Modélisation et inférence du lien entre deux variables à partir d'observations géoréférencées et hétérotopes

Nicolas Desassis

UNIVERSITÉ MONTPELLIER II
SCIENCES ET TECHNIQUES DU LANGUEDOC

THÈSE

pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ MONTPELLIER II

Discipline : Mathématiques Appliquées
Formation doctorale : Biostatistique
Ecole doctorale : Information, Structures, Systèmes

présentée et soutenue publiquement

par

Nicolas DESASSIS

le 23 octobre 2007

Titre :

MODÉLISATION ET INFÉRENCE DU LIEN ENTRE DEUX VARIABLES À PARTIR D'OBSERVATIONS GÉORÉFÉRENCÉES ET HÉTÉROTOPES

JURY

Gilles Ducharme, Université Montpellier II	Président
Olivier Perrin, Université de Toulouse	Rapporteur
Carlo Gaëtan, Université de Venise	Rapporteur
Liliane Bel, AgroParisTech	Examinatrice
Michel Goulard, INRA Toulouse	Examinateur
Jean-Noël Bacro, Université Montpellier II	Directeur de thèse
Pascal Monestiez, INRA Avignon	Co-directeur de thèse

Remerciements

Je voudrais tout d'abord exprimer ma profonde gratitude à Pascal Monestiez qui a encadré cette thèse au quotidien. Il a su me transmettre une partie de sa riche expérience, tout en me laissant apprendre par moi même. J'espère avoir "hérité" de son sceptissisme qui est pour moi le garant d'une recherche de qualité. Je suis également conscient de la patience dont il a su faire preuve à mon égard et je lui en suis très reconnaissant.

Je tiens également à remercier Jean-Noël Bacro, le directeur officiel de cette thèse qui a toujours été très présent (malgré la distance géographique) notamment dans les moments les plus difficiles. J'espère que cette thèse n'est que le début de collaborations futures.

Je tiens à exprimer ma profonde gratitude aux membres du Jury, tout d'abord aux rapporteurs de cette thèse Carlo Gaétan et Olivier Perrin, mais également à Liliane Bel et Michel Goulard qui ont examiné ce travail avec beaucoup d'attention, ce dont je leur suis très reconnaissant. Enfin j'ai été très heureux que Gilles Ducharme préside ce jury. Il a été mon premier professeur de statistiques et est très largement responsable de mon choix d'orientation en troisième cycle de biostatistique. J'ai également apprécié enseigner sous sa direction en tant qu'ATER à l'université Montpellier 2. J'ai retrouvé dans les supports annexes à ses cours, les qualités pédagogiques que j'avais tant appréciées quelques années plus tôt.

Je voudrais remercier les membres (extérieurs) de mon comité de thèse : Avner Bar-Hen, Sylvie Huet, Christian Lantuejoul et Eric Parent. Leurs nombreux conseils m'ont été très utiles pour avancer dans mon cheminement.

Je tiens à remercier Denis Allard qui a beaucoup contribué lors du DEA, à ma venue dans l'unité qu'il dirige aujourd'hui. Nos nombreuses discussions furent pour moi un réel plaisir. Je lui suis également reconnaissant de m'avoir permis de goûter aux joies de l'enseignement en me proposant dès mon arrivée, d'assurer des TD de statistique.

Je veux également exprimer ici ma profonde gratitude et mon immense reconnaissance (je sais qu'il appréciera) à Joël Chadœuf qui, notamment par ses précieux conseils, a grandement contribué à l'aboutissement de cette thèse.

Ce fut pour moi un grand bonheur de vivre ces 4 années au sein de l'équipe de l'unité Biostatistique et Processus Spatiaux (la "Biométrie") de l'INRA d'Avignon. Parmi les permanents de cette unité, je tiens tout particulièrement à remercier Malika Nassif pour tout ce qu'elle y apporte au quotidien. Je remercie également Franck Ariès qui m'a initié au calcul numérique et à toujours été disponible quand Maple faisait des siennes et Etienne Klein avec qui j'ai pris beaucoup de plaisir à parler d'écologie. J'ai également partagé d'excellents moments en marge du laboratoire avec Lionel Roques et Rachid Senoussi. J'espère sincèrement que ces contacts amicaux donneront lieu demain, à des collaborations.

Je tiens également à adresser mon amitié aux futurs docteurs Julien Fayard et Florence Carpentier qui m'ont beaucoup apporté lors de nos nombreux échanges.

Je tiens à exprimer toute ma sympathie à Samuel Soubeyrand avec qui j'ai tant partagé : le bureau déjà, par un heureux hasard, mais également les thématiques scientifiques, les visions de la recherche et je l'espère, une solide amitié.

Je tiens également à remercier Remy Beaudouin d'avoir été si agréable à héberger.

Certaines rencontres vous marquent plus encore que les autres, et il est désagréable d'essayer d'imaginer ce qu'aurait été son existence sans elles. Ces rencontres sont très rares, et j'ai eu l'immense bonheur d'en vivre une pendant ma thèse. Je remercie donc André Kretzschmar.

Pour terminer, je tiens ici à saluer mes proches. Mes amis, Manon, Jean-Philippe, Paul et Ibnou ; j'ai une pensée toute particulière pour Jacques Wadier, qui a largement contribué, par son envie irréfrénée de partager ses immenses connaissances, à m'ouvrir certaines barrières "symboliques" du monde universitaire.

Je remercie bien évidemment ma famille pour son soutien sans faille et son affection, Marie, Hélène, Thomas et bien sûr Guillaume ; mon père pour m'avoir permis de ne jamais trop m'égarer, et ma mère...pour tout le reste.

Enfin, mes derniers mots iront à Carole que j'aime tant.

Nicolas Desassis
21 décembre 2007

Table des matières abrégée

Table des matières

1

Introduction

Dans de nombreux domaines des sciences de la nature, on a très souvent pris l'habitude de collecter des données géoréférencées, c'est-à-dire d'associer aux variables mesurées, les coordonnées géographiques des sites d'observation. L'inférence du lien entre deux (ou plusieurs) variables dans ce contexte d'observations spatialisées requiert d'adapter les modèles classiques (comme le modèle de régression) aux particularités du contexte spatial. En effet, dans ce cadre de travail, certaines hypothèses des modèles classiques deviennent irréalistes - par exemple l'indépendance des résidus dans un modèle linéaire.

La caractéristique la plus commune aux données géoréférencées est la présence de dépendances spatiales[1]. Dans une optique de prédiction d'une variable en des sites où celle-ci n'est pas observée à partir d'un nombre fini d'observations en d'autres sites, la connaissance de ces dépendances spatiales est un préalable nécessaire. La géostatistique, initialement développée pour prédire les réserves minières à partir de forages, fournit de nombreux outils pour mesurer et modéliser la structuration spatiale éventuelle des observations. Une fois cette structure spatiale analysée, elle est utilisée dans une procédure d'interpolation connue sous le nom de krigeage. On peut quantifier la variance du prédicteur ainsi obtenu afin d'obtenir une mesure de la qualité de l'interpolation. Cette variance est d'autant plus faible que les corrélations spatiales de la variable à interpoler sont fortes (et que les sites d'observations sont proches du site de prédiction). En d'autres termes, les corrélations spatiales sont profitables quand l'objectif est l'interpolation.

Lorsque l'objectif de l'analyse est d'inférer le lien entre plusieurs variables spatialisées au sens classique du terme (par exemple en utilisant un modèle linéaire ou par le calcul du coefficient de corrélation), les dépendances spatiales sont plutôt considérées comme une nuisance par le statisticien. En effet, une procédure classique qui ignore ces dépendances spatiales conduit généralement à une sous-estimation des variances des estimateurs puisque la redondance de l'information apportée par deux données très proches est ignorée. La prise en compte dans la modélisation des dépendances spatiales peut se faire en intégrant aux modèles classiques, des éléments comme les fonctions de covariances spatiales de la géostatistique ; on devra alors en estimer les paramètres, simultanément avec les paramètres de liens entre variables. Moins parcimonieux que dans le contexte d'échantillons i.i.d, ces modèles sont techniquement plus difficiles à estimer, la vraisemblance des observations

[1] Par dépendances spatiales, nous entendrons toujours la tendance qu'ont deux observations d'une même variable à être d'autant plus similaires que leurs sites d'observation sont proches géographiquement. Des types de dépendances spatiales plus riches peuvent être envisagés. Dans le cas des séries temporelles, un effet saisonnier pourrait induire une moins grande dissimilarité entre des observations mesurées à la même date mais à un an d'intervalle que celle que l'on observerait entre deux dates séparées de seulement 6 mois. De tels phénomènes peuvent également être observés dans le contexte spatial.

ne s'écrivant plus comme un produit de vraisemblance. De même, la convergence asymptotique de l'estimateur du vecteur de paramètres devient plus difficile à établir que dans le cas i.i.d car contrairement à ce dernier cas, il ne suffit pas que le nombre de données tende vers l'infini. La manière dont le nombre de données croît doit être précisée.

Mardia et Marshall (1984) posent les premières pierres du pont entre la géostatistique et les statistiques classiques. Ils introduisent le modèle linéaire à résidus spatialement corrélés, proposent l'utilisation de l'algorithme numérique des scores de Fisher (*Fisher scoring*) pour maximiser la vraisemblance et donnent des conditions suffisantes à la normalité asymptotique des estimateurs. Les idées apportées dans cet article ont considérablement enrichi la géostatistique mais elles ont surtout le mérite de l'avoir plongée dans le champ de la statistique classique. Dans la même optique, Cressie et Lahiri (1996) adaptent ces idées au cas du REML (*Restricted Maximum Likelihood*) et précisent certains éléments de l'asymptotique qui avaient été évacués par Mardia et Marshall, mais il faut attendre l'article de Diggle *et al.* (1998) pour une nouvelle avancée significative dans le rapprochement entre les deux champs. En effet si Mardia et Marshall (1984) avaient ouvert une porte, leurs travaux se cantonnaient au contexte gaussien, seul cas où la vraisemblance des observations est disponible pour des données spatiales. L'apport majeur de Diggle *et al* (1998) est d'avoir construit un modèle hiérarchique dans lequel un champ gaussien est modélisé au niveau d'une couche latente conditionnellement à laquelle les observations sont supposées indépendantes spatialement. De cette manière, on peut caractériser entièrement la distribution spatiale de données pour lesquelles l'hypothèse gaussienne est inappropriée, comme par exemple des observations binaires ou de comptage. Cette nouvelle classe de modèles abondamment utilisée par la suite dans la littérature est nommée Modèles Linéaires Généralisés Mixtes (GLMM) à effets aléatoires spatialement corrélés.

Une conséquence collatérale de l'article de Diggle *et al* (1998) est d'avoir popularisé l'utilisation des méthodes bayésiennes pour estimer les modèles spatiaux. En effet, une maximisation directe de la vraisemblance n'est plus possible dans le cas des GLMM, et *a fortiori* quand on intègre à ces modèles des corrélations spatiales (la vraisemblance n'est pas disponible sous une forme analytique calculable et fait intervenir des intégrales de grandes dimensions), et l'utilisation d'algorithmes stochastiques semble être l'une des meilleures stratégies pour effectuer l'inférence de tels modèles. A cet égard, les algorithmes MCMC (*Monte Carlo Markov Chain*) dans un contexte bayésien sont des outils particulièrement pratiques qui permettent d'envisager des modèles de plus en plus riches, tant sur la complexité des liaisons entre variables, que sur les niveaux de hiérarchie. Notons que dans le cadre fréquentiste, un algorithme Monte Carlo EM (MCEM) a été proposé par Zhang (2002) pour maximiser la vraisemblance dans les GLMM à effets aléatoires corrélés.

Dans ce travail, nous nous sommes intéressés au cas où les variables dont on cherche à modéliser la relation, ne sont pas observées aux mêmes sites. On parle d'hétérotopie des données ou de jeux de données hétérotopes. En statistique classique, établir une liaison statistique entre plusieurs variables requiert de mesurer celles-ci sur les mêmes individus. Des données peuvent être manquantes mais il paraîtrait incongru de chercher à établir le lien entre deux variables en observant l'une sur un groupe d'individus, et l'autre sur un groupe entièrement distinct du premier. Lorsque l'on travaille avec des données spatialisées, les individus statistiques sont les sites d'observation. Dans ce cas, l'existence de corrélations spatiales permet d'envisager de caractériser le lien entre des variables à partir de donnnées hétérotopes.

Savoir travailler en condition d'hétérotopie permet donc d'ouvrir un important champ d'hypothèses à tester à partir des jeux de données existants, puisqu'à partir de deux jeux de données géoréférencés différents, on peut toujours construire un jeu de données hétérotopes en les compilant. La qualité de l'échantillonnage spatial au regard des liaisons à estimer est bien sûr très variable.

Estimer le lien entre deux variables, chacune observée sur deux zones distantes paraît par exemple absurde. De plus, comme dans le cas du krigeage, les structures spatiales des variables en présence doivent être suffisamment fortes pour "compenser" le fait que les sites d'observations des deux variables ne coïncident pas.

Dans la pratique, le problème de l'hétérotopie est souvent évacué par des méthodes *ad hoc*. On peut par exemple affecter en chaque site où une seule variable est observée, les valeurs des autres variables observées aux sites les plus proches du site considéré (méthode du plus proche voisin). Pour être plus précis, on peut affecter les valeurs interpolées (par krigeage par exemple) en utilisant toutes les observations au lieu du plus proche voisin. Une fois le problème de l'hétérotopie ainsi résolu, le jeu de donnée est analysé classiquement, en occultant les traitements préliminaires. La principale limite de ces approches est qu'elles ne permettent pas de prendre en compte l'incertitude due à l'affectation et de traiter les données modifiées comme si elles ne l'avaient pas été. Tous les individus auront par exemple le même poids dans l'analyse alors que certains sont manifestement plus fiables que d'autres (ceux dont le plus proche voisin est très proche par exemple). Dans le cas du krigeage, nous avons déjà évoqué le fait que l'incertitude de l'interpolation est quantifiable dans une certaine mesure (par la variance de krigeage) ; il semble donc possible de la prendre en compte dans l'analyse. Pour traiter ce problème, il faut considérer toutes les variables en présence comme aléatoires et modéliser la structure spatiale de chacune ainsi que les structures spatiales croisées. La géostatistique multivariée (Wackernagel (2003)) fournit des outils qui permettent de modéliser ces relations, soit dans un but de prédiction spatiale d'une variable d'intérêt en utilisant l'information apportée par des variables auxiliaires mesurées aux mêmes sites ou non (cokrigeage), soit pour effectuer une analyse factorielle krigeante (AFK), c'est à dire explorer les relations entre différentes variables en fonction des échelles spatiales. De même que l'ACP, l'AFK est uniquement un outil exploratoire et, en dehors de tout contexte distributionnel, la significativité des liaisons ainsi mises en évidence ne peut être avérée. Dans un contexte distributionnel, Banerjee et Gelfand (2002) utilisent un des modèles de la géostatistique multivariée, plongé dans un contexte gaussien. Il en resulte un modèle de régression spatiale que l'on peut estimer à partir de données hétérotopes. Banerjee et Gelfand proposent un algorithme MCMC et l'utilisent sur un jeu de données composé de deux variables partiellement hétérotopes, c'est-à-dire contenant des sites d'observation communs aux deux variables. Dans cet article, Banerjee et Gelfand proposent également une extension de leur modèle pour travailler avec une variable d'intérêt binaire, mais leur choix de modélisation est particulier (voir chapitre 3) et ne permet pas de travailler avec des variables de comptages (de type binômial ou de Poisson).

Dans le premier chapitre de la thèse, nous faisons un rappel des bases théoriques de la géostatistique univariée et multivariée et nous présentons les principaux outils nécessaires à la compréhension des chapitres suivants. Dans le second chapitre, nous présentons les modèles que l'on cherchera par la suite à inférer. Ces modèles se classent dans deux catégories : les modèles gaussiens et les modèles hiérarchiques. Nous montrons comment, par conditionnement, on peut facilement assembler différents éléments de modélisation (variables observées, champs aléatoires latents, hypothèses distributionnelles et de structures spatiales) pour traiter différents types de variables (gaussiennes ou non) et différents types de relations spatiales entre variables. Dans le chapitre 3, nous présentons les méthodes d'estimation pour les différents modèles. Pour les modèles de type gaussiens, nous proposons un algorithme hybride entre le Fisher scoring et l'algorithme de Newton Raphson et nous détaillons les choix de paramétrisation qui permettent efficacité et robustesse. Pour les modèles hiérarchiques, nous adaptons les récentes avancées des algorithmes MCEM pour les modèles GLMM aux différents modèles hiérarchiques spatiaux présentés dans le chapitre 2.

Dans le chapitre 4, nous rappelons les formalismes utilisés pour les développements asymptotiques dans le contexte de données spatiales. Nous précisons certains éléments des démonstrations disponibles : nous généralisons par exemple à toute dimension, un résultat asympatotique sur les champs aléatoires gussien univariés obtenu en dimension une et nous montrons une propriété qui avait été postulée par Mardia et Marshall (1984) et Cressie et Lahiri (1996)). Nous appliquons ensuite ces résultats généraux pour montrer comment l'estimateur du coefficient de corrélation se comporte dans un modèle gaussien pour deux variables, et ce en fonction de la géométrie de l'échantillonnage. Nous explorons à l'aide de simulations, le comportement à taille finie de cet estimateur. Dans le chapitre 5, nous nous intéressons à l'alternative bayésienne. Après quelques rappels généraux sur les avantages comparés des deux paradigmes, nous montrons comment construire des algorithmes MCMC pour nos modèles. Enfin, dans le chapitre 6, nous illustrons nos résultats sur des cas réels.

2

La géostatistique

L'objectif de ce chapitre est de passer en revue les fondements théoriques de la géostatistique et les outils développés dans ce cadre. Notre souci n'était pas d'être exhaustif et de se substituer à des ouvrages de référence (Cressie, 1993; Chilès et Delfiner, 1999; Wackernagel, 2003) mais simplement de fournir au lecteur le contexte théorique dans lequel se situe notre travail ainsi que les méthodes sur lesquelles nos travaux se sont appuyé.

Essentiellement développée par Matheron (1962, 1963) et les chercheurs du Centre de Géostatistique de Fontainebleau (Ecole Nationale Supérieure des Mines de Paris), la géostatistique avait pour objectif initial, la prédiction de la quantité de minerai en un site s_0 ou l'espérance de cette quantité sur un domaine, à partir de forages effectués sur un ensemble de sites $\{s_1, \ldots, s_n\}$, d'où le préfixe "géo", relatif aux sciences de la terre[1].

Au cours des 30 dernières années, la géostatistique s'est progressivement étendue du domaine minier et pétrolier à des champs d'application variés comme les ressources forestières, l'halieutique, le climat, l'écologie ou encore l'agronomie. Sous le terme de géostatistique, on retrouve un ensemble de méthodes permettant notamment de caractériser la structure spatiale d'une variable mesurée, de prédire la valeur de cette variable en un site où celle-ci n'est pas observée, et éventuellement de produire une carte. Son extension au cadre multivariable permet d'explorer les relations, parfois complexes, entre variables spatialisées.

Les méthodes de la géostatistique se basent sur un corpus d'hypothèses qui permettent de travailler avec la réalisation unique d'une variable échantillonnée spatialement.

2.1 Contexte théorique et hypothèses

On observe une variable Z sur un ensemble $\{s_1, \ldots, s_n\}$ de sites d'un domaine $\mathcal{D}$ de $\mathbb{R}^d$ où d est un entier positif.

On note $\mathbf{z} = (z(s_1), \ldots, z(s_n))'$ le vecteur d'observations et $\mathbf{Z} = (Z(s_1), \ldots, Z(s_n))'$ le vecteur aléatoire qui est supposé avoir généré $\mathbf{z}$.

2.1.1 Champ aléatoire

Modéliser les observations spatiales au sens de la théorie classique des statistiques, c'est à dire comme étant les réalisations indépendantes d'une variable aléatoire unique, ne permet pas de prendre en compte la nature spatiale des observations. La distance entre les sites d'observation est par

[1] Le terme géostatistique fut inventé par Hart (1953) et désignait à l'origine l'ensemble des techniques statistiques qui prennent en compte l'emplacement géographique des sites d'observation.

exemple un élément essentiel des données spatialisées puisqu'une part importante des phénomènes spatiaux est structurée dans l'espace. Deux mesures d'une même variable auront par exemple souvent tendance à être d'autant plus similaires que leur site d'observation sont proches. Pour nommer ces caractéristiques, nous parlerons indistinctement de structures spatiales, de corrélations ou de dépendances spatiales ou encore d'autocorrélations. En géostatstique, l'analyse de ces structures permet d'effectuer la prédiction. Du point de vue des statistiques classiques, la prise en compte des structures éventuelles est essentielle, d'une part pour améliorer l'estimation des paramètres mais surtout pour prendre en compte la redondance de l'information apportée sur les paramètres par des données spatialement corrélées. Schabenberger et Gotway (2005) relatent plusieurs exemples très convaincants sur les erreurs que l'on peut commettre en ignorant la composante spatiale.

Pour pouvoir prendre en compte le fait que les données ont été observées "quelque part", on supposera en premier lieu que les observations sont la réalisation d'un champ aléatoire. Le formalisme du champ aléatoire fournit un cadre mathématique général qui permet de prendre en compte ces structures dans la modélisation (Christakos, 1992). Rappelons la définition d'un champ aléatoire :

On se place sur l'espace probabilisé $(\Omega, \mathcal{B}, P)$.

Définition 1 *Soit $(\Omega, \mathcal{B}, P)$ un espace probabilisé et $(R, \mathcal{R})$ un espace mesurable. Un champ aléatoire $Z(.,.)$ sur $\mathbb{R}^d$ est une famille de variables aléatoires $\{Z(s,.), s \in \mathbb{R}^d\}$ où chaque variable aléatoire est définie sur $(\Omega, \mathcal{B}, P)$ et prend ses valeurs dans $(R, \mathcal{R})$.*

Dans la pratique, on observe $\{Z(s_1, \omega), \ldots, Z(s_n, \omega)\}$, une réalisation de $Z(.,.)$ sur un ensemble de n sites $\{s_1, \ldots, s_n\}$ d'un domaine $\mathcal{D} \subset \mathbb{R}^d$.

Pour tout ensemble de taille finie $\{s_1, \ldots, s_n\}$, on note $F_{s_1,\ldots,s_n}$ la mesure de probabilité associée définie par :

$$F_{s_1,\ldots,s_n}(B) = \mathbb{P}[(Z(s_1,.), \ldots, Z(s_n,.)) \in B], \forall B \in \mathcal{R}^n. \tag{2.1}$$

L'existence du champ aléatoire $Z(.,.)$ est assurée si et seulement si pour tout $B = (B_1, \ldots, B_n) \in \mathcal{R}^n$, toutes les mesures de dimensions finies définies par (2.1) vérifient les deux conditions suivantes (Kolmogorov, 1933) :

– la symétrie : pour toute permutation σ de $[\![1, n]\!]$

$$F_{s_1,\ldots,s_n}(B_1 \times \cdots \times B_n) = F_{s_{\sigma(1)},\ldots,s_{\sigma(n)}}(B_{\sigma(1)} \times \cdots \times B_{\sigma(k)}).$$

– la consistance :

$$\forall k \geq 1, F_{s_1,\ldots,s_n,s_{n+1},\ldots,s_{n+k}}(B \times R^k) = F_{s_1,\ldots,s_n}(B).$$

On notera dans la plupart des cas $Z(.) \equiv Z(.,.)$ et $Z(s) \equiv Z(s,.)$.

Pour pallier au fait que la réalisation est unique, certaines hypothèses doivent être faites.

2.1.2 La stationnarité

La stationnarité est un terme générique qui désigne la permanence dans l'espace de certaines caractéristiques du champ aléatoire, ces caractéristiques pouvant être distributionnelles, ou de manière moins restrictive, des hypothèses sur les moments.

Nous donnons ici trois types de stationnarité de la plus restrictive à la plus générale.

Stationnarité stricte

Un champ aléatoire $Z(.)$ est dit strictement stationnaire si la distribution de tout vecteur issu de $Z(.)$ est invariante par translation, c'est à dire, si $\mathcal{D}$ est un domaine de $\mathbb{R}^d$:

$$\forall n \in \mathbb{N}, \forall (s_1, \ldots, s_n) \in \mathcal{D}^n, \forall \overrightarrow{\mathbf{h}} \in \mathbb{R}^d, F_{s_1,\ldots,s_n} = F_{s_1+\overrightarrow{\mathbf{h}},\ldots,s_n+\overrightarrow{\mathbf{h}}}.$$

C'est une hypothèse très forte et invérifiable. Les géostatisticiens préfèrent en général travailler avec des hypothèses moins restrictives comme la stationnarité des deux premiers moments ou la stationnarité des accroissements.

Un champ aléatoire strictement stationnaire est parfois dit "homogène".

Stationnarité d'ordre 2 (ou stationnarité faible)

L'hypothèse de stationnarité d'ordre 2 ne dit rien des distributions multivariées. C'est une hypothèse qui porte uniquement sur les 2 premiers moments de ces distributions :

L'espérance du champ aléatoire est finie et constante sur le domaine d'étude :

$$\forall s \in \mathcal{D}, E[Z(s)] = \mu < \infty$$

La covariance entre tout couple aléatoire $(Z(s_1), Z(s_2))$ est une fonction du vecteur de séparation entre les sites d'observation s_1 et s_2.

$$\forall (s_1, s_2) \in \mathcal{D}^2, \mathrm{Cov}(Z(s_1), Z(s_2)) = C(\overrightarrow{h}),$$

où $\overrightarrow{h} = \overrightarrow{s_1 s_2}$

La fonction C sera appelée fonction de covariance du champ aléatoire $Z(.)$.

La symètrie de l'opérateur covariance implique que C est une fonction paire. De plus elle est bornée par sa valeur à l'origine (par l'inégalite de Cauchy-Schwartz). Une fonction de covariance doit également être définie-positive ; cette propriété fondamentale sera développée dans la section 2.1.4.

Stationnarité intrinsèque

La stationnarité intrinsèque permet de travailler avec des champs aléatoires dont l'espérance et/ou la variance ne sont pas définies (ex : mouvement brownien).

On suppose seulement que l'espérance des différences de tout couple est nulle et que leur variance ne dépend que du vecteur de séparation :

$$\forall (s_1, s_2) \in \mathcal{D}^2, E[Z(s_1) - Z(s_2)] = 0$$

et

$$\mathrm{Var}[Z(s_1) - Z(s_2)] = 2\gamma(\overrightarrow{h}).$$

La fonction γ est appelé variogramme[2] du champ aléatoire.

Notons que sous l'hypothèse de stationnarité intrinsèque, $\mathrm{Var}(Z(s_i))$ peut ne pas exister. Pour que la variance d'une combinaison linéaire

[2] On trouve encore dans la littérature l'appellation d'origine "semi-variogramme" devenue progressivement par abus de language, "variogramme".

$$\sum_{i=1}^{n} \lambda_i Z(s_i),$$

soit toujours définie sous l'hypothèse de stationnarité intrinsèque, il faut et il suffit que

$$\sum_{i=1}^{n} \lambda_i = 0.$$

On peut trouver la démonstration de cette propriété dans Wackernagel (2003).

Une telle combinaison est appelée combinaison linéaire autorisée.

Notons que la stationnarité stricte implique la stationnarité d'ordre 2 qui elle même implique la stationnarité intrinsèque. Sous l'hypothèse de stationnarité d'ordre 2, la relation entre le variogramme et la fonction de covariance est donnée par : $\gamma(h) = C(0) - C(h)$.

2.1.3 Isotropie

On remplace souvent l'hypothèse de stationnarité d'ordre 2 (respectivement l'hypothèse intrinsèque) par une hypothèse plus forte : l'isotropie. La covariances entre 2 points (respectivement la variance des différences entre 2 points) ne dépend plus du vecteur de séparation entre ces deux points mais seulement de leur distance.

$$\forall (s_1, s_2) \in \mathcal{D}^2, \mathrm{Cov}(Z(s_1), Z(s_2)) = C(h),$$

où $h = ||\overrightarrow{s_1 s_2}||$

2.1.4 Fonctions de covariance et variogrammes

Definie-positivité

La fonction de covariance C d'un champ aléatoire stationnaire doit être une fonction définie-positive, c'est-à-dire que pour tout ensemble de sites du domaine d'étude, la matrice de covariance théorique obtenue pour le vecteur des variables aléatoires en ces sites à partir de cette fonction, doit être une matrice définie-positive, ce qui peut s'écrire :

$$\forall n \in \mathbb{N}^\star, \forall (h_{ij})_{i,j} \in \mathbb{R}^{+n^2}, \forall (\lambda_i)_i \in \mathbb{R}^n, \sum_{i,j=1}^{n} \lambda_i \lambda_j C(h_{ij}) \geq 0. \tag{2.2}$$

Vérifier la définie-positivité d'une fonction donnée ou trouver de nouvelles fonctions définies-positives ne sont pas des problèmes triviaux. Le théorème de Bochner fournit une propriété caractéristique des fonctions définies-positives facilement vérifiable et un procédé qui permet d'en construire (voir Stein, 1999 pour plus de détails).

