be calculated at each site i by (Fraley & Raftery, 2006) :

$$u(i) = 1 - \max_k \hat{p}_{ik}$$

In this simulation the uncertainty is generally low, less than 0.1 for 90% of the sites, and 96% of the sites are correctly classified.

Spatial analysis of the classes

There is no difference in the classification results for the 3 simulated cases but for the map of the classes of assemblages, which differ by the spatial pattern of the grey class. As expected, the variograms of the black and the white class (not represented here) show the aggregation of these classes in the 3 cases. The variograms of the grey class do not detect any spatial structure in the first two cases, whereas they show the aggregation of the class in the third case. But the grey class is structured around the black class in the first case (by construction). Although this is visible on the map of the assemblages Figure 4.2, case

1, the analysis based on variograms does not suffice to describe the structure of the grey class in case 1.

To study the relationship between the grey class (colonists of species 1) and the black class (sources of migration) we computed the minimum distance between these classes (as presented in equation 4.3) for the 3 cases. Figure 4.3 shows that the minimum distance between the two classes is significantly smaller in the first case than in the other two. This is confirmed by a Wilcoxon rank-sum test (p-value $\ll 10^{-6}$). We can conclude that the grey class is structured around the black class in the first case.

One might argue that 3 datasets containing the same types of assemblages but differently organized in space are rarely available in real situations and therefore the use of point-to-set distances for comparative studies would rarely be possible. Other interpretations of this simulated example could be given in which the point-to-set distance would be used in the same way. For example the data simulated as in case 1 could be the unique available dataset, and case 2 could be constructed from case 1 to represent the null hypothesis of "no structure of the grey class".

4.4 Case-study

The CASA method, which has to be viewed as a tool for preliminary and exploratory analysis, is applied here on a previously studied dataset on host-pathogen interactions (Laine, 2004 ; Laine & Hanski, 2006 ; Soubeyrand et al., 2009). This allows us to check the relevance of the results obtained with the CASA method.

4.4.1 The dataset

This dataset deals with the large-scale spatial distribution of a fungus, the powdery mildew (*Podosphaera plantaginis*), infecting its wild host plant, *Plantago lanceolata*, in the Åland Islands in south-west Finland. Each year in September from 2001 to 2006, approximately 3000 meadows in these islands were surveyed : in each meadow, the coverage of *P. lanceolata* in square meters and the presence/absence of the powdery mildew were assessed. Based on this dataset, the spatiotemporal dynamics of the pathogen, consisting of frequent extinctions and (re-)colonizations of meadows, was analyzed using a GLMM approach (Laine & Hanski, 2006) and a mechanistic-statistical approach (Soubeyrand et al., 2009).

Here we studied the assemblages of powdery mildew and *Plantago lanceolata* and their organization in space using the CASA method and an aggregated version of the dataset. The dataset is aggregated in space and time to get (pseudo-)continuous measures of the two species of interest. We divided the area into a grid of contiguous cells with a side length of 1.5 km and calculated, for each cell, the temporal and spatial mean of the host coverage over the meadows within the cell, and the temporal and spatial mean of the presence/absence of the pathogen in the meadows. 596 cells were occupied by the host species, among which 343

were also occupied by the pathogen. Figure 4.1 shows the spatial patterns of the pathogen and the host.

4.4.2 Results

Figure 4.4 shows the results of CASA on this dataset.

Classification

The analysis of the types of abundance assemblages is restricted to the 343 cells where the host and the pathogen were both present. The clustering procedure is performed on the two species abundances with no prior transformation (unlike the simulated examples). A five-classes model is selected with the BIC, with a covariance structure whose shape depends on the class ($\boldsymbol{\Sigma}_k = \lambda_k \boldsymbol{A}_k$; see equation 4.1).

The classification is given in Figure 4.4. The dark blue class corresponds roughly to the most infected sites (with some exceptions in which the pathogen incidence is low), followed by the green class and the black one. The red and the cyan classes contain the less infected sites.

Spatial analysis of the classes

The map Figure 4.4 shows the spatial distribution of the classes over the contours of the Åland islands. The classes seem to be structured around the shorelines (the non infected meadows are generally in the inlands). Moreover, there is a clear tendency for the most infected sites to be in the north-west part of Åland (dark blue and green classes), while the south-east contains almost exclusively sites with low incidence of the pathogen (red and cyan classes).

The spatial analysis of the classes using variograms (not represented here) only reveals the spatial aggregation of the red class, which is the class containing most observation sites.

We proposed two additionnal tools for the spatial analysis of the classes for this specific dataset. These tools take into account the classification error in a way similar to the methods presented in Section 2.2 and are used to test two hypotheses : First, the existence of a north-west to south-east gradient of infection, and secondly the existence of a spatial stability within classes (at the scale of the nearest neighbours).

To test the first hypothesis we calculated the centers of gravity of each class. In order to include the classification error in this analysis, each class was drawn from a multiple Bernoulli distribution with success probability equal to the $\hat{p}_{ik}$ and its center of gravity was calculated. The center of gravity was calculated for 1000 simulations of each class. The

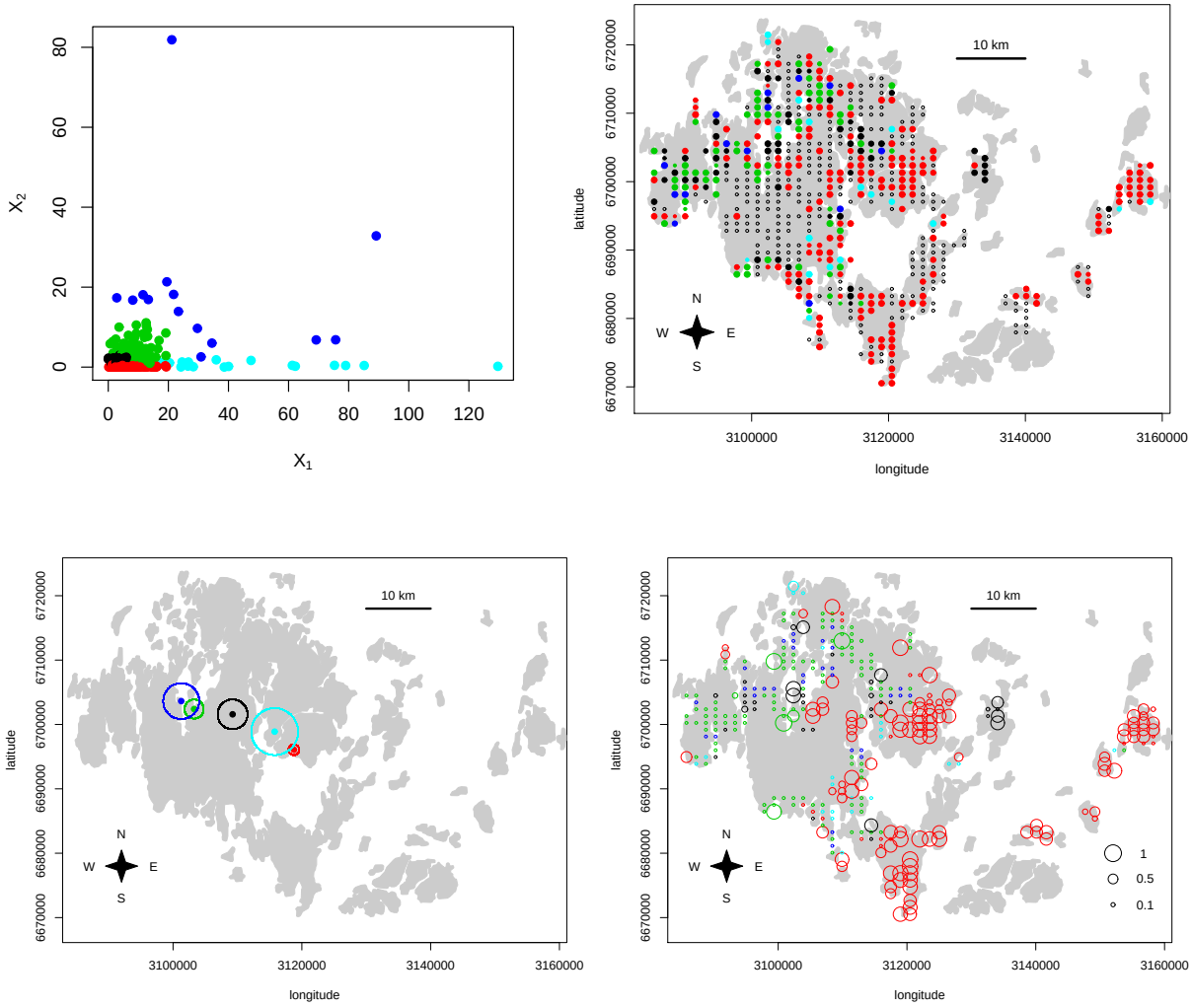


Fig. 4.4. Results of CASA on the case-study. Top left : Classification into 5 types of assemblages on the scatterplot of the abundances of two species in a host-pathogen system. (X_1 = mean cover of *P. lanceolata* (host) per site over the 6 years, X_2 = mean incidence of powdery mildew (pathogen) per site over the survey years 2001-2006.) Top right : Map of the types of assemblages of host-pathogen. The filled, colored circles represent the 5 types of assemblages. The size of the circles is inversely proportional to the uncertainty of the classification for each site. The empty circles represent sites that were not used in the clustering procedure because the pathogen is absent. Bottom left : Center of gravity of each of the 5 types of assemblages. The mean center of gravity and its 95% confidence interval are represented for each class. Bottom right : Diversity of the nearest neighbours. The color of the circles represent the most probable class for the site considering its nearest neighbours (less than 1.6 km away ; maximum 4 neighbours), and the size of the circles is proportional to the probability of being of the same color for the cell and its nearest neighbours.

mean center of gravity and the 95% confidence interval is represented on the map of Åland in Figure 4.4 (bottom left).

The infection gradient from the most infected classes (dark blue and green classes) in the North-West to the less infected (red and cyan classes) South-East appears clearly in this figure. The center of gravity of the red class is the most stable. As expected, the one of the cyan class is the most variable, since this class is the most scattered of all.

To test the spatial stability of the classes we calculated for each site i the probability for i and its nearest neighbours to belong to the same class k. This was done for all sites i and all classes k. Since we assumed that the probability for a site to belong to a class k is independent of the other sites' classes, the probability of i and its nearest neighbours, $\boldsymbol{V}_i = \{j : j \neq i, d(i,j) \leq 1.6 \text{ km}\}$, to belong to the same class k is simply the product of the probabilities :

$$\prod_{j \in \{i\} \bigcup \{\boldsymbol{V}_i\}} p_{jk}$$

The sites with no nearest neighbours are not considered in this study. In the cases where site i has less than 4 neighbours, we considered that its neighbouring sites (1, 2 or 3) are representative of the missing neighbours. Therefore we calculated the following measure of stability at each site :

$$M_i = \max_{k=1,\ldots,K} \left(\prod_{j \in \boldsymbol{V}_i} p_{jk} \right)^{4/m} p_{ik}$$

where $m \in \{1, 2, 3, 4\}$.

In Figure 4.4 (bottom right), the class k, corresponding to the maximum probability, on all classes $k \in 1, \ldots, K$, that a site and its nearest neighbours belong to k, is represented at each site. If we compare this map to the map above (the original classification Figure 4.4 top right) we can see that the classes are stable at a larger scale. In comparison to the original classification, in this map, the classes of assemblages appear more aggregated. This is especially the case for the red, green and blue class. The red class, corresponding to low infection rates, is the most stable class (probabilities M_i close to 1). Although the probabilities M_i are low for the green class, this class is more structured on this map and many sites previously affected to the red class in the North-West part of Åland become green. The gradient of infection from the North-West to the South-East appears more clearly here than in the original classification. This procedure can be seen as a sort of smoothing that gives more structure and coherence to the classes.

From an ecological and epidemiological perspective, the use of the CASA method allowed us to demonstrate that host abundance alone is not a good predictor of where the pathogen will occur. The spatial structuring suggests that correlated environmental conditions (possibly humidity and temperature) may be important for determining where the

pathogen is most likely to occur. It is also possible that host quality (i.e. resistance to the pathogen) is regionally aggregated (for example due to restricted seed dispersal of the host plant) generating areas where the pathogen incidence is low (the red and cyan classes). These results generate several interesting hypotheses to be tested in order to separate the effects of the physical environment and the effects of host resistance in determining the types of abundance assemblage of this host-pathogen system. A drawback of the CASA method here is the loss of information on the individual sizes of the meadows that results from the aggregation of the data in time and space (into 1.5 x 1.5 cells). Cells consisting of many small meadows might qualitatively be very different for the pathogen than cells with fewer, large host populations.

4.5 Discussion

The method we presented in this article can be useful to explore local spatial interactions between species, for data where the abundances of several species are available over spatially referenced observation sites. CASA starts by performing a probabilistic clustering procedure on the joint density of the species abundances resulting in a map of the various types of abundance assemblages. The relationships between these types of assemblages and their spatial structure are then analyzed by adapting specific tools for this purpose. These tools should be suited for the hypotheses to be tested and should take into account the uncertainty in classifying the observations. The advantage of using a probabilistic clustering procedure is to have an estimate of the probabilities p_{ik} of each observation i to belong to each class k. These are used to calculate a classification uncertainty which can be transferred to any further spatial analyses to obtain an accurate estimate of the overall error of CASA. However, it is to be noted that the uncertainty of the estimation of p_{ik} by the EM algorithm is not transferred to the second stage.

We chose a clustering procedure based on multivariate Gaussian mixtures for the first stage of CASA, because it is the most commonly used model for probabilistic clustering of continuous data. However, this model-based clustering procedure assumes that the observations are independent and follow a normal mixture distribution. Often these conditions do not hold in practice.

The independence assumption is generally not satisfied for spatial data. We deliberately ignore the spatial autocorrelation of the observations in the classification stage of CASA in order to gain flexibility by avoiding hypotheses about the spatial structure of the classes. An alternative to the clustering method we use would be a spatial clustering method. There are several spatial clustering methods, such as those proposed by Osaragi (2002); Guillot et al. (2005). Ambroise, Dang & Govaert (1996) even proposed a Neighbourhood Expectation Maximization algorithm (NEM) for model-based clustering. They introduce a spatial penalty term in the EM criterion in order to favor geographically homogeneous classes. However, without any information about the spatial structure of the types of species assemblages at the spatial scale we consider, we have no way of knowing to what extent adjacent

cells should be preferentially assigned to the same class and therefore no indication on the choice of the penalizing factor, which is the main difficulty in the NEM and other spatial clustering methods. The lack of spatial constraints for the classes can be very convenient for exploratory studies in which there is no prior knowledge about the structure of the assemblages. For example in the case of fragmented habitat, the types of assemblages will be defined independently of the space fragmentation.

The Gaussian mixture assumption cannot be verified *a priori* on the data. A goodness-of-fit test could be used to check *a posteriori* if the data comes from the multivariate normal mixture distribution estimated in the clustering stage of CASA, even though it is self-serving to cluster the data under the Gaussian assumption and test it afterwards, as mentioned by McLachlan & Peel (2000), page 84. The multivariate Kolmogorov-Smirnov goodness-of-fit test (Justel, Peña & Zamar, 1997), the generalized Cramér-von Mises test (Chiu & Liu, 2009) and the Chi-squared goodness-of-fit test can be used for this purpose. However, these tests will lack power if the data sample contains only a few observations for some of the component densities.

The Gaussian mixture assumption is rarely checked in practice, since any continuous distribution can be approximated arbitrarily closely by a finite mixture of Gaussian densities (Dasgupta, 2008, chapter 33). In case the distributions of the underlying groups are not Gaussian, the common assumption that the number of components of the mixture corresponds to the true underlying number of groups in the population will no longer hold (Dean & Nugent, in prep.). For example the Gaussian mixture will have the tendency to fit a curvilinear shaped cluster by several Gaussian components (Fraley & Raftery, 2002). If the shapes of the underlying groups in the data are non elliptical, asymmetric or have heavy tails this clustering procedure will most probably overestimate the number of groups in the data. In our case this would not lead to severe consequences, since the observations which have been assigned to different Gaussian components while they were in fact in the same underlying group can be spotted during the spatial analysis in the second stage of CASA (because they will probably be in the same regions).

Nevertheless, to overcome this problem, several solutions are available. A variable transformation, such as the logarithm transformation, can be helpful when the observations are very nearly colinear, or if the data is asymmetrical. Another solution is to use a model selection criterion more robust to the violation of this hypothesis than the BIC. For example Biernacki, Celeux & Govaert (2000) showed that the Integrated Classification Likelihood (ICL) is more robust than the BIC and provides better classifications in some cases (see Supplementary Material section 4.6.2). In this case, however, the uncertainty of the classification is still estimated under the Gaussian mixture. Since the propagation of the classification uncertainty to the spatial analysis is a key step in our method, clustering based on normal mixtures should not be used for data that are clearly not Gaussian.

Other mixture models can be used to classify the observation sites into types of abundance assemblages (McLachlan & Peel, 2000), especially in the case of count data, which are frequent in ecology. For discrete data, discrete mixture models should be used, such as

mixtures of multivariate multinomial distributions (Everitt, 1984) and multivariate Poisson distributions (Karlis & Meligkotsidou, 2006).

If prior knowledge about the spatial structure of the community is available, the clustering procedure using mixture distributions can be combined with an approach that takes account of the spatial nature of the data. For example Karunanayake & Laverty (2006) proposed a multivariate Poisson mixture with a hidden Markov model to model the spatial assemblages of counts of three weed species.

CASA can be used in ecological studies as a tool for a preliminary, exploratory study of spatial data of abundances of several species. CASA can contribute to the understanding of the interaction types between species, and assess whether these interaction types are linked to the environment or to the species dynamics. For example it could be used to determine which environmental variable explains best the distribution of the types of assemblages. It can also be helpful for the choice of a dynamical model, by answering questions such as : Should a spatial model be used ? Is it interesting to consider models with multiple equilibria ?

Furthermore, CASA is not specific to ecology and has potential applications beyond it, e.g. in genetics and social sciences. An illustration of CASA on image analysis is available in the Supplementary Material section 4.6.3.

Acknowledgement

We are grateful to our collegues of the "Biostatistique et Processus Spatiaux" unit at the "Institut National de la Recherche Agronomique" (INRA), and in particular to Rachid Senoussi, Etienne Klein and Florence Carpentier for useful discussions and comments. We also thank Leonhard Held, an associate editor and an anonymous referee for their constructive suggestions on an earlier version of this manuscript. This work was supported by the INRA and the French region Provence Alpes Côte d'Azur.

Conflict of Interests Statement

The authors have declared no conflict of interest.

4.6 Supplementary Material

This supplementary material provides additional information on the Classification And Spatial Analysis (CASA) method. CASA is applied here on two different datasets. An application on a simulated dataset is presented. Then an extension of CASA in the field of image analysis shows that CASA can give coherent results on different type of data, and on trivariate assemblages. We give some details on the covariance models used in the classification stage of CASA in an Appendix. Finally, the simulated datasets and the computer code of CASA implemented in R are given.

4.6.1 Simulation Study : A Model with a Hidden Interacting Species

In this example the distribution of two species is constrained by :
- a heterogeneous environment, which defines unfavorable areas for both species,
- the interaction with a third species, whose repartition is hidden and the presence of which defines favorable zones for species 1 and unfavorable zones for species 2

Model

Let $\boldsymbol{X}_i = (X_{i1}, X_{i2})$ be the abundances of the two species observed at site i ($i \in \{1, \ldots, n\}$, $n = 625$). Conditional on the states $\{E_i : i = 1, \ldots, n\}$ of the environment and the states $\{H_i : i = 1, \ldots, n\}$ of a hidden species interacting with the two species of interest, $\{\boldsymbol{X}_i,\ i = 1, \ldots, n\}$, are mutually independent and drawn from the lognormal distributions :

$$\boldsymbol{X}_i | E_i, H_i \sim \mathscr{LN}\left(\boldsymbol{\mu}(E_i, H_i), \sigma^2 \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right) \tag{4.4}$$

The variance component of the model does not depend on the environment and the hidden species. The variance parameter σ^2 is fixed at 0.25. The mean component of the model depends on the environment and on the hidden species in the following way : E_i is either 0 or 1, corresponding respectively to unfavorable and favorable environmental conditions, and H_i takes values in $\{0, 1, 2\}$, which correspond respectively to low/intermediate/high abundances of the hidden species. The mean vector $\boldsymbol{\mu}(e, h)$ defined over $\{0, 1\} \times \{0, 1, 2\}$ satisfies :

$$\boldsymbol{\mu}(0, h) = (0, 0) \quad h \in \{0, 1, 2\}$$

$$\boldsymbol{\mu}(1, h) = \begin{cases} (1, 1) & \text{if } h = 0 \text{ and } B = 0 \\ (1, 2) & \text{if } h = 0 \text{ and } B = 1 \\ (1, 0) & \text{if } h = 1 \\ (2, 0) & \text{if } h = 2 \end{cases}$$

where B is a random variable drawn from a Bernoulli distribution with success probability $1/2$.

In this model, the environmental variable defines zones that are unfavorable for both species ($\boldsymbol{\mu}(0,h) = (0,0)$). When the environment is favorable ($e = 1$), the hidden species has a deleterious effect on species 2 and a positive effect on species 1. Indeed, the second component of $\boldsymbol{\mu}(1,h)$, which corresponds to the mean distribution of species 2, is low when $h \geq 1$, whereas the first component of $\boldsymbol{\mu}(1,h)$, which corresponds to the mean distribution of species 1, is highest when $h = 2$.

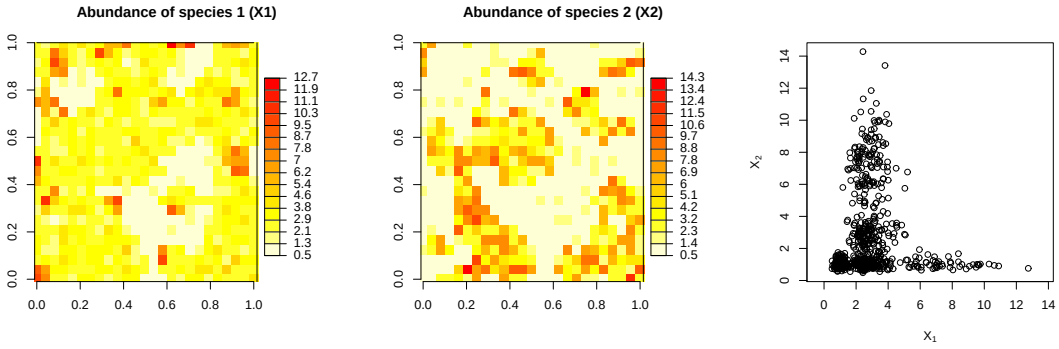


Fig. 4.5. Simulated abundances $\boldsymbol{X}_1$, $\boldsymbol{X}_2$ of the two species on a spatial grid and scatterplot of the abundances.

E_i is a binary variable obtained by truncating a spatial Gaussian field :

$$E_i = \mathbb{I}\{E_0(i) > -0.8\}$$

where $\mathbb{I}$ is the indicator function, E_0 is a Gaussian field with a mean vector of zeros of length n, $\boldsymbol{\mu_0}$, and a covariance matrix $\boldsymbol{\Sigma_0} = 0.5\, e^{-\boldsymbol{D}^2/5}$, and $\boldsymbol{D}$ is the matrix of the distances among sites : $E_0 \sim \mathcal{N}(\boldsymbol{\mu_0}, \boldsymbol{\Sigma_0})$.

The variable H_i is defined by :

$$H_i = \mathbb{I}\{H_0(i) > 0\} + \mathbb{I}\{H_0(i) > 0.8\}$$

where H_0 is a Gaussian field independent from E_0 but drawn like E_0 : $H_0 \sim \mathcal{N}(\boldsymbol{\mu_0}, \boldsymbol{\Sigma_0})$. (See below for a visual support of the model construction).

The maps and scatterplot of the species abundances obtained with a simulation of this model can be seen Figure 4.5.

Results

Classification

The clustering procedure using multivariate normal mixtures is based on two assumptions. The observations that we wish to classify have to be independent and follow a normal mixture distribution. In this simulated example, the logarithm of the species abundances

validate these assumptions, so we used the clustering procedure on these transformed observations. The model corresponding to the maximum BIC is selected.

Here, a 5 classes model with covariance parameterization $\boldsymbol{\Sigma} = \lambda \boldsymbol{I}$, where $\boldsymbol{I}$ is the identity matrix and λ a scalar estimated by the EM algorithm, is selected. The classification on the

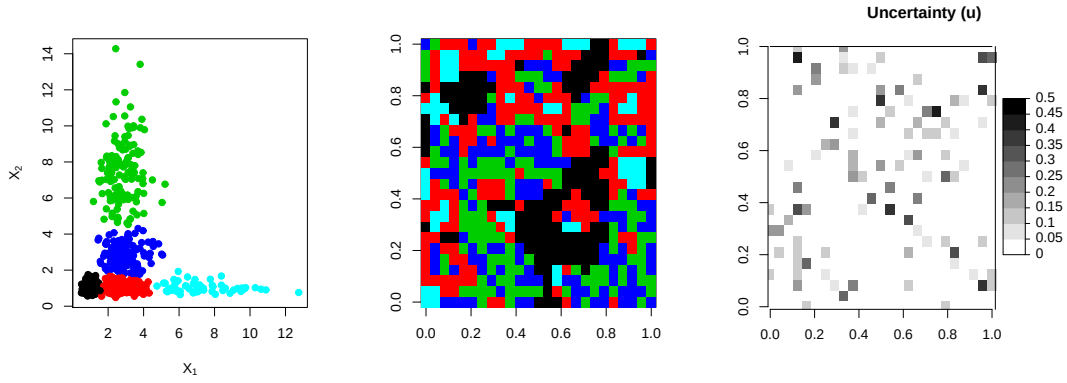


Fig. 4.6. Classification into 5 types of abundance assemblages. Left : Classification on the scatterplot of $(\boldsymbol{X}_1, \boldsymbol{X}_2)$, with each color corresponding to a given class. Center : Map of the types of assemblages. Right : Uncertainty of the classification.

scatterplot of the abundance variables, as well as the map of these classes are represented in Figure 4.6. The covariance model chosen by the BIC corresponds to the covariance used for the simulation in (4.4). The clustering procedure identifies correctly the five different interaction types simulated with the mean vector :

- very low values for both species, because of an unfavorable environment : $\boldsymbol{\mu} = (0, 0)$ (black class),
- species 1 alone, because of the negative interaction between species 2 and a third interacting species : $\boldsymbol{\mu} = (1, 0)$ (red class), $\boldsymbol{\mu} = (2, 0)$ (cyan class),
- coexistence of the two species, due to the absence of the third species : $\boldsymbol{\mu} = (1, 1)$ (dark blue class) and $\boldsymbol{\mu} = (1, 2)$ (green class).

The uncertainty u of the classification can be calculated at each site i by (Fraley & Raftery, 2006) :

$$u(i) = 1 - \max_k \hat{p}_{ik}.$$

In this simulation, the uncertainty is generally low, less than 0.11 for 90% of the sites. The uncertainty can be represented on the scatterplot of $(\boldsymbol{X}_1, \boldsymbol{X}_2)$, which shows that it is highest at the borders of the classes (not represented here), and it can also be mapped (cf. Figure 4.6). In this simulation 97% of the sites are correctly classified.

Spatial analysis of the classes

The study of the spatial structure of the classes using variograms show that all the classes are structured (see Figure 4.7, left columns). This is not surprising for the black class, which

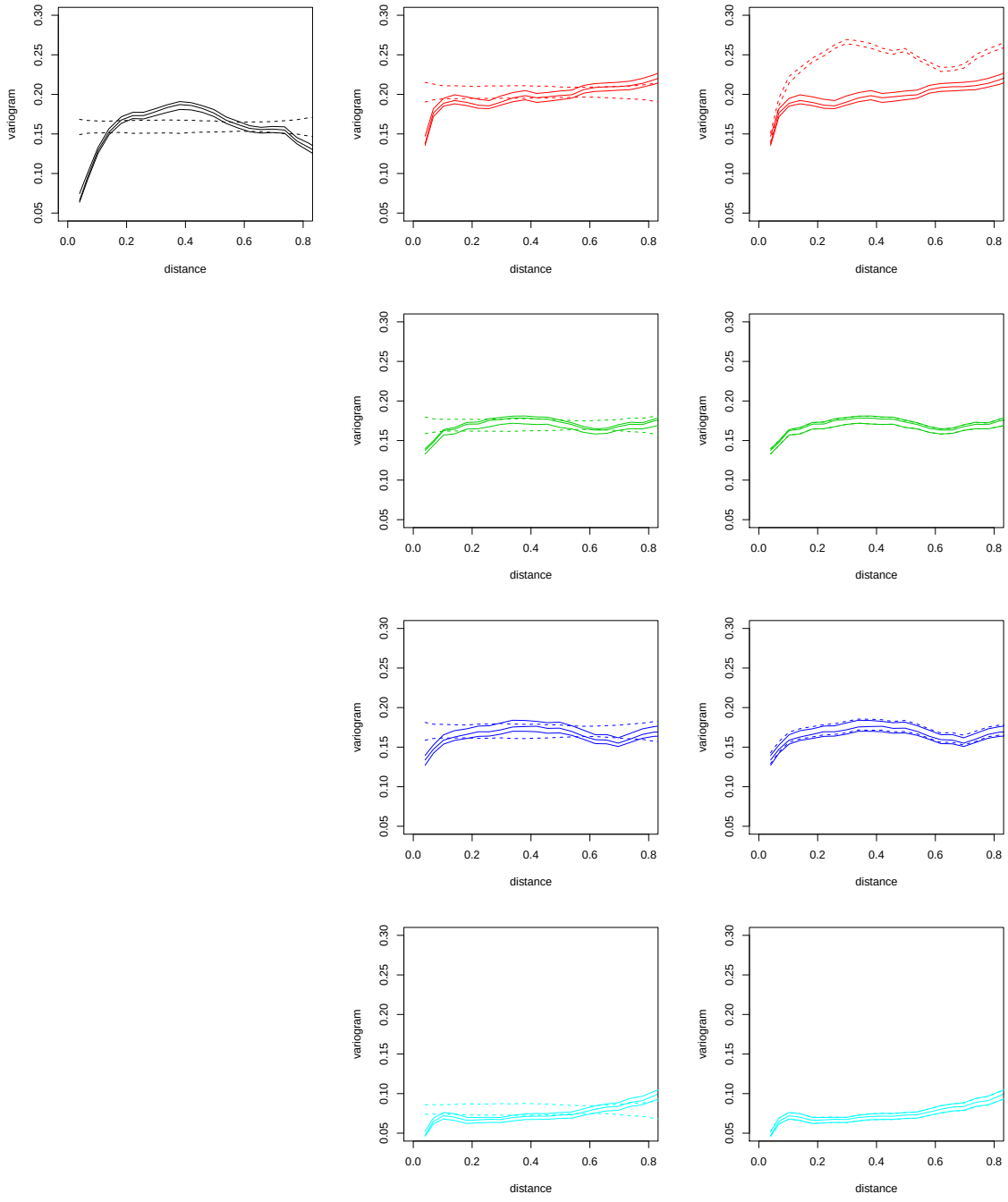


Fig. 4.7. Left and middle columns : Variograms of the 5 classes. Right column : Variograms of the classes conditional on the black class. Full lines : variogram of the $\{\hat{Z}_{ik}\}_i$ and 95% pointwise confidence interval of the class. Dashed lines : 95% pointwise confidence interval of the randomized pattern.

corresponds to the areas where the environment is unfavorable for both species, because of the construction of the environment variable E.

The red and cyan classes, which correspond to the areas where a third interacting species is present, are also expected to be aggregated, by construction of the H variable. This is more surprising for the dark blue and green classes, because of the random variable B introduced in the model. For that matter, the map in Figure 4.6 shows clearly that the black class is aggregated, whereas the blue and green classes appear unstructured.

These classes might appear structured in the analysis based on the variograms only because of the black class, which forms large patches that create a hole in the spatial pattern of the green and of the blue class. We can analyze the spatial structure of the blue and green class conditionally on the structure of the black one, by leaving the cells corresponding to the black class untouched by the permutation procedure while building the randomized pattern. We obtain the variograms in Figure 4.7 (right column), in which the cells assigned to the black class have not been permuted. The dark blue and green classes appear spatially unstructured, while the red class remains structured. The cyan class does not appear structured any more, probably because it forms too small aggregates (see below for a more detailed explanation).

Simulation procedure

Figure 4.8 illustrates the simulation procedure used to build the dataset on a one dimensional space.

In this example the repartitions of the two species are constrained by an environmental variable, obtained by truncating the Gaussian vector E_0 at the threshold value -0.8. This defines the unfavorable areas where both species have low abundances : $\boldsymbol{\mu} = (0, 0)$ (black class). This class is therefore aggregated by construction.

The Gaussian vector H_0, independent from E_0, represents the spatial distribution of a third species, which has a deleterious effect on species 2. Above the threshold value 0, species 2 has low abundance values (mean 0), whereas species 1 is positively correlated to the abundance of the third species : $\boldsymbol{\mu} = (1, 0)$ (red class), $\boldsymbol{\mu} = (2, 0)$ (cyan class). These classes are aggregated by construction, but the black class can create holes in their structure, which make the spatial pattern of the red class and especially of the cyan class appear random at a certain scale. The position of the highest threshold can also have that effect on the cyan class. The dark blue and the green classes are not structured, they are sampled from a Bernoulli distribution with probability 1/2 for each cell of the remaining space.

4.6.2 Extension

We presented in this article a method for the detection, mapping and pattern analysis of different types of assemblages of species abundances.

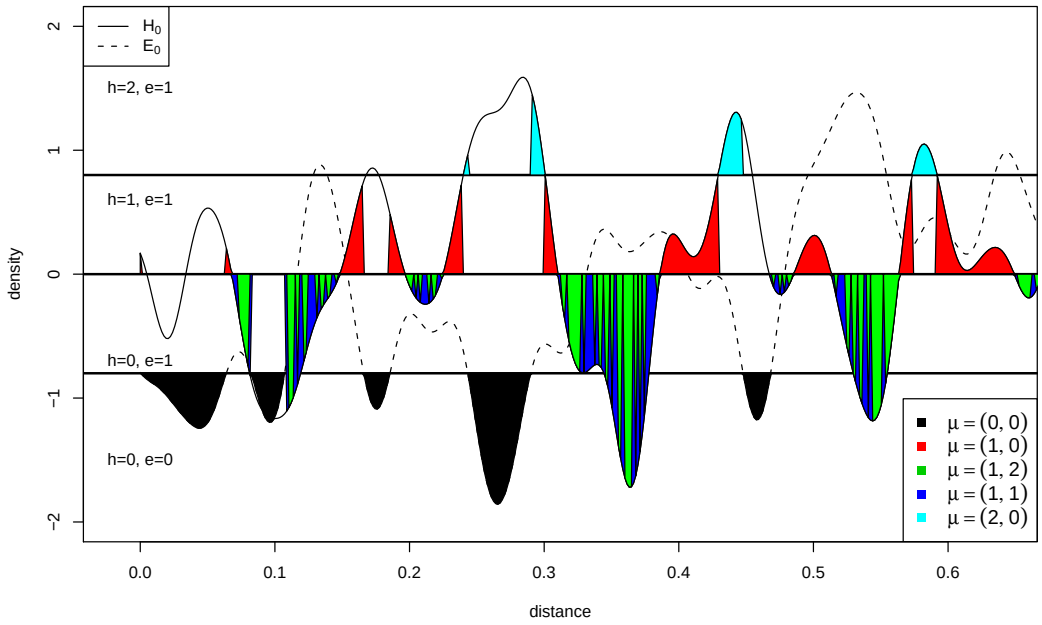


Fig. 4.8. Example 1. Simulation model illustrated on a one dimensional space. The mean vector of the distributions of the two species $\boldsymbol{\mu}$ depends on the values of the two Gaussian vectors E_0 and H_0 with respect to the 3 thresholds -0.8, 0, 0.8.

A possible extension of the CASA method could be the analysis of an image, for example an aerial landscape image, in terms of associations of colors. A color is a combination of red, green and blue in different proportions. These proportions of red, green and blue color could be used instead of the species abundances to obtain a classification of color assemblages. The identified classes will correspond most probably to the different colors that can be found in a landscape, which correspond to different soil types or plant covers. The relationship between these classes can be studied as we did for the types of species abundance assemblages.

For this purpose, we considered an aerial image of an agricultural landscape in New York State, USA. (cf. picture in Figure 4.9). The original image (of 12×12 km^2) was partitioned into 3481 cells (of approximatively 200×200 m^2). The mean proportions of red, green and blue were estimated for each cell (using ImageJ software program to split the color image into RGB components).

The clustering procedure, as implemented by Fraley & Raftery (2006), was performed on the vectors $(\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X}_3)$ of mean proportions of red, green and blue at each observation site. The BIC approximation did not provide conclusive results here. When calculated on models containing 1 to 20 components, a 10 components model with covariance parameterization $\boldsymbol{\Sigma}_k = \lambda_k \boldsymbol{D}_k \boldsymbol{A}_k \boldsymbol{D}_k^T$ was found by maximizing the BIC, but the value of the BIC remains roughly stable after 6 classes.

In cases where the normal mixture hypothesis is a poor fit to the data, an alternative criterion can be used. Biernacki, Celeux & Govaert (2000) showed that the Integrated Classification Likelihood (ICL) criterion is more robust than the BIC to the violation of some

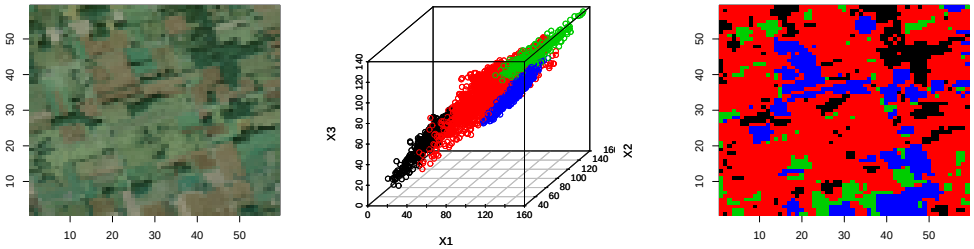


Fig. 4.9. Classification into 4 types of color assemblages with the ICL criterion. left : classification on the scatterplot of $(\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X}_3)$ of the types of assemblages, right : Aerial image of the considered region in NY, USA after the partition into cells.

of the mixture model assumptions, and provides better classifications. The ICL criterion for the model m with K components can be seen as a BIC criterion penalized by an entropy term :

$$ICL_{m,K} = BIC_{m,K} + 2\sum_{i=1}^{n}\sum_{k=1}^{K}\hat{Z}_{ik}\log\hat{p}_{ik} \tag{4.5}$$

The ICL criterion chooses 4 classes with the covariance model $\boldsymbol{\Sigma} = \lambda_k \boldsymbol{D}_k \boldsymbol{A}_k \boldsymbol{D}_k^T$. The classification results can be seen Figure 4.9, and the map of the considered region can be compared to the map of the 4 types of coulour assemblages obtained by this procedure. We can identify the 4 structures corresponding to the assemblages on the map of the region. These assemblages correspond to forests (black class), pasture lots (red class) and 2 types of cultivated lots (green and blue) corresponding maybe to different types of crops, but which are indeed distinguishable on the map by their color (beige and brown).

This example shows that this method is able to identify existing structures quite accurately, even though no spatial information is used in the clustering procedure. We can also see that the generalization to more than two species does not arise any difficulties.

4.6.3 Clustering Models in CASA

The models considered in the clustering stage of the CASA method are those used in `MCLUST` (Fraley & Raftery, 2006). A list is given in table 1.

`MCLUST` assumes a Gaussian mixture model, in which, for each component k, $\boldsymbol{\Sigma}_k$ is parameterized by eigenvalue decomposition :

$$\boldsymbol{\Sigma}_k = \lambda_k \boldsymbol{D}_k \boldsymbol{A}_k \boldsymbol{D}_k^T$$

where :

- $\boldsymbol{D}_k$ is the orthogonal matrix of eigenvectors. It determines the orientation of the principal components of $\boldsymbol{\Sigma}_k$,
- $\boldsymbol{A}_k$ is the diagonal matrix whose elements are proportional to the eigenvalues of $\boldsymbol{\Sigma}_k$. It determines the shape of the density contours,
- λ_k is a scalar, which determines the volume of the ellipsoid.

The characteristics of shape, volume and orientation are usually estimated from the data.

Model	Distribution	Volume (λ)	Shape ($\boldsymbol{A}$)	Orientation ($\boldsymbol{D}$)
$\lambda \boldsymbol{I}$	Spherical	equal	equal	
$\lambda_k \boldsymbol{I}$	Spherical	variable	equal	
$\lambda \boldsymbol{A}$	Diagonal	equal	equal	coordinate axis
$\lambda_k \boldsymbol{A}$	Diagonal	variable	equal	coordinate axis
$\lambda \boldsymbol{A}_k$	Diagonal	equal	variable	coordinate axis
$\lambda_k \boldsymbol{A}_k$	Diagonal	variable	variable	coordinate axis
$\lambda \boldsymbol{D}\boldsymbol{A}\boldsymbol{D}^T$	Ellipsoidal	equal	equal	variable
$\lambda \boldsymbol{D}_k \boldsymbol{A} \boldsymbol{D}_k^T$	Ellipsoidal	equal	equal	variable
$\lambda_k \boldsymbol{D}_k \boldsymbol{A} \boldsymbol{D}_k^T$	Ellipsoidal	variable	equal	variable
$\lambda_k \boldsymbol{D}_k \boldsymbol{A}_k \boldsymbol{D}_k^T$	Ellipsoidal	variable	variable	variable

Tab. 4.1. Parameterizations of the covariance matrix $\boldsymbol{\Sigma}_k$ for multidimensional data used in `MCLUST` (Fraley & Raftery, 2006). For example in the model $\lambda_k \boldsymbol{A}$ all clusters have diagonal covariances, with a volume that can vary between clusters, equal shapes and orientation parallel to the coordinate axes.

5

Une extension de CASA aux données de comptages : Utilisation de mélanges de lois de Poisson multivariées

La méthode que nous proposons dans le chapitre 4 permet de définir des assemblages d'espèces lorsque les données d'abondances des espèces sont continues ou pseudo-continues, à l'aide d'une classification basée sur des modèles de mélange de lois gaussiennes multivariées. D'autres méthodes de classification doivent être envisagées lorsque les données ne sont pas de ce type.

Pour traiter des données de comptages d'espèces, telles les données d'assemblages de pucerons sur des feuilles (présentées en 1.4.1), nous avons envisagé d'utiliser des modèles de mélange de lois de Poisson multivariées. C'est également l'approche qui avait été choisie par Karunanayake & Laverty (2006) pour définir des assemblages de trois espèces d'adventices, à la différence qu'ils avaient opté pour une classification spatiale, en introduisant explicitement l'autocorrélation des données dans l'algorithme d'estimation du mélange, par une chaîne de Markov cachée. Dans notre cas l'estimation du modèle est plus simple, comme le présentent Karlis & Meligkotsidou (2006), car on n'introduit pas d'hypothèses sur la corrélation spatiale des données dans la classification (nous argumentons ce choix en 1.3).

5.1 Modèle de mélange de lois de Poisson bivariées

Nous nous plaçons dans un cas simple où les données sont formées de comptages de deux espèces. On suppose que le vecteur aléatoire des abondances $\boldsymbol{Y}$ suit un modèle de mélange de lois de Poisson bivariées :

$$\boldsymbol{Y} \sim \sum_{k=1}^{K} \pi_k \mathscr{BP}(\lambda_{1k}, \lambda_{2k}, \lambda_{3k})$$

où $\mathscr{BP}$ désigne la loi de Poisson bivariée définie pour une composante k par :

$$\boldsymbol{Y}_k \sim \mathscr{BP}(\lambda_{1k}, \lambda_{2k}, \lambda_{3k}) \Longleftrightarrow \begin{cases} Y_{1k} = X_{1k} + X_{3k} & & X_{1k} \sim \mathscr{P}(\lambda_{1k}) \\ & \text{avec} & X_{2k} \sim \mathscr{P}(\lambda_{2k}) \\ Y_{2k} = X_{2k} + X_{3k} & & X_{3k} \sim \mathscr{P}(\lambda_{3k}) \end{cases}$$

avec $k \in \{1, \ldots, K\}$, $\mathscr{P}$ la loi de Poisson univariée, et $\lambda_{1k}, \lambda_{2k}, \lambda_{3k}$ les paramètres de la loi de Poisson. On peut noter que, pour chaque classe k, le terme X_{3k}, qui introduit une

corrélation entre les variables Y_{1k} et Y_{2k}, est toujours positif ou nul par définition. La densité de $\boldsymbol{Y}$ s'écrit alors :

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = \sum_{k=1}^{K} \pi_k f_k(\boldsymbol{y}; \lambda_{1k}, \lambda_{2k}, \lambda_{3k})$$

où $\boldsymbol{y}$ désigne la réalisation du vecteur aléatoire $\boldsymbol{Y}$ et f_k désigne la densité de la loi de Poisson bivariée de la $k^{\text{ème}}$ composante du mélange. Comme $X_3 \geq 0$, celle-ci peut être écrite sous la forme :

$$f_k(y_1, y_2) = e^{-(\lambda_{1k}+\lambda_{2k}+\lambda_{3k})} \frac{\lambda_{1k}^{y_1}}{y_1!} \frac{\lambda_{2k}^{y_2}}{y_2!} \sum_{r=0}^{min(y_1,y_2)} C_{y_1}^r C_{y_2}^r r! \Big(\frac{\lambda_{3k}}{\lambda_{1k}\lambda_{2k}}\Big)^r.$$

avec la notation $C_n^r = \frac{n!}{r!(n-r)!}$.

5.2 Estimation par l'algorithme EM

Comme dans le cadre de l'estimation des paramètres d'un mélange de lois normales multivariées, l'estimation des paramètres du mélange de Poisson multivarié peut se faire grâce à l'algorithme d'espérance-maximization (EM) présenté dans le paragraphe 2.2.1.

A nombre de classes K fixé, la log-vraisemblance des données observées $\boldsymbol{y}$ complétées par les appartenances inconnues des observations aux classes z_{ik}, et par les variables latentes X_{3ik} s'écrit :

$$l_c(\boldsymbol{\Phi}; \boldsymbol{y}, \boldsymbol{z}, \boldsymbol{x}_3) = \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \log(\pi_k f_k(\boldsymbol{y}_i; \lambda_{1k}, \lambda_{2k}, \lambda_{3k})) \tag{5.1}$$

Etape E : L'espérance des variables latentes sachant les paramètres courants est calculée. Dans ce cas, ceci revient à mettre à jour les probabilités a posteriori d'appartenance aux classes p_{ik} et les variables latentes X_{3ik} :

$$\hat{p}_{ik} = \mathbb{E}(Z_{ik}|\boldsymbol{Y}, \boldsymbol{\Phi}) = \frac{\pi_k f_k(y)}{\sum_{j=1}^{K} \pi_j f_j(y)}$$

$$\mathbb{E}(X_{3ik}|\boldsymbol{Y}, Z_{ik} = 1, \boldsymbol{\Phi}) = \sum_{r=0}^{\min(y_1,y_2)} r \ \mathbb{P}(X_{3ik} = r|\boldsymbol{Y}, \boldsymbol{\Phi})$$

Etape M : Les paramètres $\boldsymbol{\Phi} = (\pi_1, \ldots, \pi_K; \lambda_{11}, \ldots, \lambda_{1K}; \lambda_{21}, \ldots, \lambda_{2K}; \lambda_{31}, \ldots, \lambda_{3K})$ qui maximisent la vraisemblance (5.1) sont calculés :

$$\hat{\pi}_k = \frac{\sum_{i=1}^{n} \hat{p}_{ik}}{n} \quad \text{et} \quad \hat{\lambda}_{lk} = \frac{\sum_{i=1}^{n} \hat{p}_{ik} X_{lik}}{\sum_{i=1}^{n} \hat{p}_{ik}} \quad \text{avec} \quad l \in \{1, 2, 3\} \quad \text{et} \quad k \in \{1, \ldots, K\}.$$

5.3 Application à l'étude d'assemblages d'espèces de pucerons sur des feuilles de clémentinier

Nous avons appliqué cette méthode de classification sur le jeu de données de comptages de deux espèces de pucerons sur des feuilles de clémentinier présenté en 1.4.1.

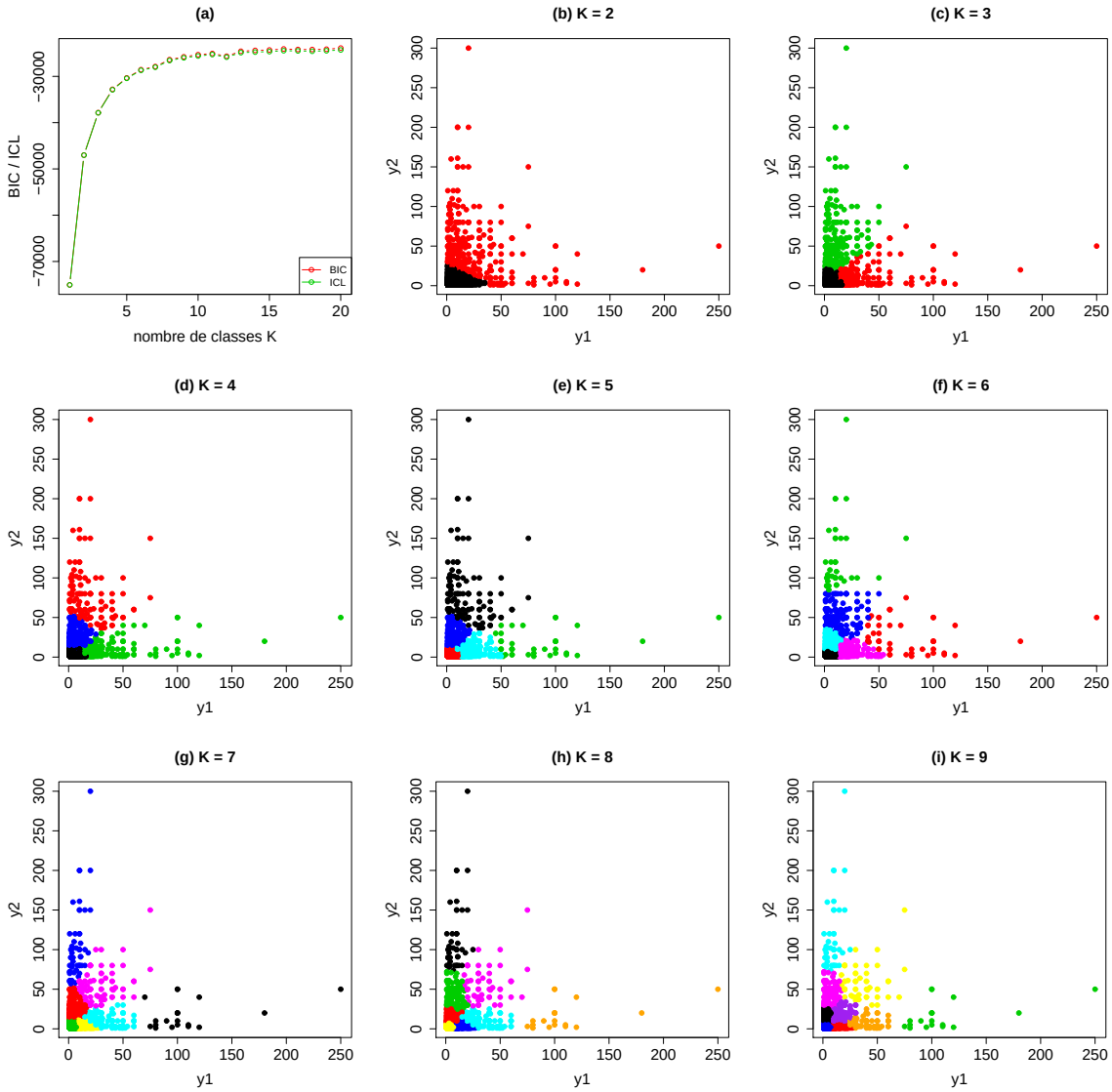


Fig. 5.1. Classification, basée sur des mélanges de lois de Poisson bivariées, des vecteurs d'abondances de deux espèces de pucerons sur 1731 feuilles de clémentinier. (a) Les critères de sélection BIC et ICL atteignent leur maximum à 20 classes, qui correspond au $K_{\max}$, donc ne permettent pas de choisir un nombre de classes. (b)–(i) Classification pour $K = 1, \ldots, 10$. Un point sur le graphique correspond à une feuille.

L'algorithme a été lancé pour un nombre maximal de classes $K_{\max} = 20$. Pour chaque nombre de classes $K \in \{1, \ldots, K_{\max}\}$, nous avons estimé le vecteur des paramètres $\boldsymbol{\Phi}$ et les probabilités a posteriori d'appartenance aux classes p_{ik} par l'algorithme EM présenté plus haut. Les paramètres $\lambda_{1k}, \lambda_{2k}, \lambda_{3k}$ ont été initialisés au hasard pour chaque classe k, et les proportions π_k ont été initialisées à $1/K$. Par la règle du maximum a posteriori (MAP), chaque observation a été affectée à la classe la plus probable (de $\hat{p}_{ik}$ le plus grand). Enfin, pour sélectionner le nombre de classes, nous avons utilisé deux critères de sélection de modèles, le BIC et l'ICL, classiques dans le cadre des modèles de mélange (voir 2.2.1 et équations 4.2 et 4.5).

Les résultats sont présentés en figure 5.1. Les critères BIC et ICL ne permettent pas de choisir un nombre de classes optimal. En effet, ils atteignent leur maximum à $K_{\max} = 20$. Or, il serait difficile de donner sens à un nombre plus grand d'assemblages d'abondances d'espèces. Nous remarquons également que la différence entre BIC et ICL est infime. En effet, l'ICL peut être vu comme une pénalisation du BIC par un terme qui décrit l'entropie au sein des classes. Or dans ce cas les classes sont bien séparées, ce qui explique la faible différence entre les deux critères.

En conclusion, le mélange de lois de Poisson bivariées ne semble pas adapté à la modélisation de cette distribution, du moins avec un nombre de composantes "raisonnable". Ceci est probablement dû aux limites des lois de Poisson multivariées, à savoir le fait qu'elles permettent seulement une corrélation positive ou nulle d'une part, et une dispersion faible des données (qui vient de l'égalité entre moyenne et variance de la loi de Poisson) d'autre part. Or, nous pouvons remarquer que ces données de comptages sont fortement surdispersées par rapport à une loi de Poisson (voir 1.4.1). En principe, contrairement à une distribution multivariée de Poisson, un mélange de distributions multivariées de Poisson peut s'ajuster à des données surdispersées ou corrélées négativement (Karlis & Meligkotsidou, 2006). Néanmoins, les contraintes citées plus haut demeurrent au sein de chaque composante, ce qui contraint la forme des classes.

Pour permettre de la surdispersion au sein des classes on pourrait envisager un modèle de mélange de lois binomiales négatives. Cependant, ceci serait difficile à mettre en œuvre dans le cas multivarié, car les distributions binomiales négatives multivariées ont été peu étudiées. De plus, de même que les lois de Poisson multivariées, ces dernières n'offrent pas la possibilité de décrire des corrélations négatives entre variables. Nous avons donc choisi une approche plus générale et flexible pour décrire des assemblages d'abondances non gaussiennes, qui est exposée dans les parties III et IV de cette thèse.

Troisième partie

Un modèle hiérarchique pour données multivariées de types différents

6

Problématique : un cadre de modélisation et d'estimation général pour des types de données variés

Dans la partie II, un assemblage d'espèces a été modélisé par une loi gaussienne multivariée. Nous avons montré les limites de cette hypothèse sur l'exemple du jeu de données présenté en 1.4.2. En effet, la perte d'information induite par la transformation des données pour les rendre pseudo-continues, peut dissimuler l'effet potentiel de la taille des parcelles de plantain sur la probabilité d'occurrence de l'oïdium. Ceci nous amène à nous intéresser à des modèles capables de prendre en compte des données corrélées de types différents.

6.1 Une classe de modèles qui s'adaptent aux types des données

Dans le chapitre 7, nous proposons une classe de modèles qui s'adaptent à des données de types différents (continu, discret). Il s'agit de modèles hiérarchiques à deux couches. La première couche (couche des observations) est basée sur des lois conditionnellement indépendantes choisies en fonction du type de données (e.g. Poisson pour des données discrètes). La corrélation est exprimée dans la couche cachée à travers une distribution multivariée classique. Nous avons choisi la loi gaussienne pour sa flexibilité (forme, gamme de corrélations) et pour sa facilité de manipulation et d'interprétation dans le cas multivarié.

On définit le modèle suivant pour un vecteur aléatoire $\boldsymbol{Y} = (Y_1, \ldots, Y_d)$:

$$\begin{cases} Y_j|\boldsymbol{\theta} \sim \mathscr{L}_j(g_j^{-1}(\boldsymbol{\theta})) \\ (\theta_1, \ldots, \theta_d) \sim \mathscr{N}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \end{cases}$$

où $j \in \{1, \ldots, d\}$, $\mathscr{N}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ est la distribution gaussienne de dimension d du vecteur aléatoire latent $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)$, de vecteur moyenne $\boldsymbol{\mu}$ et de matrice de covariance $\boldsymbol{\Sigma}$, $\mathscr{L}_j$ est la distribution univariée de la variable Y_j, dont le paramètre $g_j^{-1}(\boldsymbol{\theta})$ est relié à la réalisation du vecteur gaussien latent par la fonction de lien g_j. $\mathscr{L}_j$ est une distribution de la famille exponentielle. $\mathscr{L}_j$ et g_j sont choisis en fonction du type de données et peuvent être différents pour des variables j différentes. Ce type de modèle est parfois appelé modèle latent gaussien dans la littérature et peut être représenté schématiquement par un DAG (voir figure 6.1).

Ce modèle généralise un certain nombre de modèles existants, tels que le modèle multivarié Poisson log-normal (noté ci-après PLN ou MPLN dans le cas multivarié) pour des

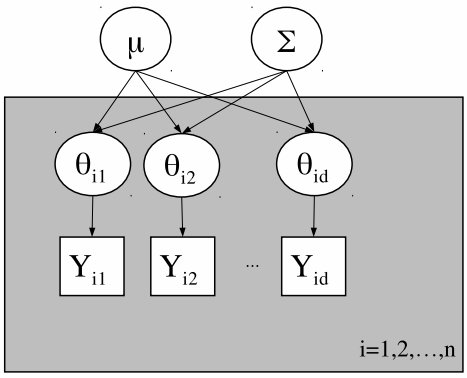


Fig. 6.1. Graphe orienté acyclique (DAG) du modèle gaussien latent

données discrètes, introduit par Aitchison & Ho (1989). Ce dernier est obtenu en choisissant, pour toutes les variables $\boldsymbol{Y}_j$, la loi de Poisson pour $\mathscr{L}_j$ et la fonction exponentielle pour g_j^{-1}. Dans le cadre de la modélisation de données discrètes, comparé au modèle de Poisson multivarié que nous avons utilisé dans le chapitre 5, les principaux avantages du MPLN sont de permettre des corrélations négatives entre variables et de pouvoir modéliser des données surdispersées par rapport à la loi de Poisson.

Remarquons également que ce modèle représente un mélange continu de lois. Dans ce cadre, la distribution conditionnelle de $\boldsymbol{Y}$ sachant $\boldsymbol{\theta}$ sera appellée la distribution mélangée (*mixed distribution*) et la distribution de son paramètre sera la distribution mélangeante (*mixing distribution*). Les modèles de mélange continus sont utilisés pour prendre en compte la surdispersion des données. Par exemple la loi binomiale négative est une loi de mélange Gamma-Poisson (c'est-à-dire une loi de Poisson dont le paramètre λ est distribué selon une loi Gamma) qui permet de modéliser des données discrètes surdispersées par rapport à la loi de Poisson (qui ne respectent pas le rapport moyenne/variance= 1 de la loi de Poisson). Karlis (2005) présente d'autres membres de la famille des mélanges de Poisson (qui se distinguent par le choix de la distribution mélangeante pour le paramètre de Poisson). Le modèle PLN fait partie de cette famille de distributions.

Le modèle hiérarchique que nous présentons dans cette partie est une étape préliminaire dans notre démarche d'étude des assemblages. L'objectif final est de l'intégrer et de l'estimer au sein d'un modèle de mélange à nombre de composantes fini, afin de pouvoir l'utiliser dans un contexte de classification automatique. Il permettra ainsi de définir des assemblages dans la première étape de CASA lorsque les données d'abondances ne sont pas continues et approximables par des gaussiennes, et lorsqu'elles sont de types différents.

6.2 Estimation par maximum de vraisemblance et discussion

Nous avons choisi une méthode d'estimation par maximum de vraisemblance pour ce modèle. Dans le cadre de l'estimation du modèle PLN, Aitchison & Ho (1989); Munkin

& Trivedi (1999) (en multivarié) et Karlis (2005) (en univarié) utilisent également des méthodes d'estimation par maximum de vraisemblance.

Si le vecteur aléatoire $\boldsymbol{Y}$ suit une distribution MPLN, sa densité non conditionnelle $f_{\boldsymbol{Y}}$ s'écrit :

$$\begin{aligned} f_{\boldsymbol{Y}}(\boldsymbol{y};\boldsymbol{\mu},\boldsymbol{\Sigma}) &= \int_{\mathbb{R}^d} f_{\boldsymbol{Y}|\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{\theta}) f_{\boldsymbol{\theta}}(\boldsymbol{\theta}|\boldsymbol{\mu},\boldsymbol{\Sigma})\ d\boldsymbol{\theta} \\ &= \int_{\mathbb{R}^d} \prod_{j=1}^{d} \frac{e^{-e^{\theta_j}} e^{\theta_j y_j}}{y_j!} \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} e^{-\frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}-\boldsymbol{\mu})}\ d\boldsymbol{\theta}, \end{aligned} \tag{6.1}$$

où $f_{\boldsymbol{Y}|\boldsymbol{\theta}}$ désigne la distribution conditionnelle de $\boldsymbol{Y}$ sachant $\boldsymbol{\theta}$ (dans le cas du MPLN c'est le produit de lois de Poisson univariées indépendantes) et $f_{\boldsymbol{\theta}}$ est la loi gaussienne multivariée.

La méthode d'Aitchison & Ho (1989) combine un algorithme d'optimisation de montée de gradient (*steepest ascent*) et un algorithme de Newton-Raphson, et utilise une intégration numérique pour calculer (6.1) après une transformation de cette intégrale, qui est spécifique à la fonction de lien et à la distribution de Poisson. Leur méthode n'est donc pas directement généralisable à d'autres choix de distributions $\mathcal{L}_j$. Munkin & Trivedi (1999) se basent sur un algorithme SML (*simulated maximum likelihood*) pour estimer un modèle MPLN bivarié. Mais McCulloch (1997) a montré dans le cadre des modèles linéaires généralisés à effets mixtes (GLMM) que les algorithmes Monte Carlo EM (MCEM) ou le Monte-Carlo Newton Raphson (MCNR) donnent de meilleurs résultats.

Karlis (2005) propose d'utiliser l'algorithme itératif d'espérance-maximisation (EM) présenté en 2.2.1 pour l'estimation du modèle PLN univarié. L'étape E de l'EM, qui consiste à calculer l'espérance conditionnelle de la vraisemblance des données complètes sachant les données observées, fait intervenir la densité non conditionnelle $f_{\boldsymbol{Y}}$ de $\boldsymbol{Y}$ (6.1). Or dans le cas du MPLN, celle-ci ne permet pas des calculs explicites et n'admet pas de solution analytique simple. Il faut par conséquent évaluer cette intégrale de dimension d à chaque itération de l'EM. Karlis (2005) propose d'employer les variantes de l'EM où l'étape E est basée sur des techniques simulatoires, telles que l'EM stochastique (SEM) de Celeux & Diebolt (1985) et le Monte Carlo EM de Wei & Tanner (1990). Pour évaluer l'espérance des variables $\boldsymbol{\theta}$ sachant les données observées $\boldsymbol{y}$, ces méthodes proposent de générer des échantillons aléatoires à partir de la loi conditionnelle $f_{\boldsymbol{\theta}|\boldsymbol{Y}}$ et d'approcher l'espérance par la moyenne empirique de ces échantillons. Plusieurs techniques existent pour échantillonner dans la distribution $f_{\boldsymbol{\theta}|\boldsymbol{Y}}$ lorsqu'on ne peut pas le faire directement, telles que l'algorithme d'acceptation rejet, l'échantillonnage d'importance ou échantillonnage pondéré (*importance sampling*) et des techniques MCMC (*Markov Chain Monte Carlo*) (voir par exemple Tanner, 1996 et Boreux, Parent & Bernier, 2010). Pour estimer les paramètres du modèle PLN univarié, Karlis (2005) utilise un MCEM avec un algorithme d'acceptation rejet. Les deux autres alternatives pour échantillonner dans la distribution conditionnelle de $\boldsymbol{\theta}$ sachant $\boldsymbol{y}$, à savoir l'algorithme MCMC et l'échantillonnage pondéré, sont plus efficaces et ont été étudiées respectivement par McCulloch (1997) et Booth & Hobert (1999) dans le cadre des GLMM.

Nous utilisons également un algorithme MCEM pour l'estimation du modèle général présenté en 6.1. Ce choix a été guidé en partie par l'extension qui sera présentée dans la partie IV dans laquelle le modèle hiérarchique présenté ici définit les distributions des composantes d'un mélange fini. En effet l'EM est la méthode d'estimation la plus communément utilisée pour l'estimation de mélanges finis de lois paramétriques dans le contexte de la classification automatique (McLachlan & Peel, 2000). Dans cette perspective nous avons besoin d'une méthode qui puisse s'adapter aux différents sous-modèles et converger efficacement même pour des modèles plus compliqués, comme ceux envisagés dans la partie IV.

Plus précisément, nous choisissons l'approche MCEM proposée par Booth & Hobert (1999) basée sur l'échantillonnage pondéré, car elle permet de converger plus efficacement qu'un algorithme d'acceptation rejet simple (comme le fait remarquer Karlis, 2005). Comparé à la variante de MCEM proposé par McCulloch (1997), qui utilise un algorithme de Metropolis-Hastings pour obtenir un échantillon de la distribution $f_{\boldsymbol{\theta}|\boldsymbol{Y}}$, l'algorithme d'échantillonnage pondéré a l'avantage de ne pas rajouter une boucle à chaque itération de l'EM. De plus la convergence de la chaîne de Markov reste difficile à établir. L'algorithme de Booth & Hobert (1999) possède deux avantages supplémentaires. D'une part, il propose une solution au problème du choix de la taille de l'échantillon pondéré, qui est augmentée au cours des itérations selon une procédure automatique (ce qui permet d'obtenir un bon compromis entre rapidité initiale et précisison finale des estimateurs) et d'autre part il permet d'estimer les écarts-types asymptotiques des estimateurs (ce qui n'est pas possible dans la version de McCulloch, 1997). Cette méthode est présentée en détail dans le chapitre 7.

Dans la suite nous allons présenter quelques limites et des extensions de ce modèle, dont quelques-unes ont déjà été traitées notamment dans le cadre des GLMM et des statistiques spatiales. Nous verrons que plusieurs de ces auteurs ont choisi une approche bayésienne. Nous n'avons pas abordé l'approche bayésienne ici, bien qu'il semble qu'elle soit préférable et préférée dans des cas à grande dimension, ou pour des modèles plus complexes et plus généraux (Dunson, 2000; Chagneau et al., 2010). Entre approche bayésienne et fréquentiste, il est difficile de savoir quelle approche est plus judicieuse car il y a souvent un fort parti pris dans la littérature et les études comparées sont rares. Hobert (2001) compare une approche bayésienne par MCMC et une approche fréquentiste par MCEM dans le cadre de l'estimation des paramètres d'un modèle hiérarchique. Selon lui, les désavantages de l'approche bayésienne sont surtout la difficulté du choix des distributions a priori des paramètres et l'absence d'estimations d'erreurs standards fiables. Cette dernière est due à la difficulté de prouver un théorème central limite pour la plupart des estimateurs de Monte Carlo basés sur des chaînes de Markov (ce qui nécessiterait la plupart du temps de prouver que la chaîne est ergodique). L'avantage du MCEM basé sur l'échantillonnage pondéré est qu'il permet une bonne estimation de l'erreur. Cependant l'approche bayésienne peut être utilisée pour un nombre de variables plus important que le MCEM.

6.3 Limites et extensions

6.3.1 Tests d'adéquation

Une première limite de ce travail est l'absence de tests d'adéquation du modèle aux données. Une perspective importante de ce travail serait d'explorer si les tests existants dans la littérature sont bien adaptés et suffisamment puissants pour ce modèle, tels que le test du χ^2 de Pearson, le ratio de log-vraisemblances, ou encore, pour des distributions continues multivariées, le test de Kolmogorov-Smirnov (Justel, Peña & Zamar, 1997) et le test de Cramer-von-Mises (Chiu & Liu, 2009). On pourrait également envisager de construire un test général qui puisse s'adapter à la diversité des données que ce modèle peut traiter.

6.3.2 Décrire la dépendance entre variables de types différents

L'un des avantages du modèle présenté ici est qu'il permet d'introduire de la dépendance entre variables de types différents d'une manière assez simple. Dans ce travail nous avons utilisé uniquement le coefficient de corrélation linéaire de Pearson, noté r, pour décrire la dépendance entre deux variables. Celui-ci est le plus adapté au cas où les lois marginales des variables sont gaussiennes et que la dépendance entre ces variables est linéaire. Lorsque la corrélation est non linéaire, celui-ci peut inférer un degré de dépendance linéaire plus faible que la dépendance réelle. Rappelons ici que le coefficient de corrélation décrit complètement la structure de dépendance dans le cas où deux variables X_1 et X_2 suivent une loi normale bivariée (que leurs densités marginales suivent une loi normale ne suffit pas) et dans ce cas $r = 0$ implique l'indépendance des variables X_1 et X_2. Dans les autres cas, une corrélation nulle n'implique pas que les variables X_1 et X_2 sont indépendantes, mais seulement qu'elles sont non corrélées. Il existe de nombreux autres indices pour décrire la dépendance de deux variables pour ces autres cas (Drouet-Mari & Kotz, 2001, chapitre 6).

Indices pour décrire la dépendance

Des indices non paramétriques existent pour mesurer la corrélation entre variables lorsque celles-ci sont ordinales, discrètes, contiennent des données extrêmes (outliers), ou ne suivent pas une loi normale : les corrélations des rangs. Les plus connues sont le coefficient de rang de Spearman ρ_s et le coefficient de concordance de Kendall τ. Ce dernier mesure la concordance (ou la discordance) entre les rangs de deux variables. Ces indices de corrélation sont indépendants des distributions marginales des variables. Les corrélations des rangs mesurent à quel point, lorsqu'une variable augmente, l'autre tend à augmenter, sans que cette croissance soit nécessairement représentée par une relation linéaire.

Dans le chapitre 7, nous nous intéressons à la gamme de corrélations entre variables Y_j disponibles après la transformation des variables latentes θ_j, et pour cela nous comparons les coefficients de corrélation linéaire respectifs des variables observées et des variables latentes. Or la dépendance entre les Y_j n'est plus linéaire, puisque les espérances conditionnelles ne

sont pas linéaires. Une extension possible de ce travail consisterait à essayer d'autres indices pour quantifier le degré de dépendance des variables observées.

Interprétation de la corrélation sousjacente

Dans le modèle hiérarchique présenté en 6.1 et dans le chapitre 7, la corrélation est introduite dans la couche latente gaussienne. Nous utilisons cette dernière pour définir certains types de corrélations entre les variables $\boldsymbol{Y}_j$, via les fonctions de lien.

Il faudrait approfondir le lien entre la corrélation des variables $\boldsymbol{Y}_j$ et la corrélation des variables latentes $\boldsymbol{\theta}_j$. En effet, dans un modèle où le vecteur de variables latentes décrivent un phénomène d'intérêt qu'on ne peut pas mesurer directement, mais qu'on peut observer à travers des variables $\boldsymbol{Y}_j$, c'est la corrélation des variables latentes qui va nous intéresser. Supposons par exemple que l'on veut modéliser les données d'abondances de deux espèces, qui correspondent toutes les deux à la surface occupée par chaque espèce, mais que l'une a été récoltée comme une donnée continue et l'autre comme une donnée ordinale. L'utilisation d'un modèle gaussien sousjacent se justifie alors par la volonté de modéliser les variables latentes d'intérêt, c'est-à-dire les surfaces occupées par chacune des deux espèces, et dans ce cas c'est bien la corrélation de la loi gaussienne sousjacente qui nous intéresse. L'interprétation de la corrélation de la loi gaussienne latente n'est pas aussi directe lorsqu'on n'a pas d'interprétation du vecteur de variables latentes.

Aller plus loin dans la description de la dépendance : les copules

Pour décrire des structures de dépendance plus complexes entre variables, les copules peuvent être utilisées (Hoff, 2007; Nikoloulopoulos & Karlis, 2008; Drouet-Mari & Kotz, 2001; Genest & Favre, 2007). Les copules sont des distributions multivariées avec des marginales uniformes. L'utilisation des copules permet de décrire les effets de la dépendance entre variables indépendemment des distributions marginales. La formulation d'une distribution multivariée en utilisant une copule se base sur l'idée qu'il existe une transformation simple pour chaque variable marginale telle que la variable transformée suive une distribution uniforme. La structure de dépendance peut ensuite être exprimée par une copule, comme une distribution multivariée ayant des marginales uniformes. Il est possible ainsi de générer des distributions multivariées avec des distributions marginales données. La forme de la dépendance vient du choix de la copule. De nombreuses copules ont été proposées et étudiées (voir Drouet-Mari & Kotz, 2001; Nikoloulopoulos & Karlis, 2008).

6.3.3 Extensions du modèle non explorées dans cette thèse

Dans cette thèse nous avons choisi d'étendre ce modèle à un mélange fini pour la classification automatique de données de types différents. De nombreuses autres extensions de ce modèle sont possibles, dont plusieurs ont déjà été proposées, notamment dans la littérature des modèles linéaires généralisés à effets mixtes (GLMM ou *Generalized Linear Mixed Models*).

Les données ordinales

Nous n'avons pas présenté d'exemple de modélisation de données ordinales dans le chapitre 7. Plusieurs modélisations pour données ordinales et données mixtes continues et ordinales ont été proposées dans la littérature. Everitt (1988) propose de créér des variables ordinales à partir de variables latentes continues par seuillage. Cette méthode est directement utilisable dans notre cas.

Une approche différente a été proposée par Olkin & Tate (1961) qui présentent le "modèle de localisation" pour données mixtes catégorielles et continues (*location model for mixed data*), qui consiste à remplacer les données catégorielles observées à un site par la réalisation d'une multinomiale à m cellules, avec $m = \sum_{i=1}^{q} m_i$, où la $i^{\text{ème}}$ variable catégorielle possède m_i modalités. Dans ce cas les variables catégorielles ne sont pas nécessairement ordonnées, et cette méthode ne tient pas compte de l'ordre.

Enfin Chagneau et al. (2010) s'appuient sur les travaux de Chib & Greenberg (1998) et Chaubert et al. (2010) pour définir des variables ordinales.

L'hypothèse de normalité

Une extension possible du modèle hiérarchique présenté consisterait à relâcher l'hypothèse de normalité pour la distribution de la variable latente. Ceci peut amener plus de flexibilité quant aux formes des lois finales obtenues et aboutirait dans certains cas à des modèles déjà existants (par exemple dans le cas d'une loi Gamma en couche latente, combinée avec des lois de Poisson en couche d'observation). Cependant, mis à part la loi normale, dont les propriétés sont remarquables et bien connues dans le cas multivarié, la plupart des autres lois usuelles ne sont pas aussi interprétables dans le cas multivarié, et leurs propriétés sont souvent méconnues et ont rarement fait l'objet d'études approfondies. Cette généralisation demanderait donc une étude approfondie des propriétés d'autres lois potentielles dans le cas multivarié (telle que la loi Gamma), et nous devrions faire face à des difficultés d'estimation et d'interprétation.

Pour toutes ces raisons, nous nous sommes limités ici à la loi normale en couche latente car ce modèle, déjà assez riche, n'est qu'une première étape, dont le but est d'être intégré dans un modèle plus riche encore de mélange fini de telles lois hiérarchique, pour être utilisé en classification.