Théorème de Bochner

Une fonction C sur $\mathbb{R}^d$ à valeurs complexes est une fonction de covariance pour un champ aléatoire stationnaire d'ordre 2 (à valeurs complexes et continu en moyenne quadratique) si et seulement si il existe une mesure finie et positive F telle que :

$$C(h) = \int_{\mathbb{R}^d} \exp(i\omega' h) F(d\omega). \tag{2.3}$$

Si F a une densité par rapport à la mesure de Lebesgue on la notera f, et on appellera la quantité $f/C(0)$ la densité sprectrale du champ aléatoire. Quand elle existe, on a la formule d'inversion suivante (Yaglom, 1987, p.332) :

$$f(\omega) = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \exp(-i\omega' h) C(h) dh. \tag{2.4}$$

Notons enfin que si F est une mesure symétrique autour de 0, le terme complexe disparaît dans l'équation (2.3) et on a alors une fonction de covariance d'un champ aléatoire à valeurs dans $\mathbb{R}$.

Propriétés des fonctions de covariance à l'origine et régularité du champ aléatoire

Un champ aléatoire $Z(.,.)$ sur $\mathbb{R}^d$ est dit continu en moyenne quadratique si pour tout s de $\mathbb{R}^d$:

$$\lim_{h \to 0} E[(Z(s+h) - Z(h))^2] = 0.$$

Un champ aléatoire sur $\mathbb{R}$ est dit différentiable en moyenne quadratique si pour tout s de $\mathbb{R}$:

$$\lim_{h \to 0} E\left[\left(\frac{Z(s+h) - Z(s)}{h}\right)^2\right]$$

existe et est finie.

Le champ aléatoire alors défini par $\{\lim_{h \to 0} \frac{Z(s+h)-Z(s)}{h}, s \in \mathcal{D}\}$ est le champ dérivé de $Z(.)$.

De manière similaire, on peut définir la différentiabilité à n'importe quel ordre, en moyenne quadratique d'un champ aléatoire. Notons également que ces définitions s'étendent facilement aux champs aléatoires sur $\mathbb{R}^d$.

Les propositions suivantes font le lien entre la régularité (en moyenne quadratique) des champs aléatoires faiblement stationnaires et la régularité de leur fonction de covariance (Stein, 1999).

- Pour un champ aléatoire $Z(.)$ sur $\mathbb{R}^d$ faiblement stationnaire de fonction de covariance C_Z, $Z(.)$ est continu en moyenne quadratique si et seulement si C_Z est continue en 0.
- Pour un champ aléatoire $Z(.)$ sur $\mathbb{R}$ faiblement stationnaire de fonction de covariance C_Z, $Z(.)$ est différentiable à l'ordre k si C_Z est différentiable à l'ordre $2k$ à l'origine.

Ce dernier résultat s'étend aux champs aléatoires sur $\mathbb{R}^d$.

Exemples de fonctions définies-positives

L'utilisateur de la géostatistique dispose d'une gamme de fonctions vérifiant la définie-positivité. Ces modèles sont en général suffisants pour l'application, d'autant que toute combinaison linéaire positive de fonctions définie-positives est définie-positive, ce qui permet de combiner ces fonctions de base pour en obtenir des plus flexibles si nécessaire.

Voici les plus utilisées :

- Le modèle sphérique :

$$C(h) = \begin{cases} \sigma^2 \left(1 - \frac{3}{2}\frac{h}{\phi} + \frac{1}{2}\frac{h^3}{\phi^3}\right) & \text{pour } 0 \leq h \leq \phi \\ 0 & \text{pour } \quad h \geq \phi \end{cases}$$

  ϕ est la portée du champ aléatoire. Deux observations d'une variable séparée d'une distance supérieure à ϕ ne sont plus corrélées.

– Le modèle de Matèrn :

$$C(h) = \frac{\pi^{1/2}\phi}{2^{\nu-1}\Gamma(\nu+1/2)\alpha^{2\nu}}(\alpha h)^{\nu}\mathcal{K}(\alpha h), \text{ avec } \nu > 0, \phi > 0 \text{ et } \alpha > 0, \tag{2.5}$$

et $\mathcal{K}_\nu$ est la fonction de Bessel modifiée de deuxième espèce :

$$\mathcal{K}_\nu(x) = \frac{\pi}{2}\frac{I_{-\nu}(x) - I_\nu(x)}{\sin(\pi\nu)}, \text{ avec } I_\nu(x) = \left(\frac{x}{2}\right)^\nu \sum_{k=0}^{\infty} \frac{1}{k!\Gamma(\nu+k+1)}\left(\frac{x}{2}\right)^{2k}.$$

Le modèle de Matèrn trouve une nouvelle popularité dans la "géostatistique méthodologique" (qui ne s'est pas vraiment encore confirmée dans l'utilisation de la géostatistique) car elle autorise tous les degrés de régularité du champ aléatoire sous-jacent au travers du paramètre ν. En effet, le champ aléatoire $Z(.)$ est k fois différentiable si et seulement si ν est supérieur à k.

Le modèle de Matèrn inclut le modèle exponentiel ($\nu = 1/2$) et le modèle gaussien ($\nu \to \infty$) dont les paramétrisations les plus souvent utilisées sont les suivantes :

– Le modèle exponentiel :

$$C(h) = \sigma^2 \exp(-h/\phi),$$

très souvent utilisé, certainement pour sa simplicité. 3ϕ est la portée pratique du champ aléatoire, c'est à dire la distance approximative à partir de laquelle les observations entre deux points séparés de cette distance ont une corrélation de 0.05.

– Le modèle gaussien :

$$C(h) = \sigma^2 \exp(-h^2/\phi),$$

peu utilisé du fait de sa trop grande régularité. En effet, si on observait une réalisation d'un champ ayant pour fonction de covariance le modèle gaussien sur n'importe quelle boule ouverte de $\mathbb{R}^d$, on pourrait prédire parfaitement la réalisation sur tout $\mathbb{R}^d$. Le modèle gaussien est donc considéré comme peu réaliste[3].

– L'effet de pépite :

$$C(h) = \begin{cases} \sigma^2 \text{ pour } h = 0 \\ 0 \text{ pour } h \geq 0 \end{cases}$$

Un champ aléatoire dont la fonction de covariance serait un effet de pépite est un bruit blanc. L'effet de pépite est souvent utilisé en conjonction avec un modèle continu pour modéliser la part d'erreur de mesures dans la variance totale, ou les variations spatialisées mais à trop petite échelle (au regard de l'échantillonnage spatial) et qui ne sont, par conséquent, pas détectables, par manque de couples d'observations proches.

Les modèles de covariance que nous utiliserons seront toujours une combinaison linéaire entre un effet de pépite et un modèle exponentiel. Contrairement à Stein (1999), nous pensons que l'utilisation du modèle de Matèrn ne doit pas être systématique et doit dépendre de la qualité de l'échantillonnage. En effet, capturer la régularité de la fonction de covariance à l'origine ne peut se faire sans un échantillonnage comportant un grand nombre de sites d'observation proches [4].

[3] L'anecdote suivante illustre cette affirmation : projetant lors d'un exposé des cartes de réalisations simulées de champs gaussiens selon différents modèles de covariance, un conférencier eut la surprise de constater que le retro-projecteur qui effectuait la mise au point de manière automatique, n'arrivait pas à se stabiliser lorsque le modèle de covariance ayant servi à la simulation était gaussien ; sans doute du fait de la trop grande régularité ce ce modèle.

[4] C'est d'ailleurs un des résultats de l'étude de Zhu et Stein (2006) qui montrent par simulations que les échantillonnages D-optimaux pour l'estimation des paramètres du modèle de covariance Matèrn sont ceux pour lesquels les sites d'observation sont souvent très proches, ce qui est assez rare en pratique.

2.1.5 L'ergodicité

L'ergodicité est une hypothèse qui permet de travailler avec une réalisation unique d'un champ aléatoire. En termes pratiques, cette hypothèse assure que si l'on observait une réalisation d'un champ aléatoire sur un domaine dont la taille est grande relativement aux structures spatiales du phénomène observé, l'inférence sur les paramètres de structure se passerait comme si on travaillait avec plusieurs réalisations.

Au sens strict, l'ergodicité assure de bonnes propriétés de mélange des champs aléatoires strictement stationnaires. Une présentation formelle de cette propriété nous éloignerait trop de notre propos (voir Adler, 1981, pour plus de détails). En géostatistique traditionnelle, on s'intéresse essentiellement à l'estimation des moments d'ordre 1 et 2 et l'usage est donc de travailler avec des hypothèses plus faibles, telles que l'ergodicité en moyenne et l'ergodicité en covariance.

L'ergodicité en moyenne

Elle assure essentiellement que si l'on pouvait observer partout une réalisation du champ aléatoire, on pourrait estimer de manière consistante l'espérance de ce champ. Formalisons cette idée :

Soit $Z(.,.)$ faiblement stationnaire et isotrope, d'espérance μ et de fonction de covariance C_Z.

Soit $Z(\mathcal{D},.)$ la variable aléatoire dite de moyenne du champ $Z(.,.)$ sur le domaine $\mathcal{D}$ et définie comme suit :

$$Z(\mathcal{D},.) = \frac{1}{|\mathcal{D}|}\int_{\mathcal{D}} Z(s,.)ds,$$

où $|\mathcal{D}|$ représente le volume du domaine $\mathcal{D}$; et soit

$$\mathrm{Var}(Z(\mathcal{D},.)) = \frac{1}{|\mathcal{D}|^2}\int_{\mathcal{D}}\int_{\mathcal{D}} C_Z(||s-s'||)dsds',$$

sa variance.

Définition 2 *Un champ aléatoire sera dit ergodique en moyenne si* $Var(Z(\mathcal{D}),.))$ *tend vers* 0 *quand le volume de* $\mathcal{D}$ *tend vers l'infini.*

Il est important de noter que $|\mathcal{D}|$ ne doit pas tendre vers l'infini dans une direction particulière de $\mathbb{R}^d$ pour que la limite ne dépende pas du choix de cette direction. On pourra considérer les ensembles de la forme $[-t,t]^d$ et faire tendre t vers l'infini pour éviter tout problème.

On montre (Gabriel, 2004, voir par exemple) que l'ergodicité en moyenne implique que

$$\lim_{|\mathcal{D}|\to\infty} Z(\mathcal{D},\omega) = \mu \text{ p.s.}$$

Cette propriété des champs aléatoires faiblement stationnaires et ergodique en moyenne est à mettre en relation avec la loi des grands nombres pour un échantillon.

L'ergodicité en covariance

En reprenant les mêmes notations que ci-dessus, on définit l'ergodicité en covariance :

Définition 3 *Un champ aléatoire faiblement stationnaire et isotrope est dit ergodique en covariance si*

$$\lim_{|\mathcal{D}|\to\infty} \frac{1}{|\mathcal{D}|^2}\int_{\mathcal{D}}\int_{\mathcal{D}} C_Z^2(||s-s'||)dsds' = 0$$

Faire l'hypothèse d'ergodicité en covariance assure essentiellement que l'on peut estimer de manière consistante la fonction de covariance quand le domaine d'observation devient grand au regard des structures spatiales.

La micro-ergodicité

Le terme de micro-ergodicité fut introduit, semble-t-il, par Matheron (voir par exemple Matheron, 1978) pour désigner la consistance de certains paramètres du champ aléatoire à partir d'observations sur un domaine fini, dans lequel les sites d'observation se densifient. Matheron montre par exemple que le comportement du variogramme à l'origine est micro-ergodique pourvu que celui ci ne soit pas trop régulier à l'origine. L'espérance du champ aléatoire n'est pas micro-ergodique. Pour le comprendre, il suffit de considérer que l'on observe une réalisation d'un champ aléatoire sur un petit domaine relativement à la portée des dépendances spatiales ; le domaine d'échantillonnage étant petit, il peut se trouver dans une zone où la réalisation du champ aléatoire est par exemple négative (du fait des grandes dépendances relatives au domaine) bien que l'espérance du champ aléatoire soit nulle et malgré l'ergodicité. On pourra densifier l'échantillonnage sur $\mathcal{D}$, mais l'estimation de l'espérance du champ ne tendra pas vers sa "vraie" valeur. L'estimation tendra vers ce que Matheron nomme la moyenne régionale $Z(\mathcal{D}, \omega)$.

On peut faire le même constat pour la variance du champ aléatoire.

En pratique

Les hypothèses d'ergodicité sont invérifiables à partir d'un nombre fini d'observations sur un domaine de taille finie et surtout à partir d'une réalisation unique. En pratique, ceci n'est pas un problème pour le géostatisticien qui est concerné par la prédiction ou l'estimation d'une grandeur régionale (valeur en un site, espérance sur une zone...) à partir d'observations sur cette région d'intérêt. L'affirmation suivante " les observations sont issues d'un champ aléatoire stationnaire" n'est alors qu'un modèle théorique utile pour la cohérence mathématique. L 'espérance de ce champ aléatoire n'est qu'un paramètre **conventionnel**, qui n'a pas d'intérêt pour le géostatisticien. Une autre valeur de l'espérance du champ aurait également pu permettre la même réalisation du champ aléatoire sur $\mathcal{D}$. Par opposition, la moyenne régionale $Z(\mathcal{D}, \omega)$ de la réalisation (dont l'estimation permettra par exemple de quantifier ce que l'on pourra espérer obtenir d'un puit de pétrole installé dans la zone échantillonnée) qui est une quantité quantifiable à partir des observations est un paramètre **objectif**. Les épithètes "conventionnels" ou "objectifs" sont à prendre ici au sens de Matheron (1978, p 73) pour qui les seules quantités "objectives" sont celles qui peuvent être calculées (estimées) à partir d'une seule réalisation sur un domaine borné.

Nous reviendrons sur ces notions dans les chapitres 4 et 5. Pour plus de considérations théoriques sur l'utilisation des champs aléatoires en géostatistique, nous recommandons la lecture de Matheron (1978).

2.1.6 Un cadre distributionnel spatial : les champs gaussiens

Dans les sections précédentes, aucune hypothèse distributionnelle n'a été faite sur les variables aléatoires mises en jeu. Ce cadre de travail, fréquemment adopté en géostatistique rend les outils d'estimation et de prédiction robustes. La spécification complète du modèle peut néanmoins être nécessaire pour simuler des données spatiales ou quantifier la variabilité des estimateurs au sens classique des statistiques (intervalles de confiance par exemple) par des résultats asymptotiques ou

par simulations (voir chapitres 5 et 6). L'hypothèse distributionnelle gaussienne est naturelle pour modéliser des champs spatiaux, du fait de l'existence de la loi gaussienne multivariée.

Définition

Un champ aléatoire $X(.) = \{X(s), s \in \mathcal{D}\}, \mathcal{D} \subset \mathbb{R}^d$ est dit gaussien si pour toute configuration de sites $\{s_1, \ldots, s_n\} \subset \mathcal{D}, n \in \mathbb{N}^\star$, le vecteur aléatoire $\mathbf{X} = (X(s_1), \ldots, X(s_n))'$ est un vecteur multigaussien. On notera

$$\mathbf{X} \sim \mathcal{N}(\mu, \Sigma),$$

où $\mu = E[\mathbf{X}]$ et $\Sigma = \mathrm{Var}[\mathbf{X}]$.

Propriétés

1. L'hypothèse de stationnarité d'ordre 2 pour un champ gaussien implique l'hypothèse de stationnarité stricte. Plus généralement, une simple spécification sur l'espérance (espérance constante, espérance comme fonction donnée des coordonnées géographiques,...) et sur la covariance (par exemple la donnée d'un modèle de covariance spatiale) revient à définir toutes les distributions fini-dimensionnelles puisque la distribution d'un vecteur multigaussien est caractérisée par ses deux premiers moments.
2. Si $\mathbf{W} = (\mathbf{X}', \mathbf{Y}')'$ est un vecteur multigaussien avec $\mathbf{X} = (X_1, \ldots, X_n)'$ et $\mathbf{Y} = (Y_1, \ldots, Y_p)'$ d'espérance

   $$\mu = \begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix},$$

   où μ_X et μ_Y sont les vecteurs d'espérance des vecteurs respectifs $\mathbf{X}$ et $\mathbf{Y}$, et de variance

   $$\Sigma = \begin{pmatrix} \Sigma_{XX} & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_{YY} \end{pmatrix},$$

   où les matrices Σ_{XX}, Σ_{XY}, Σ_{YX}, et Σ_{YY} sont les matrices de covariance respectives $\mathrm{Cov}(\mathbf{X}, \mathbf{X})$, $\mathrm{Cov}(\mathbf{X}, \mathbf{Y})$, $\mathrm{Cov}(\mathbf{Y}, \mathbf{X})$ et $\mathrm{Cov}(\mathbf{Y}, \mathbf{Y})$, alors :
   la distribution de $\mathbf{Y}$ conditionnellement à $\mathbf{X}$ est la distribution multigaussienne d'espérance

   $$\mu_{Y|X} = \mu_Y + \Sigma_{YX}\Sigma_{XX}^{-1}(\mathbf{X} - \mu_X) \tag{2.6}$$

   et de variance

   $$\Sigma_{Y|X} = \Sigma_{YY} - \Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY}. \tag{2.7}$$

3. Si $Z(.)$ est un champ aléatoire gaussien stationnaire et isotrope de fonction de covariance C, alors $Z(.)$ est ergodique si et seulement si (Adler,1981)

   $$\lim_{h \to +\infty} C(h) = 0.$$

Simulation

Pour simuler la réalisation d'un champ gaussien stationnaire $Z(.)$ d'espérance μ et de fonction de covariance C, sur un ensemble de sites $\{s_1, \ldots, s_n\}$, nous procèderons de la manière suivante :

1) calculer Σ la matrice de covariance du vecteur $\mathbf{Z} = (Z(s_1), \ldots, Z(s_n))'$ dont le $(i,j)^{\text{ème}}$ terme est donné par $C(||s_i - s_j||)$.

2) calculer $\Sigma^{1/2}$ en utilisant la décomposition en valeurs singulières de Σ.

3) simuler un échantillon gaussien $\mathbf{X} = (X_1, \ldots, X_n)'$ centré réduit.

4) On obtient $\mathbf{Z}$ par

$$\mathbf{Z} = \mu \mathbb{1}_n + \Sigma^{1/2}\mathbf{X}.$$

D'autres méthodes sont utilisées lorsque n devient grand ($n > 1000$) afin d'éviter les problèmes numériques. Pour notre travail, la méthode décrite ci-dessus sera toujours suffisante.

2.2 La prédiction d'une variable régionalisée : une procédure en 2 étapes

Revenons à l'objectif pratique de la géostatistique : la prédiction et la cartographie.

Pour prédire une variable en un site où celle-ci n'est pas observée à partir d'un nombre fini d'observations, l'utilisateur choisira ses hypothèses parmi les précédentes. Puis il caractérisera les dépendances spatiales par une analyse variographique (estimation de la fonction de covariance ou du variogramme). Enfin il interpolera la variable par krigeage en supposant que la fonction de covariance est connue.

Nous présentons cette procédure en détail sous l'hypothèse de stationnarité d'ordre 2 et d'isotropie.

2.2.1 Analyse variographique

Estimateur non paramétrique du variogramme

La première étape de l'analyse est le calcul du variogramme expérimental qui est un estimateur du variogramme. Ce calcul nécessite deux étapes :

1) calcul de la nuée variographique :

C'est le graphe $(h_{ij}, \gamma_{ij})_{i,j=1,\ldots,n}$ où $h_{ij} = ||s_i - s_j||$ et $\gamma_{ij} = \frac{1}{2}(Z(s_i) - Z(s_j))^2$.

2) calcul du variogramme expérimental :

Le variogramme expérimental (estimateur de Matheron) est un lissage de la nuée variographique. On choisit p distances $d_k, k = 1, \ldots, p$ comprises entre la distance mimimum et environ un tiers[5] de la distance maximum et dans chaque intervalle $[\![d_k, d_{k+1}[\![$, on calcule la moyenne des distances ainsi que la moyenne des γ_{ij}. Il résulte un estimateur du variogramme $(h_k, \hat{\gamma}_k)_{k=1,\ldots,p-1}$:

$$h_k = \sum_{i,j \in \mathcal{C}_k} \frac{1}{\sharp(\mathcal{C}_k)} h_{i,j},$$

$$\hat{\gamma}_k = \sum_{i,j \in \mathcal{C}_k} \frac{1}{\sharp(\mathcal{C}_k)} \gamma_{i,j}$$

où $\mathcal{C}_k = \{(i,j) \in [\![1,n]\!]; ||s_i - s_j|| \in [\![d_k, d_{k+1}[\![\}, k = 1, \ldots p-1$, et $\sharp(\mathcal{C}_k)$ est le cardinal de $\mathcal{C}_k$.

A titre d'exemple, la figure (2.1) montre la représentation graphique de la nuée variographique et du variogramme expérimental de la teneur en nickel mesurée en 259 sites du Jura suisse.

Le calcul du variogramme expérimental est l'étape préalable à toute analyse de données spatiales. Il est à la géostatistique, ce que l'histogramme est à la statistique. Notons que si la fonction

[5] Les grandes distances entre les sites d'observation sont plus rares et les points de la nuée variographique correspondant aux grandes distances sont essentiellement mesurés sur les bords du domaine d'observation. L'estimateur du variogramme est donc plus variable pour les grandes distances et moins robuste aux non stationnarités éventuelles.

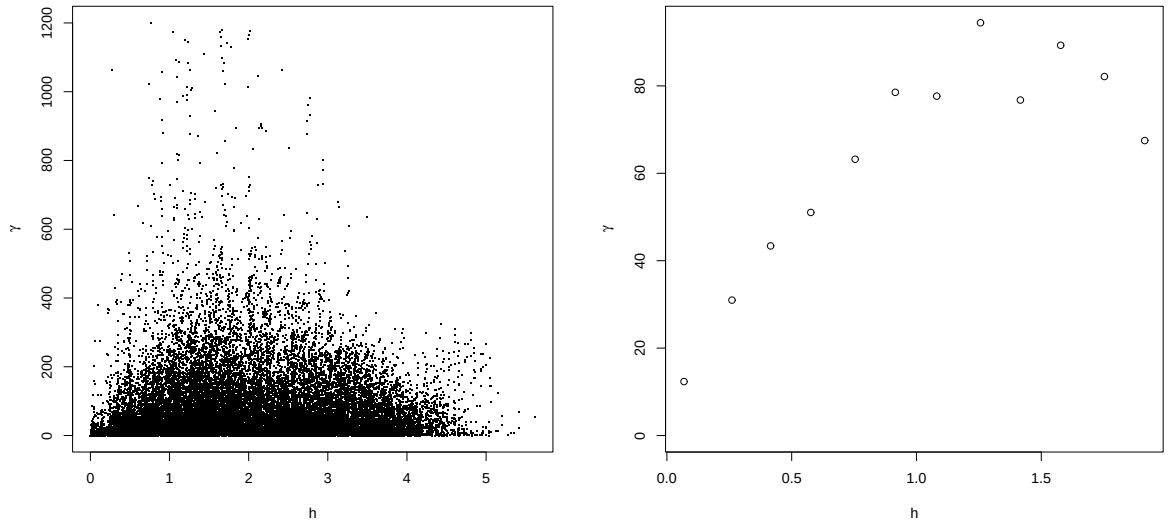

Fig. 2.1. Gauche : nuée variographique. Droite : variogramme expérimental

de variogramme n'a de sens que dans la mesure où une hypothèse de stationnarité (au minimum intrinsèque) et d'isotropie ont été faites, l'estimation du variogramme peut permettre de détecter des non-stationnarités éventuelles.

Notons que l'on peut également construire la fonction de covariance expérimentale sous l'hypothèse de stationnarité d'ordre 2 mais celle-ci est peu utilisée en pratique. Elle nécessite une hypothèse plus forte que le variogramme et par conséquent est moins robuste que celui-ci. Elle requiert une estimation de l'espérance (qui devrait dans l'idéal, tenir compte des corrélations spatiales que l'on cherche justement à caractériser) du champ sous-jacent, et elle est sans doute moins porteuse de sens que le variogramme pour l'utilisateur.

Le variogramme expérimental est un outil descriptif visuel pour caractériser les dépendances spatiales d'une variable mais en aucun cas il ne saurait constituer un modèle valide pour le champ aléatoire sous-jacent. En effet, rappelons que sous l'hypothèse de stationnarité d'ordre 2, la fonction de covariance C doit être définie-positive. Le variogramme γ doit donc être choisi de telle sorte que la fonction de covariance C obtenue par $C(h) = C(0) - \gamma(h)$ soit définie-positive.

Il convient de préciser que sous l'hypothèse de stationnarité intrinsèque, la fonction de covariance n'existe pas en général mais le variogramme γ doit quand même vérifier une condition particulière pour être valide : il doit être conditionnellement défini-négatif. C'est-à-dire que pour tout ensemble de sites $\{s_1, \ldots, s_n\}$, et pour toute combinaison linéaire autorisée

$$L = \sum_{i=1}^{n} \lambda_i Z(s_i),$$

la variance de L obtenue à partir du variogramme γ doit être positive. Travaillant sous l'hypothèse de stationnarité d'ordre 2, nous utiliserons toujours des modèles de variogramme construits à partir des modèles de covariance présentés plus haut.

Modélisation et estimation du variogramme

Disposant d'un ensemble de modèles de variogramme, le géostatisticien en choisit un[6] γ_θ en fonction de l'apparence du variogramme expérimental ou de connaissances préalables sur la régularité spatiale du phénomène étudié. Puis il estime les paramètres θ en se basant sur le variogramme expérimental :

-par moindres carrés ordinaires :

$$\hat{\theta} = \text{argmin}_\theta \sum_{k=1}^{p-1} (\gamma_\theta(h_k) - \hat{\gamma}_k)^2$$

à la manière de la régression non linéaire,

-par moindres carrés pondérés :

$$\hat{\theta} = \text{argmin}_\theta \sum_{k=1}^{p-1} \sharp(\mathcal{C}_k)(\gamma_\theta(h_k) - \hat{\gamma}_k)^2$$

pour donner plus de poids aux classes de distances pour lesquelles la précision du variogramme expérimental est la plus grande,

-par moindres carrés modifiés :

$$\hat{\theta} = \text{argmin}_\theta \sum_{k=1}^{p-1} \frac{(\gamma_\theta(h_k) - \hat{\gamma}_k)^2}{\gamma(h_k)^2}$$

ce qui donne plus d'importance aux points du variogramme expérimental à courtes distances.

D'autres méthodes d'estimation par moindres carrés peuvent être envisagées, pour prendre en compte les corrélations entre les points du variogramme expérimental par exemple. Schabenberger et Gotway (2005) présentent et comparent différentes alternatives.

Sous l'hypothèse gaussienne, on peut estimer les paramètres du variogramme par maximum de vraisemblance. Nous présenterons en détail cette alternative dans le chapitre 4.

2.2.2 Prédiction

Une fois la structure spatiale des observations caractérisée, celle ci est tradititionnellement supposée connue. La seconde étape consiste alors à interpoler la variable d'intérêt en un site où celle-ci n'a pas été mesurée, à partir des observations et de la structure spatiale modélisée.

Le krigeage est le nom de l'estimateur linéaire sans biais et de variance minimale parmi les sans biais (c'est un BLUP).

Nous traitons en détail le cas du krigeage de la moyenne.

Krigeage de la moyenne

On observe $\mathbf{z} = (z(s_1), \ldots, z(s_n))'$. Et on suppose que $\mathbf{z}$ est la réalisation d'un vecteur $\mathbf{Z}$ aléatoire issu d'un champ aléatoire faiblement stationnaire d'espérance μ à estimer et de fonction de covariance C_Z supposée connue.

L'estimateur du krigeage de la moyenne $\mu^\star$ est une combinaison linéaire des composantes de $\mathbf{Z}$:

[6] Une combinaison linéaire (on parle de modèle gigogne) de fonctions simples peut être utilisée, les scalaires de la combinaison étant des paramètres supplémentaires à estimer. On utilise souvent la combinaison linéaire $\sigma^2(\alpha \mathbb{1}_{h=0} + (1-\alpha)\rho(h))$ où ρ est une fonction d'autocorrélation continue à l'origine.

$$\mu^\star = \sum_{i=1}^{n} \lambda_i Z(s_i)$$

où $\lambda_i, i = 1, \ldots, n$ sont les poids de krigeage.

La condition de non-biais s'écrit :

$$E[\mu^\star] = E\left[\sum_{i=1}^{n} \lambda_i Z(s_i)\right] = \mu,$$

ce qui est équivalent à

$$\sum_{i=1}^{n} \lambda_i = 1. \tag{2.8}$$

Pour que $\mu^\star$ soit de variance minimale, il faut que la quantité

$$\begin{aligned} \mathrm{Var}(\mu^\star) &= \sum_{i,j=1}^{n} \lambda_i \lambda_j \mathrm{Cov}(Z(s_i), Z(s_j)) \\ &= \sum_{i,j=1}^{n} \lambda_i \lambda_j C_Z(||s_i - s_j||) \end{aligned} \tag{2.9}$$

soit minimale.