Les covariables

Une autre extension utile de ce modèle consiste à prendre en compte des covariables observées, dans le cadre de modèles linéaires généralisés à effets mixtes (GLMM). Ceci permettrait de modéliser les variables observées (appelées variables réponses ou variables à expliquer dans le cadre GLMM) en prenant en compte des effets aléatoires corrélés de types différents et des effets fixes (covariables observées).

Plusieurs auteurs ont traité ce problème et des modèles basés sur des variables latentes ont été proposés par Sammel, Ryan & Legler (1997); Dunson (2000) et Chib & Winkelmann (2001). Sammel, Ryan & Legler (1997) ont proposé un modèle à variables latentes qui permet de modéliser des variables réponses de type continu et discret. Avec les notations définies en 6.1 et pour une observation$d$-dimensionnelle i, leur modèle s'écrit de la manière suivante :

$$\begin{cases} Y_{ij}|\theta_i \sim \mathscr{L}_j(g_j^{-1}(\theta_i)) \quad j = 1, \ldots, d \\ \theta_i \sim \mathscr{N}(\mu, \sigma^2), \end{cases}$$

où :

$$g_j^{-1}(\theta_i, u_i) = \boldsymbol{x}_i^T \boldsymbol{\beta}_j = \beta_{0j} + \theta_i \beta_{1j} + u_{i1}\beta_{3j} + \ldots + u_{ip}\beta_{(p+1)j}.$$

et $\boldsymbol{u}_i = (u_{i1}, \ldots, u_{ip})^T$ est le vecteur des p covariables fixées, $\boldsymbol{x}_i = (1, \theta_i, u_{i1}, \ldots, u_{ip})^T$ et $\boldsymbol{\beta}_j = (\beta_{0j}, \ldots, \beta_{(p+1)j})^T$.

Par exemple pour une variable binaire Y_1 et une variable continue Y_2, ils supposent respectivement une fonction de lien logistique et un lien identité :

$$\begin{cases} Y_{i1}|\theta_i \sim \mathscr{B}ernoulli(1/(1 + e^{-\boldsymbol{x}_i^T \boldsymbol{\beta}_1})) \\ Y_{i2}|\theta_i \sim \mathscr{N}(\boldsymbol{x}_i^T \boldsymbol{\beta}_2, \sigma_2^2), \end{cases}$$

En fait, ils étudient seulement le cas univarié, avec une seule variable latente, et se limitent aux distributions $\mathscr{L}$ de la famille exponentielle à un paramètre. Conditionnellement à la variable latente, les variables observées sont supposées indépendantes. Ils illustrent leur modèle par une étude concernant les effets de l'utilisation d'un médicament anticonvulsant pendant la grossesse sur le développement de défauts à la naissance. Les données observées (variables d'intérêt) contiennent des mesures de taille (variables continues) et des indicateurs binaires d'anomalies physiques (variables discrètes). Ils considèrent que ces variables réponses sont dues à un score de sévérité/gravité non observé qui correspond à la variable continue latente de leur modèle. Ils estiment les paramètres grâce à un algorithme EM et proposent pour l'étape E, qui n'admet pas toujours de solution analytique, soit une intégration numérique (par la quadrature de Gauss-Hermite), soit une étape de Monte Carlo.

Dunson (2000) propose des modèles bayésiens latents pour des variables réponses de types différents. Le travail de Dunson (2000) est présenté dans un cadre très général qui inclut le GLMM et le modèle de Sammel, Ryan & Legler (1997). Le choix des distributions est laissé libre pour chaque couche du modèle. Il propose une estimation dans un cadre bayésien avec des procédures MCMC. Chib & Winkelmann (2001) présentent un modèle pour l'analyse de données de comptages corrélés. Leur modèle s'inspire du modèle MPLN de Aitchison & Ho (1989) qu'ils adaptent au cadre GLMM en introduisant des covariables. Dans ce cas également l'estimation se fait par MCMC.

L'autocorrélation spatiale

Enfin ce modèle pourrait être généralisé pour des données spatiales en introduisant explicitement des modèles d'autocorrélation spatiale. Ce type de modèles a déjà été exploré

dans un cadre GLMM. Diggle, Tawn & Moyeed (1998) ont été les premiers à proposer des modèles linéaires généralisés mixtes (GLMM) à un effet aléatoire spatialement corrélé. Il s'agit de modèles hiérarchiques, dans lesquels la couche latente est modélisée par un champ gaussien. Conditionnellement à ce champ gaussien les observations sont considérées indépendantes spatialement. Dans ce cadre, Desassis (2007) utilise la même méthode d'estimation que nous (MCEM de Booth & Hobert, 1999). Très récemment, Chagneau et al. (2010) ont généralisé ces modèles au cas multivarié. Ils utilisent des processus gaussiens latents corrélés pour introduire la dépendance entre variables tout en tenant compte de l'autocorrélation spatiale. Comme dans notre cas, leur modèle est prévu pour prendre en compte des données de types différents. Ils proposent d'estimer les paramètres de leur modèle par une approche bayésienne par MCMC.

7

A Hierarchical Model for Multivariate Data of Different Types and Maximum Likelihood Estimation

By Vera Georgescu[1,*], Nicolas Desassis[2], Samuel Soubeyrand[1], André Kretzschmar[1], Rachid Senoussi[1]

submitted to Statistics and Computing

[1] INRA, UR546 Biostatistique et Processus Spatiaux, Agroparc, F-84914 Avignon, France

[2] École Nationale supérieure des Mines de Paris, Centre de Géosciences, F-77300 Fontainebleau, France

[*] Corresponding author : vera.georgescu@avignon.inra.fr – FAX : +33(0)432722182

A general model for analyzing multivariate data, which can contain data attributes of different types (e.g. continuous, discrete, binary), is presented in this paper. This model supports a wide range of correlation structures and can accommodate overdispersed data. Our model is a two-level hierarchical model where the multivariate data is obtained from conditionally independent univariate distributions, whose parameters, possibly transformed by a set of link functions, are drawn from a multivariate normal distribution. The choices of the univariate distributions and of the link functions depend on the type of data. Maximum likelihood fitting of the model is achieved by an automated Monte Carlo Expectation Maximization (MCEM) algorithm. Our method is tested in a simulation study in the bivariate case and an application to a dataset concerning beehive activity is presented.

Key words : Continuous data ; Count data ; Mixed mode data ; Monte Carlo EM ; Overdispersion ; Poisson-log normal distribution.

7.1 Introduction

Data with attributes of different types are encountered in many fields. For instance in ecological studies, abundance data of several species measured at different sites can be counts (discrete), species coverage, weights (continuous), occurrence (binary). Nevertheless, there is a lack of models (classes of distributions) which can take into account these different types of data and are easy to adapt to different situations, while allowing a large correlation structure for the variables.

The scope of this article is to develop a general model for analyzing multiple response data which can contain variables of different types (discrete, continuous, binary) and which

supports a wide range of correlation structures. We propose here a model that generalizes the multivariate Poisson log normal (MPLN) model studied by Aitchison & Ho (1989). The MPLN model is a multivariate log normal mixture of independent Poisson distributions. This model provides a parametric class of distributions for the analysis of multivariate count data, that is able to describe a wide range of correlation and overdispersion situations. Unlike other multivariate discrete distributions, such as the multivariate Poisson distribution (first proposed by McKendrick and Wicksell), the MPLN model supports negative correlation between counts. Moreover, it can fit overdispersed data, whereas in the multivariate Poisson model the marginal mean and variance coincide (see Aitchison & Ho, 1989 for a detailed comparison between the MPLN model and the multivariate Poisson model). It seems therefore better suited to model multivariate count data such as species count data in ecological studies, which is generally overdispersed and can be negatively correlated.

The general model we propose is a two-layer hierarchical model, in which the hidden layer is a multivariate Gaussian distribution, and the observed layer is a multivariate distribution formed by independent univariate distributions chosen according to the type of variable. We chose the multivariate normal distribution for the hidden layer because it has been extensively studied and it provides a full range of correlations between variables (including negative correlation). The hierarchical structure of the model allows overdispersion in the marginal distributions.

Very recently, Chagneau et al. (2010) proposed a spatial model for random variables of different types with a Bayesian estimation procedure based on MCMC simulations. The principle of our approach is similar in that the dependence between variables is expressed at the hidden level of a hierarchical model and the obtention of different types of variables is achieved by using different conditionally independent univariate distributions and link functions. In this paper, we formalize these multivariate hierarchical models for conditional distributions belonging to the exponential class and present some of their properties in a non spatial framework. We also provide a maximum likelihood estimation procedure for these models easy to adapt to different distributions from the exponential family.

This very general model and estimation procedure can be easily adapted to different data types, such as discrete, continuous, ordinal, binary and 0-inflated data, which are frequent for example in ecological studies.

In the next section we present the general model and its properties for given conditional distributions in the observed layer. The maximum likelihood based estimation procedure is presented in Section 7.3. In Section 7.4 we test our method on simulated bivariate data of different types. An application on a real dataset concerning the honey-bee hive activity in the South of France is presented in Section 7.5. Our results and perspectives are discussed in Section 7.6.

7.2 Multivariate hierarchical model

7.2.1 Definition of the general model

Let $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ denote a random sample of size n of the d-dimensional random vector $\boldsymbol{Y} = (Y_1, \ldots, Y_d)$. In practice $\boldsymbol{Y}_i$ could correspond to the abundances of d species observed at location i. Throughout this article i labels the observation and j the variable.

We define the following hierarchical model for $\boldsymbol{Y}$:

$$\begin{cases} \boldsymbol{Y}_i|\boldsymbol{\theta}_i \sim \mathscr{L}(g^{-1}(\boldsymbol{\theta}_i)) \\ \quad \boldsymbol{\theta}_i \sim \mathscr{N}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \end{cases}$$

where $\mathscr{N}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the d-dimensional normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, g is a set of link functions and $\mathscr{L}$ is a multivariate distribution with parameters $g^{-1}(\boldsymbol{\theta}_i)$.

In this article only the case where $\mathscr{L}$ is formed by d independent univariate distributions $\mathscr{L} = \mathscr{L}_1 \times \ldots \times \mathscr{L}_d$ is considered. The choice of $\mathscr{L}_j$ and g_j^{-1}, for $j \in \{1, \ldots, d\}$, depends on the type of data (discrete, continuous, ordinal, binary). Note that the variables are not necessarily of the same type, different univariate distributions and different link functions can be used for the d variables. In the remainder of the article, the choice of $\mathscr{L}$ will be restricted to exponential families. This is not a very restrictive choice since the exponential family encompasses a broad set of parametric distributions including the most commonly-used (such as Gaussian, Poisson, Bernoulli, Binomial, Gamma).

The probability density $f_{\boldsymbol{Y}}$ of $\boldsymbol{Y}$ is defined by :

$$f_{\boldsymbol{Y}}(\boldsymbol{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \int_{\mathbb{R}^d} f_{\boldsymbol{Y}|\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{\theta}) f_{\boldsymbol{\theta}}(\boldsymbol{\theta}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \, d\boldsymbol{\theta}, \tag{7.1}$$

where $f_{\boldsymbol{Y}|\boldsymbol{\theta}}$ denotes the conditional probability density function of $\boldsymbol{Y}$, given the variable $\boldsymbol{\theta}$, $f_{\boldsymbol{\theta}}$ is the multivariate Gaussian density and a realization of a random vector is denoted by the corresponding lower-case letter. Since the distributions of the d variables are conditionally independent given $\boldsymbol{\theta}$, we have :

$$f_{\boldsymbol{Y}|\boldsymbol{\theta}}(y_1 \ldots y_d|\boldsymbol{\theta}) = \prod_{j=1}^{d} f_{Y_j|\boldsymbol{\theta}}(y_j|\boldsymbol{\theta}). \tag{7.2}$$

Unless conjugate distributions are chosen, there is no simplification of the multiple integral in equation (7.1) for most choices of $\mathscr{L}$ and g^{-1}. Nevertheless, in some cases, as will be seen in Section 7.2.3, its first two moments can be obtained in terms of the moments of $\boldsymbol{\theta}$ by using conditional expectation results and properties of the chosen distributions. Conversely, the moments of the hidden variable $\boldsymbol{\theta}$ can then be written in terms of the moments of the data $\boldsymbol{Y}$ and used to initialize the parameters in the estimation procedure presented in Section 7.3.

Remark : A more general form of the model could be defined by allowing the dimension of $\boldsymbol{\theta}_i$ to differ from the dimension of $\boldsymbol{Y}_i$, but for the sake of simplicity we chose the same dimension d. This could be used to introduce spatial correlations.

7.2.2 Exponential families

If the density of $\mathscr{L}_j$, the conditional distribution of the jth variable given $\boldsymbol{\theta}$, belongs to an exponential family, it can be written in the form :

$$f_{Y_j|\boldsymbol{\theta}}(y_j|\boldsymbol{\theta}) = c_j(y_j)\exp(\sum_{l=1}^{r_j}\eta_{jl}(\boldsymbol{\theta})T_{jl}(y_j) - b_j(\boldsymbol{\eta}_j)),$$

where :

- r_j is the number of parameters of $\mathscr{L}_j$,
- $\boldsymbol{\eta}_j = (\eta_{j1}, \ldots, \eta_{jr_j})$ is the vector of natural parameters of $\mathscr{L}_j$, which can be expressed in terms of $\boldsymbol{\theta}$ and the link funtion g_j,
- $\boldsymbol{T}_j = (T_{j1}, \ldots, T_{jr_j})$ is the vector of minimal sufficient statistics of $\mathscr{L}_j$, which can be written in terms of Y_j,
- and $b_j(\boldsymbol{\eta}_j)$ a normalization factor.

We have the following conditional moments for $l \in \{1, \ldots, r_j\}$:

$$\mathbb{E}(T_{jl}|\boldsymbol{\theta}) = \frac{\partial b(\boldsymbol{\eta}_j)}{\partial \eta_{jl}}, \tag{7.3}$$

$$\mathbb{V}(T_{jl}|\boldsymbol{\theta}) = \frac{\partial^2 b(\boldsymbol{\eta}_j)}{\partial \eta_{jl}^2}. \tag{7.4}$$

Suppose for example that $\mathscr{L}_j$ is the Poisson distribution with mean $\lambda_\theta = g_j^{-1}(\theta_j)$ and g_j^{-1} is the exponential function :

$$Y_j|\boldsymbol{\theta} \sim \mathscr{P}(e^{\theta_j}).$$

Then $r_j = 1$, $\eta_j = \log(\lambda_\theta) = \log e^{\theta_j} = \theta_j$, $T_j = Y_j$ and $b_j(\eta_j) = e^{\eta_j}$. We can verify that equation (7.3) and (7.4) hold :

$$\mathbb{E}(T_j|\boldsymbol{\theta}) = \mathbb{E}(Y_j|\boldsymbol{\theta}) = \frac{\partial b(\eta_j)}{\partial \eta_j} = e^{\eta_j} = e^{\log(\lambda_\theta)} = \lambda_\theta,$$

$$\mathbb{V}(T_j|\boldsymbol{\theta}) = \mathbb{E}(Y_j|\boldsymbol{\theta}) = \frac{\partial^2 b(\eta_j)}{\partial \eta_j^2} = \lambda_\theta.$$

Suppose now that $\mathscr{L}_j$ is a two-parameter distribution, say the Gamma distribution with the usual shape parameter $k_{\boldsymbol{\theta}}$ and scale parameter $\lambda_{\boldsymbol{\theta}}$, which can be expressed in terms of $\boldsymbol{\theta}$ by using two link functions, g_{j1} and g_{j2} :

$$k_{\boldsymbol{\theta}} = g_{j1}^{-1}(\boldsymbol{\theta}),$$
$$\lambda_{\boldsymbol{\theta}} = g_{j2}^{-1}(\boldsymbol{\theta}).$$

Then $r_j = 2$, $\boldsymbol{\eta}_j = (k_{\boldsymbol{\theta}} - 1, -\frac{1}{\lambda_{\boldsymbol{\theta}}})$, $\boldsymbol{T}_j = (\log(Y_j), Y_j)$ and $b(\boldsymbol{\eta}_j) = -(\eta_1 + 1)\log(-\eta_2) + \log(\eta_1\Gamma(\eta_1))$. We can verify that the first moments of Y obtained using equation (7.3) and (7.4) are indeed the mean and variance of the Gamma distribution :

$$\mathbb{E}(T_{j2}|\boldsymbol{\theta}) = \mathbb{E}(Y_j|\boldsymbol{\theta}) = \frac{\partial b(\eta_j)}{\partial \eta_{j2}} = -\frac{\eta_{j1}+1}{\eta_{j2}} = k_{\boldsymbol{\theta}}\lambda_{\boldsymbol{\theta}},$$

$$\mathbb{V}(T_{j2}|\boldsymbol{\theta}) = \mathbb{E}(Y_j|\boldsymbol{\theta}) = \frac{\partial^2 b(\eta_j)}{\partial \eta_{j2}^2} = k_{\boldsymbol{\theta}}\lambda_{\boldsymbol{\theta}}^2.$$

7.2.3 Submodel examples

Multivariate Poisson-Log Normal model (MPLN)

The MPLN model (Aitchison & Ho, 1989) is obtained when $\mathscr{L}$ is formed by d independent Poisson distributions and g^{-1} is the exponential function. For all observations $i \in \{1, \ldots, n\}$ and variables $j \in \{1, \ldots, d\}$ we write :

$$\begin{cases} Y_{ij}|\theta_{ij} \sim \mathscr{P}(e^{\theta_{ij}}) \\ (\theta_{i1}, \ldots, \theta_{id})^T \sim \mathscr{N}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \end{cases}$$

where the superscript T denotes the transpose of the matrix and $\mathscr{P}$ is the univariate Poisson distribution.

The unconditional moments of this distribution can be calculated by using properties of the lognormal distribution and of the conditional expectation (Aitchison & Ho, 1989; Tunaru, 2002) :

$$\begin{aligned} \mathbb{E}(Y_j) &= \mathbb{E}[\mathbb{E}(Y_j|\theta_j)] = e^{\mu_j + \frac{1}{2}\sigma_{jj}} \overset{def}{=} m_j, \\ \mathbb{V}(Y_j) &= \mathbb{E}[\mathbb{V}(Y_j|\theta_j)] + \mathbb{V}[\mathbb{E}(Y_j|\theta_j)] \\ &= m_j + m_j^2 \ (e^{\sigma_{jj}} - 1), \\ \operatorname{cov}(Y_j, Y_{j'}) &= \mathbb{E}[\operatorname{cov}(Y_j, Y_{j'}|\boldsymbol{\theta})] + \operatorname{cov}[\mathbb{E}(Y_j|\theta_j), \mathbb{E}(Y_{j'}|\theta_{j'})] \\ &= m_j \ m_{j'} \ (e^{\sigma_{jj'}} - 1), \\ \operatorname{cor}(Y_j, Y_{j'}) &= \frac{e^{\sigma_{jj'}} - 1}{\sqrt{(e^{\sigma_{jj}} - 1 + m_j^{-1})(e^{\sigma_{j'j'}} - 1 + m_{j'}^{-1})}}, \end{aligned}$$

where $\boldsymbol{\Sigma} = (\sigma_{jj'})$ and $j, j' \in \{1, \ldots, d\}$ for $j \neq j'$. Some interesting features of the model appear :

(i) $\mathbb{V}(Y_j) \geq \mathbb{E}(Y_j)$, so there is overdispersion for the marginal distributions with respect to the Poisson distribution,

(ii) the signs of the correlation between observed variables and the correlation between the hidden normally distributed variables $\boldsymbol{\theta}$ correspond,

(iii) the range of correlation is not as wide as that of the corresponding normal distribution :

$$|\operatorname{cor}(Y_j, Y_{j'})| < |\operatorname{cor}(\theta_j, \theta_{j'})|.$$

Aitchison & Ho (1989) studied the regions of count correlation and overdispersion attainable by the bivariate Poisson-log normal model for different mean counts m.

Bivariate Poisson-Normal model

Data of different types (e.g. continuous and discrete) can be obtained by using different distributions and link functions for the variables. We define the bivariate Poisson-Normal model by :

$$\begin{cases} Y_{i1}|\theta_{i1} \sim \mathscr{P}(e^{\theta_{i1}}) \\ Y_{i2}|\theta_{i2} = g_2^{-1}(\theta_{i2}) \\ (\theta_{i1}, \theta_{i2})^T \sim \mathscr{N}_2(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \end{cases}$$

where g_2^{-1} could be for instance the exponential function, if a positive continuous variable is needed. Notice that the likelihood is easier to compute here than for the MPLN model, because the variable $\boldsymbol{\theta}_2$ is observed in this model.

The moments of this distribution can be calculated in a similar way to the previous model.

$$\begin{aligned} \mathbb{E}(Y_1) &= \mathbb{E}[\mathbb{E}(Y_1|\theta_1)] = e^{\mu_1 + \frac{1}{2}\sigma_{11}} \stackrel{def}{=} m_1, \\ \mathbb{V}(Y_1) &= \mathbb{E}[\mathbb{V}(Y_1|\theta_1)] + \mathbb{V}[\mathbb{E}(Y_1|\theta_1)] \\ &= m_1 + m_1^2\ (e^{\sigma_{11}} - 1), \\ \text{cov}(Y_1, Y_2) &= \mathbb{E}[\theta_2\ \mathbb{E}(Y_1|\theta_1)] - \mathbb{E}[\mathbb{E}(Y_1|\theta_1)]\ \mathbb{E}(\theta_2) \\ &= m_1 \sigma_{12}, \\ \text{cor}(Y_1, Y_2) &= \frac{\sigma_{12}}{\sqrt{(e^{\sigma_{11}} - 1 + m_1^{-1})\ \sigma_{22}}}. \end{aligned}$$

The same properties hold : overdispersion for the count variable $\boldsymbol{Y}_1$ and large correlation range between variables, but smaller than the correlation range between θ_{i1} and θ_{i2}.

Bivariate Binomial-Poisson model

We define the bivariate Binomial-Poisson model by :

$$\begin{cases} Y_{i1}|\theta_{i1} \sim \mathscr{B}(n_b, \text{logit}^{-1}(\theta_{i1})) \\ Y_{i2}|\theta_{i2} \sim \mathscr{P}(e^{\theta_{i2}}) \\ (\theta_{i1}, \theta_{i2})^T \sim \mathscr{N}_2(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \end{cases}$$

where $\mathscr{B}$ denotes the univariate binomial distribution, with parameters n_b the number of Bernoulli trials and success probability $\text{logit}^{-1}(\theta_{i1})$, where $\text{logit}^{-1}(x) = \frac{1}{1+e^{-x}}$.

The moments of this distribution cannot be written in a closed form (see Appendix 7.7.1) but their properties can be studied by simulation or numerical computation. We studied the range of the count correlation coefficient $\text{cor}(Y_1, Y_2)$ for given values of $\mu_1, \mu_2, \sigma_{11}, \sigma_{22}$. The results given in Figure 7.1 show once more that there is a direct correspondence between the signs of $\text{cor}(Y_1, Y_2)$ and r=$\text{cor}(\theta_1, \theta_2)$, while the range of $\text{cor}(Y_1, Y_2)$ is smaller.

Bivariate Gamma-Poisson model

This is another example of model combining variables of different types (continuous vs discrete), which uses an exponential family with two parameters, the Gamma distribution. We define the bivariate Gamma-Poisson model by :

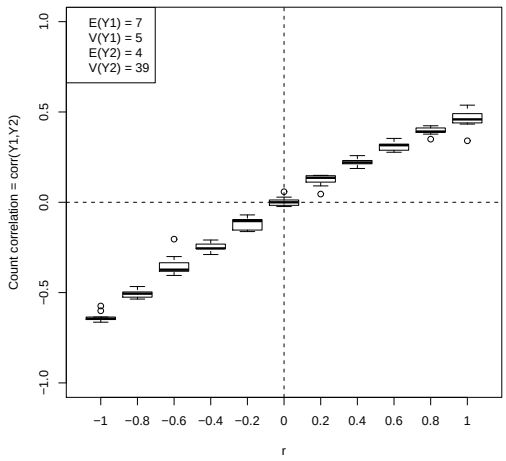


Fig. 7.1. Evolution of the correlation between Y_1, Y_2 with r, (the correlation between θ_1, θ_2) for the Binomial-Poisson model, obtained by simulating 100 samples of size $n = 1000$ with parameters $\mu_1 = \mu_2 = \sigma_{11} = \sigma_{22} = 1$ and $n_b = 10$

$$\begin{cases} Y_{i1}|(\theta_{i1}, \theta_{i2}) \sim \mathscr{G}(k_{\boldsymbol{\theta}}, \lambda_{\boldsymbol{\theta}}) \\ Y_{i2}|\theta_{i1} \sim \mathscr{P}(e^{\theta_{i1}}) \\ (\theta_{i1}, \theta_{i2})^T \sim \mathscr{N}_2(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \end{cases}$$

where the shape parameter $k_{\boldsymbol{\theta}}$ and the scale parameter $\lambda_{\boldsymbol{\theta}}$ of the Gamma distribution $\mathscr{G}$ depend on $\boldsymbol{\theta}$, by defining the mean and variance of the Gamma distribution by :

$$\mathbb{E}(Y_{i1}|\theta_{i1}, \theta_{i2}) = k_{\boldsymbol{\theta}}\lambda_{\boldsymbol{\theta}} = e^{\theta_{i1}},$$
$$\mathbb{V}(Y_{i1}|\theta_{i1}, \theta_{i2}) = k_{\boldsymbol{\theta}}\lambda_{\boldsymbol{\theta}}^2 = e^{\theta_{i2}}.$$

This yields :

$$k_{\boldsymbol{\theta}} = g_{11}^{-1}(\boldsymbol{\theta}) = e^{2\theta_{i1} - \theta_{i2}},$$
$$\lambda_{\boldsymbol{\theta}} = g_{12}^{-1}(\boldsymbol{\theta}) = e^{\theta_{i2} - \theta_{i1}}.$$

This model could be interpreted in the following way in an ecological framework : Suppose two species abundances were observed over n locations. Y_1 denotes the weight or surface occupied by species 1, and Y_2 counts of species 2. This model assumes that the expected values of these two variables depend on an unobserved variable θ_1, say resource availability. This unobserved factor θ_1 is linked to another unobserved variable θ_2, which only influences the variance of the abundance of species 1. θ_2 could be for instance a third species which is a competitor of species 1 but has no influence on species 2.

7.3 Maximum likelihood estimation via the MCEM algorithm

The model we propose has several interesting properties : it is easy to adapt to different types of data and provides a wide correlation range between variables. The price to pay for

these advantages is the increased computational complexity required for parameter estimation. It is therefore important to have a generic estimation procedure that is easy to adapt to different distributions $\mathscr{L}$ and link functions g and thus does not depend on their specific properties. For the MPLN model, Aitchison & Ho (1989) used a maximum likelihood estimation procedure (mix of Newton Raphson and steepest ascent) based on a numerical integration procedure which depends on the specific form of the MPLN likelihood. Tunaru (2002) and Chagneau et al. (2010) use a Bayesian estimation procedure (MCMC algorithm). We use a maximum likelihood estimation procedure based on the EM algorithm.

Let $\boldsymbol{\Phi}$ denote the unknown parameter vector $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Since $\boldsymbol{\theta}$ is not observed, the Expectation-Maximization (EM) algorithm of Dempster, Laird & Rubin (1977) is well suited for the maximum likelihood estimation of $\boldsymbol{\Phi}$. The idea behind the EM algorithm is to complete the observed data $\boldsymbol{y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)$ with the latent variable vector $\boldsymbol{\theta}$, write the complete-data loglikelihood :

$$l_c(\boldsymbol{\Phi}; \boldsymbol{y}, \boldsymbol{\theta}) = \sum_{i=1}^{n} (\log f_{\boldsymbol{Y}|\boldsymbol{\theta}}(\boldsymbol{y}_i|\boldsymbol{\theta}_i) + \log f_{\boldsymbol{\theta}}(\boldsymbol{\theta}_i; \boldsymbol{\Phi})) \tag{7.5}$$

and maximize the conditional likelihood expectation, given the observed data $\boldsymbol{y}$, in terms of $\boldsymbol{\Phi}$.

The EM algorithm is a two-step iterative algorithm that proceeds as follows : At iteration $t+1$, the current parameter $\boldsymbol{\Phi}^{(t)}$ is known.

– E-step (Expectation) : the conditional expectation of the complete-data log-likelihood given the observed data $\boldsymbol{y}$ and the current parameter estimates is computed :

$$Q(\boldsymbol{\Phi}, \boldsymbol{\Phi}^{(t)}) = \mathbb{E}_{\boldsymbol{\Phi}^{(t)}}[l_c(\boldsymbol{\Phi}; \boldsymbol{Y}, \boldsymbol{\theta})|\boldsymbol{Y} = \boldsymbol{y}]. \tag{7.6}$$

– M-step (Maximization) : the parameter estimates are updated by :

$$\boldsymbol{\Phi}^{(t+1)} = \arg\max_{\boldsymbol{\Phi}} Q(\boldsymbol{\Phi}, \boldsymbol{\Phi}^{(t)}).$$

The algorithm is stopped if some convergence criterion is satisfied. In our case, the expression of the Q-function (7.6) contains an integral over the $\boldsymbol{\theta}$-values in $\mathbb{R}^d$, so the E-step cannot be solved analytically. We use therefore an extension of the EM to approximate Q in the E-step, namely the Monte Carlo EM (Wei & Tanner, 1990), and more specifically the automated MCEM version proposed by Booth & Hobert (1999).

7.3.1 Expectation step :

Since the first term of equation (7.5) does not depend on the parameters $\boldsymbol{\Phi}$, the Q-function can be written :

$$Q(\boldsymbol{\Phi}, \boldsymbol{\Phi}^{(t)}) = \mathbb{E}_{\boldsymbol{\Phi}^{(t)}}[\log f_{\boldsymbol{\theta}}(\boldsymbol{\theta}; \boldsymbol{\Phi})|\boldsymbol{Y} = \boldsymbol{y}] + c(\boldsymbol{y}), \tag{7.7}$$

where c is independent of $\boldsymbol{\Phi}$.

To calculate the expectation term in equation (7.7) the density $f_{\boldsymbol{\theta}|\boldsymbol{Y}}(\boldsymbol{\theta}|\boldsymbol{y};\boldsymbol{\Phi}^{(t)})$, given by :

$$f_{\boldsymbol{\theta}|\boldsymbol{Y}}(\boldsymbol{\theta}|\boldsymbol{y};\boldsymbol{\Phi}) = \frac{f_{\boldsymbol{Y}|\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{\theta}) f_{\boldsymbol{\theta}}(\boldsymbol{\theta};\boldsymbol{\Phi})}{f_{\boldsymbol{Y}}(\boldsymbol{y};\boldsymbol{\Phi})}, \tag{7.8}$$

has to be evaluated. The evaluation of $f_{\boldsymbol{Y}}(\boldsymbol{y};\boldsymbol{\Phi})$ is difficult because of the integral in equation (7.1). The solution offered by Monte Carlo EM is to simulate at each EM iteration t and for each observation $\boldsymbol{y}_i$ a random sample $\boldsymbol{\theta}_{i1}^{(t)}, \ldots, \boldsymbol{\theta}_{iN}^{(t)}$ from the distribution $f_{\boldsymbol{\theta}|\boldsymbol{Y}}$ and to replace Q_i, the conditional expectation of the complete-data log-likelihood at observation site i, with a Monte Carlo approximation of the expectation :

$$Q_i(\boldsymbol{\Phi}, \boldsymbol{\Phi}^{(t)}) \simeq \frac{1}{N} \sum_{k=1}^{N} \log f_{\boldsymbol{\theta}|\boldsymbol{Y}}(\boldsymbol{\theta}_{ik}^{(t)}|\boldsymbol{y}_i;\boldsymbol{\Phi}) + c(\boldsymbol{y}_i).$$

Since the observations $\boldsymbol{y}_i$ are independent, we have :

$$Q(\boldsymbol{\Phi}, \boldsymbol{\Phi}^{(t)}) = \sum_{i=1}^{n} Q_i(\boldsymbol{\Phi}, \boldsymbol{\Phi}^{(t)}).$$

In our case it is difficult to sample from $f_{\boldsymbol{\theta}|\boldsymbol{Y}}$ so we use an alternative of the MCEM algorithm based on importance sampling proposed by Booth & Hobert (1999).

Student importance sampling

The random sample $\boldsymbol{\theta}_{i1}^{(t)}, \ldots, \boldsymbol{\theta}_{iN}^{(t)}$ is simulated from the importance density h_t, which has the same support as $f_{\boldsymbol{\theta}|\boldsymbol{Y}}$. The importance sampling Monte Carlo estimate of Q is defined for a given observation i by the following expression :

$$Q_i(\boldsymbol{\Phi}, \boldsymbol{\Phi}^{(t)}) \simeq \frac{1}{N} \sum_{k=1}^{N} w_{ik} \log f_{\boldsymbol{\theta}}(\boldsymbol{\theta}_{ik}^{(t)}|\boldsymbol{\Phi}) + c(\boldsymbol{y}_i),$$

where w_{ik} are the importance weights defined by :

$$w_{ik} = \frac{f_{\boldsymbol{\theta}|\boldsymbol{Y}}(\boldsymbol{\theta}_{ik}^{(t)}|\boldsymbol{y};\boldsymbol{\Phi}^{(t)})}{h_t(\boldsymbol{\theta}_{ik}^{(t)})} \propto \frac{f_{\boldsymbol{Y}|\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{\theta}) f_{\boldsymbol{\theta}}(\boldsymbol{\theta}_{ik}^{(t)};\boldsymbol{\Phi}^{(t)})}{h_t(\boldsymbol{\theta}_{ik}^{(t)})}$$

and evaluated up to the normalizing constant $f_{\boldsymbol{Y}}(\boldsymbol{y};\boldsymbol{\Phi}^{(t)})$ (which does not depend on $\boldsymbol{\Phi}$ and therefore has no effect on the M-step).

The importance density h_t we use is a multivariate Student t-distribution, as Booth & Hobert (1999) suggested. This has proved to be a very efficient choice when the unknown distribution is approximately ellipsoidal and has a mode (Evans & Swartz, 1996). Its expectation m_t and covariance matrix Σ_t are re-evaluated at each step in order to be approximately :

$$m_t = \mathbb{E}_{\boldsymbol{\Phi}^{(t)}}[\boldsymbol{\theta}|\boldsymbol{y}],$$
$$\Sigma_t = \mathbb{V}_{\boldsymbol{\Phi}^{(t)}}[\boldsymbol{\theta}|\boldsymbol{y}].$$

These quantities are obtained at each MCEM iteration by an iterative algorithm, which corresponds to the procedure used to obtain Penalized Quasi Likelihood (PQL) estimators in GLMM models (Breslow & Clayton, 1993). This algorithm is provided in Appendix 7.7.2 and has to be adapted to the distributions $\mathscr{L}_j$ and link functions g_j used.

Size of the importance sample

The size N of the importance sample is re-evaluated at each step in order to obtain a compromise between the speed of the first iterations and the final precision of the estimation. N is increased with the number of iterations by using the automatic procedure proposed by Booth & Hobert (1999) based on a normal approximation of the Monte Carlo error. The algorithm is initialized with a small value of N, in order to allow a fast evolution at the start when the current estimator of the parameter may be far from the true value, and N is increased if $\|\boldsymbol{\Phi}^{(t+1)} - \boldsymbol{\Phi}^{(t)}\|$ is small compared to the Monte Carlo error, which means that the $(t+1)$th iteration was useless because it was "swamped" by Monte Carlo error.

7.3.2 Maximization step

In our case, the M-step has an explicit solution. The parameter estimates are obtained by the following expressions :

$$\boldsymbol{\mu}^{(t+1)} = \frac{\sum_{i=1}^{n}\sum_{k=1}^{N} w_{ik}\boldsymbol{\theta}_{ik}^{(t)}}{\sum_{i=1}^{n}\sum_{k=1}^{N} w_{ik}}, \tag{7.9}$$

$$\boldsymbol{\Sigma}^{(t+1)} = \frac{\sum_{i=1}^{n}\sum_{k=1}^{N} w_{ik}(\boldsymbol{\theta}_{ik}^{(t)} - \boldsymbol{\mu}^{(t+1)})(\boldsymbol{\theta}_{ik}^{(t)} - \boldsymbol{\mu}^{(t+1)})^T}{\sum_{i=1}^{n}\sum_{k=1}^{N} w_{ik}}, \tag{7.10}$$

which represent the weighted average and the weighted empirical variance of the importance sample simulated at the final iteration. These expressions are obtained by deriving Q_N with respect to $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ and solving the equation $\partial Q_N/\partial\boldsymbol{\mu} = 0$ and $\partial Q_N/\partial\boldsymbol{\Sigma} = 0$ respectively. The reader is referred to Appendix 7.7.3 for a detailed proof of equation (7.10) in the bivariate case. The proof of equation (7.9) does not pose any difficulty.

7.3.3 Stopping rule

The following stopping rule is used :

$$\max_l \frac{|\boldsymbol{\Phi}_l^{(t+1)} - \boldsymbol{\Phi}_l^{(t)}|}{\sqrt{\mathbb{V}(\hat{\boldsymbol{\Phi}}_l)} + \delta_1} < \delta_2, \tag{7.11}$$

where $\delta_1 = 0.001$, $\delta_2 = 0.01$ and l labels the parameters. The asymptotic variance of the parameter estimates $\mathbb{V}(\hat{\boldsymbol{\Phi}})$ is obtained by using an estimate of the observed Fisher information evaluated at the current parameter estimate (Booth & Hobert, 1999; Tanner, 1991).

Dividing by $\sqrt{\mathbb{V}(\hat{\boldsymbol{\Phi}}_l)}$ instead of $|\boldsymbol{\Phi}_l^{(t)}|$, which is used in standard convergence criterions for deterministic EM algorithms, avoids unnecessary iterations when the estimate is very small compared to its standard error. The algorithm is stopped when rule (7.11) is satisfied for 3 consecutive iterations, in order to "reduce the risk of stopping the algorithm prematurely because of an unlucky Monte Carlo sample" (Booth & Hobert, 1999).

7.4 Simulation studies

Results for two of the submodels presented in Section 7.2, namely the bivariate Poisson-lognormal model (BPLN) and the Binomial-Poisson model, are shown in this section. The Poisson-Normal model is in fact easier to estimate than the bivariate Poisson-lognormal model, since there is only one hidden variable $\boldsymbol{\theta}_1$ instead of two. The results of this submodel are very similar to those of the BPLN submodel and therefore not presented here.

The range of parameters which can be estimated and problems of "practical" identifiability of parameters are studied briefly in the context of the BPLN model. In the Binomial-Poisson case we discuss the precision of the asymptotic standard deviation of the parameter estimates for different sample sizes.

Computer code (in R) is available from the authors upon demand.

7.4.1 BPLN model

A result of a single run of our estimation procedure is given in Figure 7.2. Our algorithm converged in about 60 iterations. The size N of the importance sample was plotted to illustrate the automatic increase of N with the number of iterations. N increased from an initial value of 10 to 10000 at convergence of the algorithm. The boxplot at the final iteration indicates the results obtained on $n_s = 100$ datasets of size $n = 400$ simulated with the same parameters ($\mu_1 = 1$, $\mu_2 = 0$, $\sigma_1^2 = 0.5$, $\sigma_2^2 = 2$, $r = -0.3$). These results show that our estimation is centered on the true value of the parameters, and the variance of the estimators is reasonably small.

	μ_1	μ_2	σ_1^2	r	σ_2^2
true value φ	1	0	0.5	-0.3	2
mean estimate $\bar{\varphi}$	1.00	0.00	0.49	-0.30	1.98
mASD	0.05	0.10	0.07	0.10	0.27
ESD	0.05	0.09	0.07	0.08	0.23
% IC95	96	99	97	95	97

Tab. 7.1. Results for the BPLN model obtained on $n_s = 100$ runs with samples of size $n = 400$.

Each run of our algorithm provides an estimation of the asymptotic standard deviation (ASD) of the parameter estimates. Let φ denote an element of the set of parameters $\boldsymbol{\Phi}$. The final estimates of the parameters ($\hat{\varphi}$), the mean asymptotic standard deviation (mASD), the empirical standard deviation (ESD) and the percentage of ASD leading to a 95% confidence interval ($\hat{\varphi} \pm 1.96$ ASD) containing the true value φ (% IC95) are given in Table 7.1. The

	$\mathbb{E}(e^\theta)$	$\mathbb{V}(e^\theta)$	$\mu = \mathbb{E}(\theta)$	$\sigma^2 = \mathbb{V}(\theta)$	r
true value φ	2	2	0.49	0.41	-0.9
estimate $\bar{\varphi} \pm 1.96$ ESD			0.475 $\pm$0.2	0.41 $\pm$0.23	$\mathbf{-0.78 \pm 0.35}$
true value φ	2	100	-0.94	3.26	-0.9
estimate $\bar{\varphi} \pm 1.96$ ESD			-0.97 $\pm$0.18	3.35 $\pm$0.2	$\mathbf{-0.88 \pm 0.2}$
true value φ	100	2	4.61	$2 \cdot 10^{-4}$	-0.9
estimate $\bar{\varphi} \pm 1.96$ ESD			4.6 $\pm$ 0.02	$10^{-3} \pm 2 \cdot 10^{-3}$	$\mathbf{-0.09 \pm 1.8}$

Tab. 7.2. Identifiability of the correlation coefficient r for the BPLN model. Results obtained over 100 datasets of size $n = 400$ simulated with parameters $\mu_1 = \mu_2 = \mu$, $\sigma_1^2 = \sigma_2^2 = \sigma^2$ and $r = -0.9$.

ESD was calculated for each parameter φ by the following formula :

$$\mathrm{ESD}(\hat{\varphi}) = \sqrt{\frac{\sum_{s=1}^{n_s} (\hat{\varphi}_s - \bar{\varphi})^2}{n_s - 1}},$$

where n_s is the number of runs of our algorithm, $\hat{\varphi}_s$ is the estimate of φ obtained at run s and $\bar{\varphi}$ is the mean calculated over the n_s simulations.

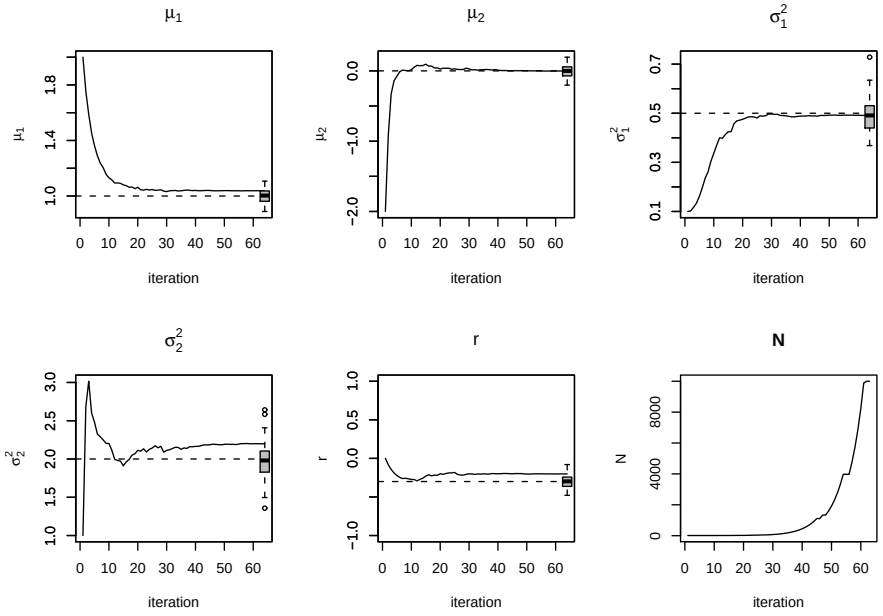


Fig. 7.2. MCEM simulation result for the BPLN model. The true values of the parameters are represented by the horizontal dashed line. The boxplots at the last iteration were obtained on 100 datasets simulated with the same parameters with $n = 400$

Parameter identifiability and estimation limits :

Due to the properties of the Poisson distribution, the correlation coefficient r of the BPLN model is not identifiable for some parameter values : a large mean associated with a small variance for the variable e^{θ} leads to a bad estimation of r (see Table 7.2), because a Poisson distribution with high mean will have a high variance, so the variance of the Poisson will erase or "swamp" the correlation between variables.

7.4.2 Binomial-Poisson model

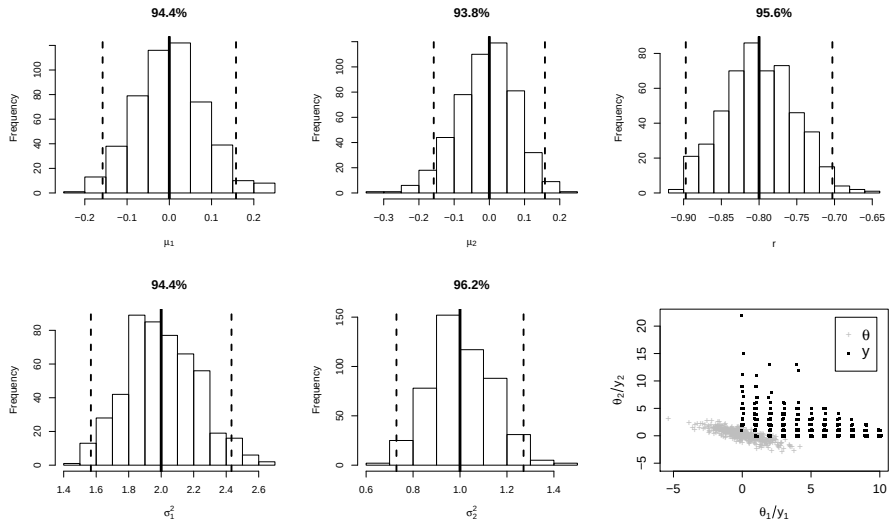


Fig. 7.3. Histograms of the parameter estimates of the Binomial-Poisson model over 500 runs and true values (bold line). The 95% confidence interval was computed using the mean asymptotic standard deviation (dashed lines). The percentage of ASD leading to a 95% confidence interval which contains the true parameter value is given above each plot. Bottomright : example of a dataset simulated with these parameter values (both $\boldsymbol{\theta}$ and $\boldsymbol{Y}$ are represented)

The final estimates obtained on $n_s = 500$ data samples of size $n = 400$ simulated according to the Binomial-Poisson model with parameters $\mu_1 = 0$, $\mu_2 = 0$, $\sigma_1^2 = 2$, $\sigma_2^2 = 1$, $r = -0.8$ and $n_b = 10$ are given in Figure 7.3 and in Table 7.3.

One of the advantages of this estimation procedure is the possibility to obtain with a single run of our algorithm the asymptotic standard deviation (ASD) of the parameter estimates. The ASD is very reliable when the data size is large, so our algorithm can be run only once to obtain the parameter estimates with an accurate confidence interval. To show this, we compared the precision of the mean ASD (mASD) with the empirical standard deviation (ESD) calculated over 500 runs. To take into account the variability of the ASD estimates over the n_s simulations, we tested for each run if the true value of the parameter was in the 95% confidence interval obtained using the corresponding ASD estimate (% IC95). The percentage of positive tests are given above each plot and in Table 7.3.

	μ_1	μ_2	σ_1^2	r	σ_2^2
true value	0.0	0.0	2.0	-0.8	1.0
mean estimate	0.00	-0.01	2.00	-0.80	1.00
mASD	0.08	0.08	0.22	0.05	0.14
ESD	0.08	0.08	0.22	0.05	0.14
% IC95	94.4	93.8	94.4	95.6	96.2

Tab. 7.3. Results for the Binomial-Poisson model obtained on 500 runs with samples of size $n = 400$.

		μ_1	μ_2	σ_1^2	r	σ_2^2
size n	true value	2.0	1.0	2.0	0.8	1.0
400	mean estimate	2.00	1.00	1.99	0.81	1.00
200		2.00	0.99	2.03	0.81	1.00
100		1.98	0.99	1.92	0.82	1.04
50		2.00	0.96	2.11	0.82	1.02
30		1.98	1.01	1.88	0.82	0.91

Tab. 7.4. Mean estimates of the parameters of the Binomial-Poisson model over 100 runs for different sample sizes. The Standard deviation of the estimates are given in Table 7.5.

Influence of sample size on the ASD estimates :

To see if the ASD remains reliable with smaller sample sizes, we performed a similar procedure on 100 runs of our algorithm on data samples simulated according to a Binomial-Poisson model with parameters $\mu_1 = 2$, $\mu_2 = 1$, $\sigma_1^2 = 2$, $\sigma_2^2 = 1$, $r = 0.8$ and different sample sizes. The resulting mean parameter estimates are given in Table 7.4 and the evolution of the ASD with the sample size is provided in Table 7.5 for $n \in \{30, 50, 100, 200, 400\}$. As expected, the mean estimates and the ASD become worse when the sample size becomes smaller, but they remain meaningful to test for example for a positive or negative correlation.

7.5 An application to beehive data

In this section we illustrate our model on two bivariate datasets extracted from a survey of the activity of honey-bee colonies over a large observatory in the south of France.

7.5.1 Beehive dataset

300 hives nested in 20 apiaries were weighed every two days during 24 days in June 2009. The variation with time of the weight of individual hives was modeled by a logistic curve in order to estimate the maximum weight gain over this period for each hive. This weight gain is a continuous variable that corresponds mainly to the production of honey during the 24 day period (in kg) and is denoted WG. The number of capped brood cells was measured in each hive at day 0, 12 and 24 (D0, D12, D24) and is used as a proxy for new bee recruitment (C0, C12, C24). Since the development from a capped cell to an emerging bee is 12 days for a working bee, the counts at a 12-day interval are considered non-overlapping, and thus

size n		μ_1	μ_2	σ_1^2	r	σ_2^2
400	mASD	0.10 (94%)	0.06 (96%)	0.26 (96%)	0.05 (94%)	0.10 (95%)
	ESD	0.10	0.06	0.26	0.04	0.09
200	mASD	0.14 (91%)	0.09 (96%)	0.37 (92%)	0.07 (93%)	0.15 (95%)
	ESD	0.15	0.09	0.40	0.06	0.15
100	mASD	0.19 (97%)	0.13 (92%)	0.50 (90%)	0.09 (90%)	0.22 (93%)
	ESD	0.16	0.14	0.57	0.12	0.25
50	mASD	0.28 (94%)	0.18 (96%)	0.79 (94%)	0.14 (90%)	0.31 (95%)
	ESD	0.29	0.19	0.80	0.12	0.31
30	mASD	0.36 (93%)	0.23 (95%)	0.98 (83%)	0.22 (90%)	0.39 (86%)
	ESD	0.35	0.22	1.05	0.17	0.35

Tab. 7.5. Evolution of the mean asymptotic standard deviation (mASD) and the empirical standard deviation (ESD) with the sample size for the Binomial-Poisson model estimated in Table 7.4. The percentage of ASD leading to a 95% confidence interval which contains the true parameter value are given between brackets. These results were computed over 100 runs of the algorithm.

the sum of C0, C12 and C24 is used to estimate the new bee recruitment over the whole period.

The number of ectoparasite mite *Varroa jacobsoni* was measured for each honey-bee colony on a sample of 20 g of adult bees (corresponding to approximately 150 bees) at D0 and D24 (V0, V24). The Varroa mite has an economic impact on the beekeeping industry and may be a contributing factor to colony collapse disorder (CCD). A recent study by Guzmán et al. (2010) shows that it is the main factor for collapsed colonies in Ontario, Canada.

We considered 3 variables from this dataset for each hive : the total number of capped brood cells (C=C0+C12+C24 divided by 100), the total number of parasites (V=V0+V24) and the weight gain (WG), which is a positive continuous variable.

7.5.2 Results

The BPLN model was used to fit the bivariate distribution of capped brood cell number (C) and bee mite number (V) given in figure 7.4. The Poisson-Normal model was used for the count variable C and the continuous variable WG (figure 7.5). An exponential inverse link function was used for the continuous variable, in order to satisfy the condition WG > 0, so WG $= Y_2 = e^{\theta_2}$. The estimation results for the two models are given in table 7.6. As expected, the number of capped brood cells C is negatively correlated to the number of parasites V, whereas the hive weight gain is positively correlated to C. The estimates of the parameters μ_1 and σ_1^2, which correspond to the count variable C in both models, are very close, so the estimation procedure is stable. In the Poisson-Normal model the maximum

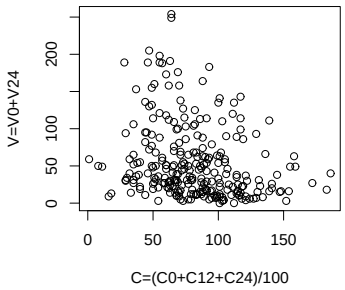


Fig. 7.4. Number of capped cells (C) and number of Varroa mite (V) for 259 beehives (hives with missing data were excluded)

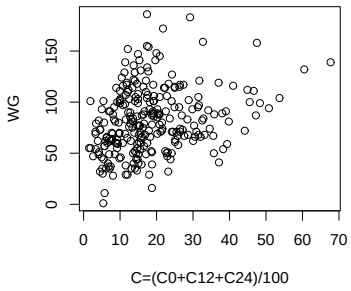


Fig. 7.5. Number of capped cells (C) and weight gain in kg (WG) for 269 beehives (hives with missing data were excluded)

likelihood estimators of μ_2 and σ_2^2 are obtained in 1 iteration and correspond to the empirical marginal mean and variance of the continuous variable $\log(\text{WG}) = \theta_2$.

	C–V		C–WG	
Parameter	Estimate	ASD	Estimate	ASD
μ_1	4.32	0.03	4.33	0.02
σ_1^2	0.18	0.02	0.16	0.01
r	**-0.26**	0.06	**0.33**	0.06
μ_2	3.59	0.06	2.71	0.04
σ_2^2	0.95	0.09	0.43	0.06

Tab. 7.6. Estimation results and asymptotic standard deviation (ASD) of the BPLN model (variables C and V) and of the Poisson-Normal model (variables C and WG). Index 2 for the parameters refers respectively to the number of Varroa mites V (count variable) in the first case and to the weight gain WG (continuous variable) in the second case.

7.6 Discussion

In this article a general parametric model for multivariate data of various types is presented and a maximum likelihood estimation method, based on a variant of the Monte Carlo EM proposed by Booth & Hobert (1999), is provided. The hierarchical structure of the model and the estimation procedure are easy to adapt to different distributions and link functions which are used to obtain data of different types in our model. It also provides the asymptotic standard deviation of the estimators, which are a useful indication of the precision of the estimation for a small computational effort.

Limits of this method are first due to the model, some parameters cannot be estimated and there may be identifiability issues for some parts of the parameter domain, as shown in Section 7.4. Estimation issues can arise when the variance of the hidden model is too high.

Possible extensions of this model include spatial studies, GLMM models with multivariate random effects of different types (McCulloch & Searle, 2001), and mixture models. A spatial autocorrelation model could be introduced in the data, in a similar way to the work of Chagneau et al. (2010), but in a maximum likelihood framework without prior distributions on the parameters. This model could also be used in the context of GLMM models, the field in which this estimation procedure was proposed by Booth & Hobert (1999), when multivariate random effects can be defined (see for example Lai & Yau, 2008; Wang et al., 2006). Finally, the class of distributions we defined in this paper could be used in a multivariate finite mixture model in a clustering context. It would allow clustering of all kinds of variables and even to data containing attributes of different types, which, to our knowledge and according to Fraley & Raftery (2002), has not been achieved yet. Moreover, our MCEM estimation procedure would be a direct extension of the EM procedure which is traditionally used to estimate multivariate mixture models.

Acknowledgement

This work was supported by the "Institut National de la Recherche Agronomique " (INRA) and the French region Provence Alpes Côte d'Azur. The beehive dataset was provided by an INRA - ADAPI ("Association pour le développement de l'apiculture provençale") project, funded by the "Fonds européen agricole de garantie" (FEAGA).

7.7 Appendices

7.7.1 Unconditional moments of the bivariate Binomial-Poisson model

Let us recall the bivariate Binomial-Poisson model given in Section 7.2.3 :

$$\begin{cases} Y_1|\theta_1 \sim \mathscr{B}(n_b, \frac{1}{1+e^{-\theta_1}}) \\ Y_2|\theta_2 \sim \mathscr{P}(e^{\theta_2}) \\ (\theta_1, \theta_2)^T \sim \mathscr{N}_2(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \end{cases}$$

For sake of simplicity, we omit the index i in this appendix section and write θ_1 instead of θ_{i1}. Let $\boldsymbol{\Phi}$ denote the vector of unknown parameters $(\mu_1, \mu_2, \sigma_{11}, \sigma_{22}, \sigma_{12})$. To initialize our MCEM estimation procedure, an approximation of the parameter vector $\boldsymbol{\Phi}$ can be obtained by the method of moments.

The moment estimators of μ_2, σ_{22} are obtained directly by using the following equations (given in Section 7.2.3) :

$$\mathbb{E}(Y_2) = e^{\mu_2 + \frac{1}{2}\sigma_{22}} \overset{def}{=} m_2, \tag{7.12}$$

$$\mathbb{V}(Y_2) = m_2 + m_2^2 \, (e^{\sigma_{22}} - 1) \overset{def}{=} v_2. \tag{7.13}$$

Equations (7.12) and (7.13) yield :

$$\hat{\mu}_2 = 2\log(m_2) - \frac{1}{2}\log(v_2 - m_2 + m_2^2),$$
$$\hat{\sigma}_{22} = \log(v_2 - m_2 + m_2^2) - 2\log(m_2).$$

To obtain moment estimators for $\mu_1, \sigma_{11}, \sigma_{12}$, we write the following statistics :

$$\mathbb{E}(Y_1) = n_b \mathbb{E}\Big(\frac{1}{1+e^{-\theta_1}}\Big) = n_b \int_{\mathbb{R}} \frac{1}{1+e^{-\theta_1}} \frac{1}{\sqrt{2\pi}} \, e^{-\frac{1}{2}\frac{(\theta_1-\mu_1)^2}{\sigma_{11}}} d\theta_1 = n_b F_1(\boldsymbol{\Phi}),$$
$$\mathbb{E}(Y_1^2) - \mathbb{E}(Y_1) = n_b(n_b-1)\mathbb{E}\Big(\frac{1}{(1+e^{-\theta_1})^2}\Big) = n_b(n_b-1)F_2(\boldsymbol{\Phi}),$$
$$\mathbb{E}(Y_1 Y_2) = \mathbb{E}(\mathbb{E}(Y_1 Y_2|\theta_1, \theta_2)) = \mathbb{E}(\frac{n_b}{1+e^{-\theta_1}} e^{\theta_2}) = n_b F_3(\boldsymbol{\Phi}),$$

where :

$$F_1(\boldsymbol{\Phi}) = \int_{\mathbb{R}} \frac{1}{1+e^{-\theta_1}} \frac{1}{\sqrt{2\pi}} \, e^{-\frac{1}{2}\frac{(\theta_1-\mu_1)^2}{\sigma_{11}}} d\theta_1$$
$$F_2(\boldsymbol{\Phi}) = \int_{\mathbb{R}} \frac{1}{(1+e^{-\theta_1})^2} \frac{1}{\sqrt{2\pi}} \, e^{-\frac{1}{2}\frac{(\theta_1-\mu_1)^2}{\sigma_{11}}} d\theta_1$$
$$F_3(\boldsymbol{\Phi}) = \iint_{\mathbb{R}^2} \frac{e^{\theta_2}}{1+e^{-\theta_1}} \, \frac{1}{2\pi|\Sigma|^{1/2}} \, e^{-\frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\mu})\Sigma^{-1}(\boldsymbol{\theta}-\boldsymbol{\mu})} d\theta_1 d\theta_2.$$

With the change of variable $Z_1 = \frac{\theta_1-\mu_1}{\sqrt{\sigma_{11}}}$, we have $dZ_1 = \frac{d\theta_1}{\sqrt{\sigma_{11}}}$, $\theta_1 = \sqrt{\sigma_{11}} Z_1 + \mu_1$ and $F_1(\boldsymbol{\Phi})$ becomes :

$$F_1(\boldsymbol{\Phi}) = \int_{\mathbb{R}} \frac{1}{1+e^{-(\sqrt{\sigma_{11}} Z_1 + \mu_1)}} \frac{1}{\sqrt{2\pi}} \, e^{-\frac{1}{2}Z_1^2} dZ_1$$

With the change of variable $\boldsymbol{Z}^T = (\boldsymbol{\theta} - \boldsymbol{\mu})^T |\boldsymbol{\Sigma}^{-1/2}|$, we have $\boldsymbol{\theta} = |\boldsymbol{\Sigma}^{1/2}|\boldsymbol{Z} + \boldsymbol{\mu}$. We have to find :

$$\boldsymbol{\Sigma}^{1/2} = \begin{pmatrix} a & b \\ b & c \end{pmatrix}.$$

such that $a > 0$, $c > 0$ and $ac - b^2 \geq 0$ ($\Sigma^{1/2}$ has to be positive definite). The solution is :

$$\begin{aligned} a &= \frac{\sigma_{11} + \Delta}{H} \\ c &= \frac{\sigma_{22} + \Delta}{H} \\ b &= \frac{\sigma_{12}}{H}. \end{aligned}$$

where $\Delta = \sqrt{\sigma_{11}\sigma_{22} - \sigma_{12}^2}$ and $H = \sqrt{\sigma_{11} + \sigma_{22} + 2\Delta}$. Thus we have :

$$\begin{aligned} \theta_1 &= aZ_1 + bZ_2 + \mu_1 \\ \theta_2 &= bZ_1 + cZ_2 + \mu_2 \end{aligned}$$

and $F_3(\boldsymbol{\Phi})$ becomes :

$$F_3(\boldsymbol{\Phi}) = \iint_{\mathbb{R}^2} \frac{e^{bZ_1+cZ_2+\mu_2}}{1 + e^{-(aZ_1+bZ_2+\mu_1)}} \frac{1}{2\pi} e^{-\frac{1}{2}\boldsymbol{Z}^T\boldsymbol{Z}} d\theta_1 d\theta_2.$$

We use the Newton-Raphson iterative procedure to obtain moment estimators for $\mu_1, \sigma_{11}, \sigma_{12}$. We need to compute the derivatives :

$$\begin{aligned} \frac{\partial F_1}{\partial \mu_1} &= \int_{\mathbb{R}} \frac{e^{-(\sqrt{\sigma_{11}}Z_1+\mu_1)}}{(1 + e^{-(\sqrt{\sigma_{11}}Z_1+\mu_1)})^2} \Phi_0(dZ_1) \\ \frac{\partial F_1}{\partial \sigma_{12}} &= 0 \\ \frac{\partial F_1}{\partial \sigma_{11}} &= \int_{\mathbb{R}} \frac{Z_1}{2\sqrt{\sigma_{11}}} \frac{e^{-(\sqrt{\sigma_{11}}Z_1+\mu_1)}}{(1 + e^{-(\sqrt{\sigma_{11}}Z_1+\mu_1)})^2} \Phi_0(dZ_1) \\ \frac{\partial F_2}{\partial \mu_1} &= \int_{\mathbb{R}} \frac{2e^{-(\sqrt{\sigma_{11}}Z_1+\mu_1)}}{(1 + e^{-(\sqrt{\sigma_{11}}Z_1+\mu_1)})^3} \Phi_0(dZ_1) \\ \frac{\partial F_2}{\partial \sigma_{12}} &= 0 \\ \frac{\partial F_2}{\partial \sigma_{11}} &= \int_{\mathbb{R}} -\frac{Z_1}{\sqrt{\sigma_{11}}} \frac{e^{-(\sqrt{\sigma_{11}}Z_1+\mu_1)}}{(1 + e^{-(\sqrt{\sigma_{11}}Z_1+\mu_1)})^3} \Phi_0(dZ_1) \\ \frac{\partial F_3}{\partial \mu_1} &= \iint_{\mathbb{R}^2} \frac{R_1 R_2}{(1 + R_1)^2} \Phi_0(dZ_1, dZ_2) \\ \frac{\partial F_3}{\partial \sigma_{12}} &= \iint_{\mathbb{R}^2} \frac{(b_{12}Z_1 + c_{12}Z_2)R_1(1 + R_2) + R_1(a_{12}Z_1 + b_{12}Z_2)R_2}{(1 + R_2)^2} \\ \frac{\partial F_3}{\partial \sigma_{11}} &= \iint_{\mathbb{R}^2} \frac{(b_{11}Z_1 + c_{11}Z_2)R_1(1 + R_2) + R_1(a_{11}Z_1 + b_{11}Z_2)R_2}{(1 + R_2)^2}, \end{aligned}$$

with the notations :

$$
\begin{aligned}
R_1 &= e^{-(aZ_1+bZ_2+\mu_1)} \\
R_2 &= e^{(bZ_1+cZ_2+\mu_2)} \\
a_{11} &= \frac{\partial a}{\partial \sigma_{11}} = \frac{(\sigma_{22}+2\Delta)H^2-(\sigma_{11}+\Delta)(\sigma_{22}+\Delta)}{2\Delta H^3} \\
a_{12} &= \frac{\partial a}{\partial \sigma_{12}} = \frac{\sigma_{12}(\sigma_{11}-H^2+\Delta)}{\Delta H^3} \\
b_{11} &= \frac{\partial b}{\partial \sigma_{11}} = -\frac{\sigma_{12}(\sigma_{22}+\Delta)}{2\Delta H^3} \\
b_{12} &= \frac{\partial b}{\partial \sigma_{12}} = \frac{\sigma_{12}^2+\Delta H^2}{\Delta H^3} \\
c_{11} &= \frac{\partial c}{\partial \sigma_{11}} = \frac{\sigma_{22}H^2-(\sigma_{22}+\Delta)^2}{2\Delta H^3} \\
c_{12} &= \frac{\partial c}{\partial \sigma_{12}} = \frac{\sigma_{12}(\sigma_{22}-H^2+\Delta)}{\Delta H^3}.
\end{aligned}
$$

7.7.2 PQL estimators of the conditional moments for different distributions and link functions

Distribution	$p = g^{-1}(\theta)$	$\lambda = \mathbb{E}(Y_j\|\theta)$	$g(\lambda)$	$\partial g(\lambda)/\partial\lambda$	$v = \mathbb{V}(Y_j\|\theta)$
Poisson(p)	e^{θ}	$p = e^{\theta}$	$\log(\lambda)$	$e^{-\theta}$	e^{θ}
Binomial(n_b,p)	$\frac{1}{1+e^{-\theta}}$	$n_b p = \frac{n_b}{1+e^{-\theta}}$	$\log\frac{\lambda}{1-\lambda}$	$\frac{(1+e^{-\theta})^2}{n_b e^{-\theta}}$	$n_b p(1-p) = \frac{n_b e^{-\theta}}{(1+e^{-\theta})^2}$

Tab. 7.7. Expressions of λ, v and $g_\lambda(\lambda)$ for the Poisson distribution and the Binomial distribution with fixed number of trials n_b.

The parameters $\boldsymbol{m}_t$ and $\boldsymbol{\Sigma}_t$ of the multivariate Student t distribution used in the E-step of the MCEM to obtain an importance sample, are re-evaluated at each step in order to be approximately :

$$
\begin{aligned}
\boldsymbol{m}_t &= \mathbb{E}_{\boldsymbol{\Phi}^{(t)}}[\boldsymbol{\theta}|\boldsymbol{y}], \\
\boldsymbol{\Sigma}_t &= \mathbb{V}_{\boldsymbol{\Phi}^{(t)}}[\boldsymbol{\theta}|\boldsymbol{y}].
\end{aligned}
$$

The Penalized Quasi Likelihood (PQL) estimators of these conditional moments are obtained by using a Laplace approximation of the likelihood and a Fisher scoring maximization procedure.

At iteration $(t+1)$ of the MCEM algorithm, the Fisher scoring iterative algorithm is used, given by :

$$
\boldsymbol{m}_t^{(k+1)} = \boldsymbol{m}_t^{(k)} + (\boldsymbol{W}(\boldsymbol{m}_t^{(k)}) + \boldsymbol{\Sigma}^{-1})^{-1}\Big(\boldsymbol{W}(\boldsymbol{m}_t^{(k)})\boldsymbol{\Delta}(\boldsymbol{m}_t^{(k)})(\boldsymbol{y} - \boldsymbol{\lambda}(\boldsymbol{m}_t^{(k)})) - \boldsymbol{\Sigma}^{-1}(\boldsymbol{m}_t^{(k)} - \boldsymbol{\mu})\Big),
$$

$$
\boldsymbol{\Sigma}_t^{(k+1)} = \Big(\boldsymbol{W}(\boldsymbol{m}_t^{(k+1)}) + \boldsymbol{\Sigma}^{-1}\Big)^{-1},
$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are the current estimators of the mean and variance of $\boldsymbol{\theta}$ (from the MCEM iteration t), and the matrices $\boldsymbol{\lambda}(\boldsymbol{m}_t^{(k)})$, $\boldsymbol{W}(\boldsymbol{m}_t^{(k)})$ and $\boldsymbol{\Delta}(\boldsymbol{m}_t^{(k)})$ are defined (for a single-parameter exponential distribution) by :

$$\begin{aligned}
\lambda_i &= \frac{\partial b(\eta)}{\partial \eta} = \mathbb{E}(Y_j|\theta),\\
v_i &= \frac{\partial^2 b(\eta)}{\partial \eta^2} = \mathbb{V}(Y_j|\theta),\\
g_\lambda(\lambda_i) &= \frac{\partial g(\lambda_i)}{\partial \lambda_i},\\
\boldsymbol{W}(\boldsymbol{\lambda}) &= \mathrm{diag}\Big(\frac{1}{v_i g_\lambda^2(\lambda_i)}\Big),\\
\boldsymbol{\Delta}(\boldsymbol{\lambda}) &= \mathrm{diag}(g_\lambda(\lambda_i)),
\end{aligned}$$

where $\mathrm{diag}(x_i)$ denotes a diagonal matrix with diagonal element x_i (at row and column i) and with the notations of section 7.2.2. The PQL estimators of $\boldsymbol{m}_t$ and $\boldsymbol{\Sigma}_t$ should thus be adapted to the model by calculating these expressions for the chosen distributions $\mathscr{L}$ and link functions g (cf table 7.7).

7.7.3 Maximum likelihood estimators of the multivariate normal parameters in the M-step of the MCEM

The maximum likelihood estimator of $\boldsymbol{\Sigma}$ is obtained by deriving Q_N with respect to $\boldsymbol{\Sigma}$ and solving the equation $\partial Q_N/\partial \boldsymbol{\Sigma} = 0$. This is equivalent to solving $\partial Q_N/\partial \boldsymbol{\Gamma} = 0$ where $\boldsymbol{\Gamma} = \boldsymbol{\Sigma}^{-1}$. The result, provided in equation (7.10), is recalled here :

$$\hat{\boldsymbol{\Sigma}} = \frac{\sum_{i=1}^{n}\sum_{k=1}^{N} w_{ik}(\boldsymbol{\theta}_{ik} - \hat{\boldsymbol{\mu}})(\boldsymbol{\theta}_{ik} - \hat{\boldsymbol{\mu}})^T}{\sum_{i=1}^{n}\sum_{k=1}^{N} w_{ik}}$$

Proof of (7.10) in the bivariate case :

$$\begin{aligned}
Q_N(\boldsymbol{\Phi}, \boldsymbol{\Phi}^{(t)}) &= \sum_{i=1}^{n} \mathbb{E}_{\boldsymbol{\Phi}^{(t)}}[\log(f_{\boldsymbol{\theta}}(\boldsymbol{\theta}_i; \boldsymbol{\mu}, \boldsymbol{\Sigma}))|\boldsymbol{Y}_i = \boldsymbol{y}_i]\\
&= \frac{1}{N}\sum_{i=1}^{n}\sum_{k=1}^{N} w_{ik}\Big(\log(f_{\boldsymbol{\theta}}(\boldsymbol{\theta}_{ik}|\boldsymbol{\mu}, \boldsymbol{\Sigma}))\Big)\\
&= c + \frac{1}{2N}\sum_{i=1}^{n}\sum_{k=1}^{N} w_{ik}\Big(-\log|\boldsymbol{\Sigma}| - (\boldsymbol{\theta}_i - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}_i - \boldsymbol{\mu})\Big)\\
&= c + u(\boldsymbol{\Sigma})
\end{aligned}$$

where $c = -\frac{1}{2N}\sum_{i=1}^{n}\sum_{k=1}^{N} w_{ik}\log(2\pi)$ is a constant.

Let $\boldsymbol{\Gamma} = \boldsymbol{\Sigma}^{-1}$. Then $|\boldsymbol{\Sigma}| = |\boldsymbol{\Gamma}|^{-1}$ and $u(\boldsymbol{\Sigma})$ is replaced by :

$$v(\boldsymbol{\Gamma}) = \frac{1}{2N}\sum_{i=1}^{n}\sum_{k=1}^{N} w_{ik}\Big(\log|\boldsymbol{\Gamma}| - (\boldsymbol{\theta}_i - \boldsymbol{\mu})^T\boldsymbol{\Gamma}(\boldsymbol{\theta}_i - \boldsymbol{\mu})\Big).$$

Solving $\partial u(\boldsymbol{\Sigma})/\partial \boldsymbol{\Sigma} = 0$ is equivalent to solving $\partial v(\boldsymbol{\Gamma})/\partial \boldsymbol{\Gamma} = 0$. We denote $\boldsymbol{\Gamma} = (\gamma_{ll'})$ where $l, l' \in \{1, 2\}$, $\boldsymbol{\mu} = (\mu_1, \mu_2)$ and $\boldsymbol{\theta} = (\theta_1, \theta_2)$. We have :

$$\begin{aligned}
\frac{\partial v(\boldsymbol{\Gamma})}{\partial \gamma_{11}} &= \frac{1}{2N} \sum_{i=1}^{n} \sum_{k=1}^{N} w_{ik} \Big(\frac{\gamma_{22}}{|\boldsymbol{\Gamma}|} - (\theta_1 - \mu_1)^2 \Big) \\
\frac{\partial v(\boldsymbol{\Gamma})}{\partial \gamma_{22}} &= \frac{1}{2N} \sum_{i=1}^{n} \sum_{k=1}^{N} w_{ik} \Big(\frac{\gamma_{11}}{|\boldsymbol{\Gamma}|} - (\theta_2 - \mu_2)^2 \Big) \\
\frac{\partial v(\boldsymbol{\Gamma})}{\partial \gamma_{12}} &= \frac{1}{2N} \sum_{i=1}^{n} \sum_{k=1}^{N} w_{ik} \Big(- \frac{\gamma_{12}}{|\boldsymbol{\Gamma}|} - 2(\theta_1 - \mu_1)(\theta_2 - \mu_2) \Big)
\end{aligned}$$

and $\gamma_{12} = \gamma_{21}$. $\partial v(\boldsymbol{\Gamma})/\partial \boldsymbol{\Gamma} = 0$ is equivalent to :

$$\begin{aligned}
\frac{\gamma_{11}}{|\boldsymbol{\Gamma}|} &= \frac{\sum_{i=1}^{n} \sum_{k=1}^{N} w_{ik} (\theta_2 - \mu_2)^2}{\sum_{i=1}^{n} \sum_{k=1}^{N} w_{ik}} \\
\frac{\gamma_{22}}{|\boldsymbol{\Gamma}|} &= \frac{\sum_{i=1}^{n} \sum_{k=1}^{N} w_{ik} (\theta_1 - \mu_1)^2}{\sum_{i=1}^{n} \sum_{k=1}^{N} w_{ik}} \\
\frac{\gamma_{12}}{|\boldsymbol{\Gamma}|} &= \frac{-2 \sum_{i=1}^{n} \sum_{k=1}^{N} w_{ik} (\theta_1 - \mu_1)(\theta_2 - \mu_2)}{\sum_{i=1}^{n} \sum_{k=1}^{N} w_{ik}}.
\end{aligned}$$

Since :

$$\boldsymbol{\Sigma} = \boldsymbol{\Gamma}^{-1} = \frac{1}{|\boldsymbol{\Gamma}|} \begin{pmatrix} \gamma_{22} & -\gamma_{12} \\ -\gamma_{12} & \gamma_{11} \end{pmatrix}$$

we have :

$$\begin{aligned}
\sigma_{11} &= \frac{\gamma_{22}}{|\boldsymbol{\Gamma}|} \\
\sigma_{22} &= \frac{\gamma_{11}}{|\boldsymbol{\Gamma}|} \\
\sigma_{12} &= -\frac{\gamma_{12}}{|\boldsymbol{\Gamma}|}.
\end{aligned}$$

It follows that :

$$\hat{\boldsymbol{\Sigma}} = \frac{\sum_{i=1}^{n} \sum_{k=1}^{N} w_{ik} (\boldsymbol{\theta} - \hat{\boldsymbol{\mu}})(\boldsymbol{\theta} - \hat{\boldsymbol{\mu}})^T}{\sum_{i=1}^{n} \sum_{k=1}^{N} w_{ik}}.$$

Quatrième partie

Utiliser des lois à structure hiérarchique pour classer des données multivariées de types différents

8

Classer des données de types différents en utilisant des modèles hiérarchiques multivariés

Dans la partie III nous avons proposé un modèle hiérarchique qui s'adapte facilement aux différents types de données et qui permet de modéliser la dépendance entre variables de types différents, ainsi qu'une méthode d'estimation par maximum de vraisemblance basée sur un algorithme MCEM. Nous proposons d'étendre ce modèle au cas où la distribution multivariée dans la couche cachée (latente) du modèle n'est plus une gaussienne multivariée mais un mélange de gaussiennes multivariées. Le but de cette démarche est d'utiliser ce modèle dans un contexte de classification automatique. Cette méthode de classification, basée sur des mélanges de lois à structure hiérarchique (modèles latents gaussiens), pourra être utilisée pour classer des données de types différents (*mixed-mode data*).