Le calcul des poids de krigeage est donc un problème d'optimisation sous contrainte. Minimiser (2.9) sous la contrainte (2.8) requiert l'utilisation d'un multiplicateur de Lagrange (η) et par suite la résolution du système linéaire suivant :

$$\begin{cases} \frac{\partial}{\partial \lambda_i} \left(\sum_{i,j=1}^{n} \lambda_i \lambda_j C_Z(||s_i - s_j||) - 2\eta \left(\sum_{i=1}^{n} \lambda_i - 1 \right) \right) = 0, \;\; i = 1, \ldots, n. \\ \sum_{i=1}^{n} \lambda_i = 1 \end{cases} \tag{2.10}$$

Les poids de krigeage sont donc solutions du système suivant :

$$\tilde{C} \Lambda = \mathbf{u} \tag{2.11}$$

avec

$$\tilde{C} = \begin{pmatrix} C & \mathbb{1}_n \\ \mathbb{1}_n' & 0 \end{pmatrix},$$

avec $C_{ij} = \mathrm{Cov}(||s_i - s_j||), i, j = 1, \ldots, n,$

$$\Lambda = \begin{pmatrix} \lambda_1 \\ \vdots \\ \lambda_n \\ -\eta \end{pmatrix} \text{ et } \mathbf{u} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix}.$$

L'expression du krigeage de la moyenne est alors donnée par

$$\mu^\star = \Lambda' \tilde{Z} = \mathbf{u}' \tilde{C}^{-1} \tilde{Z},$$

avec $\tilde{Z} = \begin{pmatrix} \mathbf{Z} \\ 0 \end{pmatrix}$.

La variance de krigeage est donnée par :

$$\mathrm{Var}(\mu^\star) = \Lambda' \tilde{C} \Lambda = \mathbf{u}' \tilde{C}^{-1} \mathbf{u}' \tag{2.12}$$

En appliquant à $\tilde{C}$, la formule d'inversion des matrices par bloc suivante (voir par exemple Rao et Toutenburg, 1995) :

$$\begin{pmatrix} A & B \\ B' & G \end{pmatrix}^{-1} = \begin{pmatrix} -A^{-1}(I + BE^{-1}B'A^{-1}) & -A^{-1}BE^{-1} \\ -E^{-1}B'A^{-1} & E^{-1} \end{pmatrix} \tag{2.13}$$

avec $E = G - B'A^{-1}B$, nous donnons une forme au krigeage de la moyenne qui nous semble plus explicite :

$$\mu^\star = \frac{\mathbb{1}_n' C^{-1} \mathbf{Z}}{\mathbb{1}_n' C^{-1} \mathbb{1}_n} \tag{2.14}$$

De même, la variance du krigeage peut s'écrire :

$$\mathrm{Var}(\mu^\star) = \frac{1}{\mathbb{1}_n' C^{-1} \mathbb{1}_n} \tag{2.15}$$

Krigeage simple

Le krigeage simple est le prédicteur de l'écart à l'espérance du champ aléatoire en un site s_0 à partir des observations, l'espérance du champ étant supposée connue. Le prédicteur de krigeage s'écrit :

$$Z^\star(s_0) = \mu + \sum_{i=1}^{n} \lambda_i (Z(s_i) - \mu).$$

La condition de non-biais s'écrit :

$$E[Z^\star(s_0) - Z(s_0)] = 0,$$

et est toujours vérifiée.

La variance de l'erreur de prédiction à minimiser est donnée par :

$$\mathrm{Var}[Z^\star(s_0) - Z(s_0)].$$

Par une méthodologie similaire à celle utilisée dans le cas du krigeage de la moyenne (mais sans la contrainte), on montre que :

$$Z^\star(s_0) = C(s_0) C^{-1} \mathbf{Z}$$

où

$$C(s_0)_i = \mathrm{Cov}(||s_i - s_0||).$$

La variance de prédiction est donnée par

$$\mathrm{Var}(Z^\star(s_0) - Z(s_0)) = C_Z(0) - C'(s_0) C^{-1} C(s_0). \tag{2.16}$$

où $C_Z(0) = \sigma_Z^2$ est la variance du champ aléatoire $Z(.)$.

Le krigeage simple peut se réécrire :

$$Z^\star(s_0) = \mu + C(s_0)' C^{-1} (\mathbf{Z} - \mu \mathbb{1}_n), \tag{2.17}$$

Krigeage ordinaire

Le krigeage ordinaire est le prédicteur de $Z(s_0)$ dans le cas où μ est inconnu.

Minimiser la variance de $Z(s_0)^\star - Z(s_0)$ sous la contrainte de non-biais conduit à l'expression suivante pour le krigeage ordinaire :

$$Z^\star(s_0) = \tilde{C}(s_0)\tilde{C}^{-1}\tilde{Z}$$

où

$$\tilde{C}(s_0) = \begin{pmatrix} C(s_0) \\ 1 \end{pmatrix},$$

avec

$$C(s_0)_i = \mathrm{Cov}(||s_i - s_0||).$$

On montre que le krigeage ordinaire est égal à la somme du krigeage simple (avec $\mu = 0$) et du krigeage de la moyenne, ou de manière équivalente le krigeage simple dans lequel μ est remplacé par son estimation $\mu^\star$:

$$Z^\star(s_0) = \mu^\star + C(s_0)'C^{-1}(\mathbf{Z} - \mu^\star \mathbb{1}_n).$$

Le krigeage avec dérive externe : un modèle de régression spatial

Le krigeage avec dérive externe est utilisé pour améliorer la prédiction du krigeage ordinaire en introduisant des covariables dans le terme de l'espérance, ces covariables devant également être disponibles aux sites de prédiction[7].

On note $\mathbf{X} = (\mathbf{x}(s_1), \ldots, \mathbf{x}(s_n))$ la matrice des observations des covariables ($\mathbf{x}(s_i)$ est le vecteur des covariables observées en $s_i, i = 1, \ldots, n.$) et $\mathbf{x}(s_0)$ le vecteur des covariables au site de prédiction.

Le modèle du krigeage avec dérive externe s'écrit :

$$\forall s \in \mathcal{D}, Z(s) = \beta X(s) + \varepsilon(s) \tag{2.18}$$

où $\varepsilon(.) = \{\varepsilon(s), s \in \mathcal{D}\}$ est modélisé comme un champ aléatoire faiblement stationnaire.

L'objectif est la prédiction de $z(s_0)$ en utilisant les observations de la variable d'intérêt, $z(s_1), \ldots, z(s_n)$, les observations des covariables en ces sites $\mathbf{X}$, et la valeur des covariables en s_0, $\mathbf{x}(s_0)$.

Supposons que la structure de covariance des résidus soit connue et notons C la matrice de covariance des résidus aux sites d'observations le vecteur de covariance entre les résidus aux sites d'observation et le résidu au site de prédiction s_0.

On montre que le prédicteur de krigeage avec dérive externe (BLUP) s'écrit :

$$Z^\star_{KU}(s_0) = \mathbf{x}(s_0)\hat{\beta} + C(s_0)'C^{-1}(\mathbf{Z} - \mathbf{X}\hat{\beta}),$$

avec

$$\hat{\beta} = (\mathbf{X}'C^{-1}\mathbf{X}^{-1})^{-1}\mathbf{X}'C^{-1}\mathbf{Z},$$

et $\mathbf{Z} = (Z(s_1), \ldots, Z(s_n))'$.

Le krigeage avec dérive externe est peu utilisé, sans doute du fait de sa procédure en plusieurs étapes, toutes basées sur des approximations :

[7] Dans la plupart des cas, les covariables sont des fonctions de la position, polynômiales en les coordonnées. On parle alors de krigeage universel.

1) Estimer β sans prendre en compte les corrélations spatiales éventuelles des résidus (régression linéaire classique).

2) Effectuer l'analyse variographique des résidus estimés.

3) En supposant que les corrélations spatiales du champ $\varepsilon(.)$ sont celles mises en évidence par l'analyse variographique, prédire $\mathbf{z}(s_0)$ (prédicteur linéaire, sans biais de variance minimale).

Optimalité du krigeage sous hypothèse gaussienne

La géostatistique traditionnelle ne fait en général pas d'hypothèse distributionnelle explicite sur les champs aléatoires sous-jacents mais certains résultats relatifs aux estimateurs (ou aux prédicteurs) de la géostatistique prennent un sens particulier dans le cadre gaussien.

Notons par exemple que le prédicteur de krigeage simple d'une variable Z en un site s_0 à partir d'observations aux sites $s_1, \ldots, s_n$ (qui est un BLUP dans le cadre d'un champ aléatoire faiblement stationnaire) est l'espérance de $Z(s_0)$ conditionnellement à $Z(s_1), \ldots, Z(s_n)$ (par (2.17) et (2.6) appliquée à $\mathbf{Y} = Z(s_0)$ et $\mathbf{X} = (Z(s_1), \ldots, Z(s_n))'$) et la variance de ce prédicteur est égale à la variance de $Z(s_0)$ conditionnellement à $Z(s_1), \ldots, Z(s_n)$ (par (2.16) et (2.7)).

Le krigeage est donc un prédicteur optimal (au sens de l'estimateur de Bayes associé au coût quadratique) sous hypothèse gaussienne alors qu'il n'est optimal que parmi les prédicteurs linéaires sans hypothèse distributionnelle.

Nous verrons par la suite d'autres comparaisons de ce type.

2.2.3 Modèles hiérarchiques

Il arrive que les hypothèses de la géostatistique soient inadaptées aux besoins de l'étude.

Prenons l'exemple de l'occurence d'un caractère (comme l'apparition d'une maladie sur un individu) observée sur une population donnée dont la taille est connue, et supposons que l'unité statistique (une commune) soit suffisamment "petite" géographiquement au regard de la zone d'étude (un territoire) pour pouvoir être considérée comme un site ponctuel sur un domaine continu. Il est évident qu'interpoler l'occurence n'a aucun sens puisqu'elle dépend de la taille de la population. L'intérêt portera plutôt sur le risque d'apparition, ou sur l'intensité d'un processus sous-jacent. Interpoler la fréquence d'appararition du caractère par unité n'est que partiellement satisfaisant car la fréquence est une observation intrinsèquement non stationnaire en variance, la mesure étant d'autant plus précise que la population totale par unité est grande.

La solution apportée par exemple par Oliver *et al.* (1993) est de modéliser, à un premier niveau de hiérarchie, le champ aléatoire de risque commme un champ stationnaire de manière analogue à la section précédente, puis conditionnellement à ce champ, l'hypothèse est faite que l'occurence en un site est distribuée selon une loi binômiale dont le paramètre de taille est la taille de la population totale et le paramètre de probabilité est le risque en ce site ; les observations étant mutuellement indépendantes conditionnellement à la réalisation du champ aléatoire. On dira que le champ aléatoire "capture" toutes les dépendances spatiales des observations. Les auteurs proposent ensuite une estimation du variogramme du risque basée sur les observations puis une procédure de krigeage pour le risque (il s'agissait de cartographier le risque du cancer d'enfants dans une région d'environ 15000 km^2 autour de Birmingham). Desassis (2002)[8] étudie le cas particulier de Bernouilli (dans une étude sur le dépérissement de la vigne en Languedoc) et dans le même esprit, Monestiez *et al.* (2006) étudient le cas poissonien (pour la cartographie de l'abondance relative de baleines en Méditérannée).

[8] Mémoire de DEA

Une fois la carte de risque - ou d'intensité - produite, elle peut être mise en relation avec une carte de facteurs éventuels. Mais l'inférence sur l'influence d'un facteur potentiel (qui peut être observé en d'autres sites que les unités ayant servi à produire la carte de risque) n'est sans doute pas satisfaisante dans la mesure où l'enchaînement de procédures d'estimation (estimation des paramètres de structure spatiale, estimation du risque, puis estimation du lien risque vs. facteur(s)), chacune avec son incertitude associée, n'est pas de nature à produire une inférence de qualité, la variabilité de l'estimateur du lien facteur(s)/risque ne pouvant être déterminée correctement.

L'article d'Oliver *et al* (1993) est néanmoins notre plus ancienne référence bibliographique dans laquelle un champ aléatoire latent est modélisé pour capturer les dépendances spatiales des observations et la réalisation de ce champ (non observée donc) fournit l'espérance conditionnelle des observations dans un modèle non gaussien (binômial dans ce cas).

2.3 Géostatistique multivariée

Un des objectifs de la géostatistique multivariée est d'utiliser l'information apportée par des observations de variables auxiliaires afin d'améliorer la qualité de la prédiction d'une variable d'intérêt par une procédure appelée cokrigeage. Il est évident que ces variables auxiliaires doivent être liées d'une manière ou d'une autre à la variable d'intérêt pour pouvoir espérer améliorer la prédiction de cette dernière. Contrairement au cas du krigeage avec dérive externe (dans lequel il était également question d'utiliser de l'information apportée par des variables auxiliaires pour améliorer la prédiction), on ne suppose pas ici que les variables auxiliaires sont observées aux sites de prédiction. Plus généralement, on ne suppose rien sur les sites d'observation des variables. Ils peuvent coïncider entre variables (isotopie), coïncider partiellement (hétérotopie partielle), ou enfin ne pas coïncider du tout (hétérotopie totale).

Il faut donc modéliser la variable d'intérêt et les variables explicatives comme étant aléatoires.

L'objectif ici n'est pas de détailler la procédure de cokrigeage mais de montrer comment la modélisation d'une variable spatialisée s'étend aux cas où l'on travaille avec plusieurs variables. Soient $X_1, \ldots, X_p$, p variables mesurées sur un domaine $\mathcal{D}$.

On travaille dans le cadre des champs aléatoires multivariés. Notons :

$$W(.) = \left\{ W(s) = \begin{pmatrix} X_1(s) \\ \vdots \\ X_p(s) \end{pmatrix}, s \in \mathcal{D} \right\}$$

On peut commencer par faire les hypothèses de stationnarité, d'ergodicité et d'isotropie pour chacun des champs aléatoires $X_i(.), i = 1, \ldots, p$ sous-jacents au champ multivarié $W(.)$ mais des hypothèses supplémentaires sont nécessaires pour le lien entre les différents champs. L'hypothèse de stationarité d'ordre 2 s'étend facilement au cas multivariable. $W(.)$ est dit faiblement stationnaire si

- $\forall i \in [\![1, p]\!]$, $X_i(.)$ est faiblement stationnaire,
- $\forall (i, j) \in [\![1, p]\!]^2$, $\forall (s, s + \overrightarrow{h}) \in \mathcal{D}^2$,

$$\mathrm{Cov}[X_i(s), X_j(s + \overrightarrow{h})] = C_{ij}(\overrightarrow{h})$$

Les fonctions C_{ij} sont appellées fonctions de covariances croisées.

Elles peuvent prendre des formes plus générales que les fonctions de covariance directe. En effet, ce ne sont pas forcément des fonctions paires ($C_{ij}(\overrightarrow{h})$ peut être différent de $C_{ij}(-\overrightarrow{h})$) et leur

maximum n'est pas obligatoirement atteint pour $\overrightarrow{h} = \overrightarrow{0}$[9]. Notons enfin que C_{ij} peut également être différent de C_{ji} pour $i \neq j$.

Ces nouvelles libertés doivent être relativisées. Pour que le modèle de $W(.)$ soit valide (dans le sens où toute matrice de covariance issue de ce modèle doit être définie-positive), les fonctions C_{ij} $i, j = 1, \ldots, p$ doivent respecter certaines contraintes, les unes par rapport aux autres. Un des objectifs de la géostatistique multivariée est donc de construire $\mathbf{C}(\overrightarrow{h})$, la fonction matricielle dont le $(i,j)^{\text{ème}}$ élément est $C_{ij}(\overrightarrow{h})$, de telle sorte que le modèle soit bien défini.

La géostatistique multivariée traditionnelle propose une classe de modèles, simples à construire et qui sont valides sous certaines conditions facilement vérifiables. Ces modèles portent le nom de modèles linéaires de corégionalisation (voir par exemple Journel et Huigbregts, 1978).Souvent utilisés dans une optique de cokrigeage, ils peuvent également servir de base à une analyse factorielle krigeante, qui est une généralisation au contexte spatial, de l'analyse en composantes principales.

Avant de présenter cette classe de modèles, nous en présentons un des éléments constitutifs, le modèle de corrélation intrinsèque.

2.3.1 Modèle de corrélation intrinsèque

Le modèle de corrélation intrinsèque suppose que les corrélations des variables mises en jeu sont indépendantes de leurs corrélations spatiales (Wackernagel, 2003). Dans ce cas la fonction matricielle $\mathbf{C}(h)$ s'écrit

$$\mathbf{C}(h) = T.\rho(h) \tag{2.19}$$

où T est une matrice de taille $p \times p$ et ρ est une fonction de corrélation éventuellement paramètrée.

Détaillons le cas $p = 2$ et notons X et Y les deux variables dont on modélise la relation.

Soient $\{s_1, \ldots, s_n\}$ un ensemble de sites. Sous l'hypothèse de corrélation intrinsèque, la matrice de covariance du vecteur $\mathbf{W} = (X(s_1), \ldots, X(s_n), Y(s_1), \ldots, Y(s_n))'$ s'écrit

$$\Sigma = T \otimes H,$$

où

$$T = \begin{pmatrix} \text{Var}((X(s)) & \text{Cov}(X(s), Y(s)) \\ \text{Cov}(X(s), Y(s)) & \text{Var}(Y(s)) \end{pmatrix},$$

et

$$H_{ij} = \rho(||s_i - s_j||).$$

Notons que par stationnarité des champs $X(.)$ et $Y(.)$, T ne dépend pas de s. On notera alors

$$T = \begin{pmatrix} \sigma_X^2 & \sigma_X \sigma_Y r \\ \sigma_X \sigma_Y r & \sigma_Y^2 \end{pmatrix}.$$

Ce modèle est valide dès lors que la matrice T est définie-positive, ce qui équivaut à $|r| < 1$ avec cette dernière paramétrisation.

Ce modèle sera parfois appelé modèle séparable ou factorisable ou encore modèle à covariance proportionnelle.

[9] C'est souvent le cas pour les séries temporelles, la valeur d'une variable pouvant avoir de l'influence sur le comportement d'une autre variable mais avec "un effet retard".

2.3.2 Modèle linéaire de corégionalisation

Pour construire un modèle linéaire de corégionalisation entre p variables $Z_1, \ldots, Z_p$, on se fixe q fonctions de corrélation $\rho_1, \ldots, \rho_q$ (avec en générale $q \leq p$) et la fonction matricielle $\mathbf{C}(h)$ s'écrit

$$\mathbf{C}(h) = \sum_{i=1}^{q} T_i.\rho_i(h) \tag{2.20}$$

Les matrices T_i sont appelées matrices de corégionalisation.

Le terme "linéaire" dans le nom modèle linéaire de corégionalisation prend son origine dans le fait que l'on peut décomposer linéairement les variables $Z_i, i = 1, \ldots, p$ en combinant des variables aléatoires multivariées $\mathbf{Y}_k = (Y_{1k}, \ldots, Y_{pk})', k = 1, \ldots, q$ (les corégionalisations) dont les composantes ont une structure commune (ρ_k) et toutes les variables $Y_{ij}, i = 1, \ldots, p, j = 1, \ldots, q$ sont orthogonales (Chilès et Delfiner, 1999).

En effet, si

$$\forall s \in \mathcal{D}, \mathbf{Z}(s) = \begin{pmatrix} Z_1(s) \\ \vdots \\ Z_p(s) \end{pmatrix} = \sum_{k=1}^{q} A_k \mathbf{Y}_k(s) \tag{2.21}$$

avec

$$T_k = A_k A_k', k = 1, \ldots, q, \tag{2.22}$$

alors (2.20) est bien vérifiée.

Les décompositions des matrices B_k données par l'équation (2.22) ne sont pas uniques (il y en a même une infinité). Mais celles données par la décomposition en valeurs singulières permettent une adaptation de l'analyse en composantes principales au contexte spatial. Cette adaptation, connue sous le nom d'analyse factorielle krigeante, se décompose en plusieurs étapes comme suit :

1) Choisir les fonction ρ_k à partir des variogrammes expérimentaux et des fonctions de covariance croisées estimées.

2) Estimer les paramètres des fonctions ρ_k par moindres carrés à partir des variogrammes expérimentaux.

3) Estimer les matrices T_k par moindres carrés (sous contrainte que ces matrices restent définie-positives) en utilisant l'algorithme proposé par Goulard et Voltz (1992).

4) En déduire les matrices A_k.

5) Cartographier les corégionalisations Y_{ij}, par une procédure de krigeage.

Les deux dernières étapes sont spécifiques à l'analyse factorielle krigeante, les 3 premières sont suffisantes si l'on travaille uniquement dans une optique de cokrigeage.

Nous référons à l'article de Monestiez *et al.* (1994) pour les interprétations que l'on peut faire des résultats d'une telle analyse, dans un problème de génétique des populations.

2.3.3 Cokrigeage et autokrigeabilité

Comme dans la procédure de krigeage, une fois le modèle choisi et estimé, on le suppose connu et on calcule le prédicteur de $Z(s_0)$ à partir des observations des différentes variables sélectionnées. Ce prédicteur est un prédicteur linéaire sans-biais de variance minimale et s'obtient par la même procédure que les krigeages présentés dans la section précédente.

La variance de prédiction est toujours plus faible pour le cokrigeage que pour le krigeage (sous l'hypothèse que le modèle estimé est le bon !) ce qui est assez naturel puisque de l'information supplémentaire est utilisée. Notons cependant que dans certains cas krigeage et cokrigeage sont équivalents (le cokrigeage n'apporte rien au krigeage) :

Supposons que l'on veuille prédire Z en utilisant $X_1, \ldots, X_{p-1}$, $p-1$ variables auxiliaires. On suppose que les p variables sont observées aux mêmes sites, mais aucune n'est observée au site de prédiction. Supposons également que ces p variables sont issues d'un champ multivarié faiblement stationnaire et isotrope, alors le cokrigeage de $Z(s_0)$, (en utilisant les $p-1$ variables auxiliaires) est équivalent au krigeage de $Z(s_0)$ si et seulement si (voir Wackernagel, 2003) :

$$\forall k \in [\![1, p-1]\!], \forall (s_i, s_j) \in \mathcal{D}^2, \exists \alpha_k \in \mathbb{R}; \mathrm{Cov}(X_k(s_i), Z(s_j)) = \alpha_k \mathrm{Cov}(X_k(s_i), X_k(s_j)).$$

On dira alors que Z est autokrigeable par rapport aux variables $X_1, \ldots, X_{p-1}$.

Notons enfin que dans le cas d'un modèle de corrélation intrinsèque, toutes les variables sont autokrigeables par rapport aux autres variables.

3

Modèles

L'objectif de ce chapitre est de montrer comment rattacher les modèles classiques des statistiques (tels que le modèle de régression linéaire, les modèles linéaires généralisés et les modèles linéaires généralisés mixtes) au contexte des données géoréférencées. Il s'agit de permettre la prise en compte dans ces modèles des deux composantes du contexte spatial évoquées en introduction, à savoir les dépendances spatiales et l'hétérotopie des observations.

Par souci de concision, tous les modèles présentés dans ce chapitre ne mettent en jeu qu'une seule variable explicative. On peut facilement introduire des variables explicatives supplémentaires dans l'espérance de la variable d'intérêt à condition de travailler conditionnellement à ces variables, ce qui est naturel si ces variables sont observées en tout site où la variable d'intérêt est observée. C'est par exemple le cas pour une fonction de la position, comme une dérive externe linéaire.

3.1 Construction d'une famille de modèles linéaires spatiaux

Dans cette section, nous traitons le cas du modèle de régression linéaire adapté au contexte spatial. Nous montrons comment prendre en compte les corrélations spatiales et comment permettre le traitement des jeux de données hétérotopes. Le processus de construction des modèles par modélisation hiérarchique crée une classe de modèles peu traitée dans la littérature (à l'exception de Royle et Berliner, 1999) qui est un intermédiaire entre le modèle de corrélation intrinsèque (souvent considéré comme pauvre) et le modèle linéaire de corrégionalisation (plus riche mais difficile à inférer). La relation entre la variable d'intérêt et la variable explicative induite par le procédé de construction est à mettre en relation avec l'autokrigeabilité définie dans le chapitre précédent. Mais nous montrons que dans le cas gaussien, l'autokrigeabilité est équivalente à une relation spatiale particulière entre les deux variables, et que cette relation est caractéristique de cette nouvelle classe de modèles, parmi les modèles linéaires de corrégionalisation.

3.1.1 Le modèle de régression linéaire classique, une référence non-spatiale

Pour fixer les notations, on rappelle les deux manières de formuler le modèle de régression linéaire classique que l'on nommera par la suite **(LM1)** :

Soit $\mathbf{Y} = (Y_1, \ldots, Y_n)'$ un vecteur aléatoire d'intérêt et $\mathbf{X} = (X_1, \ldots, X_n)'$ un vecteur aléatoire explicatif :

1 ère formulation du modèle : **(LM1)**

La distribution de $\mathbf{Y}$ sachant $\mathbf{X}$ est donnée par :

$$\mathbf{Y}|\mathbf{X} \sim \mathcal{N}(\beta_0 \mathbb{1}_n + \beta_1 \mathbf{X}, \sigma^2_{Y|X} I_n).$$

où I_n désigne la matrice identité d'ordre n et $\mathbb{1}'_n = (1, \ldots, 1)$.

On utilise souvent la formulation équivalente suivante :

2 ème formulation du modèle : **(LM1)**

Soit ε un vecteur gaussien d'espérance nulle, de matrice de covariance $\sigma^2_{Y|X} I_n$ et indépendant de $\mathbf{X}$. On a :

$$\mathbf{Y} = \beta_0 \mathbb{1}_n + \beta_1 \mathbf{X} + \varepsilon.$$

Comme en général, le vecteur $\mathbf{X}$ est entièrement renseigné, on omet en général son caractère stochastique. Ce qui conduit, dans la première formulation à ne plus utiliser l'écriture conditionnelle ; et dans la deuxième formulation, à ne plus expliciter l'hypothèse d'indépendance des résidus avec la variable explicative.

3.1.2 Modifications liées au contexte spatial

La prise en compte du contexte spatial va nous conduire à modifier le modèle précédent dans deux directions distinctes.

D'une part, il sera naturel d'étendre l'hypothèse d'indépendance des résidus pour permettre la prise en compte de corrélations spatiales éventuelles.

D'autre part, le traitement de jeux de données hétérotopes nécessitera de considérer également la variable explicative comme aléatoire et de modéliser sa structure spatiale.

La prise en compte de ces deux spécificités du contexte spatial implique donc deux actions contradictoires en terme de modélisation. La première généralise le modèle linéaire tandis que la seconde le particularise par l'ajout d'une hypothèse supplémentaire.

1er Axe : Corrélation spatiale des résidus

La mise en évidence de corrélations spatiales (par le variogramme expérimental) dans la variable Y n'impose pas forcément leur présence dans les résidus. En effet, les corrélations spatiales de Y peuvent en partie être induites par celles de la variable explicative X du fait de la relation entre ces deux variables.

La raison qui conduit le modélisateur à considérer les résidus comme potentiellement spatialement corrélés, est le plus souvent la non prise en compte de variables explicatives spatialement corrélées. En effet, le terme ε a pour but de capturer dans le modèle les erreurs de mesure (mutuellement indépendantes dans la plupart des cas) et les effets de variables non prises en compte[1].

Le modèle de regression linéaire à résidus spatialement corrélés dû à Mardia et Marshall (1984) s'écrit :

[1] Notons également que les erreurs de modélisation, comme une non linéarité non prise en compte dans le modèle, pourraient induire de la corrélation spatiale dans les résidus estimés, si la variable explicative est elle même spatialement structurée.

1 ère formulation du modèle : **(LM2)**

Soit ρ_ϕ une fonction de corrélation valide paramétrée par ϕ.
Pour toute configuration de sites $\{s_1, \ldots, s_n\}$ de $\mathcal{D}$,
si $\mathbf{Y} = (Y(s_1), \ldots, Y(s_n))'$ et $\mathbf{X} = (X(s_1), \ldots, X(s_n))'$, alors la distribution de $\mathbf{Y}|\mathbf{X}$ s'écrit :

$$\mathbf{Y}|\mathbf{X} \sim \mathcal{N}(\beta_0 \mathbb{1}_n + \beta_1 \mathbf{X}, \sigma^2_{Y|X} H(\phi)),$$

où le $(i,j)^{\text{ème}}$ terme de $H(\phi)$ est donné par $\rho_\phi(||s_i - s_j||)$.

Comme pour le modèle linéaire classique, le modèle linéaire à résidus spatialement corrélés peut se reformuler comme suit :

2 ème formulation du modèle : **(LM2)**

Soit ε un vecteur gaussien d'espérance nulle, de matrice de covariance $\sigma^2_{Y|X} H(\phi)$ et indépendant de $\mathbf{X}$. On a :

$$\mathbf{Y} = \beta_0 \mathbb{1}_n + \beta_1 \mathbf{X} + \varepsilon.$$

A partir de cette formulation, il apparaît que le modèle **LM2** est très proche du modèle utilisé pour le krigeage universel (voir 2.18). Les hypothèse de stationnarité d'ordre 2 et d'isotropie ont été remplacées ici par l'hypothèse que les résidus sont la réalisation d'un champ gaussien stationnaire et isotrope.

2ème Axe : Extension naturelle du modèle classique pour échantillonnages hétérotopes

Dans les sous sections précédentes, il n'était pas nécessaire de considérer la variable explicative comme stochastique. Dans le cas données hétérotopes, le fait que les sites d'observation de X et de Y ne coïncident pas, impose que la variable explicative X soit suffisamment structurée spatialement, au regard de l'échantillonnage des deux variables. La prise en compte de cette structure spatiale nécessite de considérer X comme la réalisation d'un champ aléatoire $X(.)$.

On peut étendre naturellement le modèle linéaire classique en supposant que $X(.)$ est un champ gaussien stationnaire :

Formulation du modèle : **(LM3)**

Pour toute configuration de sites $\{s_1, \ldots, s_n\}$ de $\mathcal{D}$, on note $\mathbf{Y} = (Y(s_1), \ldots, Y(s_n))'$ et $\mathbf{X} = (X(s_1), \ldots, X(s_n))'$. Soit ρ_ϕ une fonction de corrélation paramétrée par ϕ.
On pose :

$$\mathbf{X} \sim \mathcal{N}(\mu_X \mathbb{1}_n, \sigma^2_X H(\phi))$$

où le $(i,j)^{\text{ème}}$ terme de $H(\phi)$ est donné par $\rho_\phi(||s_i - s_j||)$, et

$$\mathbf{Y}|\mathbf{X} \sim \mathcal{N}(\beta_0 \mathbb{1}_n + \beta_1 \mathbf{X}, \sigma^2_{Y|X} I_n),$$

3.1.3 A la croisée des chemins

De la même manière que l'on a généralisé le modèle linéaire classique **(LM1)** en permettant la prise en compte des corrélations spatiales des résidus (pour obtenir **(LM2)**), on peut généraliser le modèle **(LM3)** [2] (voir figure (3.1)) :

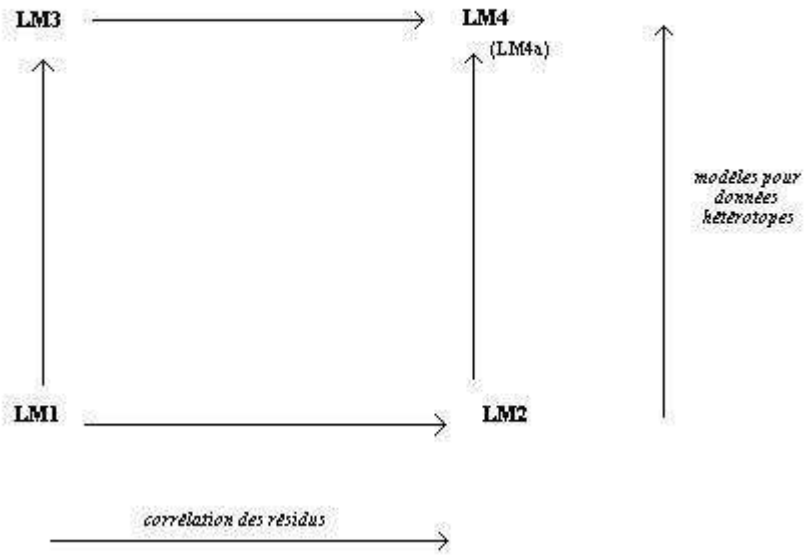

Fig. 3.1. Situations relatives des différents modèles.

1ère formulation du modèle : **(LM4)**

Pour toute configuration de sites $\{s_1, \ldots, s_n\}$ de $\mathcal{D}$, on note $\mathbf{Y} = (Y(s_1), \ldots, Y(s_n))'$ et $\mathbf{X} = (X(s_1), \ldots, X(s_n))'$.
Soit $\rho^1_{\phi_1}$ une fonction de corrélation paramétrée par ϕ_1.
On pose $\mathbf{X} \sim \mathcal{N}(\mu_X \mathbb{1}_n, \sigma^2_X H_1(\phi_1))$ où le $(i,j)^{\text{ème}}$ terme de $H_1(\phi_1)$ est donné par $\rho^1_{\phi_1}(||s_i - s_j||)$.
Soit $\rho^2_{\phi_2}$ une fonction de corrélation paramétrée par ϕ_2. On pose :

$$\mathbf{Y}|\mathbf{X} \sim \mathcal{N}(\beta_0 \mathbb{1}_n + \beta_1 \mathbf{X}, \sigma^2_{Y|X} H_2(\phi_2)),$$

où le $(i,j)^{\text{ème}}$ terme de $H_2(\phi_2)$ est donné par $\rho^2_{\phi_2}(||s_i - s_j||)$.

Comme pour le modèle de régression linéaire classique et pour le modèle linéaire à résidus corrélés, on peut reformuler le modèle **(LM4)** de la manière suivante :

[2] Cette généralisation peut également être considérée comme la spécification de la distribution de $\mathbf{X}$ dans le modèle **(LM2)** .

2ème formulation du modèle : **(LM4)**

Pour toute configuration de sites $\{s_1, \ldots, s_n\}$ de $\mathcal{D}$, $\mathbf{X} \sim \mathcal{N}(\mu_X \mathbb{1}_n, H_1(\phi_1))$
Soit $\varepsilon = (\varepsilon(s_1), \ldots, \varepsilon(s_n))'$ un vecteur gaussien d'espérance nulle de matrice de covariance $\sigma^2_{Y|X} H_2(\phi)$ et indépendant de $\mathbf{X}$. On a :

$$\mathbf{Y} = \beta_0 \mathbb{1}_n + \beta_1 \mathbf{X} + \varepsilon.$$

La variable Y s'écrit donc comme la somme d'une fonction linéaire de la variable explicative X et d'un vecteur de résidus spatialement corrélés. Ces résidus peuvent être vus, par exemple, comme la somme des effets sur Y d'autres variables, elles-mêmes spatialisées, non prises-en-compte. La distribution de ce vecteur résiduel étant donnée pour toute configuration de sites, il peut être vu comme la réalisation d'un champ aléatoire d'effets non-observés. On utilisera parfois le terme de **champ caché** pour désigner ce processus sous-jacent.

3.2 Propriétés des modèles LM4

3.2.1 Formulation des modèles (LM4) par les distributions multivariées

Une manière équivalente d'écrire les modèles **(LM4)** est de donner les distributions multivariées qui permettent d'écrire les vraisemblances. C'est à dire que pour toute configuration de sites $\{s_1, \ldots, s_n\}$ de $\mathcal{D}$, si $\mathbf{Y} = (Y(s_1), \ldots, Y(s_n))'$ et $\mathbf{X} = (X(s_1), \ldots, X(s_n))'$, on définit le modèle en donnant la distribution de $\mathbf{W} = (\mathbf{X}', \mathbf{Y}')'$.

On peut montrer que **(LM4)** est équivalent à :

$$\mathbf{W} \sim \mathcal{N}(\mu, \Sigma)$$

$$\text{avec } \mu = \begin{pmatrix} \mu_X \\ \beta_0 + \beta_1 \mu_X \end{pmatrix} \otimes \mathbb{1}_n$$

$$\Sigma = \begin{pmatrix} \sigma^2_X H_1(\phi_1) & \sigma^2_X \beta_1 H_1(\phi_1) \\ \sigma^2_X \beta_1 H_1(\phi_1) & \sigma^2_X \beta_1^2 H_1(\phi_1) + \sigma^2_{Y|X} H_2(\phi_2) \end{pmatrix}. \tag{3.1}$$

Dans le cas du modèle **LM3**, la matrice de variance s'écrit

$$\Sigma = \begin{pmatrix} \sigma^2_X H(\phi) & \sigma^2_X \beta_1 H(\phi) \\ \sigma^2_X \beta_1 H(\phi) & \sigma^2_X \beta_1^2 H(\phi) + \sigma^2_{Y|X} I_n \end{pmatrix}.$$

Cette formulation nous permet de rattacher les modèles **(LM4)** aux modèles de la géostatistique multivariée présentés au chapitre 2.

3.2.2 Positionnement des modèles (LM4) par rapport aux modèles de la géostatistique multivariée

Le modèle de corrélation intrinsèque, un cas particulier parcimonieux

Si par souci de parcimonie, on choisit dans le modèle **(LM4)** un modèle commun d'autocorrélation pour la variable explicative et les résidus, c'est à dire $\rho^1_{\phi_1} \equiv \rho^2_{\phi_2}$ et si de plus, on suppose

que $\phi_1 = \phi_2$, alors on a $H_1(\phi_1) = H_2(\phi_2)$. La matrice de variance donnée par (3.1) est donc identique à celle du modèle de corrélation intrinsèque qui s'écrit

$$\Sigma = T \otimes H(\phi).$$

Peu utilisé par les géostatisticiens du fait de sa pauvreté en terme de modélisation des structures spatiales croisées, il est souvent choisi en statistiques inférentielles pour généraliser les modèles classiques au contexte spatial (voir par exemple Banerjee et Gelfand (2002), Pascual et Zhang (2005), Banerjee *et al* (2005)) et permet d'étudier les propriétés des estimateurs dans ce contexte.

Le modèle **(LM4)** est plus flexible que le modèle de corrélation intrinsèque (que nous nommerons parfois **(LM4a)**) puisqu'il permet d'intégrer des structures spatiales différentes pour la variable explicative et la variable d'intérêt, bien que par construction, le modèle de covariance de la variable d'intérêt soit un modèle gigogne à deux structures l'une d'elle étant celle de la variable explicative.

L'argument suivant permet de comprendre la flexibilité apportée par cette nouvelle classe de modèles par rapport au modèle de corrélation intrinsèque.

Si la variable explicative est sujette à une erreur de mesure, alors la variable d'intérêt doit l'être aussi. De plus, la part de variabilité de ces erreurs de mesure dans les variances totales de chacune des variables doit être la même. C'est l'une des restrictions majeures imposée par le choix du modèle de corrélation intrinsèque. Par des choix judicieux des deux fonctions de covariance dans le modèle **(LM4)**, on peut modéliser les erreurs de mesures pour les deux variables de telle sorte qu'elles représentent chacune une part différente des variances totales. Mais ce n'est pas le seul gain apporté par ce modèle. Les structures spatiales des deux variables peuvent avoir des portées différentes, ce qui permet d'intégrer dans une modélisation commune, des phénomènes structurés à différentes échelles. Notons néanmoins que par construction, la portée de la variable d'intérêt est supérieure ou égale à celle de la variable explicative si la liaison statistique entre les deux variables est non nulle (i.e $\beta_1 \neq 0$).

Notons que le modèle **(LM4)** doit être préféré à condition que l'on conjecture l'existence d'un lien de cause à effet entre les deux variables mises en jeu, car son point de vue et sa forme ne sont pas symétriques en les variables.

Du modèle LM4 au modèle linéaire de corégionalisation (LMC)

En géostatistique multivariée traditionnelle, on généralise le modèle de corrélation intrinsèque en utilisant le modèle LMC (voir chapitre 2).

Le modèle **(LM4)** est un sous modèle du modèle LMC. En effet, la matrice de covariance du vecteur $\mathbf{W}$ s'écrit :

$$\Sigma = T_1 \otimes H_1(\phi_1) + T_2 \otimes H_2(\phi_2)$$

avec $T_1 = \begin{pmatrix} \sigma_X^2 & \beta_1\sigma_X^2 \\ \beta_1\sigma_X^2 & \beta_1^2\sigma_X^2 \end{pmatrix}$ et $T_2 = \begin{pmatrix} 0 & 0 \\ 0 & \sigma_{Y|X}^2 \end{pmatrix}$.

Les contraintes sur les deux matrices T_1 et T_2 font du modèle **(LM4)** un sous modèle des LMC. En effet, il possède 3 paramètres de moins que ce dernier.

De la même manière que nous avons tenté d'expliquer le gain qu'apportait l'utilisation du modèle **(LM4)** par rapport au modèle de corrélation intrinsèque, nous avons étudié la nature des liaisons bivariées qu'induit l'utilisation du modèle **(LM4)** par rapport au LMC en terme de modélisation de processus spatiaux. Cette étude a permis de mettre en évidence une propriété caractéristique des **(LM4)** que nous présentons dans le paragraphe suivant.

3.2.3 Caractérisation des modèles (LM4) parmi les modèles LMC

La construction du modèle **(LM4)** par l'approche conditionnelle implique que le bloc matriciel $\text{cov}(\mathbf{X}, \mathbf{Y})$ est proportionnel au bloc $\text{cov}(\mathbf{X}, \mathbf{X})$, ce qui est équivalent au fait que Y est autokrigeable par rapport à la variable X (voir chapitre 2).

Dans le cas gaussien, on peut traduire l'autokrigeabilité par des relations d'indépendance conditionnelle à l'aide de la proposition suivante :

Proposition :
Si $W(.) = \left\{ W(s) = \begin{pmatrix} X(s) \\ Y(s) \end{pmatrix}, s \in \mathcal{D} \right\}$ est un champ bivarié gaussien stationnaire, alors

$$\forall s \in \mathcal{D}, \forall \overrightarrow{h},$$
$$\text{cov}(X(s), Y(s + \overrightarrow{h})) = \beta \text{cov}(X(s), X(s + \overrightarrow{h})) \quad (3.2)$$
$$\Leftrightarrow \forall (s_i, s_j) \in \mathcal{D}^2, Y(s_i) \perp\!\!\!\perp X(s_j) | X(s_i).$$

Démonstration. Voir annexe 1.

Si pour un site s on connait $X(s)$, alors $Y(s)$ ne dépend plus de $X(s')$, quel que soit s'. La seule relation existant entre $Y(s)$ et $X(s')$ provient du fait que $Y(s)$ dépend de $X(s)$ et que $X(s)$ et $X(s')$ sont spatialement corrélés.

Cette nature particulière des relations entre les deux variables induite par l'autokrigeabilité dans le cas gaussien sera nommée propriété de dépendance ponctuelle. Cette propriété est vérifiée par le modèle **(LM4)** mais tout modèle LMC qui vérifie cette propriété peut également s'écrire sous la forme d'un modèle **(LM4)**.

Théorème 1 *Soit un modèle LMC général entre deux variables, c'est à dire dont la matrice de covariance s'écrit :*

$$\Sigma = T_1 \otimes H_1(\phi_1) + T_2 \otimes H_2(\phi_2).$$

avec $T_1 = \begin{pmatrix} a & b \\ b & c \end{pmatrix}$ *et* $T_2 = \begin{pmatrix} d & e \\ e & f \end{pmatrix}$*, où les matrices* $H_i(\phi_i)$ *sont des matrices de corrélations spatiales associées respectivement aux fonctions de corrélations spatiales* $\rho^i(\phi_i), i = 1, 2$.

Le modèle est un modèle **(LM4)** *si et seulement si* $b/a = e/d$.

Démonstration. Seule la réciproque reste à démontrer. Sa preuve est triviale. Il s'agit juste d'une reparamétrisation. On peut vérifier que les paramètres du modèle **(LM4)** défini à partir d'un modèle LMC par la relation $b/a = e/d$ s'écrivent :

- $\beta_1 = b/a$
- $\sigma^2_X = a + d$
- $\sigma^2_{Y|X} = c - \beta_1(a + d) + f$

Et les 2 fonctions de corrélations spatiales du modèle **(LM4)** s'écrivent :

$$\tilde{\rho}^1_{\tilde{\phi}_1} = \alpha_1 \rho^1_{\phi_1}(h) + (1 - \alpha_1)\rho^2_{\phi_2}(h)$$

et

$$\tilde{\rho}^1_{\tilde{\phi}_2} = \alpha_2 \rho^1_{\phi_1}(h) + (1 - \alpha_2)\rho^2_{\phi_2}(h)$$

avec

$$\alpha_1 = \frac{a}{a+d}$$

et

$$\alpha_2 = \frac{c - \beta_1^2(a+d)}{c - \beta_1^2(a+d) + f}.$$

$$\tilde{\phi}_i = (\phi_i, \alpha_i), i = 1, 2.$$

Ainsi, tout modèle de la classe LMC tel que la variable d'intérêt soit autokrigeable par rapport à la variable explicative peut s'écrire sous la forme d'un modèle **(LM4)**.

Nous avons ainsi construit une classe de modèles intermédiaires entre le modèle de corrélation intrinsèque et le modèle linéaire de corrégionalisation. Les modèles **(LM4)** autorisent de travailler avec deux structures (différentes pour chacune des variables) en restant très parcimonieux. Ils permettent également une grande flexibilité par un jeu sur la complexification (combinaisons ou mélanges) des deux fonctions de corrélation spatiale $\rho^1_{\phi_1}$ et $\rho^2_{\phi_2}$.

3.2.4 Jouons un peu avec deux structures...

Pour clore cette section sur les modèles gaussiens, nous allons montrer comment, à partir d'un modèle LMC basé sur deux fonctions de corrélations les plus simples possibles, on peut extraire différents sous-modèles (certains appartenant à la classe **(LM4)**) ayant chacun des niveaux d'interprétation différents.

L'exemple est basé sur un modèle de corrélation continu à l'origine (par exemple le modèle exponentiel) et un modèle aléatoire pur (effet de pépite).

Soit une variable d'intérêt Y liée linéairement à une variable X de la manière suivante :

$$Y = \beta X + \varepsilon$$

avec ε indépendant de X.

On suppose que X est la somme de deux variables indépendantes, la première X_p étant structurée spatialement à très petite échelle, (et donc modélisée par un bruit blanc), et la seconde X_g est structurée à plus grande échelle (structure modélisée par la fonction de covariance ρ) :

$$X = X_p + X_g.$$

Le processus d'observation soumet X à une erreur de mesure de telle sorte que l'on observe

$$X_o = X + X_e$$

où X_e est également un bruit blanc, indépendant de X.

Le résidu ε est également la somme indépendante d'une variable pépitique ε_p et d'une variable spatialement structurée ε_g :

$$\varepsilon = \varepsilon_p + \varepsilon_g.$$

Pour simplifier, on suppose que ε_g a la même fonction de covariance ρ que X.

On suppose que spatialement, toutes les variables définies ci-dessus définissent des champs gaussiens stationnaires, isotropes, d'espérance nulle et on notera pour leur variance σ^2 indicée par le nom de la variable.

Pour toutes configurations de sites $\{s_1, \ldots, s_n\}$, le modèle multivarié des observations résultant de la construction est le suivant :

$$\mathbf{W} = (X_o(s_1), \ldots, X_o(s_n), Y(s_1), \ldots, Y(s_n))' \sim \mathcal{N}(\mathbf{0}, \Sigma)$$

avec

$$\Sigma = T_1 \otimes I_n + T_2 \otimes H,$$

où

$$T_1 = \begin{pmatrix} \sigma^2_{X_p} + \sigma^2_{X_e} & \beta\sigma^2_{X_p} \\ \beta\sigma^2_{X_p} & \beta^2\sigma^2_{X_p} + \sigma^2_{\varepsilon_p} \end{pmatrix} \text{ et } T_2 = \begin{pmatrix} \sigma^2_{X_g} & \beta\sigma^2_{X_g} \\ \beta\sigma^2_{X_g} & \beta^2\sigma^2_{X_g} + \sigma^2_{\varepsilon_g} \end{pmatrix}$$

et

$$H_{ij} = \rho(||s_i - s_j||).$$

Le modèle résultant est donc un modèle LMC général à deux structures puisqu'aucune contrainte autre que la définie-positivité, n'est imposée aux matrices T_1 et T_2.

Le procédé de construction permet de donner un sens aux sous-modèles éventuels.

Reparamétrisons les deux matrices T_1 et T_2 de la manière suivante :

$$T_i = \begin{pmatrix} a_i & b_i \\ b_i & c_i \end{pmatrix}, i = 1, 2.$$

On peut par exemple considérer les sous-modèles suivants :

- $a_2 = b_2 = c_2 = 0$: ce choix supprime la composante spatiale du modèle. Il s'agit du modèle bigaussien utilisé pour l'inférence du coefficient de corrélation dans le cas i.i.d.
- $b_1 = 0$: les effets de pépite des deux variables ne sont plus corrélés. Dans l'esprit de la construction du modèle, l'effet de pépite de la variable explicative observée X_o est uniquement de l'erreur de mesure, c'est à dire $X_p = 0$. Ce modèle n'est pas dans la classe des **(LM4)**. Il reste cependant un sous-modèle simple des LMC déjà considéré dans Banerjee et Gelfand (2002).
- $b_1/a_1 = b_2/a_2$: c'est la propriété de dépendance ponctuelle. L'effet de pépite de X peut être vu comme une composante intrinsèque de la variable explicative X_o puisque sa part dans la variabilité totale de X_o est la même que celle dans la covariance croisée entre les deux variables, c'est à dire $X_e = 0$.
- $c_2/b_2 = b_2/a_2$: le résidu ε n'est pas structuré spatialement, c'est à dire $\varepsilon_g = 0$. On est dans le cas du modèle **(LM3)**.
- $b_1/a_1 = b_2/a_2$ et $c_1/b_1 = c_2/b_2$: c'est un modèle de type **(LM4a)** qui n'a pas vraiment de sens particulier pour l'utilisateur (la part de la pépite dans la variance de X_o est identique à la part de la pépite dans ε) et n'a d'intérêt que par sa parcimonie.
- ...

Ces considèrations permettent d'envisager différentes hypothèses que l'utilisateur pourra supposer s'il a des connaissances préalables, ou au contraire tester si elles traduisent une partie de ses questions sur les phénomènes étudiés. Il dispose pour cela de l'arsenal de méthodes de comparaison de modèles (par exemple un test LRT si les deux modèles considérés sont emboîtés, BIC ou AIC sinon).

Notons que pour un jeu de données totalement hétérotope, b_1 n'est pas estimable (il n'apparaît pas dans la vraisemblance) et le modèle LMC n'est par conséquent pas identifiable. En hétérotopie totale, on devra donc faire une hypothèse de nullité de b_1 ou de proportionalité ($b_1/a_1 = b_2/a_2$). Ce choix doit être guidé par les considérations évoquées ci-dessus.

3.3 Variable d'intérêt non-gaussienne

Pour certains types de données comme les variables binaires ou de comptage, le cadre gaussien devient clairement inapproprié. Comment écrire les modèles dans ces cas ?

Deux alternatives issues de la tradition des statistiques classiques s'offrent à nous :

1. Remplacer la distribution gaussienne par une autre distribution de la famille exponentielle et généraliser les modèles précédents de la même manière que le Modèle Linéaire Généralisé (GLM) généralise le modèle linéaire **(LM1)** (McCullagh et Nelder, 1989).

2. Hiérarchiser les modèles précédents à l'aide de distribution de la famille exponentielle. La variable Y des modèles précédents devient ici une variable latente. Cette approche conduit à une classe de modèles qui sont des adaptations au contexte spatial des modèles linéaires généralisés mixtes (GLMM) (Mac Culloch et Searle,).

Cette section explicite ces deux alternatives.

3.3.1 Le modèle linéaire généralisé

Le modèle linéaire se généralise en remplaçant la distribution gaussienne par une distribution de la famille exponentielle.

Formulation classique : **(GLM1)**

Soit $\mathbf{Z} = (Z_1, \ldots, Z_n)'$ un vecteur aléatoire d'intérêt et $\mathbf{X} = (X_1, \ldots, X_n)'$ un vecteur explicatif :
on suppose les composantes de $\mathbf{Z}$ mutuellement indépendantes et pour tout $i = 1, \ldots, n$:

$$Z_i | X_i \sim \mathcal{L}(\gamma_i)$$

où $\mathcal{L}$ est une loi de la famille exponentielle paramètrée par un vecteur γ_i.
Le lien entre Z_i et X_i est spécifié par une relation du type :

$$E[Z_i|X_i] = g^{-1}(\beta_0 + \beta_1 X_i) \tag{3.3}$$

où g est une fonction de lien.

Citons 2 exemples parmi les plus utilisés :
- $Z_i|X_i \sim \mathcal{P}oisson(\lambda_i)$ avec $\lambda_i = \exp(\beta_0 + \beta_1 X_i)$
- $Z_i|X_i \sim \mathcal{B}ernouilli(p_i)$ avec $p_i = \text{logit}^{-1}(\beta_0 + \beta_1 X_i)$

Corrélation des données dans le cadre du modèle linéaire généralisé

Contrairement au cas gaussien, la généralisation de l'hypothèse d'indépendance des observations dans le cas des modèles linéaires généralisés n'est pas aisée. Rappelons quelques caractéristiques qui font du contexte gaussien un cadre particulièrement adapté à la modélisation de données spatiales :

1) la donnée des deux premiers moments caractérise entièrement la distribution multivariée,

2) les composantes d'un vecteur gaussien sont indépendantes si elles ne sont pas corrélées,

3) si un vecteur aléatoire est multigaussien, les lois marginales d'un ensemble de composantes de ce vecteur, ainsi que les lois d'un ensemble des composantes conditionnellement à un autre ensemble de composantes, restent multivariées gaussiennes, et leurs paramètres sont facilement calculables à partir des paramètres du vecteur initial.

Ces trois propriétés font partie des multiples raisons qui expliquent le succès du cadre gaussien dans la modélisation de données géoréférencées. Elles permettent également d'utiliser les fonctions de covariance, pour généraliser le modèle linéaire au contexte spatial.

Qu'en est-il dans le cas non-gaussien ? La spécification d'un modèle **(GLM2)** pour des données corrélées spatialement reste un problème ouvert.

Le principal verrou pour généraliser les distributions de la famille exponentielle au contexte spatial sont les relations entre espérance et variance qui ne permettent pas de spécifier indépendamment les deux premiers moments, excepté dans le cas gaussien. Existe-t-il par exemple des distributions multivariées vérifiant des spécifications sur l'espérance telles que celle de l'équation (3.3), dont la matrice de covariance est déduite d'un modèle de covariance spatiale et dont les lois marginales sont binômiales ou poissoniennes ?

Quelques auteurs (Kaiser et Cressie, 1997, par exemple) proposent des modèles pour des données multivariées de comptage mais se limitent à des modèles sur grilles (ou réseaux). Ils utilisent des relations de type markovien entre voisins, c'est à dire qu'ils définissent en chaque site la distribution des variables conditionnellement aux voisins sans avoir à définir la distribution multivariée ou les distributions marginales en chaque site. La plupart des auteurs ayant voulu modéliser des dépendances spatiales pour des données de comptage, ont préféré la voie hiérarchique, lorsque ces dépendances n'étaient pas dues aux structures spatiales des variables explicatives ; ce que nous étudierons plus loin en développant la seconde alternative.

Extension naturelle des GLM pour le cas des données hétérotopes

Dans le cas où seules des dépendances spatiales au niveau de la variable explicative sont prises en compte dans le modèle, il est possible d'étendre le modèle **(GLM1)** de la même manière que le modèle **(LM3)** étend le modèle **(LM1)**.

Il suffit d'ajouter au GLM, une hypothèse distributionnelle pour la variable explicative :

Formulation du modèle : **(GLM3)**

Soit ρ_ϕ une fonction de corrélation paramétrée par ϕ.
Pour toute configuration de sites $\{s_1, \ldots, s_n\}$ de $\mathcal{D}$, on note $\mathbf{Z} = (Z(s_1), \ldots, Z(s_n))'$ et $\mathbf{X} = (X(s_1), \ldots, X(s_n))'$ et on pose :
$\mathbf{X} \sim \mathcal{N}(\mu_X \mathbb{1}_n, \sigma_X^2 H(\phi))$ où le $(i,j)^{\text{ème}}$ terme de $H(\phi)$ est donné par $\rho_\phi(||s_i - s_j||)$.
Conditionnellement à $\mathbf{X}$, les composantes de $\mathbf{Z}$ sont mutuellement indépendantes et pour tout $i = 1, \ldots, n$,

$$Z(s_i)|X(s_i) \sim \mathcal{L}(\gamma(s_i)),$$

avec $E[Z(s_i)|X(s_i)] = g^{-1}(\beta_0 + \beta_1 X(s_i))$.

Cette approche ne permet pas de modéliser des corrélations spatiales dans la variable d'intérêt autres que celles initialement présentes dans les variables explicatives, ce qui limite l'intérêt d'un tel modèle pour l'application.

3.3.2 Hiérarchisation des modèles gaussiens, vers une forme de modèles linéaires généralisés mixtes

Pour modéliser la relation entre une variable explicative que l'on peut supposer gaussienne, et une variable d'intérêt pour laquelle cette hypothèse est inappropriée, la seconde alternative est d'introduire un champ de variables latentes. Cette idée, formalisée par Diggle *et al* (1998) dans l'article Model-Based-Geostatistics, est une adaptation au contexte géostatistique des modèles linéaires généralisés mixtes.

Après un bref rappel de l'écriture de ces modèles, nous montrons comment les modèles gaussiens présentés ci-dessus peuvent se généraliser par hiérarchisation aux données pour lesquelles l'hypothèse gaussienne n'est pas adaptée.

Modèles GLMM

Dans les modèles linéaires généralisés mixtes (GLMM), 2 sources d'aléa sont modélisées. Au premier niveau de hiérarchie, on modélise les effets aléatoires associés aux modalités d'un des facteurs potentiellement explicatif, puis connaissant la réalisation de ces effets, on modélise la variable réponse dans une loi de la famille exponentielle.

Parmi la classe très vaste des GLMM, une utilisation courante consiste à modéliser la relation entre une variable d'intérêt Y, une variable explicative X et une variable catégorielle A. Dans ce cas le modèle s'écrit :

Formulation du modèle : **(GLMM)**

Soit $\mathbf{U}$ un vecteur aléatoire (de taille égale au nombre de modalités de A) vérifiant

$$\mathbf{U} \sim \mathcal{N}(\mathbf{0}, \sigma^2 I)$$

Pour toute réalisation possible $\mathbf{u}$ du vecteur $\mathbf{U}$, on suppose que conditionnellement à $\mathbf{U} = \mathbf{u}$, les composantes de $\mathbf{Y}$ sont mutuellement indépendantes et vérifient :

$$Y_i | X_i = x_i, \mathbf{U} = \mathbf{u} \sim \mathcal{L}(\gamma_i)$$

On note $\mu_i = E[Y_i | X_i = x_i, \mathbf{U} = \mathbf{u}]$
Le lien entre Y_i et la variable explicative et les facteurs aléatoires est donnée par une relation du type :

$$g(\mu_i) = x_i'\beta + a_i'\mathbf{u}$$

pour une fonction de lien g.

Pour simplifier, on remplacera la notation de l'évènement $\mathbf{U} = \mathbf{u}$ par $\mathbf{U}$
Le lecteur trouvera plus de détails sur les GLMM dans l'ouvrage de Mc Cullogh et Searle (2001).