La classification d'observations lorsque les variables sont dépendantes et de types différents reste un défi actuel (Fraley & Raftery, 2002). Une telle méthode de classification permettrait d'étendre le champ actuel d'utilisation de la classification basé sur les modèles de mélange de la même manière que les modèles linéaires généralisés (GLM) ont permis d'étendre les modèles linéaires (LM) à des cas où la variable réponse n'est pas gaussienne et la dépendance entre variables explicatives et variable réponse n'est pas linéaire (en permettant de spécifier un lien non linéaire). De la même manière, les modèles linéaires généralisés à effets mixtes (GLMM) ont permis d'incorporer des effets aléatoires, qui peuvent modéliser la surdispersion et la corrélation. Notons enfin que d'autres auteurs s'intéressent actuellement à ce sujet et explorent d'autres pistes, notamment Karlis & Arakelian (2010) qui envisagent l'utilisation de mélanges de copules pour classer des données mixtes.

8.1 Introduction au modèle de mélange gaussien latent

Les méthodes de classification basées sur les modèles de mélange de lois paramétriques fournissent des solutions élégantes à plusieurs problèmes inhérents à la classification automatique, tels que le choix du nombre de classes, qui se réduit à un problème classique de sélection de modèle.

L'utilisation de modèles de mélange finis dans le contexte de la classification a été étudiée par de nombreux auteurs (pour des synthèses sur le sujet voir McLachlan & Peel, 2000;

Fraley & Raftery, 2002). Plusieurs lois de probabilités ont été proposées pour modéliser la distribution des composantes du mélange. Ces lois doivent être choisies en fonction du type des données. Les plus fréquemment utilisées sont les lois gaussiennes pour classer des données continues et les lois multinomiales pour classer les données catégorielles (McLachlan & Peel, 2000). La généralisation au cas multivarié n'a été étudiée que pour certaines distributions, telles que la loi gaussienne, les lois de Poisson et de Student. Pour permettre à un large public l'utilisation de ces techniques de classification, plusieurs auteurs ont proposé un cadre général d'estimation et des logiciels, tels que `MIXMOD` de Biernacki et al. (2006) pour classer des données multivariées continues ou discrètes en utilisant respectivement des lois gaussiennes multivariées et des lois multinomiales, et `MCLUST` de Fraley & Raftery (2006) pour classer des données multivariées continues grâce à des mélanges de gaussiennes multivariées. La méthode de classification la plus flexible et la mieux étudiée reste celle basée sur des mélanges de gaussiennes.

Quelques limites de ces méthodes sont :

- pour certaines distributions multivariées, telles que les lois de Poisson multivariées (étudiées par Karlis & Meligkotsidou, 2006), la forme des classes est limitée : la corrélation négative entre variables n'est pas autorisée par construction, et la contrainte d'égalité entre moyenne et variance limite la forme des classes (voir chapitre 5),
- le choix de la distribution de la composante dépend du type de données et est limité dans le cas multivarié aux quelques distributions déjà étudiées,
- l'algorithme d'estimation des paramètres du modèle est difficile à adapter d'une distribution à une autre,
- peu de méthodes capables de classer des données contenant des attributs de types différents (*mixed-mode data*) ont été proposées et celles-ci sont généralement limitées à des cas particuliers (e.g. données continues et catégorielles).

Notre objectif est de proposer une méthode générale de classification, facile à adapter à des données de types différents (discrètes, continues, ordinales, 0-inflatées, qui sont fréquentes en écologie). Ici nous présentons un modèle de mélange fini généralisé pour la classification de données de types différents. Ce modèle constitue une extension du modèle hiérarchique présenté dans le chapitre 7, au cas où la distribution de la variable latente est un mélange fini de gaussiennes multivariées.

La suite de ce chapitre est organisée de la manière suivante :

Comme dans le chapitre 7, nous commençons par présenter le modèle général et quelques sous-modèles. La méthode d'estimation est détaillée en insistant sur les différences par rapport au modèle sans mélange du chapitre 7. Nous insistons également sur un enjeu fondamental de cette méthode : l'initialisation, qui est en quelque sorte le facteur limitant de cette méthode, et proposons quelques solutions qui ne sont pour l'instant que des compromis. Une étude de quelques sous-modèles par simulation est présentée ensuite, puis une discussion sur les possibilités offertes par cette méthode de classification, ainsi que sur ses limites.

8.2 Modèle général

Soit $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ un échantillon aléatoire de taille n, où $\boldsymbol{Y}_i$ est un vecteur aléatoire multivarié de dimension d. En pratique $\boldsymbol{Y}_i = \{\boldsymbol{Y}_{i1}, \ldots, \boldsymbol{Y}_{id}\}$ pourrait correspondre aux abondances de d espèces observées sur le site i.

Nous supposons que $\boldsymbol{Y}$ suit un mélange fini à K composantes, dont les distributions multivariées sont définies par le modèle hiérarchique présenté dans le chapitre 7. En d'autres termes, $\boldsymbol{Y}$ suit un modèle de mélange gaussien latent :

$$\begin{cases} \boldsymbol{Y}_i|\boldsymbol{\theta}_i \sim \mathscr{L}(g^{-1}(\boldsymbol{\theta}_i)) \\ \quad \boldsymbol{\theta}_i \sim \sum_{k=1}^{K} \pi_k \mathscr{N}_d(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \end{cases}$$

où $\mathscr{N}_d(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ est la distribution gaussienne de dimension d, de vecteur moyenne $\boldsymbol{\mu}_k$ et de matrice de covariance $\boldsymbol{\Sigma}_k$, $\mathscr{L} = \mathscr{L}_1 \times \ldots \times \mathscr{L}_d$ est une distribution multivariée formée de d distributions univariées indépendantes de la famille exponentielle, et $g = \{g_1, \ldots, g_d\}$ un ensemble de fonctions de lien. La densité de probabilité $f_{\boldsymbol{Y}}$ de $\boldsymbol{Y}$ est définie par :

$$f_{\boldsymbol{Y}}(\boldsymbol{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \int_{\mathbb{R}^d} f_{\boldsymbol{Y}|\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{\theta}) f_{\boldsymbol{\theta}}(\boldsymbol{\theta}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \, d\boldsymbol{\theta},$$

avec :

$$f_{\boldsymbol{Y}|\boldsymbol{\theta}}(y_1 \ldots y_d|\boldsymbol{\theta}) = \prod_{j=1}^{d} f_{Y_j|\boldsymbol{\theta}}(y_j|\boldsymbol{\theta}),$$

car les variables Y_j, $j \in \{1, \ldots, d\}$, sont supposées indépendantes conditionnellement à $\boldsymbol{\theta}$ et :

$$f_{\boldsymbol{\theta}}(\boldsymbol{\theta}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{k=1}^{K} \pi_k f_k(\boldsymbol{\theta}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k).$$

Comme dans la partie III, le choix de $\mathscr{L}$ et g dépend du type de données (discrètes, continues, ordinales, binaires). Ce modèle peut donc prendre en compte des variables de type différent. Le DAG de la figure 8.1 schématise les relations de dépendance de ce modèle.

Dans le paragraphe suivant quelques exemples de sous-modèles sont donnés. Pour plus d'exemples voir également le chapitre 7.

8.3 Exemples de sous-modèles

8.3.1 Modèle pour données de comptage : Mélange de MPLN

Pour les données de comptages, si on choisit d lois de Poisson univariées indépendantes pour $\mathscr{L}$ et la fonction exponentielle pour g^{-1}, on obtient un mélange de MPLN (*multivariate Poisson log-normal model*). Pour toutes les observations $i \in \{1, \ldots, n\}$ et variables $j \in \{1, \ldots, d\}$ on écrit :

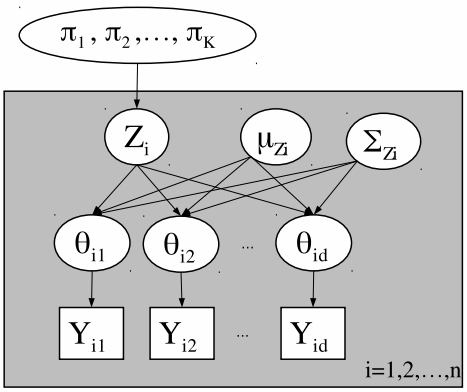


Fig. 8.1. Graphe orienté acyclique (DAG) du modèle de mélange gaussien latent

$$\begin{cases} Y_{ij}|\theta_{ij} \sim \mathscr{P}(e^{\theta_{ij}}) \\ (\theta_{i1},\ldots,\theta_{id})^T \sim \sum_{k=1}^{K} \pi_k \mathscr{N}_d(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \end{cases}$$

Contrairement au mélange de lois de Poisson multivariées considéré dans pour modéliser les assemblages de pucerons dans le chapitre 5, ce modèle permet au sein de chaque classe une surdispersion des données par rapport à la loi de Poisson et une corrélation négative des variables de comptage.

8.3.2 Modèles pour données mixtes continues et discrètes : mélange Normal-Poisson, mélange Normal-Binomial

Pour plus de simplicité, et sans perte de généralité, nous nous plaçons dans un cadre bivarié. Pour modéliser une variable continue et une variable discrète, nous proposons de garder la variable continue telle quelle, ce qui reviendrait à choisir une loi conditionnelle déterministe, et de choisir pour la variable discrète selon les cas soit une loi de Poisson avec un lien logarithmique (mélange Normal-Poisson), soit une loi binomiale avec un lien logit (mélange Normal-Binomial). La fonction de lien qui permet d'obtenir la variable continue peut être soit l'identité, soit un lien logarithmique pour avoir une variable positive.

Le mélange Normal-Poisson est défini ainsi :

$$\begin{cases} Y_{i1}|\theta_{i1} = g_1^{-1}(\theta_{i1}) \\ Y_{i2}|\theta_{i2} \sim \mathscr{P}(e^{\theta_{i2}}) \\ (\theta_{i1},\theta_{i2})^T \sim \sum_{k=1}^{K} \pi_k \mathscr{N}_2(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \end{cases}$$

avec par exemple $g_1^{-1}(\theta_{i1}) = e^{\theta_{i1}}$.

Le mélange Normal-Binomial s'écrit :

$$\begin{cases} Y_{i1}|\theta_{i1} = g_1^{-1}(\theta_{i1}) \\ Y_{i2}|\theta_{i2} \sim \mathscr{B}(n_b, \text{logit}^{-1}(\theta_{i2})) \\ (\theta_{i1},\theta_{i2})^T \sim \sum_{k=1}^{K} \pi_k \mathscr{N}_2(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k). \end{cases}$$

8.4 Estimation par maximum de vraisemblance avec l'algorithme MCEM

Nous nous plaçons dans le cadre classique d'estimation des modèles de mélange, à savoir l'estimation par maximum de vraisemblance grâce à l'algorithme EM présenté en 2.2.1. Nous avons vu que cet algorithme est adapté à l'estimation des paramètres des modèles de mélange, car il s'agit d'un problème de données incomplètes, dans lequel la composante dont provient chaque observation n'est pas connue, et qui est résolu en introduisant explicitement les indicatrices des appartenances aux composantes en tant que variables latentes. Dans ce cas, la variable $\boldsymbol{\theta}$ qui suit le modèle de mélange est également latente, ce qui complique l'estimation par rapport à l'approche classique. Nous adaptons l'algorithme MCEM présenté dans le chapitre 7 au cas où la variable latente $\boldsymbol{\theta}$ suit un mélange gaussien.

On suppose le nombre de composantes K fixé. Le modèle que nous venons de définir dans la sous-partie 8.2 peut s'écrire comme un modèle hiérarchique à trois couches :

$$\boldsymbol{Z}\ (\pi_k) \rightarrow \boldsymbol{\theta}\ (\mu_k, \boldsymbol{\Sigma}_k) \rightarrow \boldsymbol{Y}$$

où $\boldsymbol{Y}$ est le vecteur aléatoire qui correspond aux données observées $\boldsymbol{y}$, $\boldsymbol{Z}$ et $\boldsymbol{\theta}$ sont deux vecteurs aléatoires latents. On suppose que $\boldsymbol{Z} = (\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n)$ tel que le vecteur aléatoire $\boldsymbol{Z}_i$ est distribué selon une loi multinomiale $\mathscr{M}_K(1, \boldsymbol{\pi})$ correspondant à un tirage dans K catégories avec probabilités $\pi_1, \ldots, \pi_K$:

$$\boldsymbol{Z}_i \sim \mathscr{M}_K(1, \boldsymbol{\pi}).$$

Ainsi, le vecteur aléatoire $\boldsymbol{Z}_i$ indique la composante de laquelle provient $\boldsymbol{\theta}_i$:

$$z_{ik} = \begin{cases} 1 & \text{si } \boldsymbol{\theta}_i \text{ provient de la } k^{\text{ ème}} \text{ composante,} \\ 0 & \text{sinon.} \end{cases}$$

Soit $\boldsymbol{\Phi}$ le vecteur des paramètres à estimer :

$$\boldsymbol{\Phi} = (\pi_1, \ldots, \pi_K, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K, \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_K).$$

La log-vraisemblance des données complètes, c'est-à-dire des données observées $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ complétées par les données non observées indicatrices des composantes $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$, s'écrit :

$$\begin{aligned} l_c(\boldsymbol{\Phi}; \boldsymbol{y}, \boldsymbol{\theta}, \boldsymbol{z}) &= \log f_{\boldsymbol{Y},\boldsymbol{\theta},\boldsymbol{Z}}(\boldsymbol{Y}, \boldsymbol{\theta}, \boldsymbol{Z}; \boldsymbol{\Phi}) \\ &= \log f_{\boldsymbol{Y}|\boldsymbol{\theta},\boldsymbol{Z}}(\boldsymbol{Y}|\boldsymbol{\theta}, \boldsymbol{Z}; \boldsymbol{\Phi}) + \log f_{\boldsymbol{\theta},\boldsymbol{Z}}(\boldsymbol{\theta}, \boldsymbol{Z}; \boldsymbol{\Phi}). \end{aligned}$$

Or $\boldsymbol{Y}$ est indépendant de $\boldsymbol{Z}$ conditionnellement à $\boldsymbol{\theta}$ donc :

$$f_{\boldsymbol{Y}|\boldsymbol{\theta},\boldsymbol{Z}}(\boldsymbol{Y}|\boldsymbol{\theta}, \boldsymbol{Z}; \boldsymbol{\Phi}) = f_{\boldsymbol{Y}|\boldsymbol{\theta}}(\boldsymbol{Y}|\boldsymbol{\theta}).$$

La log-vraisemblance de $\boldsymbol{Y}$ s'écrit par conséquent :

$$l_c(\boldsymbol{\Phi}; \boldsymbol{y}, \boldsymbol{\theta}, \boldsymbol{z}) = \log f_{\boldsymbol{Y}|\boldsymbol{\theta}}(\boldsymbol{Y}|\boldsymbol{\theta}) + \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \log(\pi_k f_k(\boldsymbol{\theta}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)). \tag{8.1}$$

8.4.1 Etape E

L'étape E de l'algorithme EM consiste à calculer l'espérance conditionnelle de la log-vraisemblance des données complètes sachant les données observées et les valeurs des estimateurs des paramètres à l'itération t, la fonction $Q(\boldsymbol{\Phi}, \boldsymbol{\Phi}^{(t)})$:

$$Q(\boldsymbol{\Phi}, \boldsymbol{\Phi}^{(t)}) = \mathbb{E}[l_c(\boldsymbol{\Phi}; \boldsymbol{Y}, \boldsymbol{\theta}, \boldsymbol{Z})|\boldsymbol{Y} = \boldsymbol{y}, \boldsymbol{\Phi}^{(t)}] = \mathbb{E}_{\boldsymbol{\Phi}^{(t)}}[l_c(\boldsymbol{\Phi}; \boldsymbol{Y}, \boldsymbol{\theta}, \boldsymbol{Z})|\boldsymbol{Y} = \boldsymbol{y}]. \tag{8.2}$$

Ici elle s'écrit :

$$Q(\boldsymbol{\Phi}, \boldsymbol{\Phi}^{(t)}) = \mathbb{E}_{\boldsymbol{\Phi}^{(t)}}\Big[\sum_{i=1}^{n}\sum_{k=1}^{K} p_{ik}^{\theta} \log(\pi_k f_k(\boldsymbol{\theta}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k))\Big|\boldsymbol{Y} = \boldsymbol{y}\Big] + c(\boldsymbol{y}), \tag{8.3}$$

où $c(\boldsymbol{y}) = \log f_{\boldsymbol{Y}|\boldsymbol{\theta}}(\boldsymbol{Y}|\boldsymbol{\theta})$ ne dépend pas de $\boldsymbol{\Phi}$, donc n'intervient pas dans la maximisation de $Q(\boldsymbol{\Phi}, \boldsymbol{\Phi}^{(t)})$ et p_{ik}^{θ} est la probabilité a posteriori que $\boldsymbol{\theta}_i$ provienne de la composante k :

$$p_{ik}^{\theta} = \mathbb{P}_{\boldsymbol{\Phi}^{(t)}}(Z_{ik} = 1|\boldsymbol{\theta}_i) = \frac{\pi_k^{(t)} f_k(\boldsymbol{\theta}_i; \boldsymbol{\mu}_k^{(t)}, \boldsymbol{\Sigma}_k^{(t)})}{f_{\boldsymbol{\theta}}(\boldsymbol{\theta}_i; \boldsymbol{\Phi}^{(t)})} = \frac{\pi_k^{(t)} f_k(\boldsymbol{\theta}_i; \boldsymbol{\mu}_k^{(t)}, \boldsymbol{\Sigma}_k^{(t)})}{\sum_{l=1}^{K} \pi_l^{(t)} f_l(\boldsymbol{\theta}_i; \boldsymbol{\mu}_l^{(t)}, \boldsymbol{\Sigma}_l^{(t)})}.$$

Démonstration de (8.3) :

D'après (8.1) et (8.2) on a :

$$Q(\boldsymbol{\Phi}, \boldsymbol{\Phi}^{(t)}) = \mathbb{E}_{\boldsymbol{\Phi}^{(t)}}\left[\sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik} \log(\pi_k f_k(\boldsymbol{\theta}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k))\Big|\boldsymbol{Y} = \boldsymbol{y}\right] + c(\boldsymbol{y}),$$

où $c(\boldsymbol{y}) = \sum_{i=1}^{n} c(\boldsymbol{y}_i)$ ne dépend pas de $\boldsymbol{\Phi}$. Pour une observation i on a :

$$\begin{aligned}
Q_i(\boldsymbol{\Phi}, \boldsymbol{\Phi}^{(t)}) - c(\boldsymbol{y}_i) &= \mathbb{E}_{\boldsymbol{\Phi}^{(t)}}[z_{ik} \log(\pi_k f_k(\boldsymbol{\theta}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k))|\boldsymbol{Y}_i = \boldsymbol{y}_i] \\
&= \iint_{\boldsymbol{\theta}_i, z_{ik}} z_{ik} \log(\pi_k f_k(\boldsymbol{\theta}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)) f_{\boldsymbol{Z},\boldsymbol{\theta}|\boldsymbol{Y}}(z_{ik}, \boldsymbol{\theta}_i|\boldsymbol{Y}_i = \boldsymbol{y}_i) d\boldsymbol{\theta}_i dz_{ik} \\
&= \iint_{\boldsymbol{\theta}_i, z_{ik}} z_{ik} \log(\pi_k f_k(\boldsymbol{\theta}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)) f_{\boldsymbol{Z}|\boldsymbol{\theta},\boldsymbol{Y}}(z_{ik}|\boldsymbol{\theta}_i, \boldsymbol{y}_i) f_{\boldsymbol{\theta}|\boldsymbol{Y}}(\boldsymbol{\theta}_i|\boldsymbol{y}_i) d\boldsymbol{\theta}_i dz_{ik} \\
&= \int_{\boldsymbol{\theta}_i} \log(\pi_k f_k(\boldsymbol{\theta}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)) f_{\boldsymbol{\theta}|\boldsymbol{Y}}(\boldsymbol{\theta}_i|\boldsymbol{y}_i) \int_{z_{ik}} z_{ik} f_{\boldsymbol{Z}|\boldsymbol{\theta},\boldsymbol{Y}}(z_{ik}|\boldsymbol{\theta}_i, \boldsymbol{y}_i) dz_{ik} d\boldsymbol{\theta}_i \\
&= \int_{\boldsymbol{\theta}_i} \log(\pi_k f_k(\boldsymbol{\theta}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)) f_{\boldsymbol{\theta}|\boldsymbol{Y}}(\boldsymbol{\theta}_i|\boldsymbol{y}_i) \mathbb{P}(Z_{ik} = 1|\boldsymbol{\theta}_i, \boldsymbol{y}_i) d\boldsymbol{\theta}_i \\
&= \mathbb{E}_{\boldsymbol{\Phi}^{(t)}}[\mathbb{P}_{\boldsymbol{\Phi}^{(t)}}(Z_{ik} = 1|\boldsymbol{\theta}_i) \log(\pi_k f_k(\boldsymbol{\theta}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k))|\boldsymbol{Y}_i = \boldsymbol{y}_i] \\
&= \mathbb{E}_{\boldsymbol{\Phi}^{(t)}}[p_{ik}^{\theta} \log(\pi_k f_k(\boldsymbol{\theta}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k))|\boldsymbol{Y}_i = \boldsymbol{y}_i].
\end{aligned}$$

Utilisation d'un processus de Monte-Carlo dans l'étape E (MCEM)

Nous sommes confrontés au même problème que dans le chapitre 7, où on doit calculer une intégrale de dimension d à chaque étape de l'EM pour calculer la fonction Q définie en (8.3). Nous utilisons la même démarche. Une approximation de Monte Carlo de la fonction Q est réalisée par un algorithme d'échantillonnage d'importance (*importance sampling*) en simulant à chaque itération t de l'EM et pour chaque observation $\boldsymbol{y}_i$ un échantillon aléatoire $\boldsymbol{\theta}_{i1}^{(t)}, \ldots, \boldsymbol{\theta}_{iN}^{(t)}$ à partir d'une "densité d'importance" h :

$$Q_i(\boldsymbol{\Phi}, \boldsymbol{\Phi}^{(t)}) \simeq Q_i^N(\boldsymbol{\Phi}, \boldsymbol{\Phi}^{(t)}) = \frac{1}{N} \sum_{r=1}^{N} w_{ir} \left[\sum_{k=1}^{K} p_{ik}^{\theta r} \log(\pi_k f_k(\boldsymbol{\theta}_{ir}^{(t)} | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)) \right] + c(\boldsymbol{y}_i),$$

avec $p_{ik}^{\theta r} = \mathbb{P}_{\boldsymbol{\Phi}^{(t)}}(Z_{ik} = 1 | \boldsymbol{\theta}_{ir})$. Les poids w_{ir} sont définis pour $r = 1, \ldots, N$ par :

$$w_{ir} = \frac{f_{\boldsymbol{\theta}|\boldsymbol{Y}}(\boldsymbol{\theta}_{ir}^{(t)} | \boldsymbol{y}; \boldsymbol{\Phi}^{(t)})}{h(\boldsymbol{\theta}_{ir}^{(t)})} \propto \frac{f_{\boldsymbol{Y}|\boldsymbol{\theta}}(\boldsymbol{y}_i | \boldsymbol{\theta}_{ir}^{(t)}) f_{\boldsymbol{\theta}}(\boldsymbol{\theta}_{ir}^{(t)}; \boldsymbol{\Phi}^{(t)})}{h(\boldsymbol{\theta}_{ir}^{(t)})}$$

et calculés à la constante multiplicative $1/f_{\boldsymbol{Y}}(\boldsymbol{y}_i; \boldsymbol{\Phi}^{(t)})$ près (qui ne dépend pas de $\boldsymbol{\Phi}$). Dans le chapitre 7, la densité d'importance utilisée pour échantillonner les $\boldsymbol{\theta}$ était une loi de Student multivariée, dont les paramètres de moyenne et matrice de covariance étaient réestimées à chaque itération pour correspondre à :

$$\begin{aligned} m_t &= \mathbb{E}_{\boldsymbol{\Phi}^{(t)}}[\boldsymbol{\theta} | \boldsymbol{y}], \\ \Sigma_t &= \mathbb{V}_{\boldsymbol{\Phi}^{(t)}}[\boldsymbol{\theta} | \boldsymbol{y}]. \end{aligned}$$

Pour s'adapter au cas où $\boldsymbol{\theta}$ provient d'un mélange de lois gaussiennes, nous proposons d'utiliser un mélange de lois de Student multivariées, en réévaluant les paramètres à chaque itération de l'EM de manière à avoir :

$$\begin{aligned} m_{tk} &= \mathbb{E}_{\boldsymbol{\Phi}^{(t)}}[\boldsymbol{\theta}_k | \boldsymbol{y}], \\ \Sigma_{tk} &= \mathbb{V}_{\boldsymbol{\Phi}^{(t)}}[\boldsymbol{\theta}_k | \boldsymbol{y}]. \end{aligned}$$

Les estimateurs PQL (quasi-vraisemblance pénalisée) de ces moments conditionnels sont obtenus par un algorithme itératif d'optimisation présenté en 7.7.2.

De même que dans le chapitre 7, la taille de l'échantillon d'importance augmente au cours des itérations par la procédure automatique proposée par Booth & Hobert (1998), afin d'optimiser les performances de l'algorithme MCEM.

8.4.2 Etape M

L'étape M de l'algorithme EM consiste à maximiser la fonction $Q(\boldsymbol{\Phi}, \boldsymbol{\Phi}^{(t)})$, que nous avons approchée par $Q_N(\boldsymbol{\Phi}, \boldsymbol{\Phi}^{(t)}) = \sum_{i=1}^{n} Q_i^N(\boldsymbol{\Phi}, \boldsymbol{\Phi}^{(t)})$, par rapport à $\boldsymbol{\Phi}$. Dans notre cas elle admet une solution explicite.

Calcul des estimateurs du maximum de vraisemblance (MV) :

L'estimateur du MV de π_k est obtenu en dérivant Q_N par rapport à π_k et en résolvant l'équation $\partial Q_N/\partial \pi_k = 0$. La résolution de cette équation nous donne :

$$\hat{\pi}_k = \frac{\sum_{i=1}^{n}\sum_{r=1}^{N} w_{ir} p_{ik}^{\theta r}}{nN}. \tag{8.4}$$

L'estimateur du MV de $\boldsymbol{\mu}_k$ est obtenu en dérivant Q_N par rapport à $\boldsymbol{\mu}_k$ et en résolvant l'équation $\partial Q_N/\partial \boldsymbol{\mu}_k = 0$. La résolution de cette équation nous donne :

$$\hat{\boldsymbol{\mu}}_k = \frac{\sum_{i=1}^{n}\sum_{r=1}^{N} w_{ir} p_{ik}^{\theta r} \boldsymbol{\theta}_{ir}^{(t)}}{\sum_{i=1}^{n}\sum_{r=1}^{N} w_{ir} p_{ik}^{\theta r}}. \tag{8.5}$$

L'estimateur du MV de $\boldsymbol{\Sigma}_k$ est obtenu en dérivant Q_N par rapport à $\boldsymbol{\Sigma}_k$ et en résolvant l'équation $\partial Q_N/\partial \boldsymbol{\Sigma}_k = 0$. Ceci revient à résoudre $\partial Q_N/\partial \boldsymbol{\Gamma}_k = 0$ où $\boldsymbol{\Gamma} = \boldsymbol{\Sigma}^{-1}$. On obtient :

$$\hat{\boldsymbol{\Sigma}}_k = \frac{\sum_{i=1}^{n}\sum_{r=1}^{N} w_{ir} p_{ik}^{\theta r} (\boldsymbol{\theta}_{ir}^{(t)} - \hat{\boldsymbol{\mu}}_k)(\boldsymbol{\theta}_{ir}^{(t)} - \hat{\boldsymbol{\mu}}_k)^T}{\sum_{i=1}^{n}\sum_{r=1}^{N} w_{ir} p_{ik}^{\theta r}}. \tag{8.6}$$

Démonstration de (8.6) :

Nous reprenons ici la démonstration vue en 7.7.3 pour l'adapter au cas avec mélange.

$$\begin{aligned}
Q_N(\boldsymbol{\Phi}, \boldsymbol{\Phi}^{(t)}) - c(\boldsymbol{y}) &= \sum_{i=1}^{n}\sum_{k=1}^{K} \mathbb{E}_{\boldsymbol{\Phi}^{(t)}}[p_{ik}^{\theta} \log(\pi_k f_{\boldsymbol{\theta}}(\boldsymbol{\theta}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)) | \boldsymbol{Y}_i = \boldsymbol{y}_i] \\
&= \frac{1}{N}\sum_{i=1}^{n}\sum_{r=1}^{N}\sum_{k=1}^{K} w_{ir} p_{ik}^{\theta r} \Big(\log(\pi_k f_{\boldsymbol{\theta}}(\boldsymbol{\theta}_{ir}^{(t)} | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)) \Big) \\
&= c + \frac{1}{N}\sum_{i=1}^{n}\sum_{r=1}^{N}\sum_{k=1}^{K} w_{ir} p_{ik}^{\theta r} \Big(\log \pi_k - \frac{1}{2}\log|\boldsymbol{\Sigma}_k| - \frac{1}{2}(\boldsymbol{\theta}_{ir}^{(t)} - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{\theta}_{ir}^{(t)} - \boldsymbol{\mu}_k) \Big) \\
&= c + g(\boldsymbol{\Sigma}_k) + \frac{1}{N}\sum_{i=1}^{n}\sum_{r=1}^{N}\sum_{k=1}^{K} w_{ir} p_{ik}^{\theta r} \log(\pi_k)
\end{aligned}$$

où $c = -\frac{1}{2} n \log(2\pi)$ est une constante et $g(\boldsymbol{\Sigma}_k)$ est définie par :

$$g(\boldsymbol{\Sigma}_k) = \frac{1}{2N}\sum_{i=1}^{n}\sum_{r=1}^{N}\sum_{k=1}^{K} w_{ir} p_{ik}^{\theta r} \Big(-\log|\boldsymbol{\Sigma}_k| - (\boldsymbol{\theta}_{ir}^{(t)} - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{\theta}_{ir}^{(t)} - \boldsymbol{\mu}_k) \Big). \tag{8.7}$$

On pose $\boldsymbol{\Gamma}_k = \boldsymbol{\Sigma}_k^{-1}$. Alors $|\boldsymbol{\Sigma}_k| = |\boldsymbol{\Gamma}_k|^{-1}$ et $g(\boldsymbol{\Sigma}_k)$ est remplacée par :

$$g'(\boldsymbol{\Gamma}_k) = \frac{1}{2N}\sum_{i=1}^{n}\sum_{r=1}^{N}\sum_{k=1}^{K} w_{ir}p_{ik}^{\theta r}\Big(\log|\boldsymbol{\Gamma}_k| - (\boldsymbol{\theta}_{ir}^{(t)} - \boldsymbol{\mu}_k)^T\boldsymbol{\Gamma}_k(\boldsymbol{\theta}_{ir}^{(t)} - \boldsymbol{\mu}_k)\Big) \tag{8.8}$$

Résoudre $\partial g(\boldsymbol{\Sigma}_k)/\partial\boldsymbol{\Sigma}_k = 0$ revient à résoudre $\partial g'(\boldsymbol{\Gamma}_k)/\partial\boldsymbol{\Gamma}_k = 0$. On note $\boldsymbol{\Gamma}_k = (\gamma_k^{ll'})$ avec $l, l' \in \{1, 2\}$, $\boldsymbol{\mu}_k = (\mu_k^1, \mu_k^2)$ et $\boldsymbol{\theta}_{ir}^{(t)} = (\theta_1, \theta_2)$. Alors :

$$\begin{aligned}
\frac{\partial g'(\boldsymbol{\Gamma}_k)}{\partial\boldsymbol{\gamma}_k^{11}} &= \frac{1}{2N}\sum_{i=1}^{n}\sum_{r=1}^{N}\sum_{k=1}^{K} w_{ir}p_{ik}^{\theta r}\Big(\frac{\gamma_k^{22}}{|\boldsymbol{\Gamma}_k|} - (\theta_1 - \mu_k^1)^2\Big)\\
\frac{\partial g'(\boldsymbol{\Gamma}_k)}{\partial\boldsymbol{\gamma}_k^{22}} &= \frac{1}{2N}\sum_{i=1}^{n}\sum_{r=1}^{N}\sum_{k=1}^{K} w_{ir}p_{ik}^{\theta r}\Big(\frac{\gamma_k^{11}}{|\boldsymbol{\Gamma}_k|} - (\theta_2 - \mu_k^2)^2\Big)\\
\frac{\partial g'(\boldsymbol{\Gamma}_k)}{\partial\boldsymbol{\gamma}_k^{12}} &= \frac{1}{2N}\sum_{i=1}^{n}\sum_{r=1}^{N}\sum_{k=1}^{K} w_{ir}p_{ik}^{\theta r}\Big(-\frac{\gamma_k^{12}}{|\boldsymbol{\Gamma}_k|} - 2(\theta_1 - \mu_k^1)(\theta_2 - \mu_k^2)\Big)
\end{aligned}$$

et $\gamma_k^{12} = \gamma_k^{21}$. $\partial g'(\boldsymbol{\Gamma}_k)/\partial\boldsymbol{\Gamma}_k = 0$ équivaut à :

$$\begin{aligned}
\frac{\gamma_k^{11}}{|\boldsymbol{\Gamma}_k|} &= \frac{\sum_{i=1}^{n}\sum_{r=1}^{N}\sum_{k=1}^{K} w_{ir}p_{ik}^{\theta r}(\theta_2 - \mu_k^2)^2}{\sum_{i=1}^{n}\sum_{r=1}^{N}\sum_{k=1}^{K} w_{ir}p_{ik}^{\theta r}}\\
\frac{\gamma_k^{22}}{|\boldsymbol{\Gamma}_k|} &= \frac{\sum_{i=1}^{n}\sum_{r=1}^{N}\sum_{k=1}^{K} w_{ir}p_{ik}^{\theta r}(\theta_1 - \mu_k^1)^2}{\sum_{i=1}^{n}\sum_{r=1}^{N}\sum_{k=1}^{K} w_{ir}p_{ik}^{\theta r}}\\
\frac{\gamma_k^{12}}{|\boldsymbol{\Gamma}_k|} &= \frac{-2\sum_{i=1}^{n}\sum_{r=1}^{N}\sum_{k=1}^{K} w_{ir}p_{ik}^{\theta r}(\theta_1 - \mu_k^1)(\theta_2 - \mu_k^2)}{\sum_{i=1}^{n}\sum_{r=1}^{N}\sum_{k=1}^{K} w_{ir}p_{ik}^{\theta r}}.
\end{aligned}$$

Or :

$$\boldsymbol{\Sigma}_k = \boldsymbol{\Gamma}_k^{-1} = \frac{1}{|\boldsymbol{\Gamma}_k|}\begin{pmatrix}\gamma_k^{22} & -\gamma_k^{12}\\ -\gamma_k^{12} & \gamma_k^{11}\end{pmatrix}.$$

Donc :

$$\begin{aligned}
\sigma_k^{11} &= \frac{\gamma_k^{22}}{|\boldsymbol{\Gamma}_k|}\\
\sigma_k^{22} &= \frac{\gamma_k^{11}}{|\boldsymbol{\Gamma}_k|}\\
\sigma_k^{12} &= -\frac{\gamma_k^{12}}{|\boldsymbol{\Gamma}_k|}.
\end{aligned}$$

Il suit que :

$$\boldsymbol{\Sigma}_k = \frac{\sum_{i=1}^{n}\sum_{r=1}^{N}\sum_{k=1}^{K} w_{ir}p_{ik}^{\theta r}(\boldsymbol{\theta}_{ir}^{(t)} - \boldsymbol{\mu}_k)(\boldsymbol{\theta}_{ir}^{(t)} - \boldsymbol{\mu}_k)^T}{\sum_{i=1}^{n}\sum_{r=1}^{N}\sum_{k=1}^{K} w_{ir}p_{ik}^{\theta r}}.$$

8.4.3 Règle d'arrêt

Nous utilisons la règle d'arrêt suivante :

$$\max_i \frac{|\Phi_l^{(t+1)} - \Phi_l^{(t)}|}{|\Phi_l^{(t)}| + \delta_1} < \delta_2 \tag{8.9}$$

où $\delta_1 = 0.001$, $\delta_2 = 0.005$, $l \in \{1, \ldots, \alpha\}$ est l'indice des paramètres, avec α le nombre de paramètres du modèle $\alpha = K(d + d(d+1)/2) + K - 1$, donc Φ_l désigne les éléments du vecteur de paramètres $\boldsymbol{\Phi}$. L'algorithme est arrêté lorsque cette règle est atteinte pour 3 itérations consécutives.

8.4.4 Initialisation

Comme nous l'avons vu en 2.2.1, l'initialisation est un problème fondamental dans la maximisation du critère de vraisemblance par les algorithmes de type EM, car ces derniers conduisent à la construction d'une suite de solutions faisant croître la vraisemblance vers un maximum local qui dépend de la position initiale. Or, pour les modèles de mélange, la logvraisemblance est très plate avec de nombreux maxima locaux. Ceci est d'autant plus vrai dans le cas de données multivariées ou si le nombre de classes à estimer est grand. Par conséquent l'EM doit être initialisé près de la solution pour fournir de bons résultats.