GLMM dans le contexte spatial

Dans le contexte spatial, ce que l'on appelle effets aléatoires n'est en général plus attaché à un facteur mais aux sites d'observations. Dans ce cas, le modèle **(GLMM1)** consisterait à avoir un effet aléatoire par site d'observation. Pour prendre en compte le contexte spatial, on modélise les effets à partir d'un champ aléatoire de la même manière que les résidus dans le cas du modèle linéaire à résidus spatialement corrélés **(LM2)**. Puis conditionnellement à ce champ, on suppose

que les observations de la variable d'intérêt sont mutuellement indépendantes et on donne leur distribution conditionnellement à la réalisation du champ en chaque site, comme une loi de la famille exponentielle. Le champ capture toutes les dépendances spatiales de la variable d'intérêt autres que celles qui sont dues à la variable explicative.

Ce modèle s'écrit :

1ère formulation du modèle : **(GLMM2)**

Pour toute configuration de sites $\{s_1, \ldots, s_n\}$ de $\mathcal{D}$, on note $\mathbf{Z} = (Z(s_1), \ldots, Z(s_n))'$ le vecteur d'observations de la variable d'intérêt et $\mathbf{X} = (X(s_1), \ldots, X(s_n))'$ celui de la variable explicative.

On introduit $\mathbf{Y} = (Y(s_1), \ldots, Y(s_n))'$ un vecteur aléatoire latent tel que :

$$\mathbf{Y}|\mathbf{X} \sim \mathcal{N}(\beta_0 \mathbb{1}_n + \beta_1 \mathbf{X}, \sigma_Y^2 H(\phi))$$

où le $(i,j)^{\text{ème}}$ terme de $H(\phi)$ est donné par $\rho_\phi(||s_i - s_j||)$ pour une fonction de corrélation ρ_ϕ.

Conditionnellement à $\mathbf{Y}$, les $Z(s_i), i = 1, \ldots, n$ sont mutuellement indépendants, et de distributions conditionnelles :

$$Z(s_i)|Y(s_i) \sim \mathcal{L}(\gamma(s_i))$$

où $\mathcal{L}$ est une loi de la famille exponentielle paramètrée par $\gamma(s_i)$.

Le lien entre $Z(s_i)$ et $Y(s_i)$ est spécifié par une relation du type :

$$E[Z(s_i)|Y(s_i)] = g^{-1}(Y(s_i))$$

pour une fonction de lien donnée g.

Ce modèle est équivalent à celui de Diggle *et al* (1998) dont l'écriture suivante est la plus utilisée :

2ème formulation du modèle : **(GLMM2)**

Pour toute configuration de sites $\{s_1, \ldots, s_n\}$ de $\mathcal{D}$, on note $\varepsilon = (\varepsilon(s_1), \ldots, \varepsilon(s_n))'$ et on pose :

$\varepsilon \sim \mathcal{N}(\mathbf{0}, \sigma_\varepsilon^2 H(\phi))$ où le $(i,j)^{\text{ème}}$ terme de $H(\phi)$ est donné par $\rho_\phi(||s_i - s_j||)$.

$\mathbf{X}$ et ε sont indépendants et conditionnellement à $\mathbf{X}$ et ε, les composantes de $\mathbf{Z}$ sont mutuellement indépendantes et de distributions conditionnelles

$$Z(s_i)|X(s_i), \varepsilon(s_i) \sim \mathcal{L}(\gamma(s_i))$$

avec

$$E[Z(s_i)|X(s_i), \varepsilon(s_i)] = g^{-1}(\beta_0 + \beta_1 X(s_i) + \varepsilon(s_i)),$$

où g est une fonction de lien.

Sur la base de ce modèle et de la construction des modèles **(LM3)** et **(LM4)** , on peut construire une classe de nouveaux modèles **(GLMM4)** (incluant un modèle **(GLMM3)**) afin par exemple de traiter les jeux de données hétérotopes. La formulation générale est la suivante :

Formulation générale du modèle : **(GLMM4)**

Pour toute configuration de sites $\{s_1, \ldots, s_n\}$ de $\mathcal{D}$, on note $\mathbf{Z} = (Z(s_1), \ldots, Z(s_n))'$ le vecteur d'observations de la variable d'intérêt et $\mathbf{X} = (X(s_1), \ldots, X(s_n))'$ celui de la variable explicative.

On pose : $\mathbf{X} \sim \mathcal{N}(\mu_X \mathbb{1}_n, \sigma_X^2 H(\phi))$ où le $(i,j)^{\text{ème}}$ terme de $H(\phi)$ est donné par $\rho_\phi(||s_i - s_j||)$.

On introduit $\mathbf{Y} = (Y(s_1), \ldots, Y(s_n))'$ un vecteur aléatoire latent tel que :

$$\mathbf{Y}|\mathbf{X} \sim \mathcal{N}(\beta_0 \mathbb{1}_n + \beta_1 \mathbf{X}, \Sigma)$$

pour une matrice de covariance Σ dont la modélisation sera discutée dans la suite. Conditionnellement à $\mathbf{Y}$, les $Z(s_i), i = 1, \ldots, n$ sont mutuellement indépendants, et de distributions conditionnelles :

$$Z(s_i)|Y(s_i) \sim \mathcal{L}(\gamma(s_i))$$

où $\mathcal{L}$ est une loi de la famille exponentielle paramètrée par $\gamma(s_i)$.

Le lien entre $Z(s_i)$ et $Y(s_i)$ est spécifié par une relation du type :

$$E[Z(s_i)|Y(s_i)] = g^{-1}(Y(s_i))$$

pour une fonction de lien donnée g.

Les différents modèles à dépendance ponctuelle décrits dans le cadre gaussien peuvent être utilisés pour modéliser, au travers de Σ, la relation entre $\mathbf{Y}$ et $\mathbf{X}$. Rappelons les différentes alternatives :

- si l'on souhaite que la seule part de structure spatiale modélisée dans les données proviennent de la structure de la variable X, on choisira $\Sigma = \sigma^2_{Y|X} I_n$ **(GLMM3)** , le cas $\sigma^2_{Y|X} = 0$ correspondant au modèle **(GLM3)** ,

- si l'on veut que $\mathbf{X}$ et $\mathbf{Y}$ soient en corrélation intrinsèque, on choisira $\Sigma = \sigma^2_{Y|X} H(\phi)$ **(GLMM4a)** ,

-si l'on veut utiliser un modèle à dependance ponctuelle plus général, on choisira $\Sigma = \sigma^2_{Y|X} H_2(\phi_2)$ où $H_2(\phi_2)$ est associée à une fonction de corrélation paramètrée par ϕ_2 **(GLMM4)** .

Ainsi l'ensemble des modèles à dépendance ponctuelle décrits dans le cas gaussien peut être généralisé aux cas où la variable d'intérêt est non gaussienne par la hiérarchisation décrite ci-dessus (voir figure 3.3.2).

Un modèle proche : le modèle spatial de régression probit

Pour expliquer une variable d'intérêt binaire Z par une variable explicative continue X, dans un contexte d'hétérotopie, Banerjee et Gelfand (2002) utilisent un modèle hiérarchique très proche d'un modèle de type **GLMM4a**. En effet ils introduisent une variable latente Y dont ils modélisent la relation avec X en utilisant un champ bivarié gaussien en corrélation intrinsèque. Puis, ils considèrent que conditionnellement à Y les $Z(s_i)$ sont mutuellement indépendants et :

$$Z(s_i) = 1 \Leftrightarrow Y(s_i) > 0.$$

Ce modèle revient à écrire que conditionnellement à ce champ de variables latentes, les observations binaires sont définies (elles ne sont plus aléatoires).

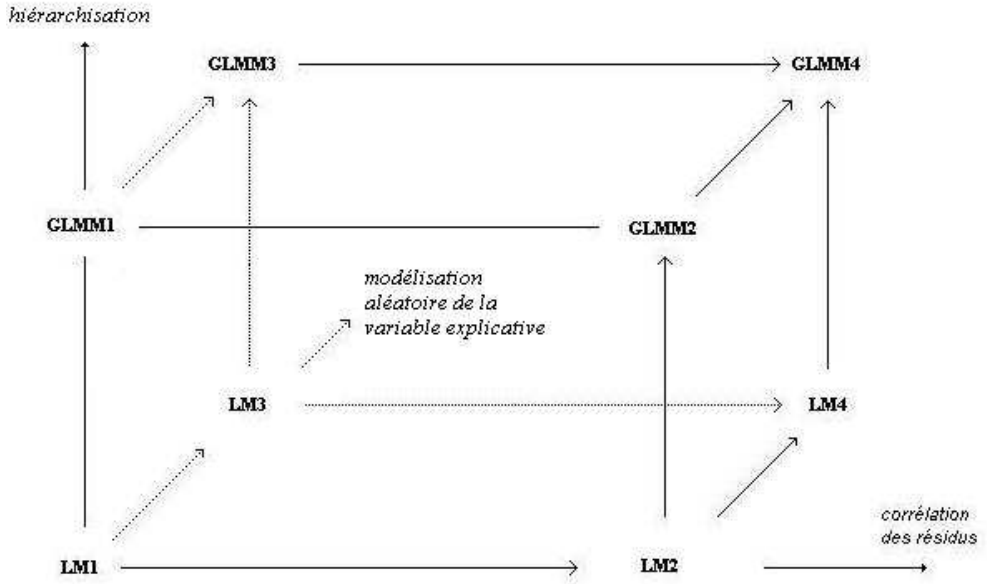

Fig. 3.2. Situations relatives des différents modèles.

La distribution marginale $Z(s_i)$ est une loi de Bernoulli de paramètre $(p(s_i))$, avec

$$p(s_i) = \text{probit}^{-1}(Y(s_i)).$$

Ce modèle ne doit pas être confondu avec son apparenté de la classe **(GLMM4a)**donné par :

$$Z(s_i)|Y(s_i) \sim \mathcal{B}ernouilli(p(s_i)),$$

avec $p(s_i) = \text{probit}^{-1}(Y(s_i))$, le modèle de liaison entre X et Y restant inchangé.

La différence fondamentale entre les deux modèles tient au fait que dans l'un, la distribution de la variable d'intérêt est définie conditionnellement à une variable latente alors que dans l'autre cette variable d'intérêt est entièrement caractérisée conditionnellement à la variable latente (elle n'est plus aléatoire).

Extension possible au LMC

On pourrait envisager de modéliser le lien entre la variable latente et la variable explicative en utilisant un modèle linéaire de corrégionalisation (LMC hiérarchique).

Mais l'inférence d'un tel modèle semble délicate en l'absence d'une quantité importante de données. En effet les modèles LMC décrivent des relations complexes entre les variables et dans un modèle hiérarchique on n'observe qu'une version très bruitée du champ latent (notamment pour des données binaires ou de comptage à faibles effectifs).

3.4 Modèles hiérarchiques bivariés

De nombreux jeux de données en écologie concernent la présence (ou l'absence) de deux espèces ou le nombre de chaque espèce en différents sites et les écologues s'intéressent aux interactions entre

les deux espèces sur la base de ces données non-gaussiennes. Dans l'esprit de la construction des modèles **(GLMM4)** , on peut construire une nouvelle classe de modèles en hiérarchisant deux champs latents, un par variable d'intérêt, ces champs latents étant conjointement modélisés.

Notons $Z_1(s)$ et $Z_2(s)$ les variables aléatoires correspondant aux observations au site s et $Y_1(s)$ et $Y_2(s)$ deux variables latentes au site s.

On modélise $Y_1(s)$ et $Y_2(s)$ comme un champ aléatoire bivarié comme dans la section 1. Puis on suppose que pour tout ensemble de sites $\{s_1, \ldots, s_n\}$, les variables $Z_j(s_i), i = 1, \ldots, n, j = 1, 2$ sont mutuellement indépendantes et de distributions conditionnelles dans une loi de la famille exponentielle avec :

$$E[Z_j(s_i)|Y_j(s_i)] = g^{-1}(Y_j(s_i)).$$

Notons que ce n'est pas forcément l'hétérotopie qui conduit à utiliser ce type de modèle. Le fait que les deux variables soient non gaussiennes impose de modéliser des champs aléatoires latents pour chacune d'entre elle, de la même manière que nous avons utilisé un champ latent pour modéliser les structures spatiales d'une variable d'intérêt non gaussienne.

3.5 Modélisations alternatives pour des données spatiales

Une alternative à la modélisation géostatistique, fréquemment rencontrée dans la littérature, et permettant de traiter des données spatialement corrélées est l'utilisation de champs markoviens. Cette approche fait l'hypothèse que les données sont observées sur un réseau (*lattice*), sur lequel est définie une notion de voisinage qui, dans une certaine mesure remplace la distance euclidienne ; les "corrélations spatiales" sont définies à partir de relations de type markoviennes entre les voisins précédemment définis. Citons par exemple les modèles conditionnellement autoregressifs (CAR) pour des champs Markoviens gaussiens mais également les modèles autologistiques, souvent utilisés en analyse d'images (Besag *et al.*,1991) , et qui généralisent les champs gaussiens markoviens (cas auto-normal).

Ce cadre de modélisation est particulièrement adapté lorsque les données sont observées sur des cellules ou des blocs, tels qu'une unité administrative (voir figure (3.5)). En effet, dans ces cas la notion de voisinage (par exemple deux cellules sont voisines si elles partagent une frontière commune) est souvent plus pertinente que l'utilisation de la distance euclidienne. Lorsque les données sont observées en des sites ponctuels la définition du voisinage opère obligatoirement un seuillage des distances et par conséquent induit une perte d'information sur les dépendances spatiales des observations par rapport à l'utilisation de la distance euclidienne.[3] Nous nous restreindrons donc dans ce travail aux modèles issus de la géostatistique pour des données à support continu.

[3] Notons toutefois que dans le cas gaussien, pour un échantillonnage régulier, dans le cas d'un modèle hiérarchique de type **GLMM2** , Hrafnkelsson et Cressie (2003) ont montré que l'approche géostatistique et l'approche utilisant des champs markoviens pouvaient mener à des inférences presque identiques en ce qui concerne la prédiction des effets aléatoires, l'approche par champs markoviens étant même beaucoup moins coûteuse numériquement pour l'estimation des paramètres.

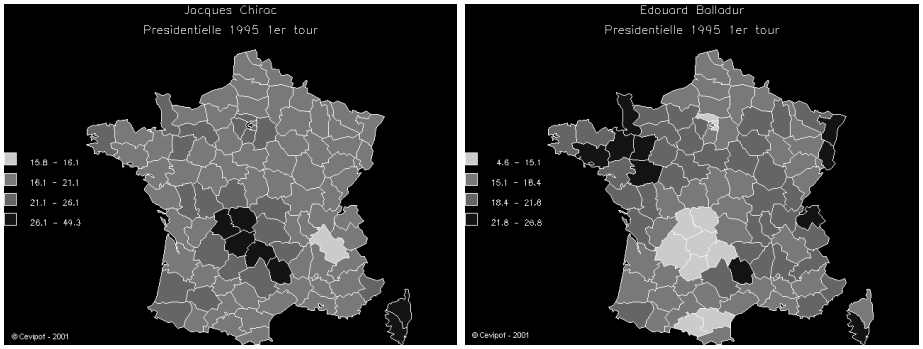

Fig. 3.3. Pourcentage des suffrages de deux candidats à l'élection présidentielle de 1995 par départements

4

Estimation

Dans ce chapitre, nous expliquons comment estimer les différents modèles présentés dans le chapitre précédent. Le chapitre comporte trois sections. Dans la première, nous expliquons comment estimer les modèles gaussiens. Dans la seconde, nous rappelons quelques résultats concernant les modèles GLM non spatiaux et leur estimation. Enfin dans la troisième section, nous présentons les outils existants pour estimer les modèles GLMM par maximum de vraisemblance ; et nous proposons les adaptations de ces outils aux modèles GLMM spatiaux présentés au chapitre précédent.

4.1 Modèles multivariés gaussiens

Les modèles gaussiens présentés dans le chapitre précédent peuvent tous s'écrire sous la forme

$$\mathbf{Y} \sim \mathcal{N}(X'\beta, \Sigma(\theta)) \tag{4.1}$$

où β et θ sont des vecteurs de paramètres.[1]

Notons $\mathbf{Y}_c = \mathbf{Y} - X'\beta$ la variable $\mathbf{Y}$ centrée.

La log-vraisemblance s'écrit à une constante additive près :

$$L(\beta, \theta, \mathbf{Y}) = -\frac{1}{2}\log|\Sigma(\theta)| - \frac{1}{2}\mathbf{Y}'_c\Sigma^{-1}(\theta)\mathbf{Y}_c \tag{4.2}$$

La forme de la vraisemblance ne permet pas d'obtenir analytiquement le vecteur $\hat{\psi} = (\hat{\beta}, \hat{\theta})$ qui maximise $L(\beta, \theta, \mathbf{Y})$. On doit donc utiliser une approche numérique. Nous présentons deux algorithmes d'optimisation parmi les plus fréquemment utilisés en statistiques : l'algorithme de Newton-Raphson (dont l'utilisation ne se cantonne pas aux statistiques) et le Fisher scoring (encore appelé algorithme des scores de Fisher). Puis nous proposons un algorithme hybride qui utilise les bonnes propriétés de convergence spécifiques à chacun de ces deux algorithmes. Enfin nous montrons comment les utiliser pour maximiser la log-vraisemblance (4.2).

[1] Pour éviter toute confusion par rapport au chapitre précédent, nous précisons qu'ici $\mathbf{Y}$ est le vecteur contenant toutes les variables aléatoires et X est le vecteur des covariables, supposées non-aléatoires. Ainsi le vecteur $\mathbf{Y}$ est à mettre en relation avec le vecteur $(\mathbf{X}', \mathbf{Y}')$ du chapitre précédent, la matrice X avec la matrice

$$\begin{pmatrix} \mathbb{1}_n & \mathbf{0} \\ \mathbf{0} & \mathbb{1}_n \end{pmatrix}$$

et β avec le vecteur (μ_X, μ_Y). Nous changeons ici le formalisme pour atteindre un niveau de généralité supplémentaire.

4.1.1 Algorithme de Newton-Raphson

Soit $\mathbf{y}$ un vecteur d'observations supposé être une réalisation d'un vecteur aléatoire $\mathbf{Y}$ de distribution $\mathcal{L}(\theta)$ où $\theta \in \Theta \subset \mathbb{R}^p$ est un vecteur de paramètres.

En pratique, on obtient généralement $\hat{\theta}$ en résolvant le système suivant :

$$\frac{\partial L}{\partial \psi_i}(\psi; \mathbf{y}) = 0, i = 1, \ldots, p. \tag{4.3}$$

Lorsque le système (4.3) ne peut être résolu analytiquement, on peut mettre en place un algorithme itératif permettant d'approcher numériquement la solution.

Sous l'hypothèse que la fonction de vraisemblance (ou la log-vraisemblance) est concave et unimodale, la séquence d'itérations définie comme suit converge vers le maximum de vraisemblance :

Soit $\psi^{(0)} \in \Theta$,

$$\psi^{(k+1)} = \psi^{(k)} - H_L^{-1}(\psi^{(k)}; \mathbf{y}) \nabla L(\psi^{(k)}; \mathbf{y})$$

où

$$\nabla L(\psi^{(k)}; \mathbf{y})$$

est le gradient de L évalué en $\psi^{(k)}$ dont le $i^{\text{ème}}$ terme est défini par :

$$\frac{\partial L}{\partial \psi_i}(\psi; \mathbf{y})\Big|_{\psi=\psi^{(k)}}$$

et

$$H_L(\psi^{(k)}; \mathbf{y})$$

est la matrice hessienne de L dont le $(i,j)^{\text{ème}}$ terme est défini par :

$$\frac{\partial^2 L}{\partial \psi_i \partial \psi_j}(\psi; \mathbf{y})\Big|_{\psi=\psi^{(k)}}$$

On parlera parfois d'information observée pour désigner la quantité $-H_L(\psi; \mathbf{y})$.

4.1.2 Fisher scoring

Lorsque la fonction à maximiser est une vraisemblance ou une log-vraisemblance, il est courant d'utiliser une version légèrement modifiée de l'algorithme de Newton-Raphson : le Fisher scoring. Il s'obtient du précédent en remplaçant l'information observée par l'information de Fisher $I(\psi)$ définie par :

$$I(\psi) = -E[H_L(\psi; \mathbf{Y})].$$

Cette modification entraîne des simplifications de calcul fort appréciables, sans doute à l'origine du succès du Fisher scoring en statistiques. Cependant cette modification de l'algorithme n'est pas anodine et les comportements des deux algorithmes peuvent être différents. Comme le relève Knight (2000), nous avons noté que l'algorithme de Fisher scoring était plus robuste au choix des valeurs initiales que le Newton-Raphson mais lorsque les deux convergent, ce dernier est plus rapide.

L'utilisation intensive que nous faisons par la suite des algorithmes d'estimation (simulations, bootstrap paramètrique) nécessite des outils performants, combinant à la fois robustesse et bonne vitesse de convergence. Nous avons donc construit un algorithme hybride entre le Fisher-scoring et l'algorithme de Newton-Raphson. Le principe est d'utiliser le Fisher scoring durant les premières

itérations puis l'algorithme de Newton-Raphson pour atteindre l'optimum. Le nombre d'itérations à partir duquel le changement d'algorithme se fait est basé sur quelques essais. Une approche moins brutale ou basée sur un critère plus objectif que le nombre d'itérations pourrait être envisagée pour décider à partir de quand changer d'algorithme mais notre approche donne des résultats très satisfaisants relativement à nos objectifs.

4.1.3 Calcul du gradient, de la matrice hessienne et de l'information de Fisher dans le cadre gaussien

En utilisant des résultats de calcul matriciel et notamment de dérivation matricielle (voir par exemple Mardia et Marshall, 1984), on montre que :

$$\frac{\partial L}{\partial \beta}(\beta, \theta, \mathbf{Y}) = -X'\Sigma(\theta)^{-1}X\beta + X'\Sigma(\theta)^{-1}\mathbf{Y} \tag{4.4}$$

$$\frac{\partial L}{\partial \theta_i}(\beta, \theta, \mathbf{Y}) = -\frac{1}{2}\mathrm{tr}(\Sigma(\theta)^{-1}\Sigma(\theta)_i) - \frac{1}{2}\mathbf{Y}'_c\Sigma(\theta)^{(i)}\mathbf{Y}_c \tag{4.5}$$

où

$$\Sigma(\theta)_i = \frac{\partial \Sigma(\theta)}{\partial \theta_i}$$

est la matrice dont le $(i,j)^{\text{ème}}$ terme est donné par la dérivée par rapport à θ_i du $(i,j)^{\text{ème}}$ terme de $\Sigma(\theta)$ et

$$\Sigma(\theta)^{(i)} = \frac{\partial \Sigma^{-1}(\theta)}{\partial \theta_i} = -\Sigma^{-1}(\theta)\Sigma(\theta)_i\Sigma^{-1}(\theta).$$

De même pour les éléments de la matrice hessienne, on montre que :

$$\frac{\partial^2 L}{\partial \beta \partial \beta'}(\beta, \theta, \mathbf{Y}) = -X'\Sigma(\theta)^{-1}X$$

$$\frac{\partial^2 L}{\partial \beta \partial \theta_i}(\beta, \theta, \mathbf{Y}) = -X'\Sigma(\theta)^{(i)}X\beta + X'\Sigma(\theta)^{(i)}\mathbf{Y}$$

$$\frac{\partial^2 L}{\partial \theta_i \partial \theta_j}(\beta, \theta, \mathbf{Y}) = -\frac{1}{2}\mathrm{tr}(\Sigma(\theta)^{-1}\Sigma(\theta)_{ij} + \Sigma(\theta)^{(i)}\Sigma(\theta)_j) - \frac{1}{2}\mathbf{Y}'_c\Sigma(\theta)^{(ij)}\mathbf{Y}_c$$

où

$$\Sigma(\theta)_{ij} = \frac{\partial^2 \Sigma(\theta)}{\partial \theta_i \partial \theta_j}$$

est la matrice dont le $(i,j)^{\text{ème}}$ terme est donné par la dérivée seconde par rapport à θ_i et θ_j du $(i,j)^{\text{ème}}$ terme de $\Sigma(\theta)$ et

$$\Sigma(\theta)^{(ij)} = \frac{\partial^2 \Sigma^{-1}(\theta)}{\partial \theta_i \partial \theta_j} = \Sigma^{-1}(\theta)(\Sigma(\theta)_i\Sigma^{-1}(\theta)\Sigma(\theta)_j + \Sigma(\theta)_j\Sigma^{-1}(\theta)\Sigma(\theta)_i - \Sigma(\theta)_{ij})\Sigma^{-1}(\theta).$$

Enfin, les éléments de la matrice d'information de Fisher s'obtiennent en calculant l'opposé des espérances des quantités précédentes, ce qui conduit aux simplifications suivantes :

$$-E\left[\frac{\partial^2 L}{\partial \beta \partial \beta'}(\beta, \theta, \mathbf{Y})\right] = X'\Sigma(\theta)^{-1}X$$

$$-E\left[\frac{\partial^2 L}{\partial\beta\partial\theta_i}(\beta,\theta,\mathbf{Y})\right]=\mathbf{0}$$

$$-E\left[\frac{\partial^2 L}{\partial\theta_i\partial\theta_j}(\beta,\theta,\mathbf{Y})\right]=\frac{1}{2}\mathrm{tr}(\Sigma(\theta)^{-1}\Sigma(\theta)_i\Sigma(\theta)^{-1}\Sigma(\theta)_j).$$

4.1.4 Difficultés numériques

Bien que la matrice d'information de Fisher soit théoriquement définie-positive, il peut arriver, à cause des erreurs d'arrondis, qu'elle soit singulière (une de ses valeurs propres est nulle) ou même que l'une de ses valeurs propres soit négative. Cela se produit souvent quand la quantité d'information apportée par les données sur les paramètres est faible (la vraisemblance est plate dans certaines directions) ou quand le paramètre courant $\psi^{(k)}$ est loin du maximum (les algorithmes sont en effet d'autant plus efficaces que $\psi^{(k)}$ est proche de la solution car l'approximation quadratique sur laquelle est basé l'algorithme de Newton-Raphson est d'autant plus juste que l'on se situe proche du mode). Quand cela se produit, on choisit la nouvelle valeur $\psi^{(k+1)}$ en maximisant la vraisemblance le long du gradient :

$$\psi^{(k+1)}=\psi^{(k)}+\hat{\delta}\nabla L(\psi^{(k)};\mathbf{y})$$

où

$$\hat{\delta}=\mathrm{argmax}_{\delta\in\mathbb{R}^+}L(\psi^{(k)}+\delta\nabla L(\psi^{(k)};\mathbf{y})$$

est obtenu par une recherche dorée (*golden search*).

Il peut également arriver que $\psi^{(k+1)}$ n'appartienne plus à Θ. Dans ce cas, on divise le pas par 2 et on repart de $\psi^{(k)}$ jusqu'à ce que $\psi^{(k+1)}$ soit correct. On peut également changer la paramétrisation pour certains paramètres comme suggéré par Zhu et Stein (2005) afin que leur espace soit $\mathbb{R}$ tout entier. Par exemple, pour le modèle de corrélation intrinsèque, nous avons choisi d'écrire la matrice de covariance entre les deux variables d'un même site T de la manière suivante :

$$T=\begin{pmatrix}\sigma_X^2 & \sigma_X\sigma_Y\tanh(r^\star)\\ \sigma_X\sigma_Y\tanh(r^\star) & \sigma_Y^2\end{pmatrix}.$$

De cette façon, T est définie-positive pour tout $r^\star$ de $\mathbb{R}$.

Ces choix ne sont peut-être pas optimaux mais il resulte un algorithme très efficace.

4.1.5 Approche par profile-vraisemblance

Dans certains cas, un sous-système du système (4.3) peut être résolu analytiquement. Il est alors commode de maximiser numériquement la vraisemblance en fonction des paramètres restants, le long de la solution analytique. En effet, la dimension de l'espace des paramètres sur lequel la fonction objectif est maximisée, est réduite. Ainsi, la convergence est plus rapide et certaines difficultés inhérentes à l'optimisation numérique de systèmes à grande dimension sont diminuées.

Pour le modèle général (4.1), on peut par exemple résoudre l'équation (vectorielle) (4.4) et obtenir, pour θ fixé le maximum de vraisemblance pour β :

$$\hat{\beta}(\theta)=(X'\Sigma(\theta)^{-1}X)^{-1}X'\Sigma(\theta)^{-1}\mathbf{Y}.$$

On remplace alors β par $\hat{\beta}$ dans l'expression de la log-vraisemblance et on maximise alors la profile-vraisemblance :

$$L_p(\theta; \mathbf{Y}) = L(\hat{\beta}(\theta), \theta; \mathbf{Y})$$

en fonction de θ.