Dans le cas du modèle de mélange gaussien latent que nous considérons ici, nous avons une difficulté supplémentaire. En effet, même en ayant une partition initiale des données en K classes, l'initialisation des paramètres n'est pas immédiate à partir des moments empiriques des classes. Dans le chapitre 7, nous proposons d'utiliser la méthode des moments pour estimer les paramètres des lois gaussiennes sousjacentes. Pour certains sous-modèles, comme le MPLN, ces calculs sont explicites, pour d'autres, tel le modèle Binomial-Poisson, nous proposons d'utiliser un algorithme itératif de Newton-Raphson (voir 7.7.1).

Dans cette thèse, notre stratégie d'initialisation consiste à effectuer une préclassification des données, combinée à la méthode des moments pour estimer les paramètres des lois gaussiennes sousjacentes, qui sont nécessaires à l'initialisation de notre algorithme. Nous avons choisi d'utiliser l'algorithme `MCLUST` (Fraley & Raftery, 2006) basé sur des mélanges de gaussiennes multivariées. Nous verrons que, dans certains cas, celui-ci peut suffire pour une bonne initialisation (parfois moyennant une transformation des données), mais dans les cas qui sont éloignés de l'hypothèse gaussienne (les lois binomiales à faible effectif en font partie), d'autres stratégies d'initialisation doivent être employées.

8.4.5 Algorithme de classification automatique

Nous utilisons une démarche classique de classification à l'aide de modèles de mélanges. Notre algorithme procède de la manière suivante :

Un nombre maximal $K_{\max}$ de classes est sélectionné. Pour $K \in \{1, \ldots, K_{\max}\}$ fixé :

1. Initialisation :
   - l'algorithme `MCLUST` est utilisé pour obtenir une classification des données en K classes, (parfois après une transformation de données, par exemple logarithmique sur les données MPLN),
   - pour chaque classe nous calculons les paramètres moyenne et matrice de covariance de la loi gaussienne sous-jacente en utilisant la méthode des moments (soit directement pour des modèles tels que le MPLN, en inversant les équations présentées dans

7.2.3, soit en utilisant un algorithme d'optimisation itératif de Newton-Raphson, de la manière décrite dans l'annexe 7.7.1 du chapitre 7, qui explicite le calcul des moments pour le modèle Binomial-Poisson).

2. Utilisation de l'algorithme MCEM pour obtenir les estimateurs du maximum de vraisemblance du vecteur des paramètres $\boldsymbol{\Phi}$ et les probabilités a posteriori d'appartenance aux classes $\hat{p}_{ik}$. Affectation des données à la classe la plus probable par la règle du MAP.
3. Calcul de deux critères de sélection de modèles : les critères *Bayesian Information Criterion* (BIC) et *Integrated Classification Likelihood* (ICL) ont été calculés.

Le nombre de classes qui maximise le critère de sélection de modèle (BIC ou ICL) est choisi. Selon une étude comparant plusieurs critères de sélection du nombre de classes menée par McLachlan & Ng (2000), le critère ICL semble être le plus performant pour sélectionner le nombre de classes (McLachlan & Peel, 2000, p. 220).

Par la suite, dans l'étude des sous-modèles par simulation, nous nous limitons au cas bivarié ($d = 2$). En effet, l'estimation devient plus difficile en plus grande dimension et il faudra recourir à des modèles de covariance contraints pour la loi gaussienne sous-jacente. A cette fin, une décomposition en valeurs propres de la matrice de covariance, telle que celle proposée par Banfield & Raftery (1993) et Celeux & Govaert (1995), pourra être utilisée. Celle-ci permettra d'implémenter les modèles de covariance contraints présentés en 4.6.3, qui sont disponibles dans le logiciel `MCLUST` dans R (2009).

8.5 Etude de trois sous-modèles par simulation

Dans cette sous-partie nous présentons les performances de l'algorithme d'estimation sur des données simulées selon les sous-modèles mélange Normal-Poisson, mélange Normal-Binomial et mélange de BPLN (c'est-à-dire MPLN bivarié). Les deux premiers sous-modèles sont plus faciles à estimer car ils ne contiennent en réalité qu'une seule variable latente. En effet, la distribution gaussienne marginale correspondant à la variable continue est observée (à une transformation près), donc, si les appartenances aux classes sont connues, les moments marginaux qui correspondent à la variable continue sont estimés directement par les moments empiriques de l'échantillon (voir 7.2.3).

Nous étudions le comportement de notre algorithme sur les points suivants :

– l'identification des classes lorsqu'elles sont plus ou moins bien séparées (en termes d'erreur de classification),
– la sélection du nombre de classes (quel critère est le plus judicieux ?),
– l'identification des classes lorsque les composantes sont fortement non gaussiennes.

L'évaluation de la performance de notre algorithme en terme d'identification des classes se fait grâce à trois critères :

– taux de mal-classés ϵ_1 : taux d'erreur d'affectation des observations aux classes,
– taux de mal-classés en probabilité ϵ_2 : cumul des probabilités de chaque observation d'être mal classée divisé par le nombre d'observations :

$$\epsilon_2 = \sum_{i=1}^{n} (1 - \hat{p}_{ik}^{\text{vrai}})/n,$$

– incertitude de la classification u : cumul pour chaque observation des probabilités d'être classée dans une autre classe que celle à laquelle elle a été affectée, divisé par le nombre d'observations :

$$u = \sum_{i=1}^{n} \Big(1 - \max_{k \in \{1,\dots,K\}} \hat{p}_{ik}\Big)/n.$$

Nous comparons ces critères entre l'initialisation obtenue avec l'algorithme `MCLUST` et l'estimation finale obtenue par l'algorithme MCEM, afin d'évaluer les améliorations apportées par notre algorithme d'estimation basé sur le vrai modèle.

Les deux premiers critères ϵ_1 et ϵ_2 peuvent être calculés seulement lorsque la vraie classification est connue, le troisième u peut être calculé dans tous les cas.

Dans tous les exemples suivants, pour chaque jeu de paramètres, l'estimation se fait sur $n_s = 100$ jeux de données simulés de taille $n = 400$.

8.5.1 Estimation à nombre de classes fixé pour un modèle de mélange Normal-Poisson à deux classes

Nous commençons par l'étude d'un cas simple du sous-modèle de mélange Normal-Poisson à deux composantes. Nous comparons la performance de notre algorithme sur un cas où les classes sont bien séparées (figure 8.2), et sur un cas où les classes sont proches (figure 8.3). Le nombre de classes est fixé au préalable à $K = 2$ (pour l'initialisation avec `MCLUST` et pour notre algorithme).

La figure 8.2 montre que, lorsque les deux classes sont bien séparées, notre modèle n'apporte presque pas d'amélioration par rapport à la procédure d'initialisation `MCLUST` fixée à deux classes. Ceci n'est pas surprenant dans ce cas, puisque la forme des classes n'est pas incompatible avec une distribution gaussienne bivairée. De plus, les effectifs obtenus par la loi de Poisson sont élevés, ce qui permet une bonne approximation par une loi normale.

Lorsque nous rapprochons les classes et que la séparation des classes est moins nette (figure 8.3), notre algorithme améliore la classification donnée par `MCLUST`, comme le montrent les boxplot de la figure 8.4. Un test unilatéral de Student sur échantillons appariés réalisé sur les 100 simulations (cf tableau 8.1), confirme que les erreurs de classification de notre algorithme sont significativement plus petites que les erreurs de classification de `MCLUST` (pour un risque α fixé à 0.05). Ceci est également vrai pour les erreurs ϵ_1 et ϵ_2 de l'exemple 8.2, mais la différence moyenne entre erreurs est négligeable dans un contexte de classification.

Les figures 8.2 (d) et 8.3 (d) représentent la classification choisie par l'algorithme `MCLUST` non contraint en termes de nombre de classes. Une tendance à surestimer le nombre de classes est constatée, qui s'explique par le fait que le BIC est calculé sur la vraisemblance du mélange gaussien et non sur la vraisemblance de la classification. Dans l'exemple suivant, nous explorons les choix du nombre de classes de notre algorithme pour les deux critères de sélection de modèle BIC et ICL.

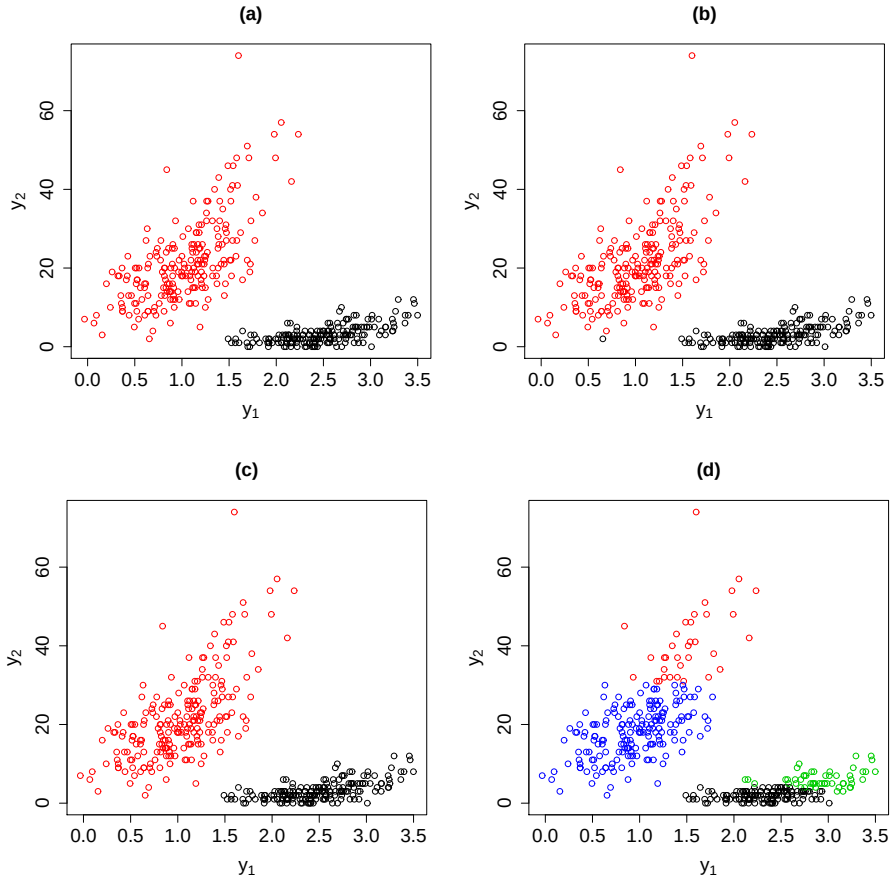


Fig. 8.2. Normal-Poisson : exemple de simulation à 2 composantes, classes bien séparées. (a) Données simulées avec les paramètres : composante 1 (en noir) : $\pi_1 = 0.5$, $\boldsymbol{\mu}_1 = (2.5, 1)^T$, $\boldsymbol{\sigma}_1^2 = (0.2, 0.2)^T$, $r_1 = 0.9$, composante 2 (en rouge) : $\pi_2 = 0.5$, $\boldsymbol{\mu}_2 = (1, 3)^T$, $\boldsymbol{\sigma}_2^2 = (0.2, 0.2)^T$, $r_2 = 0.8$ (b) Initialisation avec `MCLUST`, pour K fixé à 2 : $\epsilon_1 = 0.0025$, $\epsilon_2 = 0.0032$, $u = 0.00092$ (c) Notre classification, K fixé à 2 : $\epsilon_1 = 0$, $\epsilon_2 = 4.4\ 10^{-14}$, $u = 4.4\ 10^{-14}$ (d) Nombre de classes choisi par `MCLUST` pour cette simulation : K=4.

8.5.2 Estimation et choix du nombre de classes : modèle de mélange Normal-Poisson

Nous utilisons la procédure décrite dans la sous-partie 8.4 avec $K_{\max} = 6$, pour choisir le nombre de classes sur un exemple à 3 classes simulées selon un modèle de mélange Normal-Poisson.

Les résultats sur trois situations distinctes (classes plus ou moins bien séparées) sont présentés dans les figures 8.5, 8.6, 8.7. Le graphique 8.8 montre le nombre de classes choisi sur les 100 simulations pour chaque exemple et pour chaque critère de sélection. Le nombre de classes choisi par l'algorithme `MCLUST` non contraint (i.e. le nombre de classes est choisi automatiquement grâce à un critère BIC en comparant différents modèles) a également été calculé pour les 100 simulations. Nous pouvons constater que, pour notre algorithme, le BIC surestime le nombre de classes et a tendance à ne pas s'arrêter avant la limite fixée $K_{\max} = 6$. Le critère ICL choisit en fréquence toujours mieux le nombre de classes que le BIC. Enfin,

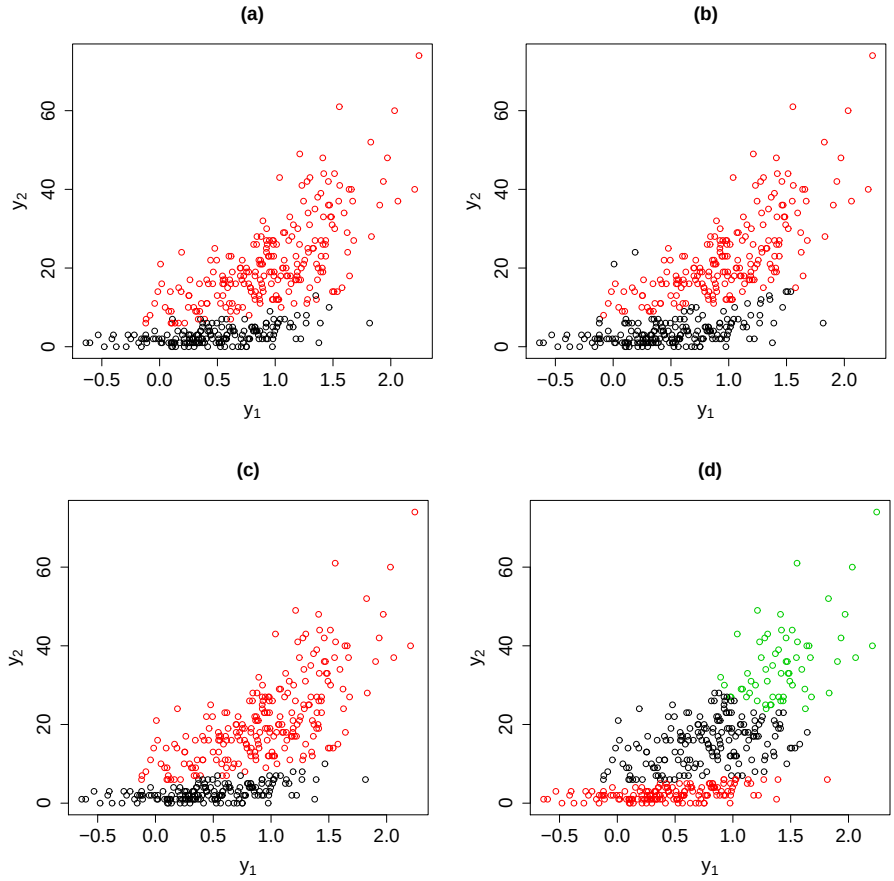


Fig. 8.3. Normal-Poisson : exemple de simulation à 2 composantes, classes rapprochées. (a) Données simulées avec les paramètres de la figure 8.2 à l'exception de $\boldsymbol{\mu}_1$ qui est remplacé par $\boldsymbol{\mu}_1 = (0.5, 1)^T$ (b) Initialisation avec `MCLUST`, pour K fixé à 2 : $\epsilon_1 = 0.08$, $\epsilon_2 = 0.12$, $u = 0.083$ (c) Notre classification, K fixé à 2 : $\epsilon_1 = 0.038$, $\epsilon_2 = 0.059$, $u = 0.038$ (d) Nombre de classes choisi par `MCLUST` : K=3.

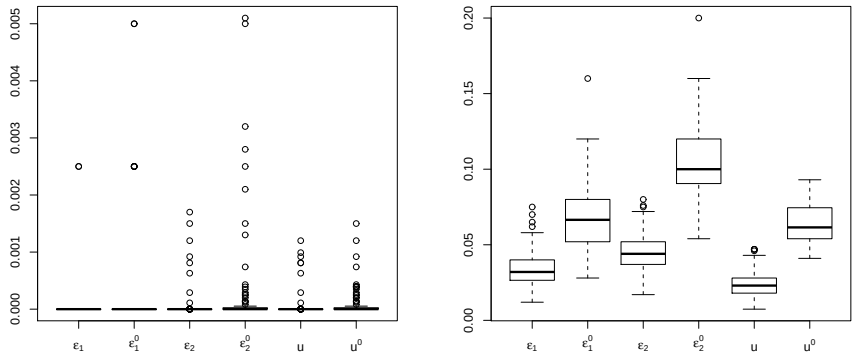


Fig. 8.4. Boxplot des erreurs ϵ_1, ϵ_2 et u sur 100 simulations des modèles Normal-Poisson à 2 composantes présentés en figure 8.2 (gauche) et 8.3 (droite). ϵ_1^0, ϵ_2^0 et u^0 sont calculés à l'initialisation, donc sur la classification effectuée par `MCLUST`.

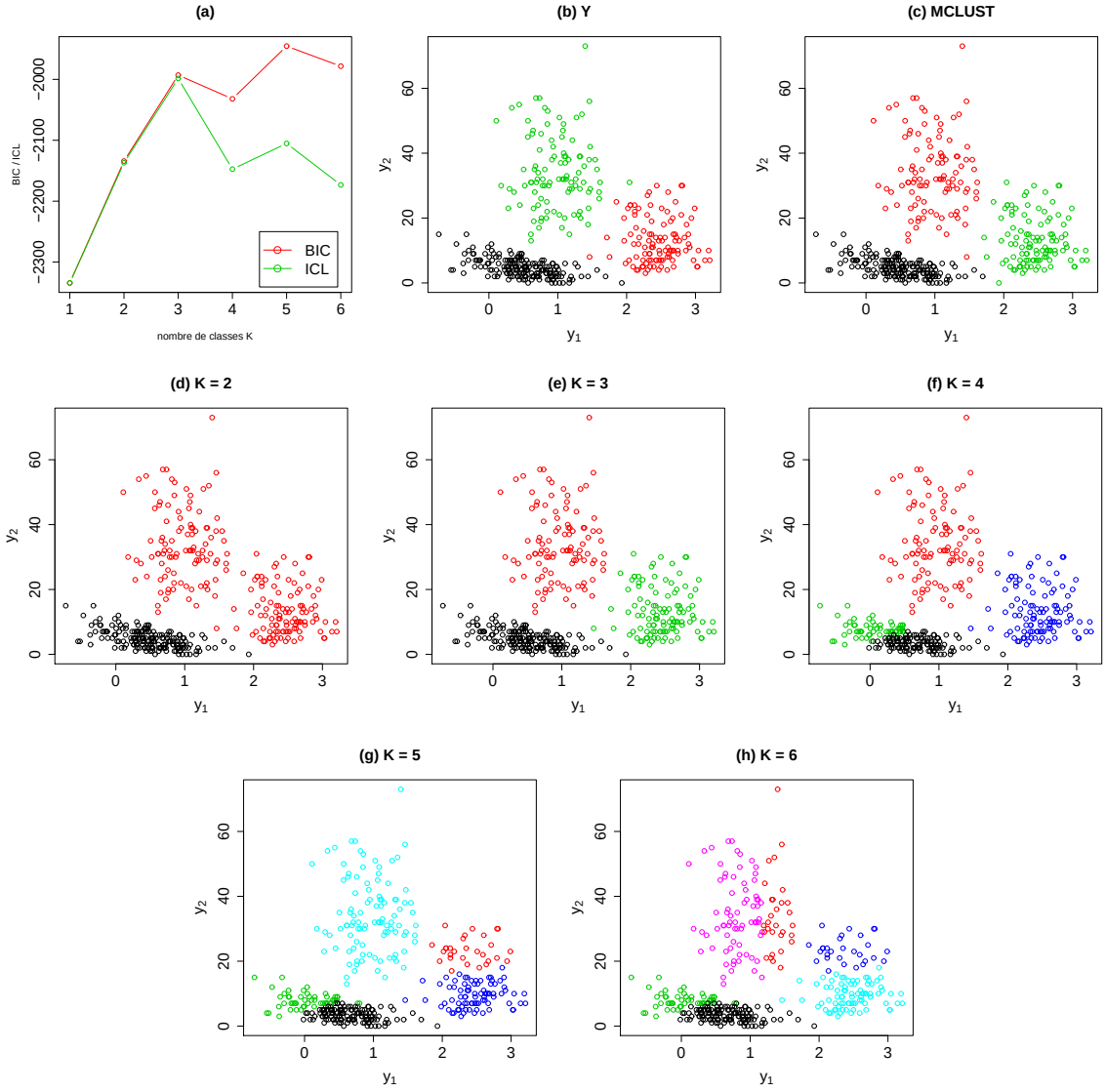


Fig. 8.5. Sélection du nombre de classes dans le cas où les classes sont bien séparées. (a) Choix BIC : $K = 5$ et ICL : $K = 3$ (b) Données simulées selon un modèle Normal-Poisson à 3 composantes de paramètres : composante 1 (en noir) : $\pi_1 = 0.4$, $\boldsymbol{\mu}_1 = (0.5, 1.5)^T$, $\boldsymbol{\sigma}_1^2 = (0.2, 0.2)^T$, $r_1 = -0.9$, composante 2 (en rouge) : $\pi_2 = 0.3$, $\boldsymbol{\mu}_2 = (2.5, 2.5)^T$, $\boldsymbol{\sigma}_2^2 = (0.1, 0.2)^T$, $r_2 = 0$, composante 3 (en vert) : $\pi_3 = 0.3$, $\boldsymbol{\mu}_3 = (1, 3.5)^T$, $\boldsymbol{\sigma}_3^2 = (0.1, 0.1)^T$, $r_3 = 0$ (c) MCLUST choisit $K = 3$ (d)–(h) Classification de notre algorithme pour $K = 2, \ldots, 6$

l'algorithme `MCLUST` non contraint choisit en fréquence mieux que notre algorithme dans les 3 cas (avec un critère BIC).

Cette étude montre que les critères BIC et ICL que nous avons considérés ici ne permettent pas de sélectionner le bon nombre de classes suffisamment souvent pour notre modèle. Cela pourrait être dû à un nombre de données insuffisant. En effet, ce modèle permet de gagner beaucoup de souplesse et de généralité par sa formulation hiérarchique, mais au détriment de son estimabilité. Ainsi, il est évident que l'estimation d'un modèle de mélange dans une couche latente demande une quantité plus importante d'information

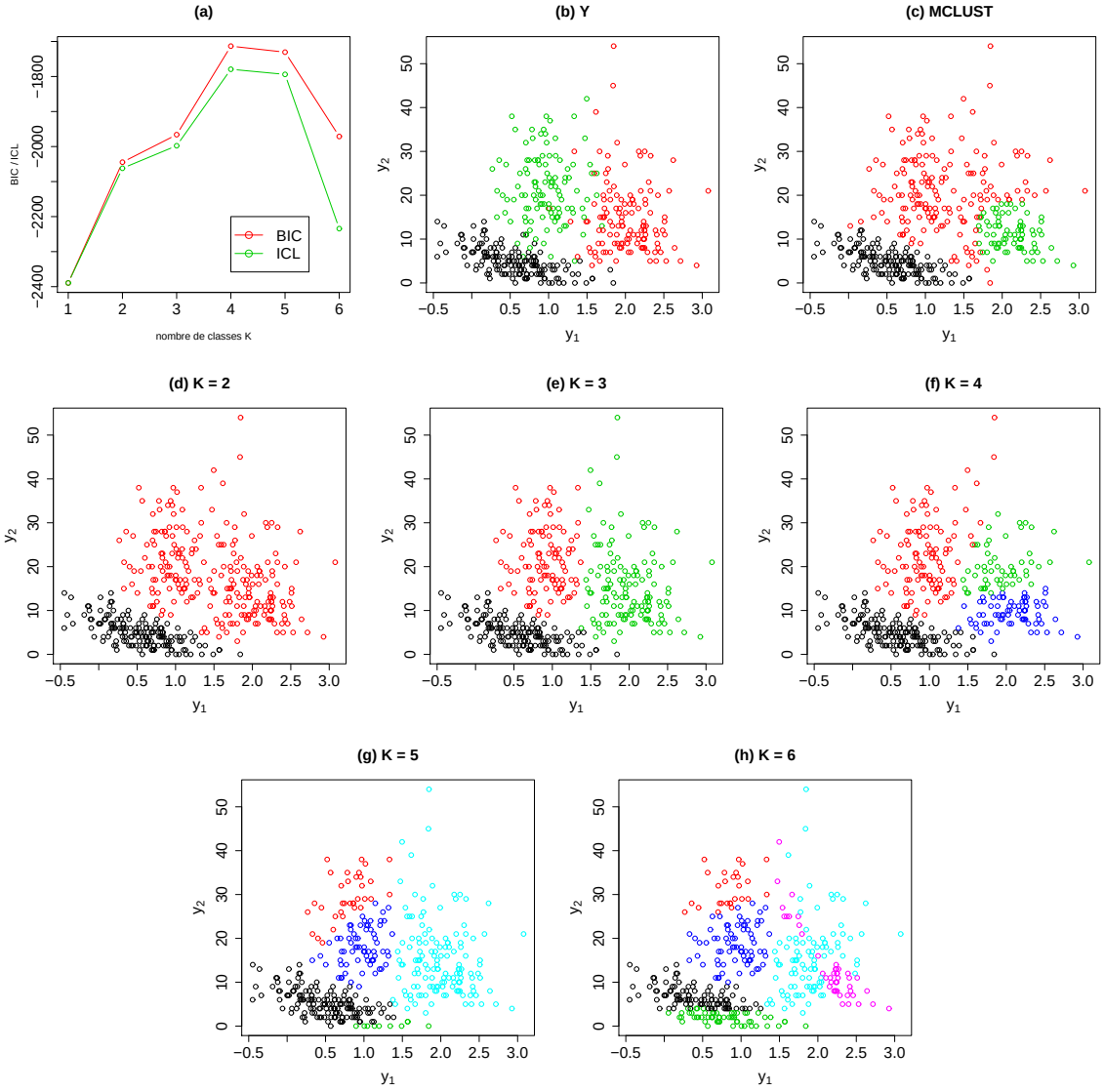


Fig. 8.6. Sélection du nombre de classes dans le cas où les classes sont proches. (a) Choix BIC : $K = 4$ et ICL : $K = 4$ (b) Données simulées selon un modèle Normal-Poisson à 3 composantes de mêmes paramètres que l'exemple figure 8.5 à l'exception de la composante 2 (en rouge) : $\boldsymbol{\mu}_2 = (2, 2.5)^T$, composante 3 (en vert) : $\boldsymbol{\mu}_3 = (1, 3)^T$ (c) MCLUST choisit $K = 3$ (d)–(h) Classification de notre algorithme pour $K = 2, \ldots, 6$

dans les données que l'estimation d'un modèle de mélange directement observé. Des recherches futures sur ce thème devraient explorer l'utilisation d'autres critères de sélection ou proposer des pénalisations supplémentaires des critères BIC et ICL explorés ici.

8.5.3 Estimation à nombre de classes fixé sur des composantes fortement non gaussiennes : modèle de mélange Normal-Binomial

Dans les exemples précédents, nous avons vu que les classifications réalisées avec `MCLUST` lorsqu'on fixe le nombre de classes sont très bonnes, même si les composantes ne sont pas gaussiennes. Ceci s'explique par le fait que la loi Poisson log-normale, lorsque les effectifs

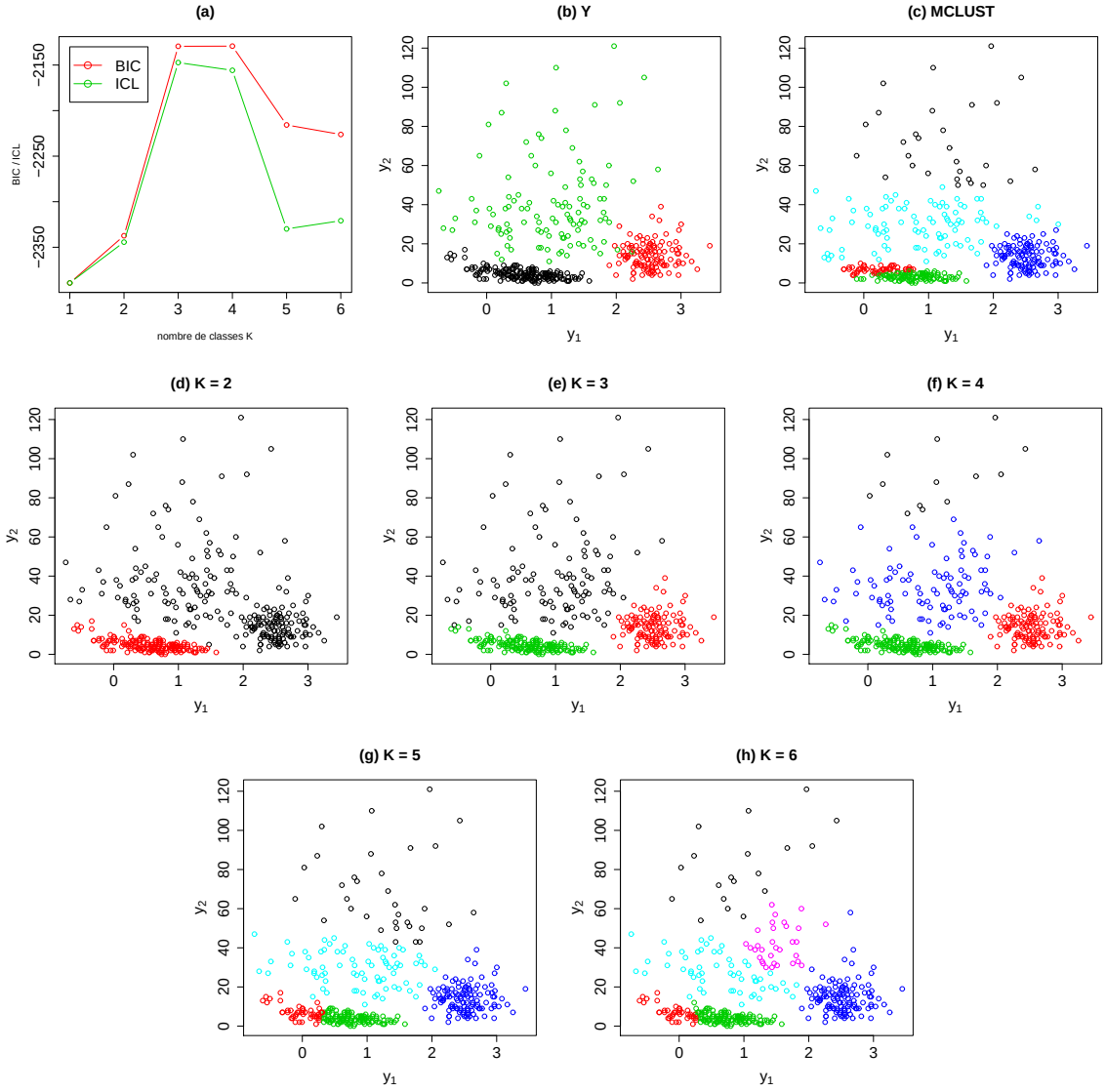


Fig. 8.7. Sélection du nombre de classes dans le cas où les classes sont proches. (a) Choix BIC : $K = 4$ et ICL : $K = 3$ (b) Données simulées selon un modèle Normal-Poisson à 3 composantes de mêmes paramètres que l'exemple figure 8.5 à l'exception de la composante 3 (en vert) : $\boldsymbol{\sigma}_3^2 = (0.5, 0.2)^T$ (c) MCLUST choisit $K = 5$ (d)–(h) Classification de notre algorithme pour $K = 2, \ldots, 6$

sont grands, s'approche d'une distribution normale. Nous considérons à présent une distribution mal adaptée à une approximation gaussienne, en simulant un modèle de mélange Normal-Binomial à 3 composantes avec un effectif faible pour la binomiale : $n_b = 10$. Les classifications réalisées sur ce modèle sont données figure 8.9.

Premièrement, nous pouvons constater que la classification réalisée avec `MCLUST` (sur les deux variables non transformées) sont bien plus éloignées des vraies classes que dans les exemples précédents (cf figure 8.9 (b) `MCLUST` contraint à $K = 3$ et (e) `MCLUST` non contraint, qui choisit $K = 8$). Nous proposons d'utiliser dans ce cas uniquement l'information apportée par la variable continue pour établir une préclassification des données en 3 classes. Le

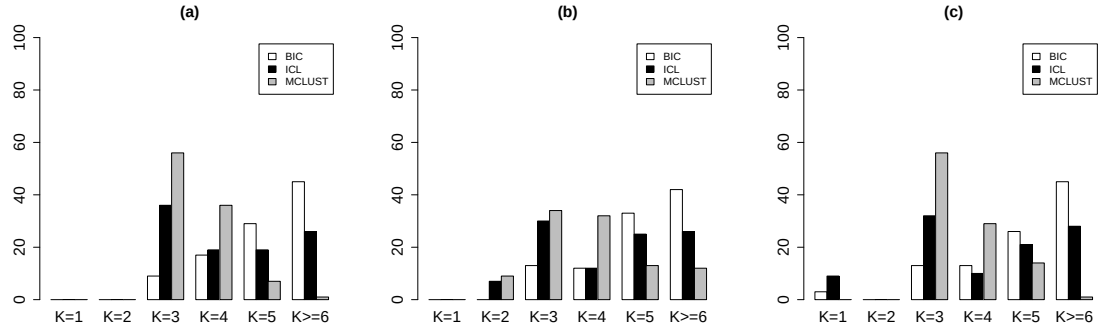


Fig. 8.8. Nombre de classes choisies par les critères BIC et ICL avec notre algorithme et BIC avec `MCLUST` sur 100 simulations pour les cas (a) figure 8.5 (b) figure 8.6 et (c) figure 8.7 . NB : $K = 6$ correspond à $K_{\max}$ pour notre algorithme, donc le critère de sélection (BIC ou ICL) pourrait atteindre leur maximum à K plus grand, d'où la notation $K \geq 6$. Les cas où `MCLUST` a choisi un nombre de classes K plus grand que 6 ont été regroupés dans cette catégorie.

résultat de `MCLUST` univarié sur la variable continue $\boldsymbol{y}_1$ fournit la classification de la figure 8.9 (c). Avec cette initialisation, nous obtenons la classification de la figure 8.9 (d), qui améliore nettement l'initialisation obtenue avec `MCLUST` univarié fixé à 3 classes (cf tableau 8.1 et boxplot figure 8.9 (f)), ainsi que la classification obtenue avec `MCLUST` bivarié contraint à 3 classes.

L'utilité du modèle de mélange gaussien latent est mise en évidence dans cet exemple, pour lequel l'hypothèse gaussienne n'est pas plausible. Il reste cependant une difficulté importante à résoudre pour pouvoir appliquer cette méthode en pratique : l'initialisation pour des composantes non gaussiennes. Le compromis que nous avons choisi ici, à savoir initialiser en classant seulement la variable continue, atteint rapidement ses limites (par exemple lorsque les composantes se séparent selon la marginale correspondant à la binomiale, ou encore si l'hypothèse gaussienne n'est plausible pour aucune des deux variables).

8.5.4 Lorsque les deux variables gaussiennes sont latentes : modèles de mélange BPLN

Dans les cas que nous avons considérés jusqu'ici, seulement l'une des deux variables gaussiennes était cachée. Dans le cas du modèle BPLN l'estimation est plus difficile, car aucune des deux variables n'est observée.

Les résultats d'estimation d'un modèle à deux composantes sont donnés figure 8.10. Notre algorithme diminue l'erreur de classification de manière significative par rapport à l'initialisation (cf tableau 8.1). De manière assez surprenante, nous pouvons constater que la classification proposée par `MCLUST` contraint à 2 classes est correcte, malgré la non adéquation des données au modèle gaussien. Par contre, celle-ci est fortement dégradée lorsqu'on enlève la contrainte du nombre de classes.

Notons que l'estimation de ce modèle se heurte dans de nombreux cas à des problèmes numériques, qui sont dûs à des singularités des matrices de covariance qui entraînent l'al-

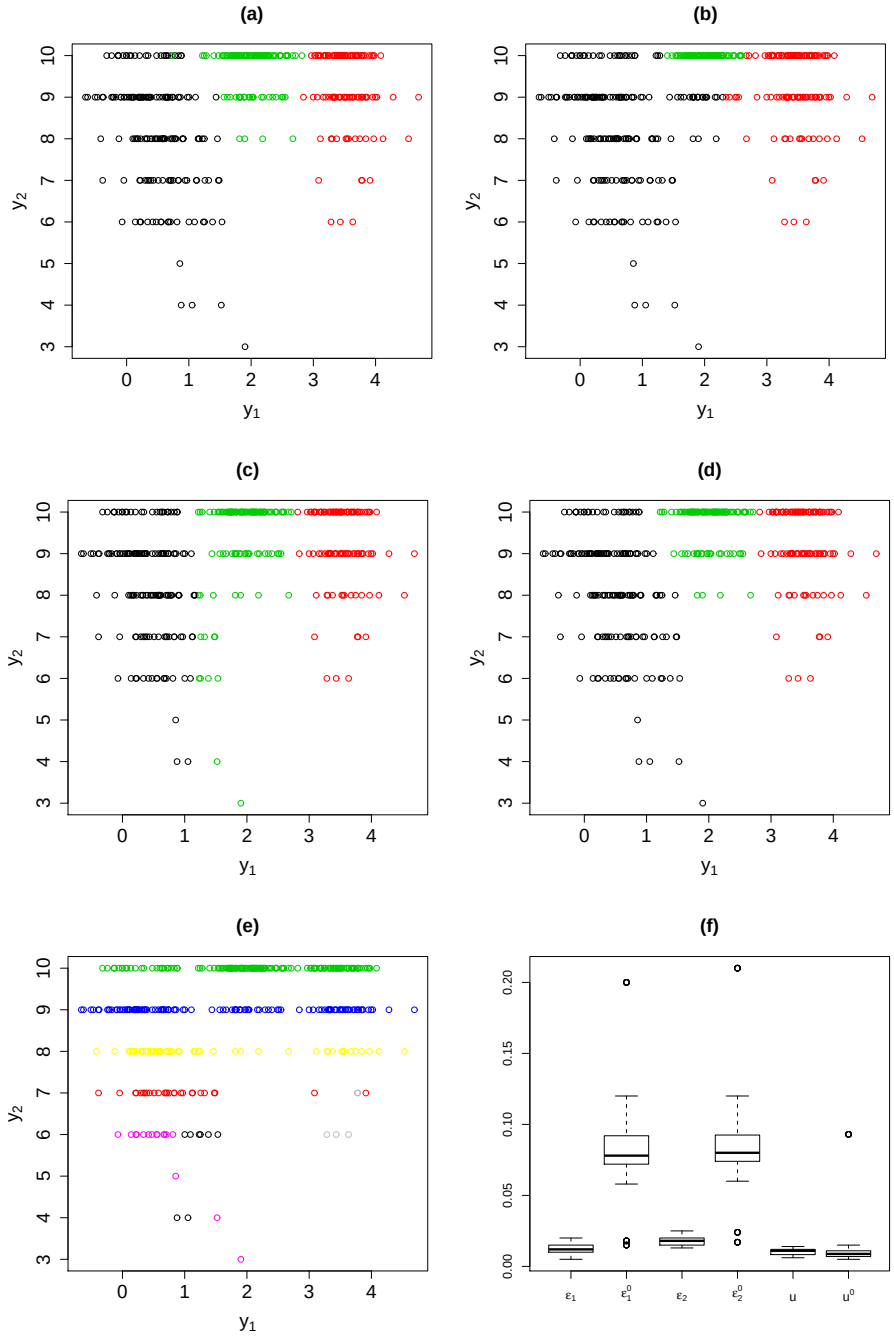


Fig. 8.9. Normal-Binomial : exemple de simulation à 3 composantes, avec $n_b = 10$. (a) Données simulées avec les paramètres : composante 1 (en noir) : $\pi_1 = 0.4$, $\boldsymbol{\mu}_1 = (0.5, 1.5)^T$, $\boldsymbol{\sigma}_1^2 = (0.2, 0.2)^T$, $r_1 = -0.9$, composante 2 (en rouge) : $\pi_2 = 0.3$, $\boldsymbol{\mu}_2 = (3.5, 2.5)^T$, $\boldsymbol{\sigma}_2^2 = (0.1, 0.2)^T$, $r_2 = 0$ composante 3 (en vert) : $\pi_3 = 0.3$, $\boldsymbol{\mu}_3 = (2, 3.5)^T$, $\boldsymbol{\sigma}_3^2 = (0.1, 0.1)^T$, $r_3 = 0$ (b) `MCLUST` bivarié, K fixé à 3 : $\epsilon_1 = 0.09$, $\epsilon_2 = 0.092$, $u = 0.011$ (c) Initialisation avec `MCLUST` univarié, K fixé à 3 (d) Notre classification, K fixé à 2 : $\epsilon_1 = 0.0075$, $\epsilon_2 = 0.016$, $u = 0.011$ (e) Nombre de classes choisi par `MCLUST` : K=8. (f) Boxplot des erreurs ϵ_1, ϵ_2, u pour 100 simulations. ϵ_1^0, ϵ_2^0, u^0 désignent les erreurs de classification de `MCLUST` bivarié avec K=3.

gorithme d'estimation vers des solutions dégénérées de vraisemblance infinie. Ceci est une conséquence de l'absence de contraintes sur les matrices de covariance. En effet, l'algorithme

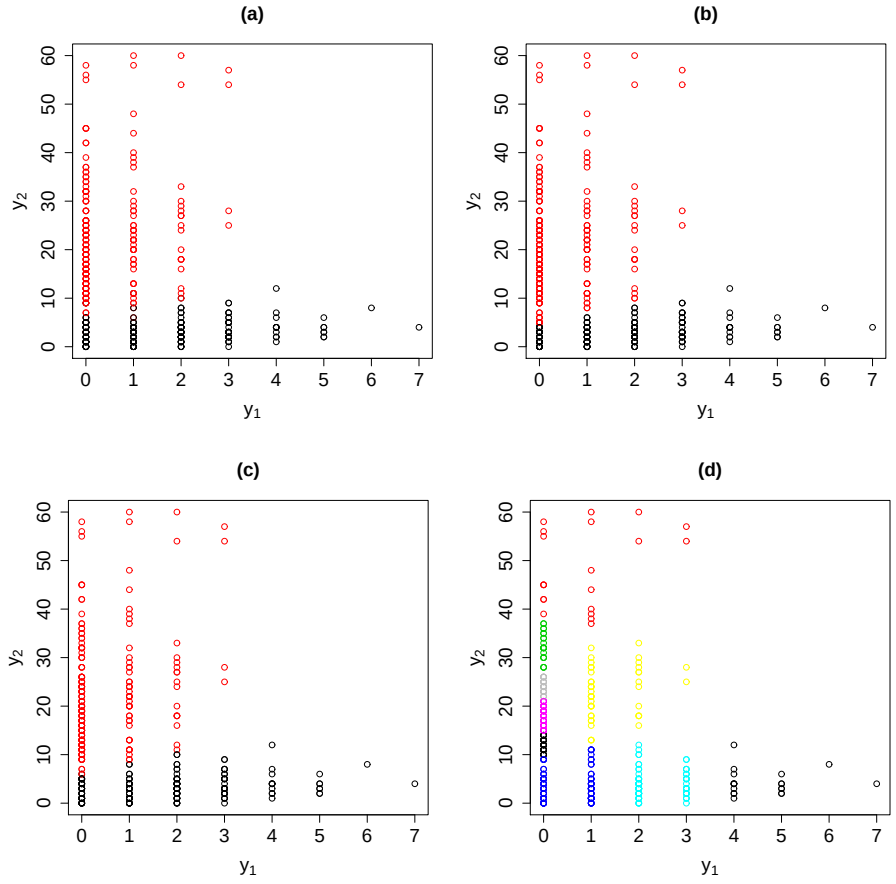


Fig. 8.10. Poisson log-normal bivarié (BPLN) : exemple de simulation à 2 composantes. (a) Données simulées avec les paramètres : composante 1 (en noir) : $\pi_1 = 0.5$, $\boldsymbol{\mu}_1 = (0.5, 1)^T$, $\boldsymbol{\sigma}_1^2 = (0.2, 0.2)^T$, $r_1 = 0.9$, composante 2 (en rouge) : $\pi_2 = 0.5$, $\boldsymbol{\mu}_2 = (-1, 3)^T$, $\boldsymbol{\sigma}_2^2 = (0.2, 0.2)^T$, $r_2 = 0.8$ (b) Initialisation avec `MCLUST` bivarié, K fixé à 3 : $\epsilon_1 = 0.038$, $\epsilon_2 = 0.076$, $u = 0.061$ (c) Notre classification, K fixé à 2 : $\epsilon_1 = 0.02$, $\epsilon_2 = 0.034$, $u = 0.024$ (d) Nombre de classes choisi par `MCLUST` : K=9.

`MCLUST` est souvent confronté au même problème, car le modèle "VVV" (cf 4.6.3), qui n'impose aucune contrainte sur les covariances des composantes, est rarement estimé dans les exemples que nous avons proposés ici, à cause des mêmes difficultés numériques. L'introduction de ces contraintes est d'autant plus nécessaire dans notre cas, que notre modèle est riche, par sa structure en couche latente.

8.6 Discussion - Perspectives - Extensions

Dans ce chapitre, nous avons proposé un modèle de mélange gaussien latent, dans le but de généraliser la classification automatique basée sur les modèles de mélanges à une grande variété de types de données, ainsi qu'à des données mixtes corrélées. Bien qu'il reste encore des difficultés à résoudre, notamment en ce qui concerne l'initialisation de notre algorithme, les résultats de notre étude simulatoire sont prometteurs, en particulier dans le cas où les composantes sont fortement non gaussiennes, tel que l'exemple de la figure 8.9, où

		différence moyenne	p-valeur	95%IC borne supérieure	test à 95% ($\alpha = 0.05$)
exemple fig 8.2	ϵ_1	$-1.75\ 10^{-4}$	0.035	$-1.61\ 10^{-5}$	*
	ϵ_2	$-2\ 10^{-4}$	0.011	$-5.7\ 10^{-5}$	*
	u	$-3.1\ 10^{-5}$	0.12	$1.24\ 10^{-5}$	
exemple fig 8.3	ϵ_1	-0.034	$< 10^{-16}$	-0.03	***
	ϵ_2	-0.06	$< 10^{-16}$	-0.057	***
	u	-0.04	$< 10^{-16}$	-0.038	***
exemple fig 8.9	ϵ_1	-0.07	$< 10^{-16}$	-0.063	***
	ϵ_2	-0.067	$< 10^{-16}$	-0.06	***
	u	$-4.2\ 10^{-3}$	0.018	$-9\ 10^{-4}$	*
exemple fig 8.10	ϵ_1	$-6.5\ 10^{-3}$	$9\ 10^{-13}$	$-5\ 10^{-3}$	***
	ϵ_2	-0.04	$< 10^{-16}$	-0.039	***
	u	-0.038	$< 10^{-16}$	-0.037	***

Tab. 8.1. Test t de Student d'égalité des moyennes sur échantillons appariés (unilatéral). Hypothèse alternative H1 : $\epsilon < \epsilon^0$, avec $\epsilon \in \{\epsilon_1, \epsilon_2, u\}$, où ϵ désigne les erreurs calculées à l'issue de notre algorithme de classification, et ϵ^0 les erreurs calculées après initialisation avec `MCLUST` bivarié.

les données sont simulées selon le modèle de mélange Normal-Binomial avec une binomiale à effectif faible ($n_b = 10$).

Plusieurs problèmes restent à résoudre avant de pouvoir proposer une méthode de classification automatique à l'aide des modèles gaussiens latents étudiés dans ce chapitre :

- Il faudrait établir l'identifiabilité des différents sous-modèles que nous considérons, ainsi que l'identifiabilité "pratique" lorsque la quantité d'information dans les données est faible (typiquement lorsque les $\boldsymbol{Y}_j$ sont obtenus par des lois de Bernoulli ou des lois binomiales à faible effectif).
- Avec l'algorithme actuel, des problèmes d'initialisation se posent lorsque les composantes du mélange ne sont pas approximables par des gaussiennes.
- Des problèmes numériques d'estimation peuvent se présenter dans certains cas, surtout dans les sous-modèles où les deux variables $\boldsymbol{\theta}_j$ sont cachées, tels que le mélange de BPLN. Comme nous l'avons mentionné plus haut, il s'agit de problèmes liés à la singularité des matrices de covariances des composantes.
- Enfin il reste à trouver un critère de sélection de modèles qui permette de choisir (dans la plupart des cas) le bon nombre de classes.

Dans le paragraphe suivant, nous proposons des perspectives d'évolution de l'algorithme d'estimation des modèles de mélange gaussiens latents présenté dans ce chapitre, qui tentent de résoudre les trois derniers problèmes que nous venons d'évoquer. Nous aborderons ensuite quelques extensions possibles de notre modèle.

8.6.1 Perspectives

L'initialisation pour des composantes non approximables par des gaussiennes

Dans l'étude par simulation nous avons vu que, malgré la non adéquation de l'hypothèse gaussienne aux types de données, `MCLUST` donnait de bons résultats pour l'initialisation de

notre algorithme. L'exemple de la figure 8.9 illustre les difficultés d'initialisation pour des composantes qui sont très mal adaptées à l'hypothèse gaussienne. Ceci est le cas notamment pour les données de comptage d'effectif maximal faible.

Nous avons vu en 2.2.1 que d'autres stratégies d'initialisation existent, notamment les initialisations au hasard. Dans le cas du modèle de mélange gaussien latent que nous proposons ici, une difficulté supplémentaire vient du fait qu'on n'observe pas une réalisation du mélange directement, celui-ci est caché. Étant données les difficultés des algorithmes de type EM à converger vers le vrai maximum de vraisemblance à partir d'une mauvaise initialisation, il est difficile dans ce cadre d'initialiser les paramètres des gaussiennes sousjacentes au hasard. De plus, la couche cachée rajoute une difficulté supplémentaire d'estimation, donc il est peu probable qu'une telle stratégie aboutisse.

Une solution alternative serait d'adapter la stratégie d'initialisation de Fraley & Raftery (2002) à notre modèle. Ces auteurs proposent une préclassification utilisant un algorithme hiérarchique ascendant basé sur le vrai modèle (qui est, dans leur cas, le mélange gaussien). Selon eux, ce type d'algorithme permet de converger plus rapidement qu'un EM en fournissant des solutions suboptimales, mais qui sont suffisamment bonnes pour une initialisation.

Introduire des modèles de covariance contraints et parcimonieux

Dans certains cas, notre algorithme est confronté à des problèmes de dégénerescence. Typiquement, une ou plusieurs proportions de mélange π_k tendent vers 0 ce qui fait disparaître des classes, ou encore des problèmes de singularités des matrices de covariance. Comme nous l'avons évoqué en 2.2.1, ceci est lié à la fonction de vraisemblance du mélange et arrive le plus souvent lorsque la covariance est autorisée à varier entre composantes et lorsqu'il y a beaucoup de composantes.

Une perspective pour améliorer l'estimabilité de notre modèle serait d'introduire des sous-modèles parcimonieux (en paramètres) pour les matrices de covariance des lois gaussiennes sousjacentes. Ainsi, des modèles parcimonieux, qui permettent de spécifier certaines caractéristiques de forme, de volume et d'orientation des classes, en passant par une décomposition en valeurs propres des matrices de covariance, ont été proposés par Banfield & Raftery (1993) et Celeux & Govaert (1995), et sont implémentés dans de nombreux logiciels d'estimation de modèles de mélange gaussiens multivariées, tels `MCLUST` et `MIXMOD`. Ceci permet de contraindre ces caractéristiques entre les différentes classes, de manière à ce qu'elles soient égales ou variables entre classes.

Cette étape facilitera les calculs en grande dimension et permettra de stabiliser l'algorithme dans certains cas. En effet, le modèle dont la matrice de covariance n'est pas contrainte entre classes ("VVV" dans `MCLUST`, cf 4.6.3), est rarement estimé pour tous les nombres de classes, alors que plus les modèles sont contraints, plus ils sont estimables en pratique.

Choisir le nombre de composantes et choisir le nombre de classes

Nous avons vu que les critères BIC et ICL ne choisissent pas systématiquement le bon nombre de classes. Ainsi, le BIC surestime ce nombre presque systématiquement, et l'ICL arrive à estimer le bon nombre de classes plus souvent, mais pas assez souvent pour être fiable (voir tableau 8.1). Comme nous l'avons évoqué dans le commentaire des résultats, ceci peut être dû à un nombre de données insuffisant par rapport aux inconnues du modèle (couche latente). Quelques solutions algorithmiques peuvent être envisagées pour améliorer ce résultat.

Dans un contexte de classification à l'aide de modèles de mélange, le nombre de composantes estimées est considéré comme étant le nombre de classes. Baudry et al. (2010) proposent un point de vue différent, en argumentant que, même si le BIC permet de choisir le bon nombre de composantes gaussiennes pour ajuster la densité de probabilité des données, celui-ci pourrait ne pas correspondre au nombre de groupes (classes) dans les données, tout simplement parce qu'une classe pourrait être mieux représentée par un mélange de composantes gaussiennes que par une seule gaussienne. Ils proposent donc de combiner les approches BIC et ICL de la manière suivante : utiliser le BIC pour sélectionner le nombre de composantes du mélange gaussien, puis, comme dans une approche de classification hiérarchique, combiner les composantes qui optimisent un critère d'entropie (calculé de la même manière que pour le critère ICL, équation 4.5). Le critère ICL peut être utilisé ensuite pour choisir le meilleur modèle parmi les solutions fournies par cette approche hiérarchique.

Cette approche pourrait être intéressante ici, car, comme nous pouvons le voir dans les figures 8.5, 8.6 et 8.7, les classes sont progressivement subdivisées entre $K = 1$ et $K = 6$, et il est possible qu'en utilisant un critère d'entropie, un nombre de classes inférieur soit choisi.

8.6.2 Extensions

Les extensions proposées pour le modèle hiérarchique sans mélange, discutées dans 6.3, peuvent également être envisagées dans ce cas. Plusieurs auteurs ont traité certaines de ces extensions.

Les données ordinales

Nous n'avons pas abordé les données ordinales dans cette thèse. Des modèles pour classer des données mixtes catégorielles et continues ont déjà été proposés. Certains sont des extensions directes des modèles que nous avons évoqués en 6.3.3 et que nous reprenons brièvement ici (voir aussi McLachlan & Peel, 2000, chapitre 5.2, p. 136).

Le modèle d'Everitt (1988), où chaque variable ordinale est obtenue à partir d'une variable continue sousjacente par seuillage, a également été proposé dans un cadre de classification, en supposant que les variables continues non observées et les variables continues observées proviennent d'un mélange de lois normales multivariées avec la même matrice de

covariance. Puisque les variables ordinales ne fournissent pas d'information sur les moyennes et variances des variables continues sous-jacentes, Everitt (1988) fixe les moyennes à 0 et les variances à 1. Les valeurs des seuils sont autorisées à varier entre classes. Les probabilités des catégories sont déterminées par les valeurs seuils.

Le modèle de localisation (*location models*) de Olkin & Tate (1961), que nous avons brièvement présenté en 6.3.3, a également été étendu pour la classification. Krzanowski (1993) utilise des mélanges finis de modèles de localisation (*location mixture model*) pour classer simultanément des données catégorielles et des données continues corrélées. Willse & Boik (1999) fixent les conditions d'identifiabilité de ce modèle (*restricted location mixture model*). Dans ce modèle, d variables sont observées, dont q sont catégorielles, et $d-q$ sont continues. Les q variables catégorielles sont remplacées par une variable aléatoire $\boldsymbol{Y}'$ issue d'une distribution multinomiale à m cellules, i.e. $(\boldsymbol{Y}')_s = 1$ si les réalisations des q variables catégorielles dans $\boldsymbol{Y}$ correspondent à la combinaison s. m est défini par $m = \sum_{i=1}^{q} m_i$, où la $i^{\text{ème}}$ variable catégorielle possède m_i modalités. Donc les associations entre variables catégorielles sont transformées en relations entre les probabilités des cellules de la multinomiale. Ce modèle suppose ensuite que, conditionnellement à $(\boldsymbol{y}')_s = 1$ et à l'appartenance à la $k^{\text{ème}}$ composante, la distribution des $d-q$ variables continues est gaussienne de moyenne $\boldsymbol{\mu}_{ks}$ et matrice de covariance $\boldsymbol{\Sigma}_k$ qui est identique pour toutes les cellules s (i.e. pour toutes combinaisons possibles des réalisations des q variables catégorielles). Cette hypothèse de normalité conditionnelle rend l'estimation par MV immédiate à implémenter dans l'EM.

L'hypothèse de normalité latente

Dans ce cas, cette hypothèse est probablement la plus souple, d'une part parce qu'elle est la plus facile à estimer dans un cadre multivarié, d'autre part car elle est assez riche en termes de formes des classes et de corrélations entre variables. En effet, la pratique du logiciel `MCLUST` permet d'illustrer la souplesse de l'hypothèse gaussienne, tant par la large variété de formes et d'orientations de classes qu'elle permet, que par sa capacité à s'adapter à des données non gaussiennes et même parfois à des données non continues.

Prise en compte de covariables observées

Ce modèle peut être étendu pour prendre en compte des covariables observées. Cette extension pourra s'appuyer sur plusieurs travaux existants. Celeux, Martin & Lavergne (2005) ont été les premiers à proposer des mélanges de modèles linéaires à effets mixtes, avec un algorithme EM qui se résout de manière explicite. Martinez, Lavergne & Trottier (2009) ont étendu cette approche à un mélange de modèles exponentiels à effets mixtes (un cas particulier de GLMM ou le lien est exponentiel au lieu d'être linéaire). Dans leur cas comme dans le notre, l'EM ne peut être utilisé directement. Ils proposent deux manières pour estimer les paramètres, une méthode basée sur une linéarisation spécifique à l'hypothèse de distribution exponentielle, et une approche générale basée sur l'algorithme de McCulloch

(1997) (MCEM avec une étape de MCMC Metropolis-Hastings dans l'étape E, que nous avons discutée en 6.2).

Prise en compte de l'autocorrélation spatiale des données

Enfin on pourrait intégrer l'autocorrélation spatiale des données dans ce modèle en utilisant un champ markovien caché. Comme nous l'avons mentionné dans le chapitre 3 (en 3.2.1), l'extension des modèles de mélange à des données dépendantes peut se faire en utilisant des champs de Markov cachés pour les variables $\boldsymbol{Z}$ indicatrices des composantes (McLachlan & Peel, 2000, p. 33 et chapitre 13).

Notons pour finir qu'un modèle qui prend en compte des covariables ou des dépendances entre observations (telles que l'autocorrélation spatiale) risque de nécessiter un nombre de données considérable pour pouvoir être estimé.

Cinquième partie

Conclusion et perspectives

9

Les assemblages : un outil d'étude des interactions entre espèces

Dans ce chapitre, nous dressons un bilan des contributions méthodologiques de cette thèse dans le cadre de l'étude de la coexistence d'espèces en écologie spatiale. Nous ne reprenons pas ici les discussions sur les limites et extensions spécifiques aux méthodes proposées dans les parties II, III et IV, qui ont été détaillées respectivement dans les chapitres 3, 6 et en section 8.6. Au lieu de cela, nous résumons la démarche scientifique de cette thèse, puis nous envisageons plusieurs perspectives d'application et d'évolution des méthodes proposées. Nous intégrons ainsi la démarche suivie dans cette thèse dans une démarche plus générale d'étude des interactions entre espèces, en utilisant la notion d'assemblage d'espèces, que nous avons définie en termes d'abondances locales des espèces et qui représente une des signatures des interactions interspécifiques. Ce chapitre s'articule ainsi autour de la méthode CASA, présentée dans la partie II, et de ses extensions, qui ont pour finalité d'appréhender les interactions entre espèces qui sont à l'origine des assemblages.

9.1 CASA : une approche de classification pour analyser la structure spatiale des assemblages d'espèces

Le point de départ de cette thèse a été un problème d'analyse de données en écologie spatiale : comment caractériser la manière dont plusieurs espèces s'assemblent en densité ? Pour répondre à cette question, nous avons introduit une notion d'assemblage d'espèces, que nous avons définie comme un ensemble de sites pour lesquels les combinaisons d'abondances des espèces observées sont similaires. Cette notion d'assemblage d'espèces est définie indépendamment de l'échelle d'étude des espèces, et par conséquent elle peut être utilisée autant pour une zone d'étude très étendue (île, comme dans le jeu de données présenté en 1.4.2) que pour une zone d'étude très petite (rameau, comme dans le jeu de données présenté en 1.4.1). L'approche que nous avons proposée dans cette thèse diffère des approches existantes, qui consistent à calculer des indices (la plupart du temps globaux, parfois locaux) d'association (ou de dissociation) spatiale entre espèces, et qui insistent sur la prise en compte de l'autocorrélation spatiale des données en se basant sur les distances entre sites (Perry & Dixon, 2002). Les différentes étapes de notre approche sont résumées dans la figure 9.1.

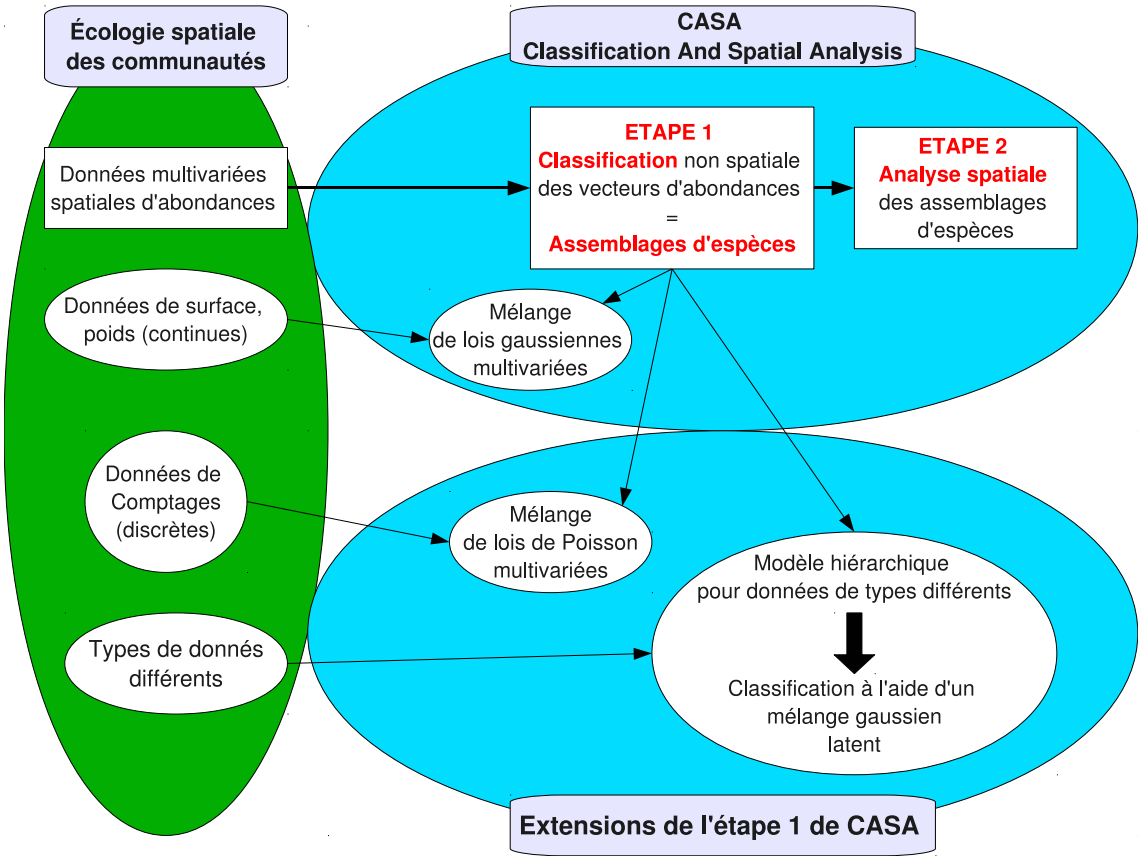


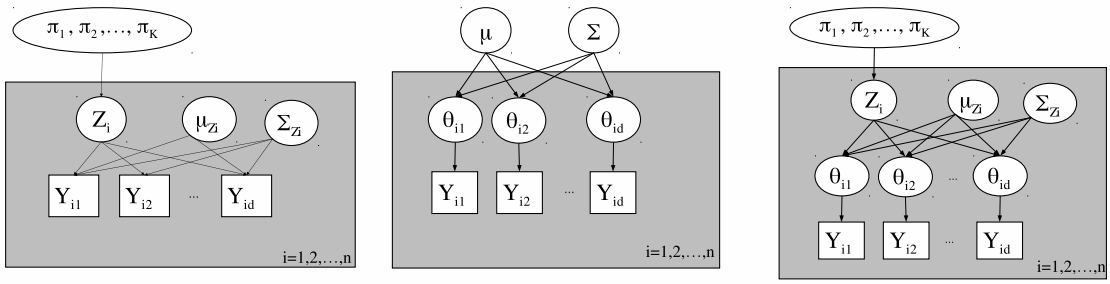
Fig. 9.1. Schéma résumant les étapes de la thèse

Nous avons défini les assemblages d'espèces en effectuant une classification des vecteurs multivariés d'abondances d'espèces mesurés aux différents sites d'observation. Les assemblages d'espèces ainsi définis sont ensuite explorés grâce à des méthodes classiques d'analyse de données spatiales, à la recherche d'une structure spatiale particulière des classes individuelles (par des variogrammes), ou d'une disposition relative particulière des classes entre elles (par des barycentres ou des calculs de distances). La méthode de classification choisie pour définir des types d'assemblages, basée sur des mélanges de lois gaussiennes multivariées (figure 9.2 (a)), présente plusieurs avantages. D'une part elle fournit pour chaque observation, en plus de sa classe, des probabilités d'appartenance à chacune des autres classes, ce qui permet d'intégrer l'erreur d'affectation des observations aux classes dans la deuxième étape d'analyse spatiale, afin d'avoir une estimation précise de l'erreur finale. D'autre part, cette méthode définit les assemblages d'espèces par des lois de probabilité multivariées, ce qui permet d'envisager d'autres lois lorsque les données d'abondances sont de types différents. Cette méthode, intitulée CASA ("*Classification And Spatial Analysis*"), a été détaillée dans la partie II.

La suite de cette thèse a consisté à élargir le champ d'application de CASA à d'autres types de données par la recherche de lois de probabilité adaptées. Une première adaptation aux données de comptage a été proposée en remplaçant les lois gaussiennes du modèle de mélange par des lois de Poisson multivariées (cf chapitre 5). Les limites de cette approche

sont étroitement liées aux limites des lois de Poisson multivariées, à savoir le fait de ne pas permettre des corrélations négatives au sein des classes et le rapport d'égalité de moyenne et de variance de la loi de Poisson, qui limite la forme des classes.

Pour avoir un modèle général qui puisse s'adapter à la diversité des types de données en écologie, ainsi qu'à des données corrélées de plusieurs types (*mixed mode data*), nous avons proposé dans la partie III un modèle latent gaussien (figure 9.2 (b)), où les divers types de données sont obtenus par différentes fonctions de lien et distributions univariées, indépendantes conditionnellement à la variable latente gaussienne. Ce modèle et sa méthode d'estimation sont adaptés à la classification dans la partie IV, en remplaçant l'hypothèse d'une loi gaussienne sousjacente par une hypothèse de mélange gaussien sousjacent. On obtient ainsi un "modèle de mélange gaussien latent" (figure 9.2 (c)). Ce modèle bénéficie des avantages des lois gaussiennes multivariées, qui sont d'une part très souples en termes de gamme de corrélations et de formes de classes permises, et d'autre part aisées à manipuler et à estimer dans un cadre multivarié. Cependant, la structure latente du modèle, qui lui confère cette richesse en termes de formes de classes et types de données, engendre dans certains cas des difficultés pratiques de mise en œuvre de cette méthode de classification. Ces difficultés, discutées en 8.6, et dont la plus limitante est l'initialisation dans les cas fortement non gaussiens, restent encore à résoudre.



(a) Mélange gaussien (b) Modèle gaussien latent (c) Mélange gaussien latent

Fig. 9.2. Graphes orientés acycliques (*Directed Acyclic Graph* ou DAG) des trois modèles

Par la suite, nous présentons quelques perspectives de la méthode CASA, d'abord sous l'angle des applications, qui permettent d'illustrer, sur les deux jeux de données écologiques présentés en introduction (en 1.4.1 et 1.4.2), l'utilité des extensions de CASA que nous envisageons, ensuite sous un angle qui intègre la dimension temporelle des données et des processus écologiques.

9.2 Extensions de CASA en vue de perspectives d'application

Nous présentons ici quelques extensions de CASA en illustrant leur utilité sur des applications, notamment sur les deux jeux de données écologiques présentés en introduction (en 1.4.1 et 1.4.2).

9.2.1 Extension de l'étape de classification de CASA

Une première extension consiste à intégrer la méthode de classification étudiée dans le chapitre 8 dans CASA, pour proposer une méthode CASA généralisée à des types de données variés.

Etude des assemblages dans le système hôte-pathogène plantain-oïdium grâce au modèle hiérarchique Normal-Binomial

Pour illustrer l'utilité de cette démarche, nous prenons l'exemple de données considérées dans le paragraphe 1.4.2. Ces données ont été analysées avec la méthode CASA dans le chapitre 4, qui a permis de mettre en évidence cinq assemblages différents des deux espèces plantain (plante hôte) et oïdium (pathogène), dont l'analyse spatiale a révélé l'existence d'un gradient d'infection du nord-est au sud-ouest des îles Åland. Nous rappelons que les données brutes d'abondances d'espèces sont des surfaces dans le cas du plantain et des occurrences (présence-absence) dans le cas de l'oïdium, et que celles-ci ont dû être agrégées (dans le temps et l'espace) afin d'obtenir des données pseudo-continues, qui puissent être classées sous l'hypothèse gaussienne. Dans le paragraphe 3.2.1, nous avons soulevé les problèmes posés par cette agrégation des données d'abondances, notamment la perte d'information sur les tailles de parcelles de plantain, qui masque ainsi l'influence potentielle de la taille des parcelles sur la fréquence d'infection.

Grâce au modèle proposé dans la partie III, nous pouvons envisager la prise en compte des données brutes. Afin de garder l'information sur la taille des parcelles de plantain, nous proposons d'agréger les données uniquement dans le temps. Le modèle qui semble le plus adapté pour ce jeu de données est le modèle Normal-Binomial, défini en 8.3.2, avec une fonction de lien logarithme ($g_1 = \log$) pour la variable continue, qui est ici la surface moyenne couverte par le plantain sur chaque parcelle au cours des 7 années, et avec l'effectif de la binomiale fixé à 7 ($n_b = 7$) pour la variable discrète, qui est ici le nombre d'occurrences de l'oïdium sur les 7 années (2001–2008). Ceci revient à supposer que l'occurrence de l'oïdium sur une parcelle est une épreuve de Bernoulli, indépendante entre parcelles conditionnellement à la variable latente gaussienne $\boldsymbol{\theta}$, et qui a la même probabilité p de succès au cours des 7 années pour une parcelle donnée. On compte ainsi le nombre de "succès d'infection" sur les 7 années de suivi, ce qui correspond à l'hypothèse d'une distribution binomiale. Enfin, la fonction de lien logarithme sert à diminuer la très forte dispersion des données de surface.

Nous avons estimé ce modèle sous l'hypothèse d'une seule classe d'assemblage, c'est-à-dire sous le modèle hiérarchique Normal-Binomial sans mélange de lois. Les résultats de convergence des paramètres de ce modèle, obtenus par l'algorithme MCEM initialisé par la méthode des moments (à l'aide d'un algorithme itératif similaire à celui présenté en 7.7.2, mais adapté au cas Normal-Binomial), sont présentés sur la figure 9.3. Une corrélation positive ($r = 0.4$ avec un écart-type asymptotique de 0.026) est estimée entre la surface

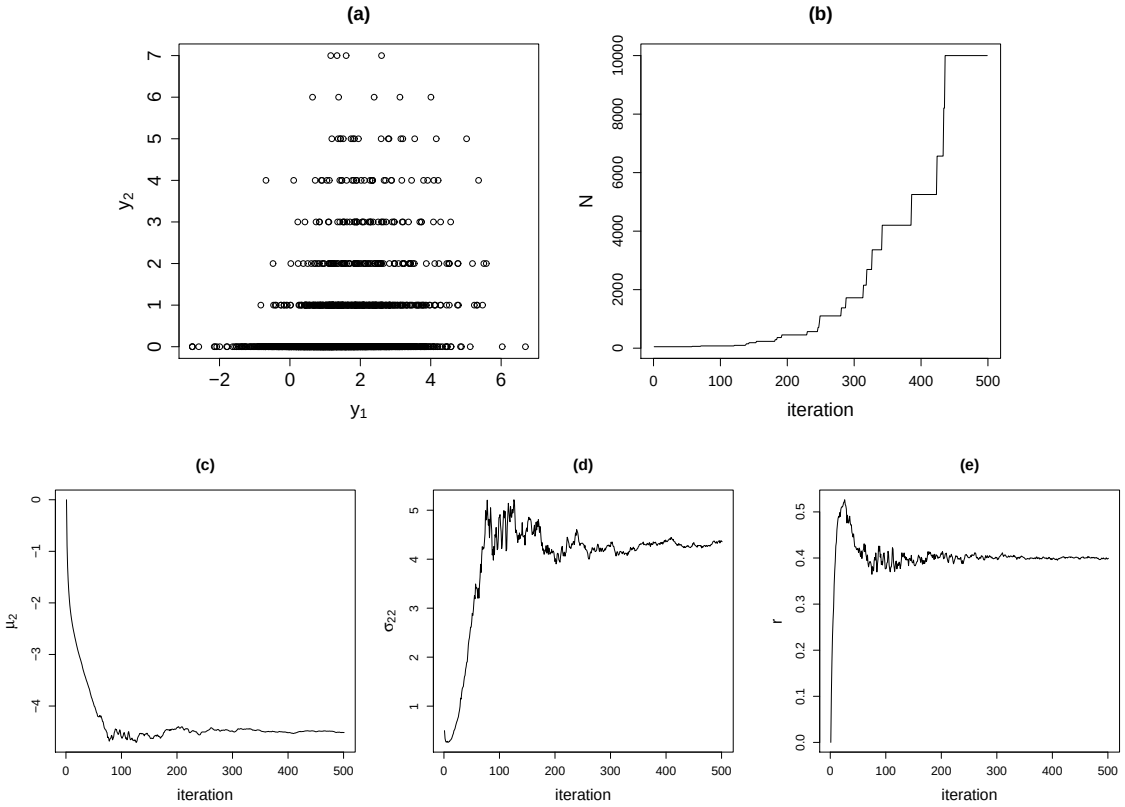


Fig. 9.3. Ajustement du modèle Normal-Binomial (sans mélange) sur les données plantain-oïdium. (a) Les données : y_1 : logarithme des surfaces de plantain observées sur chacune des 2639 parcelles et y_2 : nombre d'occurrences de l'oïdium sur les 7 années d'études (2001–2008), (b) La taille de l'échantillon d'importance N augmente de 50 à 10000, (c)–(e) Convergence des paramètres moyenne μ_2, variance σ_{22} et corrélation r atteinte sur 500 itérations.

moyenne du plantain par parcelle et le nombre d'occurrences de l'oïdium, qui montre qu'il y a une influence positive de la taille des parcelles sur la probabilité d'occurrence de l'oïdium.

Pour des contraintes de temps et à cause de problèmes d'initialisation qui restent à résoudre pour pouvoir appliquer la méthode de classification présentée dans la partie IV, nous n'avons pas pu tester l'hypothèse de l'existence de plusieurs assemblages, que nous avions mis en évidence dans le chapitre 4 sur une version transformée de ce jeu de données.

Enfin, une question qui apparaît lors de l'étude de ce jeu de données et qu'il faudrait approfondir est : quelle quantité d'information faut-il dans les données, afin de pouvoir estimer tous les paramètres du modèle (de la gaussienne sousjacente ou du mélange gaussien sousjacent) ? Pour pouvoir répondre, il faudrait établir les propriétés d'identifiabilité des différents sousmodèles qui peuvent être considérés ici, tels que le modèle Normal-Bernoulli ou le modèle Normal-Binomial. En effet, il se pourrait que le modèle Normal-Bernoulli, et le modèle Normal-Binomial à faible effectif, ne soient identifiables que sous certaines conditions (variance de la gaussienne sousjacente fixée, ou, dans le cas avec mélange, imposition de contraintes entre classes sur les matrices de covariance des composantes).