Pour certains des modèles présentés dans le chapitre précédent, on peut résoudre analytiquement un sous-système encore plus important. Pour le modèle de corrélation intrinsèque, Pascual et Zhang (2006) utilisent la séparabilité de la matrice de covariance dans le contexte où les variables sont observées aux mêmes sites pour résoudre analytiquement le système des équations normales associées à tous les paramètres exceptés ceux intervenant dans la fonction de corrélation spatiale. Pour le modèle **(LM4)** qui s'écrit, avec les notations du chapitre 3 :

$$\mathbf{X} \sim \mathcal{N}(\mu_X \mathbb{1}_n, \sigma_X^2 H_1(\theta_1))$$

et

$$\mathbf{Y}|\mathbf{X} \sim \mathcal{N}(\beta_0 \mathbb{1}_n + \beta_1 \mathbf{X}, \sigma_{Y|X}^2 H_2(\theta_2)),$$

on peut écrire la log-vraisemblance comme la somme de deux log-vraisemblances gaussiennes :

$$l(\psi; \mathbf{x}, \mathbf{y}) = l_{Y|X}(\beta_0, \beta_1, \sigma_{Y|X}^2, \theta_1, \theta_2; \mathbf{y}|\mathbf{x}) + l_X(\mu_X, \sigma_X^2, \theta_1; \mathbf{x}).$$

On suppose dans un premier temps que θ_1 et θ_2 sont connus et on obtient :

$$\hat{\mu}_X(\theta_1) = \frac{\mathbb{1}_n' H_1(\theta_1)^{-1} \mathbf{x}}{\mathbb{1}_n' H_1(\theta_1)^{-1} \mathbb{1}_n},$$

qui est l'expression du krigeage de la moyenne des $\mathbf{x}$,

$$\hat{\sigma}_X^2(\theta_1) = \frac{1}{n}(\mathbf{x} - \hat{\mu}_X(\theta_1)\mathbb{1}_n)' H_1(\theta_1)^{-1}(\mathbf{x} - \hat{\mu}_X(\theta_1)\mathbb{1}_n)$$

$$\hat{\beta}(\theta_2) = (\hat{\beta}_0(\theta_2), \hat{\beta}_1(\theta_2)) = (\mathbf{X}' H_2(\theta_2)^{-1} \mathbf{X})^{-1} \mathbf{X}' H_2(\theta_2)^{-1} \mathbf{y}$$

où $\mathbf{X}$ est la matrice $\begin{pmatrix} \mathbb{1}_n & \mathbf{x} \end{pmatrix}$;

$$\hat{\sigma}_{Y|X}^2(\theta_2) = \frac{1}{n}(\mathbf{y} - \mathbf{X}\hat{\beta}'(\theta_2))' H_2(\theta_2)^{-1}(\mathbf{y} - \mathbf{X}\hat{\beta}'(\theta_2))$$

Puis on injecte ces expressions dans les deux log-vraisemblances et on estime les paramètres θ_1 et θ_2 numériquement.

4.2 Estimation des modèles GLM

Avant de présenter en détail l'estimation des modèles GLMM spatiaux présentés dans le chapitre 3, nous donnons quelques éléments sur l'estimation des GLM qui seront utiles par la suite. Le lecteur trouvera plus de détails dans McCulloch et Searle (2001).

4.2.1 Famille exponentielle

Les modèles GLM supposent que la variable d'intérêt est distribuée selon une loi de la famille exponentielle. Nous donnons une définition de cette famille ainsi que quelques propriétés.

Définition

Y est une variable de la famille exponentielle si elle admet une densité f_Y (relativement à une mesure) qui peut s'écrire :

$$f_Y(y) = \exp\left\{\frac{y\gamma - b(\gamma)}{\tau^2} - c(y,\tau)\right\}. \tag{4.6}$$

La plupart des lois usuelles sont issues de la famille exponentielle (lois de Poisson, binômiales, gaussienne, gamma...).

Supposons par exemple que Y soit une variable de Bernoulli, alors la densité de Y (relativement à la mesure de comptage) s'écrit sous la forme (4.6) avec $\gamma = \log\left(\frac{p}{1-p}\right)$, $b(\gamma) = \log(1+e^\gamma)$, $\tau = 1$ et $c \equiv 0$.

Propriétés

Sous certaines conditions de régularité, on a

$$E[Y] = \frac{\partial b(\gamma)}{\partial \gamma},$$

$$\mathrm{Var}[Y] = \tau^2 \frac{\partial^2 b(\gamma)}{\partial \gamma^2}.$$

On note

$$\mu = E[Y]$$

l'espérance de Y et

$$v(\mu) = \frac{\mathrm{Var}[Y]}{\tau^2} = \frac{\partial^2 b(\gamma)}{\partial \gamma^2}$$

la fonction de variance qui donne le lien entre l'espérance et la variance.

Il suit facilement

$$\frac{\partial \gamma}{\partial \mu} = \frac{1}{v(\mu)}. \tag{4.7}$$

4.2.2 Modèles GLM

On rappelle l'écriture des GLM classique.

Soit $\mathbf{Y} = (Y_1, \ldots, Y_n)'$ le vecteur des variables d'intérêt et $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ la matrice de design (matrice des covariables ou matrice d'incidence selon la nature des variables explicatives).

On suppose que les Y_i sont mutuellement independantes et vérifient

$$Y_i \sim \mathcal{L}(\gamma_i)$$

où $\mathcal{L}(\gamma_i)$ est une loi de la famille exponentielle paramètrée par γ_i.

Le lien entre Y_i et $\mathbf{x}_i$ est donné par la relation :

$$E[Y_i] = g^{-1}(\mathbf{x}_i'\beta),$$

pour une fonction de lien g. Notons que si $g \equiv b$, le lien est dit canonique.

4.2.3 Fisher scoring

Notons

$$l(\beta;\mathbf{y}) = \sum_{i=1}^{n}\left(\frac{y_i\gamma_i - b(\gamma_i)}{\tau^2} - c(y_i,\tau)\right)$$

la log-vraisemblance des observations, $\mu_i = E_\beta[Y_i]$ et

$$g_\mu(\mu_i) = \frac{\partial g(\mu_i)}{\partial \mu_i}.$$

En utilisant la relation (4.7) et la règle de dérivation en chaîne, on montre que

$$\frac{\partial l(\beta;\mathbf{y})}{\partial \beta} = \frac{1}{\tau^2}\mathbf{X}'\mathbf{W}\Delta(\mathbf{y}-\mu), \tag{4.8}$$

où

$$\mu = (\mu_1,\ldots,\mu_n),$$

$\mathbf{W}$ est la matrice diagonale dont le $i^{\text{ème}}$ terme diagonal est égal à

$$\frac{1}{v(\mu_i)g_\mu^2(\mu_i)}$$

et Δ est la matrice diagonale dont le $i^{\text{ème}}$ terme diagonal est égal à

$$g_\mu(\mu_i).$$

De même, on peut montrer que

$$E\left[\frac{\partial^2 l(\beta;\mathbf{y})}{\partial\beta\partial\beta'}\right] = -\frac{1}{\tau^2}\mathbf{X}'\mathbf{W}\mathbf{X}. \tag{4.9}$$

La mise à jour des paramètres par le Fisher scoring s'effectue donc à l'aide de la relation :

$$\beta^{(t+1)} = \beta^{(t)} + (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\Delta(\mathbf{y}-\mu).$$

Notons qu'ici, μ est l'espérance du vecteur $\mathbf{Y}$ calculée pour $\beta = \beta^{(t)}$ et par suite, Δ et $\mathbf{W}$ dépendent de t et sont calculées à chaque itération. Enfin, notons que si g est le lien canonique, $\Delta = \mathbf{W}^{-1}$, ce qui entraîne des simplifications dans ce cas.

Cette méthode, utilisiée dans la plupart des logiciels de statistiques[2], est numériquement peu coûteuse et converge très rapidement.

4.3 Modèles hiérarchiques

Nous considérons ici les modèles de la classe **GLMM** de la section précédente ainsi que les modèles obtenus par double hiérarchisation.

Maximiser la vraisemblance dans le cas des modèles hiérarchiques spatiaux est un problème beaucoup plus complexe que dans le cas gaussien. En effet, la vraisemblance des observations n'a plus une forme analytique simple. Elle fait en particulier intervenir des intégrales de grandes dimensions ;

[2] Notons que dans le cas des GLM, le Fisher scoring est souvent présenté d'une autre manière dans les ouvrages de référence. On utilise des variables auxiliaires (*working variables*) que l'on met à jour à chaque itération. Mais les 2 méthodes sont strictement équivalentes.

plus précisément de dimensions égales au nombre d'observations[3]. Ce problème n'est pas spécifique aux GLMM pris dans le contexte spatial (bien que la difficulté soit accrue dans ce cas) ; on y est également confronté lorsque l'on utilise les GLMM classiques, la dimension de l'intégrale étant alors égale aux nombres de modalités de la variable catégorielle pour laquelle on modélise des effets aléatoires (dans le cas où l'on a modélisé un seul facteur à effets aléatoires).

Pour préciser notre propos, considérons un modèle de type GLMM non spatial. Pour bien comprendre la difficulté, nous nous appuyons sur un exemple concret.

Supposons que l'on veuille comparer les effets de différents traitements médicaux en effectuant des essais sur n patients pris dans p hôpitaux. On suspecte que l'hôpital dans lequel est effectué l'essai puisse avoir une influence sur le résultat pour diverses raisons. On s'intéresse à l'effet des traitements et non à l'effet de chaque hôpital mais on veut prendre en compte la part de variabilité induite par l'hôpital sur l'effet du traitement.

Après l'expérience, on observe pour chaque patient si le traitement a eu un effet positif.

Notons y_{ij} la variable qui résume l'effet du traitement (par exemple y_{ij} vaut 1 si le patient i traité dans l'hôpital j a guéri et 0 sinon, on pourait envisager d'autres résumés de l'effet du traitement) et on note x_i le traitement testé sur ce cobaye. Notons que x_i peut être un vecteur. Enfin, on note u_j l'effet non observé de l'hôpital j.

On modélise les observations y_{ij} comme les réalisations du modèle suivant :

Conditionnellement au vecteur des effets aléatoires $\mathbf{U} = (U_1 \dots, U_p)$, les variables Y_{ij} sont indépendantes et distribuées selon une loi de la famille exponentielle $\mathcal{L}(\gamma_i)$ à densité $f_i(y_i|U_j)$.

Le lien entre les facteurs, les effets aléatoires et la variable d'intérêt est donné par une relation du type :

$$E[Y_{ij}] = g^{-1}(x_i'\beta + u_j).$$

Enfin, on suppose que $\mathbf{U}$ est distribué selon une loi à densité f_U de paramètres θ_U. Notons que les U_j sont souvent supposés i.i.d gaussiens, d'espérance nulle et de variance σ^2.

La log-vraisemblance des observations s'écrit

$$l(\psi; \mathbf{y}) = \log \int_{\mathbb{R}^p} \left(\prod_{i=1}^{n} f_i(y_i|u_j) \right) f_U(\mathbf{u}; \theta_U) d\mathbf{u}. \tag{4.10}$$

Le succès des modèles GLMM a conduit de nombreux auteurs à s'intéresser à l'estimation des paramètres de ces modèles (voir l'ouvrage de McCulloch et Searle, 2001). Il existe plusieurs approches dans le contexte non spatial. Citons les principales :

- Dans les cas les plus simples, la dimension de l'espace d'intégration n'est pas trop importante et une intégration numérique (par exemple par la quadrature de Gauss-Hermite) peut être envisagée. On peut alors estimer numériquement les paramètres par maximum de vraisemblance. C'est par exemple le cas quand les effets aléatoires sont indépendants. On doit alors effectuer une intégration numérique par individu statistique.
- On peut simplifier le modèle en choisissant la loi des effets aléatoires de telle sorte que l'intégrale de l'équation (4.10) disparaisse. On dira que la loi des effets aléatoires est une loi conjuguée (relativement à la loi des y_{ij}). Par exemple si la loi des y est une loi de Poisson, on supposera que la loi des effets aléatoires est une loi gamma. [4]

[3] Le terme "nombre d'observations" doit être précisé. En reprenant les notations du chapitre 3, ce que nous désignons ici par nombre d'observation est le nombre d'observations des Z dans les modèles hiérarchiques et le nombre d'observations des Z_1 et Z_2 dans les modèles obtenus par "double hiérarchisation".

[4] Nous reviendrons sur l'utilisation des lois conjuguées, mais dans un autre contexte, dans le chapitre 7.

– L'approche par vraisemblance conditionnelle : dans cette approche, on ignore le fait que les effets aléatoires sont aléatoires et on travaille uniquement sur $\prod_{i=1}^{n} f_i(y_i|u_i)$. On estime les u_i comme les paramètres dans un modèle GLM. Cette approche ne doit pas être utilisée quand un des intérêts de l'étude porte sur la variabilité des effets aléatoires.
– L'approche PQL (*Penalized Quasi Likelihood*) que nous allons présenter plus bas.
– L'approche par algorithme EM.

Les trois premières approches sont inappropriées à notre cadre de travail :

1) La première approche nécessite l'intégration numérique dans un espace de dimension n, ce qui n'est pas sérieusement envisageable quand n est le nombre d'observations.

2) La deuxième nécessite de changer la distribution des effets aléatoires selon la vraisemblance des y et nous avons évoqué au chapitre 3 les difficultés d'utiliser d'autres distributions que la gaussienne dans le contexte spatial.

3) L'approximation donnée par l'approche conditionnelle ne permet plus de prendre en compte les corrélations spatiales.

Les estimateurs obtenus par l'approche PQL sont connus pour être peu efficaces et l'approche par l'algorithme EM nous semble la plus judicieuse pour notre problème. Cependant, comme nous allons le voir, cette dernière approche nécessite l'utilisation d'algorithmes de Monte Carlo (ce qui conduit à des algorithmes MCEM) dont l'efficacité en terme de convergence et de vitesse de calcul de la solution peut être grandement améliorée en intégrant à chaque itération une approche de type PQL. Plusieurs approches théoriques conduisent à ces estimateurs PQL. On peut par exemple les obtenir en effectuant des approximations de Laplace. Il nous semble que cette dernière approche permet la présentation la plus concise des résultats.

4.3.1 Approximation de Laplace

L'approximation de Laplace permet d'approximer des intégrales de grandes dimensions. La forme de l'approximation de Laplace est basée sur un développement à l'ordre 2 en série de Taylor et s'écrit :

$$\log \int_{\mathbb{R}^p} e^{h(\mathbf{u})} d\mathbf{u} \simeq h(\mathbf{u}_0) + \frac{p}{2}\log 2\pi - \frac{1}{2}\log \left| -\frac{\partial^2 h(\mathbf{u})}{\partial \mathbf{u} \partial \mathbf{u}'} \right|_{\mathbf{u}=\mathbf{u}_0}, \tag{4.11}$$

où $\mathbf{u}_0$ est la solution de

$$\left. \frac{\partial h(\mathbf{u})}{\partial \mathbf{u}} \right|_{\mathbf{u}=\mathbf{u}_0} = \mathbf{0}. \tag{4.12}$$

Ce résulat est utilisé pour l'approximation de la log-vraisemblance dans les GLMM donnée par exemple par (4.10) en posant $h \equiv \log f_{Y|U} + \log f_U$,

où

$$f_{Y|U} = \prod_{i=1}^{n} f_i(y_i|u_j).$$

Supposons que

$$\mathbf{U} \sim \mathcal{N}(\mathbf{0}, \Sigma),$$

on a alors

$$h(\mathbf{u}) = \sum_{i=1}^{n} \log f_i(y_i|u_j) - \frac{1}{2}\mathbf{u}'\Sigma^{-1}\mathbf{u} - \frac{p}{2}\log 2\pi - \frac{1}{2}\log|\Sigma|.$$

En utilisant l'équation (4.8) appliquée à $\mathbf{u}$ (au lieu de β) et les notations de la section 4.2, on a

$$\frac{\partial h(\mathbf{u})}{\partial \mathbf{u}} = \frac{1}{\tau^2}\mathbf{Z}'\mathbf{W}(\mathbf{u})\Delta(\mathbf{u})(\mathbf{y}-\mu(\mathbf{u})) - \Sigma^{-1}\mathbf{u}, \tag{4.13}$$

où $\mathbf{Z}$ est la matrice d'incidence [5] associée au facteur $\mathbf{u}$.

On a également

$$\frac{\partial^2 h(\mathbf{u})}{\partial \mathbf{u}\partial \mathbf{u}'} = -\frac{1}{\tau^2}\mathbf{Z}'\mathbf{W}(\mathbf{u})\Delta(\mathbf{u})\frac{\partial \mu}{\partial \mathbf{u}'} + \frac{1}{\tau^2}\mathbf{Z}'\frac{\partial \mathbf{W}(\mathbf{u})\Delta(\mathbf{u})}{\partial \mathbf{u}'}(\mathbf{y}-\mu) - \Sigma^{-1}. \tag{4.14}$$

Quand g est le lien canonique, le deuxième terme de la somme dans (4.14) est nul. Dans le cas contraire, son espérance est nulle et on considère en général ce terme comme négligeable. L'équation (4.14) devient après simplifications :

$$H(\mathbf{u}) = -\frac{\partial^2 h(\mathbf{u})}{\partial \mathbf{u}\partial \mathbf{u}'} = \frac{1}{\tau^2}\mathbf{Z}'\mathbf{W}(\mathbf{u})\mathbf{Z} + \Sigma^{-1}. \tag{4.15}$$

On obtient $\mathbf{u}_0$ la solution de l'équation (4.12), par l'algorithme de Fisher scoring :

$$\mathbf{u}^{(t+1)} = \mathbf{u}^{(t)} + \left(\frac{1}{\tau^2}\mathbf{Z}'\mathbf{W}(\mathbf{u}^{(t)})\mathbf{Z} + \Sigma^{-1}\right)^{-1}\left(\frac{1}{\tau^2}\mathbf{Z}'\mathbf{W}(\mathbf{u}^{(t)})\Delta(\mathbf{u})^{(t)}(\mathbf{y}-\mu(\mathbf{u}^{(t)})) - \Sigma^{-1}\mathbf{u}^{(t)}\right) \tag{4.16}$$

Pour obtenir l'estimateur PQL de β et Σ, on alterne une étape de Fisher scoring pour $\mathbf{u}$ avec β fixé selon l'équation (4.16) et une étape de Fisher scoring pour β (similaire aux GLM) avec $\mathbf{u}$ fixé. Il faut ensuite estimer Σ. Notre objectif n'est pas de présenter cette méthode en détail car seuls les calculs de $\mathbf{u}_0$ et de $-\frac{\partial^2 h(\mathbf{u})}{\partial \mathbf{u}\partial \mathbf{u}'}\Big|_{\mathbf{u}=\mathbf{u}_0}$ nous intéresseront par la suite. Nous renvoyons le lecteur aux articles de Breslow et Clayton (1993) ou Breslow et Lin (1995) pour plus de détails sur les estimateurs basés sur la PQL dans les GLMM, ou encore Dean *et al.* (2004) pour le cas de données spatialement corrélées (au sens des modèles conditionnellement autorégressifs). Les estimateurs fournis par la méthode PQL sont d'autant plus mauvais que la distribution des $\mathbf{y}$ conditionnellement aux effets aléatoires est loin de la loi normale. En effet, l'approximation de Laplace est exacte dans ce dernier cas.

4.3.2 L'algorithme EM

L'algorithme EM, initialement introduit par Dempster *et al.* (1977), notamment pour estimer des modèles de mélanges, a rencontré un grand succès pour estimer le maximum de vraisemblance pour des problèmes avec des données manquantes ou des modèles faisant intervenir des structures latentes. Dans le cas des GLMM, son utilisation fut semble-t-il introduite par Mc Cullogh (1994).

Rappelons-en rapidement le principe et fixons les notations.

Principe

L'objectif est la maximisation de la log-vraisemblance $l(\psi, \mathbf{y})$ en fonction d'un vecteur de paramètres ψ et à partir d'un vecteur d'observations $\mathbf{y}$. La maximisation ne pouvant être effectuée de manière simple du fait de la forme de la fonction l, on considère la log-vraisemblance complète $l_c(\psi, \mathbf{Y}, \mathbf{U})$ où $\mathbf{U}$ est le vecteur aléatoire des données manquantes ou des éléments de la structure latente. Notons que $\mathbf{Y}$ et $\mathbf{U}$ sont considérées comme aléatoires de telle sorte que la vraisemblance complète est une variable aléatoire. Dans les modèles GLMM, le passage de la log-vraisemblance à la log-vraisemblance complète revient à supprimer l'intégrale dans l'équation (4.10).

[5] On aurait pu écrire le modèle sous la forme $E[\mathbf{Y}|\mathbf{U}] = g^{-1}(\mathbf{X}'\beta + \mathbf{U}'\mathbf{Z})$.

L'algorithme EM est un algorithme itératif à 2 étapes :
à l'itération $t+1$, on dispose du paramètre courant $\psi^{(t)}$
-l'étape E (Exceptation) revient à calculer la quantité

$$Q(\psi, \psi^{(t)}) = E_{\psi^{(t)}}[l_c(\psi; \mathbf{Y}, \mathbf{U})|\mathbf{Y} = \mathbf{y}],$$

l'espérance calculée sous $\psi^{(t)}$ de la log-vraisemblance conditionnellement au fait que la variable $\mathbf{Y}$ est égale aux observations.

-l'étape M (Maximization) consiste à calculer

$$\psi^{(t+1)} = \text{argmax}_{\psi \in \Theta} Q(\psi, \psi^{(t)}).$$

La séquence $(\psi^{(t)})_t$ converge vers le maximum de vraisemblance $\hat{\psi}$ sous certaines conditions de régularité de la vraisemblance des observations (voir par exemple McLachlan et Krishnan, 1997).

EM généralisé

L'étape M de l'algorithme EM pouvait être résolue analytiquement dans les problèmes qui furent à l'origine du développement de cet algorithme. Mais l'usage croissant de cette méthodologie pour la résolution de problèmes de plus en plus complexes a conduit fréquemment à des étapes M qui devaient être effectuées numériquement, par des algorithmes eux mêmes itératifs.

Afin d'éviter d'utiliser à chaque itération de l'EM, un algorithme itératif, plusieurs algorithmes ont été développés en se basant sur la propriété suivante :

Si à l'étape M, $\psi^{(k+1)}$ est choisi de telle sorte que

$$Q(\psi^{(k+1)}, \psi^{(k)}) \geq Q(\psi^{(k)}, \psi^{(k)}),$$

alors

$$l(\psi^{(k+1)}; \mathbf{y}) \geq l(\psi^{(k)}; \mathbf{y}).$$

Dans les algorithmes EM généralisés (EMG), au lieu de maximiser globalement $Q(., \psi^{(k)})$ à l'étape M, on choisit juste $\psi^{(k+1)}$ de telle sorte que $Q(\psi^{(k+1)}, \psi^{(k)}) \geq Q(\psi^{(k)}, \psi^{(k)})$.

Citons par exemple l'algorithme du gradient EM (GEM) introduit par Lange (1995) et qui consiste à effectuer une étape de Newton-Raphson à chaque itération de l'algorithme EM :

$$\psi^{(k+1)} = \psi^{(k)} - d^{20}Q(\psi^{(k)}, \psi^{(k)})^{-1} d^{10}Q(\psi^{(k)}, \psi^{(k)}),$$

où d^{10} et d^{20} sont les opérateurs respectifs de dérivation et de dérivation seconde par rapport à la première variable de Q.

Lange (1995) démontre que l'algorithme GEM a des propriétés de convergence locale presque identiques à celui de l'algorithme EM.

4.3.3 Dans le cas des GLMM

Dans le cas des GLMM, l'algorithme EM semble à première vue particulièrement adapté puisque la log-vraisemblance complète des observations et des effets aléatoires s'écrit

$$l_c(\psi; \mathbf{Y}, \mathbf{U}) = \log f_{Y|U}(\mathbf{Y}|\mathbf{U}; \beta) + \log f_U(\mathbf{U}; \theta_U)$$

Cependant, son espérance conditionnellement aux observations

$$Q(\psi, \psi^{(k)}) = E_{\psi^{(k)}}[l_c(\psi; \mathbf{Y}, \mathbf{U})|\mathbf{Y} = \mathbf{y}]$$

n'est pas calculable analytiquement pour la même raison que dans le cas de la vraisemblance : la grande dimension de l'espace d'intégration.

Néanmoins, contrairement au cas du calcul direct de la vraisemblance, on dispose des valeurs des paramètres à l'itération k et on va donc pouvoir calculer $Q(\psi, \psi^{(k)})$ par Monte Carlo, c'est à dire en simulant un échantillon de taille N du vecteur des effets aléatoires $(\mathbf{u}_k^{(1)}, \ldots, \mathbf{u}_k^{(N)})$ selon la distribution $f_{U|Y}(\mathbf{u}|\mathbf{y}; \psi^{(k)})$ puis en utilisant la loi des grands nombres on pourra effectuer l'approximation de Monte Carlo :

$$E_{\psi^{(k)}}[l_c(\psi; \mathbf{Y}, \mathbf{U})|\mathbf{Y} = \mathbf{y}] \simeq \frac{1}{N} \sum_{i=1}^{N} l_c(\psi; \mathbf{y}, \mathbf{u}_k^{(i)})$$

L'étape M devient alors très simple. En effet, la log-vraisemblance des données complètes s'écrit comme une somme de deux termes ne faisant pas intervenir les mêmes paramètres

$$Q_N(\psi, \psi^{(k)}) = \frac{1}{N} \sum_{i=1}^{N} l_c(\psi; \mathbf{Y}, \mathbf{U}) = \sum_{i=1}^{N} \log f_{Y|U}(\mathbf{y}|\mathbf{u}_k^{(i)}; \beta) + \sum_{i=1}^{N} \log f_U(\mathbf{u}_k^{(i)}; \theta_U).$$

Le premier terme se maximise par une procédure similaire à celle employée pour l'estimation des GLM, c'est-à-dire par un algorithme Newton-Raphson :

$$\beta^{(t+1)} = \beta^{(t)} + \left(\mathbf{X}' \left(\frac{1}{N} \sum_{i=1}^{N} \mathbf{W}_\beta(\mathbf{u}_k^{(i)}, \beta^{(t)}) \right) \mathbf{X} \right)^{-1} \mathbf{X}' \left(\frac{1}{N} \sum_{i=1}^{N} \mathbf{W}_\beta(\mathbf{u}_k^{(i)}, \beta^{(t)}) \Delta_\beta(\mathbf{u}_k^{(i)}, \beta^{(t)}) (\mathbf{y} - \mu(\mathbf{u}_k^{(i)}, \beta^{(t)})) \right)$$

où

$$\mu(\mathbf{u}_k^{(i)}, \beta^{(t)}) = E_{\beta^{(t)}}[Y_j|\mathbf{U} = \mathbf{u}_k^{(j)}],$$

et$\mathbf{W}_\beta$ et Δ_β sont définis comme l'étaient $\mathbf{W}$ et Δ dans la section 2.

Le second terme s'écrit :

$$\begin{aligned}
\sum_{i=1}^{N} \log f_U(\mathbf{u}_k^{(i)}; \theta_U) &= -\frac{1}{2} \log |\Sigma(\theta_U)| - \frac{1}{2} \sum_{i=1}^{N} \mathbf{u}_k^{(i)\prime} \Sigma(\theta_U)^{-1} \mathbf{u}_k^{(i)} \\
&= -\frac{1}{2} \log |\Sigma(\theta_U)| - \frac{1}{2} \sum_{i=1}^{N} \operatorname{trace} \left(\Sigma(\theta_U)^{-1} \mathbf{u}_k^{(i)} \mathbf{u}_k^{(i)\prime} \right) \\
&= -\frac{1}{2} \log |\Sigma(\theta_U)| - \frac{1}{2} \operatorname{trace} \left(\Sigma(\theta_U)^{-1} \left(\sum_{i=1}^{N} \mathbf{u}_k^{(i)} \mathbf{u}_k^{(i)\prime} \right) \right)
\end{aligned}$$

Il se maximise par un algorithme de Newton-Raphson de manière similaire aux modèles gaussiens. Dans le cas spatial, nous avons remplacé toutes les quantités du type :

$$\sum_{i=1}^{N} \mathbf{u}_k^{(i)\prime} M(\theta_U)^{-1} \mathbf{u}_k^{(i)}$$

(qui interviennent dans la vraisemblance gaussienne et ses dérivées par rapport aux éléments de θ_U) par

$$\operatorname{trace} \left(M(\theta_U)^{-1} \left(\sum_{i=1}^{N} \mathbf{u}_k^{(i)} \mathbf{u}_k^{(i)\prime} \right) \right),$$

ce qui permet de limiter très significativement le nombre de calculs surtout quand la taille N des échantillons d'effets aléatoires $\mathbf{u}_k^{(i)}$ simulés est importante.