Etude des assemblages de pucerons grâce au modèle MPLN

Le jeu de données 1.4.1 a été étudié dans le chapitre 5 en supposant que les données étaient issues d'un mélange de lois de Poisson multivariées. Nous avons vu qu'aucun des deux critères BIC et ICL ne permettait de choisir un nombre de classes "raisonnable" (pour qu'elles puissent représenter des assemblages interprétables), et nous en avons conclu que les limites de forme de classe des lois de Poisson sont à l'origine de cette mauvaise adéquation. La classification basée sur des mélanges de MPLN, qui permet à la fois une corrélation négative et des données surdispersées par rapport à la loi de Poisson au sein de chaque classe, semble adaptée à ce jeu de données et devra être testée. Pour cela, des travaux supplémentaires sur l'initialisation et sur la stabilité de l'algorithme doivent être menés, notamment par l'introduction de modèles contraints.

9.2.2 Extensions de l'étape d'analyse spatiale de CASA

Des évolutions de l'étape d'analyse spatiale peuvent être envisagées à plusieurs niveaux : la mise au point ou l'adaptation de nouveaux outils d'analyse spatiale, notamment pour des espaces non euclidiens, l'intégration de covariables dans l'analyse spatiale, ou encore l'utilisation de la démarche de l'étape d'analyse spatiale dans d'autres domaines d'application.

Proposer des outils d'analyse spatiale pour des supports arborescents : application à l'étude des assemblages de pucerons sur rameaux

Un des arguments que nous avons donné pour montrer la flexibilité de la méthode CASA et de la notion d'assemblage d'espèces que nous avons défini, qui s'affranchit de l'autocorrélation spatiale des données, a été de dire que l'adaptation de cette méthode à des espaces non euclidiens en sera ainsi facilitée. Le jeu de données de comptages de pucerons (1.4.1) illustre ce propos. Une extension de CASA pourrait proposer une plus large gamme d'outils d'analyse spatiale, afin de permettre l'analyse spatiale d'assemblages d'abondances sur des espaces non euclidiens. Par exemple, l'adaptation de méthodes de statistiques spatiales pour des supports hiérarchiques pourrait être envisagée (Garretta, Monestiez & Ver Hoef, 2010).

Tenir compte des covariables

Comme nous l'avons mentionné dans la discussion du chapitre 8, lorsqu'on suspecte l'influence de certains facteurs environnementaux sur la distribution des assemblages, on peut envisager de prendre en compte des covariables environnementales dans la classification en proposant des mélanges de GLMM (Martinez, Lavergne & Trottier, 2009).

Une autre approche consiste à tenir compte des covariables seulement après la classification, de la même manière que l'analyse spatiale est considérée seulement dans la deuxième étape de CASA. La première étape de CASA aura ainsi servi à définir des assemblages,

alors que la deuxième étape permettra de tester des hypothèses sur le pouvoir explicatif de divers facteurs environnementaux sur la structure spatiale des assemblages, grâce à des modèles GLMM spatiaux.

Application à l'étude de la structuration génétique spatiale

Enfin, notons que d'autres champs d'application potentiels existent, notamment en génétique des populations, où l'on s'intéresse à la structuration génétique spatiale des individus. Une approche de classification probabiliste a déjà été étudiée dans ce cadre par Pritchard, Stephens & Donnelly (2000), qui proposent d'affecter les individus à des populations selon leurs génotypes sous certaines hypothèses génétiques, et d'estimer leur modèle par une méthode bayésienne, basée sur des simulations MCMC. Plus précisément, si la variable $\boldsymbol{Y}$ est le génotype observé des individus, $\boldsymbol{Z}$ représente la population d'origine (inconnue) des individus, et $\boldsymbol{P}$ les fréquences alléliques (inconnues) dans chaque population, ils supposent que chaque allèle à chaque locus dans chaque génotype est obtenu par un tirage indépendant à partir de la distribution de fréquences alléliques correspondant à sa population d'origine, ce qui permet de caractériser entièrement la distribution conditionnelle de $\boldsymbol{Y}$ sachant $\boldsymbol{Z}$ et $\boldsymbol{P}$ ($f_{\boldsymbol{Y}|\boldsymbol{Z},\boldsymbol{P}}$). Ces auteurs ont proposé un logiciel, intitulé `structure`, pour effectuer cette classification. Les résultats de cette classification, qui est pourtant probabiliste, ne sont pas exploités entièrement, car seulement les affectations finales aux classes (qui représentent ici les populations d'origine des individus) sont étudiées (Hubisz et al, 2009). L'intégration de cette méthode de classification dans une démarche en deux étapes telle que celle proposée dans CASA, pourrait s'avérer utile. En effet, CASA prendrait en compte les probabilités d'appartenance aux classes fournies par le logiciel `structure` dans l'analyse spatiale des classes obtenues, ce qui permettrait une meilleure estimation de l'erreur.

9.3 Vers une dynamique des assemblages

Dans cette thèse nous avons eu une approche uniquement spatiale des assemblages d'espèces, même lorsque les données comportaient une dimension temporelle (jeu de données 1.4.2). Or les processus écologiques sont intrinsèquement dynamiques. Nous abordons ici quelques perspectives d'étude des assemblages dans un cadre temporel, ou du moins qui tentent de tenir compte de la dimension temporelle inhérente au processus de formation des assemblages.

9.3.1 CASA temporel : suivi de données temporelles

L'introduction de données spatiotemporelles dans CASA peut être envisagée pour suivre la dynamique des assemblages au cours du temps. La méthode CASA pourrait être utilisée pour détecter les assemblages au temps initial (sur les données obtenues à la 1$^{\text{ère}}$ campagne de mesures, par exemple en 2001 pour les données plantain-oïdium de 1.4.2). L'évolution

de la distribution spatiale des assemblages au cours du temps pourra ensuite être étudiée en utilisant une classification supervisée (*discriminant analysis*, Fraley & Raftery, 2002) pour affecter les nouvelles données aux classes existantes, en supposant que les assemblages restent les mêmes au cours du temps (en termes de nombre et de caractéristiques). Cette analyse fournirait des cartes des assemblages pour chaque campagne de mesures, ce qui permettrait de mettre en évidence des éventuels déplacements des assemblages, ou au contraire une persistence des assemblages d'espèces observés (qui pourra ensuite être attribuée soit à un facteur environnemental persistant, soit à une hypothèse d'équilibre stable de la communauté d'espèces).

9.3.2 Simulation de la dynamique d'espèces en interaction : comprendre la formation des assemblages par une approche mécaniste

Une autre manière d'étudier la dynamique des assemblages est de penser en termes d'interactions simples entre individus, par un modèle mécaniste. L'utilisation d'un modèle mécaniste permettrait d'interpréter les assemblages obtenus par rapport aux types d'interactions simples à l'échelle des individus mis en entrée du modèle. Cependant, plusieurs difficultés se posent dans un tel cadre. D'une part, il faut connaître de nombreuses caractéristiques des espèces étudiées pour alimenter un modèle, d'autre part, il faut avoir des idées sur les interactions et contraintes qui aboutissent à la construction de plusieurs types d'assemblages. Rappelons enfin que des interactions différentes peuvent mener à des assemblages identiques, donc qu'il est risqué d'inférer les causes d'un processus par un modèle.

Une approche mécaniste stochastique, basée sur les modèles de branchements spatiaux, a été brièvement abordée au cours de cette thèse.

Introduction aux processus de branchements multitypes

Les processus de branchement sont des modèles de reproduction qui ont été posés au début dans un cadre non spatial (Haccou, Jagers & Vatutin, 2007).

Le processus de branchement simple (en temps discret), appellé processus de Bienaymé-Galton-Watson, est un processus markovien qui modélise la reproduction d'une population (e.g. une espèce). Il est basé sur un modèle individu-centré qui définit un comportement moyen des individus : Chaque individu i de la génération t produit ξ_i individus dans la génération $t+1$ selon une loi de probabilité fixée. A la génération $t+1$ nous obtenons Z_{t+1} individus, avec : $Z_{t+1} = \sum_{i=1}^{Z_t} \xi_i$.

Un processus de branchement multitype permet de modéliser S types d'individus $Z_t = (Z_{t1}, \ldots, Z_{tS})^T$. Ici, chaque individu i peut "produire" des individus dans chacun des S types $\xi_i = (\xi_{t1}, \ldots, \xi_{tS})^T$. La distribution de ξ_i dépend du type de l'individu i et la taille de la population finale sera : $Z_{t+1} = \sum_{s=1}^{S} \sum_{i=1}^{Z_t} \xi_i^{(s)}$. Une résolution analytique est possible sous une hypothèse forte d'indépendance (chaque individu génère de façon indépendante un processus). Il n'y a plus de résolution analytique de ces problèmes dès qu'on introduit :

– la densité-dépendance ou capacité du milieu (globale ou locale),

– des environnements qui varient dans le temps, de manière déterministe ou aléatoire (typiquement, lorsque la distribution du nombre de descendants ξ varie avec le temps).

Approche par simulation : Processus de branchement sur grille

Des études par simulation peuvent permettre de relier un assemblage d'espèces (défini avec CASA) à un type d'interaction introduite à l'échelle des individus, ou à une interaction avec le milieu.

Les processus de branchements sur grille (Haccou, Jagers & Vatutin, 2007; Révész, 1994) font partie des méthodes individu-centrées utilisées en écologie théorique pour modéliser de manière spatialement explicite la croissance d'une population, à travers des processus de reproduction et de déplacement des individus sur une grille. Au temps $t = 0$, un ou plusieurs individus sont localisés sur une grille. Ces individus se déplacent, donnent naissance à quelques descendants selon la loi d'un processus de Galton-Watson et meurent après un temps (aléatoire ou non). Les descendants évoluent indépendamment les uns des autres selon la même loi.

Pour chaque modèle de simulation, on précise :

1. la loi du champ aléatoire initial (e.g. individu unique au centre de la grille),
2. la loi qui gouverne le déplacement des individus (e.g. marche aléatoire),
3. la loi de la durée de vie d'un individu (e.g. durée de vie de 1),
4. le type de processus de Galton-Watson, qui peut dépendre de la position ($m < 1$ souscritique, $m = 1$ critique, $m > 1$ supercritique, en notant m le nombre d'individus produits à chaque temps t).

Ces simulations peuvent être un support pour comprendre comment des assemblages peuvent se former de manière dynamique, mais il reste la difficulté d'introduire plusieurs types d'assemblages. Un exemple de simulation est donné dans la figure 9.4 avec les caractéristiques suivantes :

1. Loi du champ aléatoire initial : 5 individus situés au centre de la grille.
2. Loi qui gouverne le déplacement des individus : marche aléatoire simple, symétrique sur $\mathbb{Z}^2$ avec possibilité de rester sur place (équiprobable).
3. Loi de la durée de vie d'un individu : durée de vie de 1.
4. Type de processus de Galton-Watson : de $m = 0.5$ à $m = 1.5$ selon le milieu.

Une perspective serait de combiner une telle approche mécaniste par processus de branchement avec la méthode CASA, pour aller plus loin dans l'analyse des assemblages. Deux utilisations de ces processus peuvent être envisagées :

– un processus de branchement non spatial, afin d'obtenir des familles de lois adaptées aux types de données que l'on considère, avec des paramètres ayant une interprétation biologique, pour l'étape de classification de CASA,

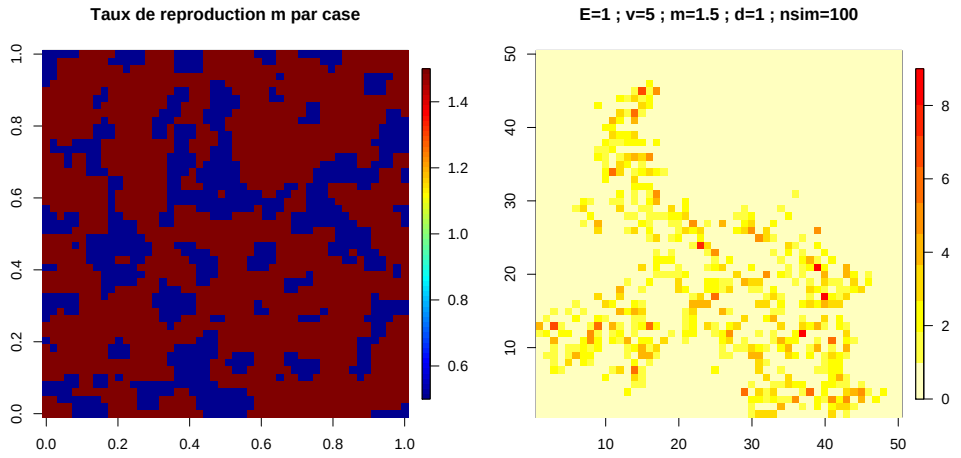


Fig. 9.4. Exemple simulé d'un processus de branchement sur grille simple (1 seul type d'individus) en milieu hétérogène.

– un processus de branchement spatial sur grille, afin de simuler la dynamique et les interactions de plusieurs espèces et d'explorer les propriétés de CASA dans la définition des assemblages sur ces simulations.

Références

Aitchison, J. & Ho, C.H.(1989) The multivariate Poisson log-normal distribution. *Biometrika* **76**, 643–653. 72, 73, 78, 82, 85, 88

Allard, D., Brix A. & Chadœuf, J. (2001) Testing local independence between two point processes. *Biometrics* **57**, 508–517. 9, 38

Ambroise, C., Dang, M. & Govaert, G. (1996) Clustering of spatial data by the EM algorithm. In : Proceeding of geoENV. 34, 54

Anselin, L. (1995) Local indicators of spatial association – LISA. *Geographical Analysis* **27**, 93–115. 8

Banfield, J.D. & Raftery, A.E. (1993) Model-based Gaussian and non-Gaussian clustering. *Biometrics* **49**, 803–821. 115, 126

Bar-Hen, A. & Picard, N. (2006) Simulation study of dissimilarity between point process. *Computational Statistics* **21**, 487–507. 8, 38

Baudry, J.-P., Raftery, A., Celeux, G., Lo, K. & Gottardo, R. (2010) Combining mixture components for clustering. *Journal of Computational and Graphical Statistics* **19** , 332–353. 127

Bellier E. (2007) Distributions et structures spatiales au sein de systèmes en patchs hiérarchiques dynamiques : mise en évidence des relations spatiales espèces-environnement à échelles multiples. Thèse de doctorat. INRA, unité Biostatique et Processus Spatiaux. Avignon, France. 4, 5

Besag, J. (1986) On the statistical analysis of dirty pictures. *Journal of the Royal Statistical Society* **48** (3), 259–302. 34

Biernacki, C., Celeux, G. & Govaert, G. (2000) Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE actions on Pattern Analysis and Machine Intelligence* **22**, 719–725. 26, 55, 62

Biernacki, C., Celeux, G. & Govaert, G. (2003) Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate Gaussian mixture models. *Computational Statistics and Data Analysis* **41**, 561–575. 24, 25

Biernacki, C., Celeux, G., Govaert G. & Langrognet, F. (2006) Model-based cluster analysis and discriminant analysis with the MIXMOD software. *Computational Statistics and Data Analysis* **51**, 587–600. 106

Booth, J.G. & Hobert, J.P. (1998) Standard errors of prediction in generalized linear mixed models. *Journal of the American Statistical Association* **93**, 262–272. 111

Booth, J.G. & Hobert, J.P. (1999) Maximizing Generalized Linear Mixed Model likelihoods with an automated Monte Carlo EM algorithm. *Journal of the Royal Statistical Society* B **61**, 265–285. 73, 74, 79, 88, 89, 90, 91, 97

Boreux, J.-J., Parent, E. & Bernier, J. (2010) *Pratique du calcul bayésien.* Springer-Verlag, Paris. 73

Breslow, N.E. & Clayton, D.G. (1993) Approximate inference in generalized linear mixed models. *Journal of the Americal Statistical Association* **88**, 9-25. 90

Celeux, G. & Diebolt, J. (1985) The SEM algorithm : a probabilistic teacher algorithm derived from the EM algorithm for the mixture problem. *Computational Statistics Quaterly* **2**, 73–82. 25, 73

Celeux, G. & Govaert, G. (1995) Parsimonious Gaussian models in cluster analysis. *Pattern Recognition* **28**, 781–793. 115, 126

Celeux, G., Hurn, M. & Robert, C.P. (2000) Computational and Inferential Difficulties with Mixture Posterior Distributions. *Journal of the American Statistical Association* **95** (451), 957–970. 27

Celeux, G., Martin, O. & Lavergne, C. (2005) Mixture of linear mixed models - Application to repeated data clustering. *Statistical Modelling* **5**, 243–267. 128

Chagneau, P., Mortier, F., Picard, N. & Bacro, J.-N. (2010) A hierarchical bayesian model for spatial prediction of multivariate non-Gaussian random fields. *Biometrics* doi :10.1111/j.1541-0420.2010.01415.x 74, 77, 79, 82, 88, 97

Chalmond, B. (1989) An iterative gibbsian technique for reconstruction of m-ary images. *Pattern Recognition* **22** (6), 747–761. 34

Chaubert, F., Mortier, F. & Saint-André, L. (2008) Multivariate dynamic model for ordinal outcomes. *Journal of Multivariate Analysis* **99**, 1717–1732. 77

Chave, J. (2004) Neutral theory and community ecology. *Ecology Letters* **7**, 241–253.

Chib, S. & Greenberg, E. (1998) Analysis of multivariate probit models. *Biometrika* **85**, 347–361. 77

Chib, S. & Winkelmann, R. (2001) Markov Chain Monte Carlo analysis of correlated count data. *Journal of Business & Economic Statistics* **19**, 428–435. 78

Chilès, J.-P. & Delfiner, P. (1999) *Geostatistics : Modeling Spatial Uncertainty.* John Wiley and Sons, New York. 8, 43

Chiu, S.N. & Liu, K.I. (2009) Generalized Cramér-von Mises goodness-of-fit tests for multivariate distributions. *Computational Statistics and Data Analysis* **53**, 3817–3834. 55, 75

Cristofoli, A. & Mahy, G. (2010) Restauration écologique : contexte, contraintes et indicateurs de suivi. *Biotechnology, Agronomy, Society and Environment (BASE)* **14** (1), 203–211. 10

Dale, M.R.T., Dixon, P., Fortin M.-J., Legendre, P., Myers, D.E. & Rosenberg M.S. (2002) Conceptual and mathematical relationships among methods for spatial analysis. *Ecography* **25**, 558–577. 7

Dasgupta, A. (2008) *Asymptotic Theory of Statistics and Probability*. Springer-Verlag. 55

Dean, N. & Nugent, R. (in prep.) Mixture model component trees : visualizing the hierarchical structure of complex groups. 55

Dempster, A.P., Laird, N.M. & Rubin, D.B. (1977) Maximum likelihood for incomplete data via the EM algorithm. *Journal of the Royal Statistical Society* **B**, **39**, 1–38. 19, 23, 42, 88

Desassis N. (2007) Modélisation et inférence du lien entre deux variables à partir d'observations géoréférencées et hétérotopes. Thèse de doctorat. INRA, unité Biostatique et Processus Spatiaux. Avignon, France. 79

Diggle, P.J. (2003) *Statistical Analysis of Spatial Point Patterns, 2nd edition*. Edward Arnold, London. 6, 7, 8, 38

Diggle, P.J., Tawn, J.A. & Moyeed, R.A. (1998) Model-based geostatistics (with discussion) *Applied Statistics* **47**, 299–350. 79

Drouet-Mari D. & Kotz, S. *Correlation and Dependence*. Imperial College Press, London. 75, 76

Dublin, H.T., Sinclair, A.R.E. & McGlade, J. (1990) Elephants and fire as causes of multiple stable states in the Serengeti-mara woodlands. *Journal of Animal Ecology* **59**, 1147–64. 11, 38

Dunson, D.B. (2000) Bayesian latent variable model for clustered mixed outcomes. *Journal of Animal Ecology* **59**, 1147–64. 74, 78

Durett, R. & Levin, S.A. (1994) Stochastic spatial models : a user's guide to ecological applications. *Philosophical Transactions of the Royal Society London B* **343**, 329–350. 9

Estes, J.A. & Duggins, D.O. (1995) Sea otters and kelp forests in Alaska : Generality and variation in a community ecological paradigm. *Ecological Monographs* **65**, 75–100. 11, 38

Evans, M., & Swartz, T.B. (1996) Bayesian integration using multivariate Student importance sampling. *Computing Science and Statistics Series* **27**, 456-461. 89

Everitt, B.S. (1984) *An Introduction to latent variable models*. Chapman and Hall, London. 56

Everitt, B.S. (1988) A finite mixture model for the clustering of mixed mode data. *Statistics & Probability Letters* **6**, 305–309. 77, 127, 128

Fauchald, P., Erikstad, K. & Skarsfjord, H. (2000) Scale-dependent predator-prey interactions : the hierarchical spatial distribution of seabirds and prey. *Ecology* **81** (3), 773–783. 4

Fraley, C. & Raftery, A.E. (2002) Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association* **97**, 611–631. 20, 25, 26, 27, 32, 35, 39, 41, 42, 55, 97, 105, 106, 126, 140

Fraley, C. & Raftery, A.E. (2006) MCLUST version 3 for R : normal mixture modeling and model-based clustering. Technical Report No. 504, Department of Statistics, University of Washington. 32, 33, 43, 49, 59, 62, 63, 64, 106, 114

Fraley, C. & Raftery, A.E. (2007) Bayesian regularization for normal mixture estimation and model-based clustering. *Journal of Classification* **24** (2), 155–181. 24, 26

Garretta V., Monestiez P. & Ver Hoef J.M. (2010) Spatial modelling and prediction on river networks : up model, down model or hybrid ? *Environmetrics.* **21** (5) 439–456. 138

Gause, G.F. (1935) *The struggle for existence.* Williams & Wilkins, Baltimore. 4

Georgescu, V., Soubeyrand, S., Kretzschmar, A. & Laine, A.-L. (2009) Exploring spatial and multitype species assemblages. *Biometrical Journal* **51** (6), 979–995. 32

Georgescu, V., Desassis, N., Soubeyrand, S., Kretzschmar, A. & Senoussi, R. (submitted) A hierarchical model for multivariate data of different types and maximum likelihood estimation.

Genest, C. & Favre, A.-C. (2007) Everything you always wanted to know about copula modeling but were afraid to ask. *Journal of Hydrologic Engineering* **12** (4), 347–368. 76

Gottwald, T.R., Gibson, G.J., Garnsey, S.M. and Irey, M. (1999) Examination of the effect of aphid vector population composition on the spatial dynamics of Citrus Tristeza Virus spread by stochastic modeling. *The American Phytopathological Society* **89**(7), 603–608. 14

Guillot, G., Estoup, A., Mortier, F. & Cosson, J.F. (2005) A spatial statistical model for landscape genetics. *Genetics* **170**, 1261–1280. 54

Guzmán-Novoa, E., Eccles, L., Calvete, Y., Mcgowan, J., Kelly, P.G. & Correa-Benítez A. (2010) Varroa destructor is the main culprit for the death and reduced populations of overwintered honey bee (Apis mellifera) colonies in Ontario, Canada. *Apidologie.* 17, 95

Haccou, P., Jagers, P. & Vatutin, V.A. (2007) *Branching processes. Variation, Growth, and Extinction of Populations.* Cambridge Studies in Adaptive Dynamics. 9, 140, 141

Hanski, I. (1995) Multiple equilibria in metapopulation dynamics. *Nature* **377**, 618–21.

Hanski, I. (1999) *Metapopulation Ecology.* Oxford University Press. Oxford. 9

Hoff, P.D. (2007) Extending the rank likelihood for semiparametric copula estimation. *Annals of Applied Statistics* **1**, 265–283. 76

Hobert, J.P. (2001) Hierarchical models : a current computational perspective. In : *Statistics in the 21st Century* (eds. Raftery, A.E. Tanner, M.A. & Wells, M.T.), Monographs on Statistics & Applied Probability, Chapman and Hall, 368–377. 74

Hubisz, M., Falush, D., Stephens, M. & Pritchard, J.K. (2009) Inferring weak population structure with the assistance of sample group information. *Molecular Ecology Resources* **9**, 1322–1332. 139

Hui, C. (2009) On the scaling patterns of species spatial distribution and association. *Journal of Theoretical Biology* **261**, 481–487. 4, 10

Hutchinson, G. (1961) The paradox of the plankton. *American Naturalist* **95**, 137–147. 3

Illian, J., Penttinen, A., Stoyan, H. & Stoyan, D. (eds) (2008) *Statistical Analysis and Modelling of Spatial Point Patterns.* John Wiley and Sons, Chichester. 38

Jain, A.K., Murty, M.N. & Flynn, P.J. (1999) Data Clustering : A Review. *ACM Computing Surveys* **31** (3), 264–323. 19

Jordano, P., García, C., Godoy, J.A. & García-Castaño, J.L. (2007) Differential contribution of frugivores to complex seed dispersal patterns. *Proceedings of the National Academy of Science of the USA* **104**, 3278–3282. 45

Justel, A., Peña, D. & Zamar, R. (1997) A multivariate Kolmogorov-Smirnov test of goodness of fit. *Statistics and Probability Letters* **35**, 251–259. 55, 75

Karlis, D. & Meligkotsidou, L. (2006) Finite mixture of multivariate Poisson distributions with application. *Journal of Statistical Planning and Inference* **137**, 1942–1960. 31, 56, 65, 68, 106

Karlis, D. (2005) An EM algorithm for mixed Poisson and other discrete distributions. *ASTIN Bulletin* **35**, 3–24. 72, 73, 74

Karlis, D. & Arakelian, V. (2010) Clustering dependencies via mixture of copulas. *Présentation orale*, 17th annual summer session of the Working Group on Model-Based Clustering. 105

Karunanayake, C.P. & Laverty, W.H. (2006) Multivariate Poisson Markov dependent finite mixture Models for analysis of weed counts. *Model Assisted Statistics and Applications* **1**, 267–277. 9, 11, 31, 34, 56, 65

Kretzschmar, A., Soubeyrand, S. & Desassis, N. (2010) Aggregation patterns in hierarchy/proximity spaces. *Ecological Complexity* **7** (1), 21–31. 12

Krzanowski, W.J. (1993) The location model for mixtures of categorical and continuous variables. *Journal of Classification* **10**, 25–49. 128

Lai, X. & Yau, K.K.W. (2008) Long-term survivor model with bivariate random effects : Applications to bone marrow transplant and carcinoma study data. *Statistics in Medicine* **27**, 5692–5708. 97

Laine, A.-L. (2004) Resistance variation within and among host populations in a plant-pathogen metapopulation : implications for regional pathogen dynamics. *Journal of Ecology* **92**, 990–1000. 14, 39, 50

Laine, A.-L. & Hanski, I. (2006) Large-scale spatial dynamics of a specialist plant pathogen in a fragmented landscape. *Journal of Ecology* **94**, 217–226. 14, 39, 50

Lawrence, C.J. & Krzanowski, W.J. (1996) Mixture separation for mixed-mode data. *Statistics and Computing* **6**, 85–92.

Lehman, C.L. & Tilman, D. (1997) Competition in spatial habitats. In : *Spatial ecology. The role of space in population dynamics and interspecific interactions.* (eds. Tilman, D. & Kareiva, P.), Princeton University Press, Princeton, New Jersey. 3

le Roux, P.C. & McGeoch, M.A. (2008) Spatial variation in plant interactions across a severity gradient in the sub-Antarctic. *Oecologia* **155**, 831–844. 11, 38

MacQueen, J. (1967) Some methods for classification and analysis of multivariate observations. In : *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* (eds. Le Cam, L.M. & Neyman, J.), University of California Press, CA, 281–297. 19

Manly, B.F.J. (1997) *Randomization, Bootstrap and Monte Carlo Methods in Biology, 2nd edition.* Chapman and Hall, London. 43

Martinez, M.J., Lavergne, C. & Trottier, C. (2009) A mixture model-based approach to the clustering of exponential repeated data. *Journal of Multivariate Analysis* **100**, 1938–1951. 128, 138

McCulloch, C.E. (1997) Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American Statistical Association* **92**, 162–170. 73, 74, 128

McCulloch, C.E. & Searle, S.R. (2001) *General, Linear and Mixed Models.* Wiley, New York. 97

McLachlan, G.J. & Peel, D. (2000) *Finite Mixture Models.* John Wiley and Sons, New York. 19, 20, 23, 24, 25, 26, 27, 34, 41, 55, 74, 105, 106, 115, 127, 129

McLachlan, G.J. & Ng, S.K. (2000) A comparison of some information criteria for the number of components in a mixture model. Technical Report. Department of Mathematics, University of Queensland, Brisbane. 115

Messier, F. (March 1994) Ungulate population models with predation : a case study with the North American moose. *Ecology* **75**, 478–488. 11, 38

Muller-Landau, H.C., Wright, S.J., Calderón, O., Condit, R. & Hubbell, S.P. (2008) Interspecific variation in primary seed dispersal in a tropical forest. *Journal of Ecology* **96**, 653–667. 45

Munkin, M.K. & Trivedi, P.K. (1999) Simulated maximum likelihood estimation of multivariate mixed-Poisson regression models, with application. *Econometrics Journal* **2**, 29–48. 72, 73

Nikoloulopoulos, A.K. & Karlis, D. (2008) Fitting copulas to bivariate earthquake data : the seismic gap hypothesis revisited. *Environmetrics* **19**, 251–269. 76

Olkin, I. & Tate, R.F. (1961) Multivariate correlation models with mixed discrete and continuous variables. *Annals of Mathematical Statistics* **22**, 92–96. 77, 128

Osaragi, T. (2002) Classification methods for spatial data representation. Working paper. CASA Working Papers (40). Centre for Advanced Spatial Analysis (UCL), London, UK. 54

Pacala, S.W., Canham, C.D., Saponora, J., Silander, J.A., Kobe, R.K. & Ribbens, E. (1996) Forest models defined by field measurements : estimation, error analysis and dynamics. *Ecological Monographs* **66**, 1–43. 10

Pearson, K. (1894) Contributions to the theory of mathematical evolution. *Philosophical Transactions of the Royal Society of London A* **185**, 71–110. 19

Perry, J. N. & Dixon, P.M. (2002) A new method to measure spatial association for ecological count data. *Ecoscience* **9**, 133–141. 9, 10, 31, 38, 133

Perry, J. N., Liebhold, A.M., Rosenberg, M.S., Dungan, J. Miriti, M., Jakomulska, A. & Citron-Pousty, S. (2002) Illustrations and guidelines for selecting statistical methods for quantifying spatial pattern in ecological data. *Ecography* **25**, 578–600. 5

Perry, J. N., Winder, L., Holland J.M. & Alston, R.D. (1999) Red-blue plots for detecting clusters in count data. *Ecology letters* **2**, 106–113. 8, 9

Peterson, C.H. (1984) Does a rigorous criterion for environmental identity preclude the existence of multiple stable points ? *The American Naturalist* **124**, 127–133. 11, 38

Petraitis, P.S. & Latham, R.E. (1999) The importance of scale in testing the origins of alternative community states. *Ecology* **80**, 429–442. 11, 38

Pritchard, J.K., Stephens, M. & Donnelly, P. (2000) Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–959. 139

R Development Core Team (2009) R : A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http ://www.R-project.org. 115

Révész, P. (1994) *Random walks of infinitely many particles.* World Scientific. 9, 141

Richardson, S. & Green P.J. (1997) On Bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society B* **59**, 731–792. 26

Rose, G.A. & Legget W. (1981) The importance of scale to predator-prey spatial correlations : an example of atlantic fishes. *Ecology* **71**, 33–43. 4

Sammel, M.D., Ryan, L.M. & Legler, J.M. (1997) Latent variable models for mixed discrete and continuous outcomes. *Journal of the Royal Statistical Society B* **59** (3), 667–678. 78

Schwarz, G. (1978) Estimating the dimension of a model. *The Annals of Statistics* **6**, 461–464. 26

Seabloom, E.W. & Richards, S.A. (2003) Multiple stable equilibria in grasslands mediated by herbivore population dynamics and foraging behaviour. *Ecology* **84**, 2891–2904. 11, 38

Soubeyrand, S., Held, L., Höhle, M. & Sache, I. (2008) Modelling the spread in space and time of an airborne plant disease. *Journal of the Royal Statistical Society* **57**, 253–272. 47

Soubeyrand, S., Laine, A.-L., Hanski, I. & Penttinen, A. (2009) Spatio-temporal structure of host-pathogen interactions in a metapopulation. *The American Naturalist* **174**, 308–320. 14, 50

Tanner, A.M. (1991) *Tools for Statistical Inference : Observed Data and Data Augmentation Methods.* Springer, New York. 23, 91

Tanner, M.A. (1996) *Tools for statistical inference : methods for the exploration of posterior distributions and likelihood functions, 3rd edition.* Springer-Verlag. 42, 73

Teicher, H. (1963) Identifiability of finite mixtures, *The Annals of Mathematical Statistics* **34**, 1265–1269. 41

Tilman, D. & Kareiva, P. (eds.) (1997) *Spatial ecology. The role of space in population dynamics and interspecific interactions.* Princeton University Press, Princeton, New Jersey. 3, 4

Tunaru, R. (2002) Hierarchical bayesian models for multiple count data. *Austrian Journal of Statistics.* **31**, 221–229. 85, 88

Ueda, N. & Nakano, R. (1998) Deterministic annealing EM algorithm. *Neural Networks.* **11**, 271–282. 25

Walker, B.H., Ludwig, D., Holling, C.S. & Peterman, R.M. (1981) Stability of semi-arid savanna grazing systems. *The Journal of Ecology* **69**, 473–498. 11, 38

Wang, K.,Yau, K.K.W., Lee, A.H. & McLachlan, G.J.(2007) Two-component Poisson mixture regression modelling of count data with bivariate random effects. *Mathematical and Computer Modelling* **46**, 1468–1476. 97

Wei, G.C.G. & Tanner, M.A. (1990) A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistical Association* **85**, 699–704. 24, 73, 88

Willse, A. & Boik, R.J. (1999) Identifiable finite mixtures of location models for clustering mixed-mode data. *Statistics and Computing* **9**, 111–121. 128

Yokomi, R.K., Lastra, R., Stoetzel, M.B., Damsteegt, V.D., Lee, R.F., Garnsey, S.M., Gottwald, T.R., Rocha-Peña, M.A. and Nibblet, C.L. (1994) Establishment of the brown citrus aphid (Homoptera : Aphididae) in Central America and the Caribbean Basin. *Journal of Economic Entomology* **88**, 1078–1085. 14

Model-based clustering for multivariate and mixed-mode data : Application to multi-species spatial ecological data.

Summary

In population ecology, species spatial patterns are studied in order to infer the existence of underlying processes, such as interactions within and between species, and species response to environmental heterogeneity. We propose to analyze spatial multi-species data by defining species abundance assemblages. Species assemblages are one of the signatures of the local spatial interactions between species and with their environment. Species assemblages are defined here by a non spatial classification of the multivariate observations of species abundances. Model-based clustering procedures using mixture models were chosen in order to have an estimation of the classification uncertainty and to model an assemblage by a multivariate probability distribution. We propose :

1. an exploratory tool for the study of spatial multivariate observations of species abundances, which defines species assemblages by a model-based clustering procedure, and then maps and analyzes the spatial structure of the assemblages. Common distributions, such as the multivariate Gaussian, are used to model the assemblages.
2. a hierarchical model for abundance assemblages which cannot be modeled with common distributions. This model can be easily adapted to mixed mode data, which are frequent in ecology.
3. a clustering procedure for mixed-mode data based on mixtures of hierarchical models,

Two ecological case-studies guided and illustrated this work : the small-scale study of the assemblages of two aphid species on leaves of *Citrus* trees, and the large-scale study of the assemblages of a host plant, *Plantago lanceolata*, and its pathogen, the powdery mildew, on the Åland islands in south-west Finland.

Field : applied mathematics

Key words : Coexistence, finite mixture models, hierarchical model, latent gaussian model, model-based clustering, Monte Carlo Expectation Maximization (MCEM) algorithm, mixed mode data, multivariate data, spatial data, species assemblages.

Résumé

En écologie des populations, les distributions spatiales d'espèces sont étudiées afin d'inférer l'existence de processus sous-jacents, tels que les interactions intra- et interspécifiques et les réponses des espèces à l'hétérogénéité de l'environnement. Nous proposons d'analyser les données spatiales multi-spécifiques sous l'angle des assemblages d'espèces, que nous considérons en termes d'abondances absolues et non de diversité des espèces. Les assemblages d'espèces sont une des signatures des interactions spatiales locales des espèces entre elles et avec leur environnement. L'étude des assemblages d'espèces peut permettre de détecter plusieurs types d'équilibres spatialisés et de les associer à l'effet de variables environnementales.

Les assemblages d'espèces sont définis ici par classification non spatiale des observations multivariées d'abondances d'espèces. Les méthodes de classification basées sur les modèles de mélange ont été choisies afin d'avoir une mesure de l'incertitude de la classification et de modéliser un assemblage par une loi de probabilité multivariée.

Dans ce cadre, nous proposons :

1. une méthode d'analyse exploratoire de données spatiales multivariées d'abondances d'espèces, qui permet de détecter des assemblages d'espèces par classification, de les cartographier et d'analyser leur structure spatiale. Des lois usuelles, telle que la Gaussienne multivariée, sont utilisées pour modéliser les assemblages.
2. un modèle hiérarchique pour les assemblages d'abondances lorsque les lois usuelles ne suffisent pas. Ce modèle peut facilement s'adapter à des données contenant des variables de types différents, qui sont fréquemment rencontrées en écologie,
3. une méthode de classification de données contenant des variables de types différents basée sur des mélanges de lois à structure hiérarchique (définies en 2.).

Deux applications en écologie ont guidé et illustré ce travail : l'étude à petite échelle des assemblages de deux espèces de pucerons sur des feuilles de clémentinier et l'étude à large échelle des assemblages d'une plante hôte, le plantain lancéolé, et de son pathogène, l'oïdium, sur les îles Åland en Finlande.

Discipline : mathématiques appliquées

Mots-clés : Assemblage d'espèces, classification basée sur des modèles de mélange, coexistence, données mixtes, données multivariées spatiales, modèle gaussien latent, modèle hiérarchique, Monte Carlo EM.