L'utilisation d'algorithmes EM pour estimer des GLMM a été introduite par McCulloch (1994) dans le cas de réponses binaires et pour une fonction de lien probit. Puis McCulloch (1997) étend cet algorithme à tous les GLMM du type de celui présenté dans l'exemple. Les effets aléatoires sont i.i.d et simulés à l'aide de l'algorithme de Metropolis-Hasting (voir chapitre 7 pour la description de cet algorithme). Dans le cas du modèle **(GLMM2)**, Zhang (2002) propose un algorithme MCEMG (gradient MCEM). C'est-à-dire que l'étape M est remplacée par une seule itération de Newton-Raphson (selon le principe de l'algorithme GEM). Les effets aléatoires, spatialement corrélés, sont simulés par un algorithme dit de Metropolis-sous-Gibbs (voir chapitre 7). L'idée d'utiliser une seule étape de Newton-Raphson à l'étape M ne semble pas la plus efficace car le coût numérique dans les MCEM provient plus de la simulation des effets que de la maximisation des paramètres de la log-vraisemblance complète. La simulation des effets aléatoires dans le cas où ceux-ci sont spatialement corrélés est également coûteuse numériquement car pour chaque $\mathbf{u}_k^{(i)}$, il faut simuler chaque composante dans sa distribution conditionnellement aux autres composantes. Enfin, cet algorithme de simulation est une chaîne de Markov (Monte Carlo Markov Chain, MCMC) ce qui pose essentiellement deux problèmes. Le premier est que seule la limite de la chaîne produite a comme distribution $f_{U|Y}(\mathbf{u}|\mathbf{y})$. En pratique, on n'a pas la garantie d'avoir simulé selon cette distribution car la chaîne produite n'a pas forcément convergé. Le second problème vient du fait que les échantillons $\mathbf{u}_k^{(i)}$ simulés sont corrélés. Ces deux difficultés inhérentes à tous les algorithmes MCMC n'empêchent pas la convergence de l'estimateur

$$Q_N(\psi, \psi^{(k)}) = \frac{1}{N} \sum_{i=1}^{N} l_c(\psi; \mathbf{y}, \mathbf{u}_k^{(i)})$$

vers

$$Q(\psi, \psi^{(k)}) = E_{\psi^{(k)}}[l_c(\psi; \mathbf{Y}, \mathbf{U})|\mathbf{Y} = \mathbf{y}]$$

mais elles rendent la tâche d'estimer l'erreur de Monte-Carlo :

$$Q_N(\psi, \psi^{(k)}) - Q(\psi, \psi^{(k)}) \tag{4.17}$$

beaucoup plus complexe (voir Booth et Hobert, 1999, pour une discussion détaillée). Or cette quantité est utile à plusieurs niveaux dans les algorithmes MCEM ; elles permettent entre autres, de contrôler l'erreur commise dans l'estimation de $\hat{\psi}$.

Booth et Hobert (1999) proposent d'utiliser l'échantillonnage d'importance pour approximer l'étape E d'un algorithme EM pour des GLMM. Cette procédure, outre le fait qu'elle s'est montrée très efficace, permet une estimation rapide de (4.17). Nous expliquons en détail cette procédure.

4.3.4 L'échantillonnage d'importance

Principe

L'échantillonnage d'importance (*importance sampling*) permet d'approximer les quantités du type $E_s[f(X)]$ pour une fonction f et une variable aléatoire X admettant pour densité une fonction s. Le principe est proche de celui du calcul d'intégrales par Monte Carlo mais ici on utilise un échantillon $u^{(1)}, \ldots, u^{(N)}$ simulé à partir d'une distribution instrumentale de densité v, ce qui conduit à estimer $E[f(X)]$ par :

$$\begin{aligned} E_s[f(X)] &= \int f(x)s(x)dx \\ &= \int \frac{s(x)}{v(x)} f(x)v(x)dx \\ &= E_v\left[\frac{s(x)}{v(x)} f(x)\right] \\ &\simeq \frac{1}{N}\sum_{i=1}^{N} w_{ik} f(\mathbf{u}_k^{(i)}), \end{aligned}$$

où les éléments

$$w_{ik} = \frac{s(\mathbf{u}_k^{(i)})}{v(\mathbf{u}_k^{(i)})}$$

sont appelés poids d'importance. La densité v selon laquelle les effets aléatoires sont simulés est la densité d'importance. La seule condition pour que l'estimateur de Monte-Carlo $\frac{1}{N}\sum_{i=1}^{N} w_{ik} f(\mathbf{u}_k^{(i)})$ converge vers $E_s[f(X)]$ est que le support de s (l'ensemble des valeurs pour lesquelles s est non nulle) soit inclus dans celui de v (Robert et Casella, 1999) ce qui rend cette méthode très attractive. Cependant la vitesse de convergence peut être très mauvaise si la distribution instrumentale v est mal choisie. Les résultats théoriques suggèrent de choisir pour v, des distributions à queues lourdes, relativement à celle de s. Quand s est une distribution approximativement ellipsoïdale et a un mode, on choisit souvent pour v la distribution de Student multivariée $\mathcal{T}(\mathbf{m}_t, \Sigma_t, d)$ où $\mathbf{m}_t$ est le mode de s, Σ_t est l'inverse de la matrice hessienne de $-\log s$ et d est le degré de liberté (choisi sur la base de quelques essais). Ce choix est connu pour être très efficace (Evans et Swartz, 1998).

Dans le cas des GLMM

A l'étape k de l'algorithme EM, on dispose de $\psi^{(k)} = (\beta^{(k)}, \theta_U^{(k)})$.

Etape E :

On simule un échantillon $(\mathbf{u}_k^{(1)}, \ldots, \mathbf{u}_k^{(N)})$ selon une distribution de Student t multivariée [6] dont les paramètres d'espérance $\mathbf{m}_t$ et Σ_t sont choisis de telle sorte que le mode et la courbure de la distribution d'importance soient proches de ceux de $f_{U|Y}$ sous $\psi^{(k)}$.

On voudrait choisir

$$\mathbf{m}_t = E_{\psi^{(k)}}[\mathbf{U}|\mathbf{y}]$$

et

$$\Sigma_t = \mathrm{Var}_{\psi^{(k)}}(\mathbf{U}|\mathbf{y}).$$

Les approximations de Laplace de $E_{\psi^{(k)}}[\mathbf{U}|\mathbf{y}]$ et de $\mathrm{Var}_{\psi^{(k)}}(\mathbf{U}|\mathbf{y})$ sont obtenues à partir de l'approximation de Laplace de l'intégrale (4.10). Elles sont données respectivement par $\mathbf{u}_0$, la solution de l'équation (4.12) et par l'inverse de la matrice hessienne de $-h$ évaluée en $\mathbf{u}_0$, $H^{-1}(\mathbf{u}_0)$ selon les notations de la section 4.3.1 (voir par exemple Booth, 1998). Dans les deux cas, β reste fixé à sa valeur courante $\beta^{(k)}$.

On obtient donc m_t par l'algorithme de Newton-Raphson suivant :

$$\mathbf{m}_t^{(r+1)} = \mathbf{m}_t^{(r)} + \left(\frac{1}{\tau^2}\mathbf{Z}'\mathbf{W}_k(\mathbf{m}_t^{(r)})\mathbf{Z} + \Sigma^{-1}\right)^{-1}\left(\frac{1}{\tau^2}\mathbf{Z}'\mathbf{W}_k(\mathbf{m}_t^{(r)})\Delta_k(\mathbf{m}_t^{(r)})(\mathbf{y} - \mu_k(\mathbf{m}_t^{(r)})) - \Sigma^{-1}\mathbf{m}_t^{(r)}\right),$$

[6] Pour la simulation des effets aléatoires selon la loi de Student multivariée et pour le calcul des poids d'importance associés, nous avons utilisé le package de R *bayesm*.

$$\Sigma_t = \left(\frac{1}{\tau^2} \mathbf{Z}' \mathbf{W}_k(\mathbf{m}_t^{(r)}) \mathbf{Z} + \Sigma^{-1} \right)^{-1}.$$

Notons que les w_{ik} n'ont besoin d'être connus qu'à une constante multiplicative près car ils ne font intervenir que $\psi^{(k)}$ et non ψ, le terme en fonction duquel on va maximiser $Q(., \psi^{(k)})$.

L'étape E consiste à estimer $Q(\psi, \psi^{(k)})$ à une constante de proportionnalité près par :

$$Q_N(\psi, \psi^{(k)}) = \frac{1}{N} \sum_{i=1}^{N} w_{ik} l_c(\psi; \mathbf{y}, \mathbf{u}_k^{(i)})$$

Pour simplifier l'étape M, on va normer les poids d'importance w_i obtenus par

$$w'_{ik} = \frac{w_{ik} N}{\sum_{i=1}^{N} w_{ik}},$$

de telle sorte que la somme des w'_{ik} soit égale à N. L'étape M devient alors similaire à celle présentée dans le cas où les $\mathbf{u}_k^{(i)}$ étaient simulés selon $f_{U|Y}$. Il suffit juste de remplacer toutes les moyennes de Monte-Carlo qui interviennent sous la forme

$$\frac{1}{N} \sum_{i=1}^{N} G(\mathbf{u}_k^{(i)})$$

par

$$\frac{1}{N} \sum_{i=1}^{N} w'_{ik} G(\mathbf{u}_k^{(i)})$$

Choix de N

Le choix de la taille des échantillons d'effets aléatoires simulés permet de contrôler l'erreur de Monte-Carlo. On a donc intérêt à choisir N grand pour être précis. Cependant, le choix de N doit être un compromis entre la précision et le coût numérique. Si on note

$$\psi^{\star(k+1)} = \text{argmax}_{\psi \in \Theta} Q(\psi, \psi^{(k)}),$$

la quantité que l'on cherche à estimer à l'étape M de la $k+1^{\text{ème}}$ itération et $\psi^{(k)}$ son estimation obtenue en maximisant

$$\frac{1}{N} \sum_{i=1}^{N} w'_{ik} l_c(\psi; \mathbf{y}, \mathbf{u}_k^{(i)}),$$

il est inutile que $\psi^{(k+1)}$ soit très proche de $\psi^{\star(k+1)}$ si $\psi^{\star(k+1)}$ est encore loin de de $\hat{\psi}$ (Wei et Tanner, 1990). Or

$$||\psi^{\star(k+1)} - \hat{\psi}||$$

est d'autant plus grand que k est petit, ce qui suggère de faire grandir N avec k. Pour le faire, plusieurs auteurs (Wei et Tanner, 1990, par exemple) ont proposé des choix arbitraires. Booth et Hobert (1999) proposent un automate qui se base sur l'erreur de Monte-Carlo.

L'idée est la suivante : si à l'étape $k+1$, $||\psi^{(k+1)} - \psi^{(k)}||$ est grand devant l'erreur de Monte Carlo, alors $N(k+2)$ reste inchangé. Dans le cas contraire, l'itération $k+1$ n'a servi à rien et donc on augmente N à l'étape $k+2$.

Booth et Hobert montrent que $\psi^{(k+1)}$ est approximativement distribué comme une loi normale d'espérance $\psi^{\star(k+1)}$ et de matrice de variance covariance

$$\text{Var}(\psi^{(k+1)}|\psi^{(k)}) \simeq (d^{20} Q_N(\psi^{\star(k+1)}, \psi^{(k)}))^{-1} \text{Var}(d^{10} Q_N(\psi^{\star(k+1)}, \psi^{(k)}))(d^{20} Q_N(\psi^{\star(k+1)}, \psi^{(k)}))^{-1}.$$

En remplaçant $\psi^{\star(k+1)}$ par $\psi^{(k+1)}$ et en estimant

$$\mathrm{Var}(d^{10}Q_N(\psi^{\star(k+1)}, \psi^{(k)}))$$

par :

$$\hat{\mathrm{Var}}(d^{10}Q_N(\psi^{\star(k+1)}, \psi^{(k)})) = \frac{1}{N^2}\sum_{i=1}^{N}\left(w_{ik}\frac{\partial}{\partial\psi}\log f(\mathbf{y}, \mathbf{u}_k^{(i)}; \psi^{(k+1)})\right)\left(w_{ik}\frac{\partial}{\partial\psi}\log f(\mathbf{y}, \mathbf{u}_k^{(i)}; \psi^{(k+1)})\right)',$$

on peut approcher l'ellipse de confiance $100(1-\alpha)$ pour $\psi^{\star(k+1)}$ autour de $\psi^{(k+1)}$. Si la valeur précédente $\psi^{(k)}$ appartient à cette ellipse, on considère que l'étape qui vient d'être effectuée n'a pas eu plus d'efficacité que de l'erreur de Monte Carlo pure. Cette dernière étant trop importante, on change N :

$$N \leftarrow N + \left[\frac{N}{j}\right]$$

où j est un entier (Booth et Hobert, 1999, suggèrent $j \in \{3, 4, 5\}$) et [.] désigne la partie entière. Booth et Hobert suggèrent de choisir $\alpha = 0.75$.

Critère d'arrêt

Nous stoppons les algorithmes lorsque pour 3 itérations successives, on a

$$\max_i \frac{||\psi_i^{(k+1)} - \psi_i^{(k)}||}{|\psi_i^{(k+1)}| + \delta_1} < \delta_2,$$

où δ_1 et δ_2 sont deux constantes que nous avons fixées respectivement à 0.001 et 0.005.

Adaptations dans le contexte spatial

Notre présentation de l'algorithme, basée sur un GLMM classique, est directement adaptable au cas du modèle **(GLMM2)**.

Qu'en est-t-il dans le cas du modèle **(GLMM4)** et dans celui du modèle hiérarchique bivarié ?

Nous reprenons les notations du chapitre 3.

Modèle GLMM4

Dans le modèle **GLMM4**, on considère pour tout ensemble de sites $\{s_1, \ldots, s_n\}$, un vecteur $\mathbf{Z} = (Z(s_1), \ldots, Z(s_n))'$ de variables d'intérêt non gaussiennes un vecteur $\mathbf{X} = (X(s_1), \ldots, X(s_n))'$. On introduit un vecteur de variables latentes $\mathbf{Y} = (Y(s_1), \ldots, Y(s_n))'$. On modélise la relation entre $\mathbf{X}$ et $\mathbf{Y}$ en utilisant un modèle **LM4** et on suppose que conditionnellement à $\mathbf{Y}$, les composantes de $\mathbf{Z}$ sont mutuellement indépendantes et vérifient

$$Z(s_i)|Y(s_i) \sim \mathcal{L}(\gamma_i),$$

et

$$E[Z(s_i)|Y(s_i)] = g(Y(s_i)).$$

Le modèle et de type GLMM et le vecteur d'effets aléatoires est $\mathbf{Y}$. Notons que si les sites d'observations de Z et de X ne coïncident pas, il est suffisant de considérer Y aux sites d'observation de Z pour définir entièrement le modèle.

La vraisemblance s'écrit

$$L(\psi; \mathbf{z}, \mathbf{x}) = \int_{\mathbb{R}^n} \prod_{i=1}^{n} f_i(z(s_i)|y(s_i)) f_{(X,Y)}(\mathbf{x}, \mathbf{y}; \psi) d\mathbf{y}$$

où f_i est la densité de la distribution $\mathcal{L}(\gamma_i)$.

Etape E :

L'approximation de Laplace de $L(\psi; \mathbf{z}, \mathbf{y}))$ est modifiée à cause du deuxième terme $f_{(X,Y)}(\mathbf{x}, y; \psi)$ qui dépend de $\mathbf{x}$. L'équation (4.13) adaptée au cas qui nous intéresse s'écrit :

$$\frac{\partial h(\mathbf{y})}{\partial \mathbf{y}} = \frac{1}{\tau^2} \mathbf{W}(\mathbf{y}) \Delta(\mathbf{y})(\mathbf{z} - \mu(\mathbf{y})) - V_{YY}(\mathbf{y} - \mu_Y) - V_{YX}(\mathbf{x} - \mu_X), \tag{4.18}$$

où les matrices V_{YX} et V_{XX} sont obtenues à partir de la matrice de covariance de $f_{(X,Y)}(\mathbf{x}, \mathbf{y}; \psi)$ par :

$$\begin{aligned} \Sigma^{-1} &= \begin{pmatrix} \Sigma_{XX} & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_{YY} \end{pmatrix}^{-1} \\ &= \begin{pmatrix} V_{XX} & V_{XY} \\ V_{YX} & V_{YY} \end{pmatrix} \end{aligned}$$

Notons que l'on a supprimé la matrice d'incidence (que nous notions $\mathbf{Z}$) dans l'expression (4.19) car elle est ici égale à l'identité puisque qu'il y a un effet aléatoire par site.

La matrice hessienne de $-h$ s'écrit alors :

$$H(\mathbf{y}) = \frac{1}{\tau^2} \mathbf{W}(\mathbf{y}) + V_{YY}$$

On obtient ensuite les paramètres de la distribution de Student multivariée $\mathcal{T}(\mathbf{m}_t, \Sigma_t, d)$ selon laquelle simuler les effets aléatoires :

- $\mathbf{m}_t$ par Fisher scoring
- $\Sigma_t = H(\mathbf{m}_t)^{-1}$.

Les poids d'importance s'écrivent

$$w_{ik} = \frac{f_{Y|X,Z}(\mathbf{y}_k^{(i)}|\mathbf{x}, \mathbf{z}; \psi^{(k)})}{v(\mathbf{y}_k^{(i)}; \mathbf{m}_t, \Sigma_t, d)}$$

où v est la densité de la distribution de Student multivariée $\mathcal{T}(\mathbf{m}_t, \Sigma_t, d)$.

$f_{Y|X,Z}(\mathbf{y}_k^{(i)}|\mathbf{x}, \mathbf{z}; \psi^{(k)})$ est proportionnelle à

$$f_{Z|Y}(\mathbf{z}|\mathbf{y}_k^{(i)}) f_{(X,Y)}(\mathbf{x}, \mathbf{y}_k^{(i)}; \psi^{(k)}),$$

avec

$$f_{Z|Y}(\mathbf{z}|\mathbf{y}_k^{(i)}) = \prod_{i=1}^{n} f_i(z_i|y_i).$$

Rappelons que les poids n'ont besoin d'être connus qu'à une constante multiplicative près.

Etape M :

L'étape M consiste à trouver :

$$\psi^{(k+1)} = \text{argmax}_{\psi \in \Theta} \sum_{i=1}^{N} \log f_{(X,Y)}(\mathbf{x}, \mathbf{y}_k^{(i)}; \psi)$$

puisque les paramètres n'interviennent que dans ce terme. Ce terme est une somme de log-vraisemblances gaussiennes et il se maximise de manière similaire aux vraisemblances gaussiennes de la section 1 (en utilisant la factorisation évoquée plus haut pour limiter le nombre de calculs).

Modèle hiérarchique bivarié

Dans le modèle hiérarchique bivarié, on considère deux vecteurs aléatoires :

$\mathbf{Z}_1 = (Z_1(s_1), \ldots, Z_1(s_n))'$ et $\mathbf{Z}_2 = (Z_2(s_1), \ldots, Z_2(s_n))'$. On suppose que les composantes du vecteur $\mathbf{Z} = (\mathbf{Z}_1', \mathbf{Z}_2')'$ sont mutuellement indépendantes conditionnellement aux réalisations d'un champ aléatoire bivarié

$$Y = \left\{ \begin{pmatrix} Y_1(s) \\ Y_2(s) \end{pmatrix}, s \in \mathcal{D} \right\}$$

aux sites $\{s_1, \ldots, s_n\}$.

On suppose de plus que la distribution de $Z_j(s_i)$ conditionnellement à $Y_j(s_i)$ est une distribution de la famille exponentielle $\mathcal{L}_j(\gamma_i)$ spécifiée par une relation du type :

$$E[Z_j(s_i) = g_j(Y_j(s_i))$$

pour une fonction de lien g_j.

Enfin, on fait l'hypothèse que le champ bivarié latent Y définit un modèle **LM4**.

La vraisemblance des observations s'écrit

$$L(\mathbf{z}_1, \mathbf{z}_2; \psi) = \int \prod_{j=1}^{2} \prod_{i=1}^{n} f_{j,i}(z_j(s_i)|y_j(s_i)) f_{(Y_1,Y_2)}(\mathbf{y}_1, \mathbf{y}_2) d\mathbf{y}_1 d\mathbf{y}_2,$$

où $f_{j,i}$ est la densité de la distribution $\mathcal{L}_j(\gamma_i)$.

Il y a peu de changements dans la mise en place de l'algorithme ; nous donnons juste les modifications entraînées à l'étape de l'approximation de Laplace.

L'équation (4.13) adaptée au cas qui nous intéresse s'écrit en notant $\mathbf{y} = (\mathbf{y}_1', \mathbf{y}_2')'$ et $\mu^\star = (\mu_{Y_1}, \mu_{Y_2})$:

$$\frac{\partial h(\mathbf{y})}{\partial \mathbf{y}} = \mathbf{B}^{-1}\mathbf{K}'\mathbf{W}(\mathbf{y})\Delta(\mathbf{y})(\mathbf{z} - \mu(\mathbf{y})) - \Sigma^{-1}(\mathbf{y} - \mu^\star) \tag{4.19}$$

avec

$$\mathbf{B} = \begin{pmatrix} \tau_1 & 0 \\ 0 & \tau_2 \end{pmatrix},$$

$$\mathbf{K} = \begin{pmatrix} \mathbb{1}_n & 0 \\ 0 & \mathbb{1}_n \end{pmatrix},$$

est la matrice d'incidence,

$$\Delta(\mathbf{y}) = \begin{pmatrix} \Delta_1(\mathbf{y}_1) & 0 \\ 0 & \Delta_2(\mathbf{y}_2) \end{pmatrix},$$

et

$$\mu(\mathbf{y}) = \begin{pmatrix} \mu_1(\mathbf{y}_1) \\ \mu_2(\mathbf{y}_2) \end{pmatrix}.$$

La matrice hessienne de $-h$ s'écrit alors :

$$H(\mathbf{y}) = \mathbf{B}^{-1}\mathbf{K}'\mathbf{W}(\mathbf{y})\mathbf{K} + \Sigma^{-1}.$$

L'adaptation au cas où les deux variables ne sont pas observées aux mêmes sites est lourde en terme de notations mais ne pose aucune difficulté supplémentaire.

Exemple

Nous concluons cette section en montrant sur un jeu de données simulé, comment fonctionne l'algorithme MCEM. Nous avons choisi de travailler avec le modèle hiérarchique bivarié.

Sur une grille rectangulaire de $\mathbb{R}^2$, de pas unité et de taille 10×10, nous avons simulé un champ bivarié gaussien selon le modèle de corrélation intrinsèque **(LM4a)** avec pour fonction de corrélation un modèle exponentiel :

$$\rho_\theta(h) = e^{-\theta h},$$

où θ est égal à 0.5. Les espérances μ_1 et μ_2 sont fixées à 0 et les variances σ_1^2 et σ_2^2 à 1. On travaille avec $r = 0.5$

On dispose ainsi de 100 valeurs pour chaque variable Y_1 et Y_2.

Pour chacune des valeurs $Y_i(s_j)$, $i = 1, 2$ et $j = 1, \ldots, 100$, on simule $Z_i(s_j)$ selon une loi binômiale d'espérance

$$\text{logit}^{-1}(Y_i(s_j)) = \frac{e^{Y_i(s_j)}}{1 + e^{Y_i(s_j)}},$$

et de paramètre de taille 2. Le jeu de données simulé est représenté à la figure (4.1).

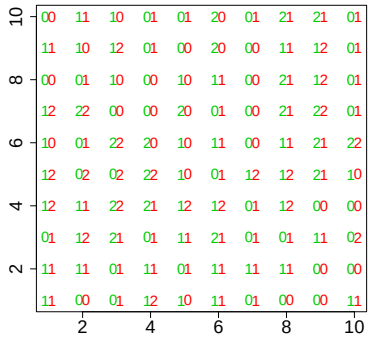

Fig. 4.1. Représentation du jeu de données simulé. En chaque site, on a l'observation de chaque variable : en rouge, la variable X, en vert la variable Y.

On lance l'algorithme en utilisant des échantillons de Monte-Carlo de taille initiale 10. On met à jour cette taille à chaque itération selon la règle que nous avons détaillée plus haut avec $j = 3$:

$$N \leftarrow N + \left[\frac{N}{3}\right],$$

et d le degré de liberté de la distribution de Student multivariée est fixé à 40.

On lance l'algorithme deux fois avec deux jeux de valeurs initiales :

-$\mu_X = \mu_Y = 0$, $r = 0$, $\sigma_X^2 = \sigma_Y^2 = 1$, et $\theta = 0.1$,

-$\mu_X = \mu_Y = 1$, $r = -0.2$, $\sigma_X^2 = \sigma_Y^2 = 0.04$, et $\theta = 1$.

Selon le critère d'arrêt, l'algorithme a convergé sur cet exemple en 117 itérations pour le premier jeu de valeurs initiales et 133 pour le second.

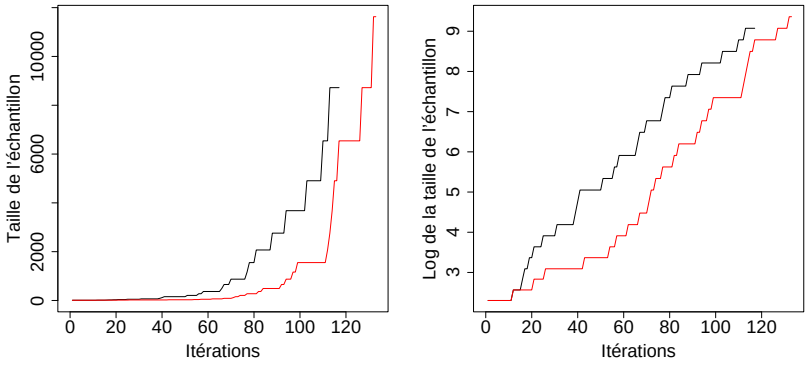

Fig. 4.2. Gauche : Taille de l'échantillon des effets aléatoires simulés en fonction de l'itération courante de l'algorithme EM. Droite : Log de la taille de l'échantillon en fonction de l'itération. Noir : premier jeu de valeurs initiales. Rouge : second jeu de valeurs initiales.

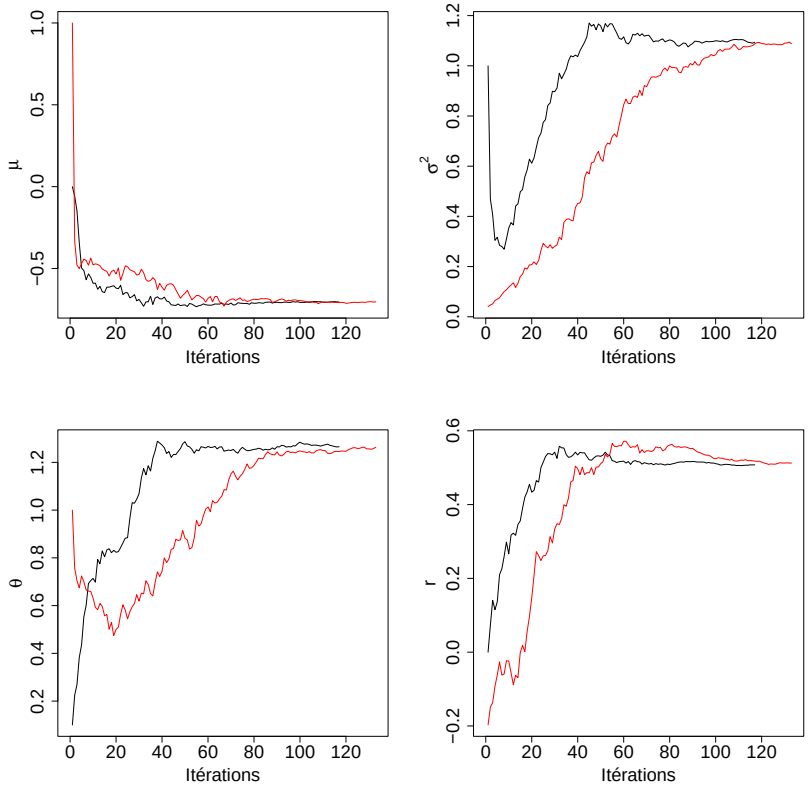

Fig. 4.3. Valeurs des paramètres en fonction de l'itération. De gauche à droite et de bas en haut : μ_1, σ_1^2, θ et r. Noir : premier jeu de valeurs initiales. Rouge : second jeu de valeurs initiales

La figure (4.2) montre l'évolution de N en fonction des itérations et la figure (4.3) montre l'évolution des paramètres.

Pour le premier jeu de valeurs initiales (respectivement le second), on constate que N évolue peu jusqu'à environ la soixantième itération (respectivement quatre-vingtième), or c'est également dans cette zone que les paramètres estimés se stabilisent.

En dépit de la nature stochastique de l'algorithme et des choix différents des valeurs initiales, l'estimation est la même dans les deux cas, ce qui conduit à penser que l'algorithme fonctionne correctement. Concernant les valeurs estimées, on note que l'espérance de X, μ_X ainsi que le paramètre de portée θ sont relativement loin de leur vraie valeur. Pour s'assurer de la bonne marche de l'algorithme, il faudrait répéter l'expérience en simulant plusieurs jeux de données selon les mêmes paramètres. Les résultats que l'on obtiendrait permettraient de s'assurer plus avant que l'algorithme converge. Ils donneraient en outre, des indications quant à la variabilité de l'estimateur. On sait cependant que dans le cas des champs gaussiens, l'estimateur du paramètre de portée à une forte variabilité (et un biais non négligeable à taille finie), or on observe ici une version très bruitée du processus gaussien, ce qui accentue la mauvaise qualité de l'estimation. L'estimation des autres paramètres semble de bonne qualité pour ce jeu de données particulier.

Notons qu'avec le logiciel R, la majeure partie du temps total pour que l'algorithme ait convergé (environ une heure sur cet exemple) est dévolu aux 20 dernières itérations. Il ne faut par exemple qu'une dizaine de secondes par itération quand $N = 10$.

5

Comportement des estimateurs

5.1 Introduction

Dans le chapitre précédent, nous avons montré comment estimer les paramètres dans les différents modèles présentés, que l'on soit en présence de données hétérotopes ou non. Bien que l'hétérotopie empêche d'utiliser la vraisemblance profilée dans le cas des modèles gaussiens à dépendance ponctuelle, elle ne pose pas vraiment de nouvelles difficultés en terme d'estimation des paramètres par rapport à l'isotopie. Cependant, il est évident que la qualité des estimateurs du (ou des) paramètre(s) de lien entre deux variables va être affectée par le fait que ces dernières ne sont pas observées aux mêmes sites. De même que l'on prédit d'autant mieux une variable en un site où celle-ci n'est pas observée que les sites d'observation de cette variable en sont proches, on estimera d'autant mieux la manière dont 2 variables covarient que celles-ci sont observées en des sites proches[1]. Le comportement de l'estimateur de la liaison entre 2 variables dépend donc de la géométrie de l'échantillonnage, l'idéal étant dans cette optique, l'isotopie. La variabilité de l'estimateur du lien doit également être fonction de la structure spatiale des variables mises en jeu, de même que la variance de krigeage dépend des structures spatiales de la variable à prédire. On comprend bien par exemple qu'estimer le lien entre deux bruits blancs toujours observés en des sites différents, même très proches, est absurde.

Ces considérations très intuitives restent qualitatives et sont donc peu utiles pour l'utilisateur qui a besoin d'une mesure quantitative de la qualité de son estimateur. Les questions qui se posent sont d'ordre pratique ou théorique et peuvent essentiellement se résumer comme suit :

avec n observations d'une variable X et p d'une variable Y, mesurées respectivement sur les ensembles de sites $\mathcal{C}_X = \{s_1, \ldots, s_n\}$ et $\mathcal{C}_Y = \{t_1, \ldots, t_p\}$, quelle confiance peut-on avoir dans l'estimation $\hat{r}$ du coefficient de corrélation dans un modèle de corrélation intrinsèque par exemple ? Dans quelles mesures les corrélations spatiales permettent-elles de compenser l'hétérotopie ? Plus généralement, peut-on écrire la distribution $\mathcal{V}$ de l'estimateur de r sous la forme :

$$\mathcal{V}(\hat{r}) = f(r, \mathcal{C}_X, \mathcal{C}_Y, C_\phi)$$

où C_ϕ est la fonction de corrélation des 2 variables ?

Ce chapitre est composé de deux parties distinctes.

La première partie concerne des résultats généraux de la théorie asymptotique pour les modèles gaussiens issus de la géostatistique. Comme cela a été évoqué dans le chapitre 2, la manière dont le

[1] Cette assertion se base sur l'hypothèse que les 2 variables sont liées par un modèle de type LM4 ou GLM4 et que l'autocorrélation est une fonction décroissante de la distance. Autrement dit, le maximum de corrélation entre les 2 variables est atteint quand celles-ci sont observées aux mêmes sites.

nombre d'observations tend vers l'infini est un élément crucial pour ce type de problèmes puisque l'on peut accroître le domaine d'observation ou au contraire densifier les sites d'observation dans un domaine de taille fixée. Après avoir rappelé ces deux grands types d'asymptotique, nous donnerons quelques éléments des formalismes utilisés dans chacun des deux cas. Nous nous concentrerons ensuite sur l'asymptotique par accroissement du domaine et nous démontrerons que les estimateurs par maximum de vraisemblance des paramètres d'un champ gaussien stationnaire et de fonction de covariance exponentielle sont asymptotiquement gaussiens sous certaines conditions. Nous nous appuierons essentiellement pour ce faire sur les articles de Mardia et Marshall (1984) et Cressie et Lahiri (1996) en précisant certains éléments de leurs démonstrations et en généralisant leurs résultats.

Dans la seconde partie, nous allons montrer comment ces résultats peuvent s'adapter aux modèles multivariés ; plus précisément, nous allons étudier comment se comporte l'estimateur par maximum de vraisemblance $\hat{r}$ du coefficient de corrélation r dans un modèle de corrélation intrinsèque, en fonction de la géométrie de l'échantillonnage et des structures spatiales et sous quelles conditions géométriques (asymptotique), cet estimateur est asymptotiquement gaussien.

5.2 Asymptotique pour les modèles gaussiens spatiaux

5.2.1 Introduction

De nombreux résultats sont disponibles dans le cas d'un champ gaussien univarié ; c'est à dire lorsque la (ou les) variable(s) explicative(s) est (sont) observée(s) en tout point où la variable d'intérêt est disponible, et que seul le processus résiduel est modélisé comme aléatoire (cas du modèle linéaire à résidus spatialement corrélés **(LM2)**). Ces résultats concernent le comportement asymptotique du maximum de vraisemblance des paramètres de covariance du champ résiduel et des paramètres de régression. Avant de se positionner par rapport à cette littérature, nous présentons les principales caractéristiques de l'étude de la convergence des estimateurs pour des données géoréférencées ainsi que quelques définitions relatives à ce problème.

Dans le cas d'un échantillon $X_1, \ldots, X_n$ i.i.d, l'asymptotique est définie naturellement : on fait grandir la taille de l'échantillon avec n en ajoutant à chaque pas une observation, indépendante des précédentes et identiquement distribuée. Dans le contexte des données géoréférencées, il faut définir le type d'asymptotique, c'est à dire la manière dont sont ajoutées les observations. En effet, une nouvelle observation ne sera *a priori* pas indépendante des précédentes. On distingue essentiellement deux types d'asymptotique :

- l'asymptotique par accroissement du domaine (*increasing domain asymptotic*) : les observations sont ajoutées à une distance de celles déjà disponibles minorée, de telle sorte que le volume du domaine d'observations (disons de son enveloppe convexe) tend vers l'infini.
- l'asymptotique par densification du domaine (*infill (ou fixed domain) asymptotic*) : les observations sont ajoutées dans un domaine $\mathcal{D}$ fixe de manière à ce que celui-ci se densifie en sites d'observation.

Ces deux grands types d'asymptotiques (on pourrait en définir d'autres) se caractérisent par des résultats très différents en terme de convergence des estimateurs. Avant de présenter les formalismes utilisés pour ces deux types d'asymptotique, nous allons voir sur un exemple comment se forger une intuition de ces différences.

5.2.2 Exemple introductif

Soit un champ aléatoire gaussien d'espérance nulle et de fonction de covariance

$$C(h) = exp(-h/10)$$

sur une droite $\mathcal{D}$. On suppose que l'espérance, la variance et le paramètre de portée sont inconnus. Le modèle, qui est un sous modèle de la classe **(LM2)**, s'écrit :

$$\forall\{s_1, \ldots, s_n\} \in \mathcal{D}, (Y(s_1), \ldots, Y(s_n)) \sim \mathcal{N}(\mu, \sigma^2 H(\phi)),$$

où le $(i,j)^{\text{ème}}$ terme de $H(\phi)$ est donné par $\exp(-|s_i - s_j|/\phi)$. On s'intéresse à l'estimation de $\psi = (\mu, \sigma^2, \phi)$. La figure (5.1) montre une réalisation du champ sur l'intervalle $I_1 = [0, 5000]$ et un zoom sur $I_2 = [0, 5]$.

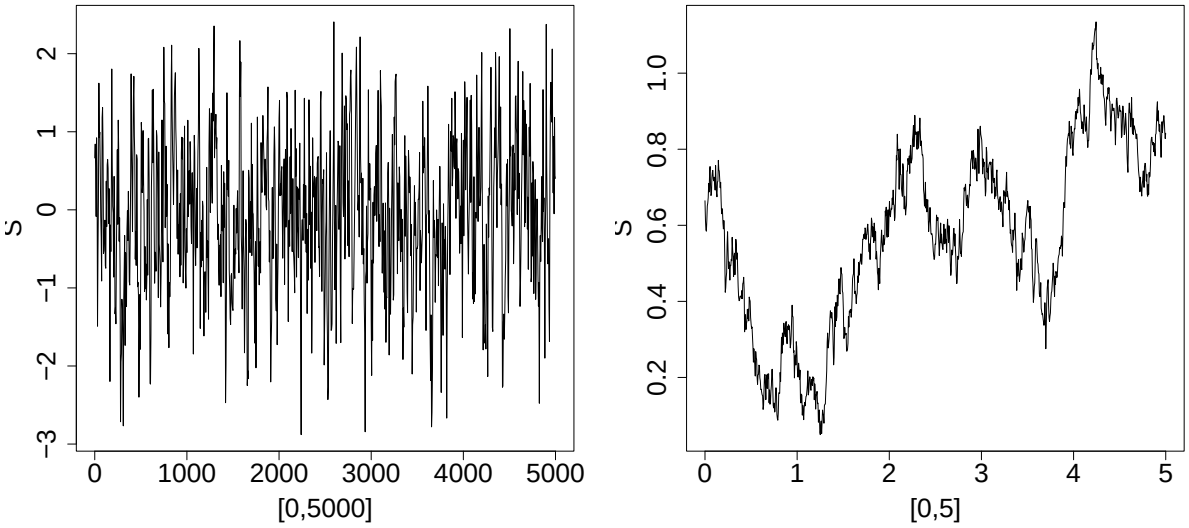

Fig. 5.1. Une réalisation d'un champ aléatoire stationnaire d'ordre 2 sur [0,5000] (à gauche) et un zoom sur [0,5] (à droite).

Il est clair que le paramètre d'espérance μ ne pourra pas être estimé de manière consistante sur I_2 puisque cet intervalle contient un ensemble d'observations corrélées qui sont dans une zone de valeurs positive de la réalisation du champ. On pourra densifier les sites d'observation sur I_2, l'estimateur de l'espérance de μ ne pourra pas tendre vers μ, même dans le cas où les paramètres σ^2 et ϕ seraient connus. Au contraire, l'intervalle I_1 est grand au regard des structures spatiales. Si le domaine s'agrandit, chaque nouvelle observation, peu corrélée avec les précédentes, amène de l'information "nouvelle" sur μ ; on s'attend à ce que l'estimateur $\hat{\mu}$ soit convergent. Pour les paramètres σ^2 et ϕ, le problème de la convergence des estimateurs est moins clair. En asymptotique par densification du domaine, de même que pour l'espérance, on "sent" déjà qu'il va être difficile d'estimer la variance du champ aléatoire ; mais qu'en est il si l'on connaît le paramètre de portée, c'est-à-dire si l'on connait les contraintes spatiales du champ ? De même si l'on connaît la variance du champ, n'est il pas possible d'en déduire les contraintes spatiales (le paramètre de portée) ? Dans ce cas, ne peut-on pas trouver une fonction des paramètres ϕ et σ^2 qui soit estimable ?

5.2.3 Asymptotique par accroissement du domaine

Nous présentons ici la démarche permettant de prouver la convergence asymptotique de l'estimateur par maximum de vraisemblance dans les modèles gaussiens spatiaux présentés dans le chapitre 3, sous un schéma d'asymptotique *increasing domain* . Le point de départ de la démonstration est un théorème général de Sweeting (1980) donnant des conditions nécessaires à la normalité asymptotique (uniforme) de l'estimateur du maximum de vraisemblance de processus. L'adaptation au contexte spatial (plus précisément au modèle **(LM2)**) est due à Mardia et Marshall (1984). Les idées originales développées dans cet article ont ensuite été reprises par Cressie et Lahiri (1993) dans le cas du REML et par Cressie et Lahiri (1996) dans cas du REML dans le contexte spatial. Ce fut l'occasion de préciser certains éléments de la démonstration de Mardia et Marshall et notamment de rajouter une hypothèse que ces derniers avaient omise, et qui, dans l'esprit de leur démonstration, est nécessaire à la convergence. Nous reprenons les éléments de ces 4 articles pour donner une démonstration complète de la normalité asymptotique de l'estimateur par maximum de vraisemblance dans le cas des modèles **(LM4)**.

Définition et notations :

Soit $(\Omega_n, \mathcal{A}_n)$ une suite d'espaces mesurables et P_ψ^n une mesure de probabilité définie sur $(\Omega_n, \mathcal{A}_n)$ dépendant du vecteur de paramètre $\psi \in \Theta$, un ensemble ouvert de $\mathbb{R}^k$. On suppose que pour tout n, et tout $\psi \in \Theta$, P_ψ^n est absolument continue par rapport à une mesure σ-finie λ_n. Soit $p_n(\psi)$ la densité de P_ψ^n par rapport à λ_n. On suppose que les dérivées secondes partielles de $p_n(\psi)$ existent et sont continues pour tout $\psi \in \Theta$. On note $l_n(\psi) = \log p_n(\psi)$ et $\mathcal{F}_n(\psi)$ la matrice d'information (aléatoire), c'est à dire la matrice $k \times k$ dont le $(i,j)^{\text{ème}}$ élément est donné par :

$$\frac{\partial^2 l_n}{\partial \psi_i \partial \psi_j}(\psi)$$

On travaille avec la norme matricielle euclidienne (dite également norme de Frobénius) définie par :

$$||A|| = \sqrt{\text{trace}(A'A)} = \left(\sum_{j=1}^{n} \sum_{j=1}^{n} a_{ij}^2 \right)^{1/2}.$$

Notons Γ la matrice $(\psi^1, \ldots, \psi^k)$ où $\psi^i \in \Theta, i = 1, \ldots, k$, et définissons $\mathcal{F}_n(\Gamma)$ qui est la matrice $\mathcal{F}_n$ où la colonne i est évaluée en ψ^i.

Pour toute suite de fonctions $(f_n)_n$, on notera $f_n \to_u f$ si $(f_n)_n$ converge uniformément vers f quand n tend vers l'infini.

Pour toute suite de vecteurs aléatoires, $(\mathbf{V}_n)_n$, on notera $\mathbf{V}_n \Rightarrow_u \mathbf{V}$ où $\mathbf{V}$ est un vecteur aléatoire, si pour toute fonction $f : \mathbb{R}^n \to \mathbb{R}$ bornée et continue $E_\phi[f(\mathbf{V}_n)] \to E_\phi f(\mathbf{V})$. Notons que par le théorème d'Helly-Bray, $\mathbf{V}_n \Rightarrow_u \mathbf{V}$ implique que $\mathbf{V}_n$ converge en loi vers $\mathbf{V}$ (Cressie et Lahiri, 1993).

Théorème 2 *Sweeting (1980)*

On suppose que $\mathcal{F}_n(\psi)$ satisfait les conditions suivantes :

C1. (accroissement de l'information et convergence) Il existe des matrices carrées non aléatoires $A_n(\psi)$, continues en ψ satifaisant

$$\{A_n(\psi)\}^{-1} \to_u 0 \tag{5.1}$$

et telles que

$$R_n(\psi) \equiv \{A_n(\psi)\}^{-1} \mathcal{F}_n(\psi) \{A_n(\psi)\}^{-1'} \Rightarrow_u R(\psi) \tag{5.2}$$

où $R(\psi)$ est définie-positive presque sûrement.

C2. (*continuité*) Pour tout $c > 0$

(i) $\sup |\{A_n(\psi)\}^{-1}\{A_n(\psi')\} - I_k| \rightarrow_u 0$ où le sup est obtenu sur l'ensemble

$|\{A_n(\psi)\}'(\psi' - \psi)| \leq c$

(ii) $|\{A_n(\psi)\}^{-1}[\mathcal{F}_n(\Gamma) - \mathcal{F}_n(\psi)]\{A_n(\psi)\}^{-1'}| \rightarrow_u 0$ en probabilité, où le sup est obtenu sur l'ensemble $|\{A_n(\psi)\}'(\psi_i - \psi)| \leq c, 1 \leq i \leq k$.

Soit $\hat{\psi}_n$, le maximum de vraisemblance. Sous les hypothèses (C1) et (C2), on a

$$A_n(\psi)'(\hat{\psi}_n - \psi) \Rightarrow_u \mathcal{N}_k(0, R(\psi)^{-1}).$$

Dans l'esprit de Mardia et Marshall (1984), et Cressie et Lahiri (1996), nous nous concentrerons exclusivement sur (C1) et nous postulerons que (C2) est vérifiée. Les conditions données par (C2) sont des conditions de régularité satisfaites dès lors que la matrice $\Sigma(\phi)$ est suffisamment régulière en ϕ (voir Cressie et Lahiri, 1993, pour plus de détails).

La première étape consiste à simplifier (C1) en donnant une condition suffisante pour que cette hypothèse soit vérifiée.

D'après lemme 4.2 de Sweeting (1980), pour montrer (5.2), il suffit de montrer que

$$\{A_n(\psi)\}^{-1}\mathcal{F}_n(\psi)\{A_n(\psi)\}^{-1'}$$

converge en probabilité vers $R(\psi)$ uniformément en ψ. Puis par l'inégalité de Chebyshev, il est suffisant de montrer que

$$E_\psi[||\{A_n(\psi)\}^{-1}\mathcal{F}_n(\psi)\{A_n(\psi)\}^{-1'} - R(\psi)||^2] \rightarrow_u 0. \tag{5.3}$$

Une suite naturelle de matrices $A_n(\psi)$ est donnée par $A_n(\psi) = E_\psi[\mathcal{F}_n(\psi)]^{1/2}$ de telle sorte que $R(\psi) = I_k$. D'autres choix peuvent permettre dans certains cas des simplifications dans la démonstration (voir par exemple Cressie et Lahiri (1993) pour la démonstration de la convergence du REML).

Dans le cas des modèles gaussiens présentés dans le chapitre 3, on peut décomposer le vecteur de paramètres de la manières suivante : $\psi = (\beta, \phi)$ où β est le vecteur de taille p des paramètres de régression et ϕ, de taille q, est celui des paramètres de la matrice de covariance des observations. On dénotera les espaces associés à β et ϕ, Θ_p et Θ_q de telle sorte que $\Theta = \Theta_p \times \Theta_q$.

On note :

$$\mathcal{F}_n(\psi) = \begin{pmatrix} L_{\beta\beta} & L_{\beta\phi} \\ L'_{\beta\phi} & L_{\phi\phi} \end{pmatrix}$$

où

$$L_{\beta\beta} = \frac{\partial l(\psi, \mathbf{W})}{\partial\beta\partial\beta'}$$

$$L_{\beta\phi} = \frac{\partial l(\psi, \mathbf{W})}{\partial\beta\partial\phi'}$$

et

$$L_{\phi\phi} = \frac{\partial l(\psi, \mathbf{W})}{\partial\phi\partial\phi'}.$$

dont les expressions sont données au chapitre 3.

et

$$A_n(\psi) = \begin{pmatrix} A_{\beta\beta} & \mathbf{0} \\ \mathbf{0} & A_{\phi\phi} \end{pmatrix} = \begin{pmatrix} E_\psi[L_{\beta\beta}]^{1/2} & \mathbf{0} \\ \mathbf{0} & E_\psi[L_{\phi\phi}]^{1/2} \end{pmatrix}$$

puisque $E_\psi[L_{\beta\phi}] = 0$.

Du fait que $L_{\beta\beta}$ est non aléatoire, pour montrer (5.3), il suffit de montrer que :

$$E_\psi[||A_{\phi\phi}^{-1}L_{\phi\beta}A_{\beta\beta}^{-1}||^2] \to_u 0 \tag{5.4}$$

$$E_\psi[||A_{\phi\phi}^{-1}L_{\phi\phi}A_{\phi\phi}^{-1} - I_q||^2] \to_u 0 \tag{5.5}$$

De ces considérations, on déduit le théorème suivant :

Théorème 3 *Sous l'hypothèse (C2), si (5.4) et (5.5) sont vérifiées et si*

$$A_{\beta\beta}^{-1} \to_u \boldsymbol{0} \text{ et } A_{\phi\phi}^{-1} \to_u \boldsymbol{0}, \tag{5.6}$$

alors

$$A_n(\psi)(\hat{\psi}_n - \psi) \Rightarrow_u \mathcal{N}_k(\boldsymbol{0}_k, I_k)$$

Notons $b_{ij}^{\psi,\phi}$ le $(i,j)^{\text{ème}}$ terme de la matrice $A_{\phi\phi}^{-2}$, $b_{js}^{\psi,\beta}$ le $(j,s)^{\text{ème}}$ terme de $A_{\beta\beta}^{-2}$ et x_j la $j^{\text{ème}}$ colonne de la matrice X. Le théorème 3 peut être reformulé de la manière suivante :

Théorème 4 *Sous l'hypothèse (C2), si (5.6) est vérifiée et si :*

$$\sum_{r,s,i,j=1}^{q} b_{ir}^{\psi,\phi}b_{js}^{\psi,\phi}tr(\Sigma\Sigma^{(rj)}\Sigma\Sigma^{(si)}) \to_u 0 \tag{5.7}$$

et

$$\sum_{r,i=1}^{q}\sum_{s,j=1}^{p} b_{ir}^{\psi,\phi}b_{js}^{\psi,\beta}x_j'\Sigma^{-1}\Sigma_r\Sigma^{-1}\Sigma_i\Sigma^{-1}x_s \to_u 0 \tag{5.8}$$

alors

$$A_n(\psi)'(\hat{\psi}_n - \psi) \Rightarrow_u \mathcal{N}_k(\boldsymbol{0}, I_k).$$

Démonstration. La démonstration est donnée en annexe.

Le théorème suivant fournit un ensemble de conditions suffisantes sur les matrices de covariance des observations et sur leurs dérivées d'ordre 1 et 2 par rapport aux paramètres pour avoir la convergence asymptotique de l'estimateur par maximum de vraisemblance vers la normalité.

Notons $\lambda_1 \leq \cdots \leq \lambda_n$, les valeurs propres de Σ ordonnées par ordre croissant, $|\lambda_1^i| \leq \cdots \leq |\lambda_n^i|$ celles de Σ_i et $|\lambda_1^{ij}| \leq \cdots \leq |\lambda_n^{ij}|$, celles de $\ddot{\Sigma}_{ij}$ ordonnées par valeurs absolues croissantes, pour $i,j = 1,\ldots,q$. De plus, on note $t_{ij} = \text{tr}(\Sigma^{-1}\Sigma_i\Sigma^{-1}\Sigma_j)$ et $g_{ij} = \frac{t_{ij}}{(t_{ii}t_{jj})^{1/2}}$ pour tout $i,j = 1,\ldots,q$ et $G_n(\phi) = ((g_{ij}))$

Théorème 5 *On suppose (C2) et*

(i) Il existe une matrice $G(\phi)$ non singulière telle que $G_n(\phi) \to_u G(\phi)$.

(ii) Pour tout compact $K \subseteq \Theta$, il existe 2 constantes $0 < C(K) \leq \infty$ et $\eta(K) > 0$ telles que

$$\forall n \in \mathbb{N}, \max\{|\lambda_n|, |\lambda_n^i|, |\lambda_n^{ij}|, i,j = 1,\ldots,q\} \leq C(K) \tag{5.9}$$

et

$$\forall n \in \mathbb{N}, |\lambda_1| \geq \eta(K) \tag{5.10}$$

uniformément en $\phi \in K$.

(iii) Pour tout compact K, il existe une suite $(r_n(K))_{n\geq 1}$ telle que $(r_n(K)) = o(n^{-1/2})$ et

$$||\Sigma_i||^{-2} \leq r_n(K) \tag{5.11}$$

pour tout $\phi \in K$

(iv) $\lim_{n\to\infty}(X'X)^{-1} = \boldsymbol{0}$

alors

$$E[\mathcal{F}_n(\psi)]^{1/2}(\psi)'(\hat{\psi}_n - \psi) \Rightarrow_u \mathcal{N}_k(\boldsymbol{0}, I_k).$$

La condition (i) assure essentiellement que la matrice $A_{\phi\phi}$ n'est pas singulière à la limite, c'est à dire que les paramètres de covariance ne sont pas asymptotiquement linéairement dépendants.

La condition (iv) se vérifie immédiatement dans les modèles **(LM)** présentés dans le chapitre 3 qui sont stationnaires en espérance. En effet, dans ce cas, $\beta = (\mu_X, \mu_Y)$ et $X = \begin{pmatrix} \mathbb{1}_{n_X} & \mathbf{0} \\ \mathbf{0} & \mathbb{1}_{n_Y} \end{pmatrix}$.

Le théorème suivant donne des conditions sur les matrices de variance et leurs dérivées qui sont suffisantes à la convergence asymptotique de $\hat{\psi}_n$. Ces conditions, bien que plus fortes que celles du théorème 5 sont parfois plus facilement vérifiables en pratique puisqu'elles ne concernent que des sommes d'éléments de ces matrices et non leurs valeurs propres.

On note $S_n = \{s_1, \ldots, s_n\}$ la suite d'ensembles de sites d'observation, dont la taille croit avec n et $\sigma_{ij}(S_n)$, $\sigma^k_{ij}(S_n)$ et $\sigma^{kl}_{ij}(S_n)$ les $(i,j)^{\text{ème}}$ éléments respectifs des matrices Σ, Σ^k et Σ^{kl}, $k, l = 1, \ldots, q$ pour les sites de S_n.

Théorème 6 *On suppose que les hypothèses (i), (iii) et (iv) du théorème 5 ainsi que (C2) sont vérifiées.*

Si pour tout compact $K \subset \Theta^q$*, il existe 2 constantes* $0 < C(K) < \infty$ *et* $\eta_1(K) \in]0,1[$ *telles que*

$$\forall n \in \mathbb{N}^\star, \quad \max_{1\le i\le n} \sum_{j}^{n} |\sigma^{(.)}_{ij}(S_n)| \le C(K) \tag{5.12}$$

pour $\sigma^{(.)}_{ij} = \sigma_{ij}, \sigma^k_{ij}$ *et* $\sigma^{kl}_{ij}; 1 \le k, l \le q$.

$$\forall n \in \mathbb{N}^\star, \quad \max_{1\le i\le n} \sum_{\substack{j=1 \\ j\neq i}}^{n} |\sigma_{ij}(S_n)| \le \eta_1(K) \min_{1\le i\le n} \sigma_{ii}(S_n) \tag{5.13}$$

alors

$$E[\mathcal{F}_n(\psi)]^{1/2}(\hat{\psi}_n - \psi) \Rightarrow_u \mathcal{N}_k(\boldsymbol{0}, I_k).$$

Démonstration. Voir annexe

Le théorème suivant donne un exemple d'utilisation du théorème 6 :

Théorème 7 *Soit* $X(.)$ *un champ gaussien stationnaire d'ordre 2, d'espérance* μ *et de fonction de covariance exponentielle donnée par*

$$C(h) = \sigma^2 \exp(-h\phi).$$

On observe une réalisation de $X(.)$ *sur* S_n*, une grille rectangulaire à mailles régulières, de taille* $n = \prod_{k=1}^{d} n_k$ *pour des entiers positifs* $n_k, k = 1, \ldots, d$*. On suppose que quand* n *tend vers l'infini, la grille se remplit de manière homogène dans toutes les directions, c'est-à-dire que pour tout* $k \in [\![1, d]\!]$*,* n_k *tend vers l'infini. On note* $\boldsymbol{h} = (h_1, \ldots, h_d)'$ *le vecteur des pas de grille dans toutes les d directions.*

On note

$$\Theta = \Theta^\mu \times \Theta^\phi \times \Theta^{\sigma^2},$$

l'ouvert de $\mathbb{R}^3$ *dans lequel varie le vecteur des paramètres à estimer* $\psi = (\mu, \phi, \sigma^2)$. *Et pour tout compact de* $K \subset \Theta$, *on note de même*

$$K = K^{\mu} \times K^{\phi} \times K^{\sigma^2}.$$

On suppose que la condition (C2) est vérifiée. Soit $\phi_m > 0$ *et posons* $\Theta_1 =]\phi_m, \infty[$. *Pour tout compact* $K = K^{\phi} \times K^{\sigma^2} \times K^{\mu}$ *de* Θ *vérifiant* $K^{\phi} \subset \Theta_1$, *si*

$$\min_{1 \leq k \leq d} h_i \geq \frac{\sqrt{d} \log 3}{\phi_m},$$

alors $\hat{\psi}_n$, *l'estimateur par maximum de vraisemblance de* ψ *calculé à partir de* $\boldsymbol{X} = (X(s_1), \ldots, X(s_n))$ *vérifie, quand* n *tend vers l'infini :*

$$E[\mathcal{F}_n(\psi)]^{1/2}(\hat{\psi}_n - \psi) \Rightarrow_u \mathcal{N}_3(\boldsymbol{0}, I_3).$$

Démonstration. Voir annexe

Le théorème 7 est une généralisation à $\mathbb{R}^d$ de résultats connus en dimension 1. En effet, en dimensions 1, les hypothèses du théorème 7 conduisent, à une reparamétrisation près, à un modèle gaussien AR(1) utilisé pour les séries temporelles. Or, la normalité asymptique du maximum de vraisemblance est connue dans ce cas (Zhang et Zimmerman, 2005). Notons également que Cressie et Lahiri (1996) traitent le cas d'un modèle auto-régressif du premier ordre en dimensions d séparable. Nous avons été confronté à des difficultés techniques supplémentaires dues à la non-séparabilité (en les dimensions de l'espace) de notre modèle. Enfin, notons que nous avons montré que dans notre cas, il n'était pas nécessaire de postuler l'hypothèse (i) du théorème 5.

Concernant les hypothèses du théorème, il apparaît clairement ici qu'elles correspondent uniquement à un schéma asymptotique de type accroissement du domaine. De même qu'il est clair que les hypothèses du théorème 6 ne sont pas vérifiées sous un schéma d'asymptotique par densification. Pour le théorème 5, cela est peut-être moins évident, mais les valeurs propres de la matrice de variance ne peuvent pas être bornées quand le domaine d'observation se densifie en sites d'observation.

Nous donnons quelques éléments sur l'asymptotique par densification dans la sous-section suivante.

5.2.4 Asymptotique par densification du domaine

Lorsque les données sont ajoutées dans un domaine borné de telle sorte que le domaine se densifie, le comportement des estimateurs est très différent du contexte précédent et le formalisme qui permet de l'apprécier également. Ce formalisme (Stein, 1999) utilise notamment les mesures gaussiennes équivalentes afin d'étudier le comportement asymptotique des paramètres de modèles gaussiens spatiaux.

Mesures de probabilités orthogonales et équivalentes

Rappellons quelques définitions :

Soient P_1 et P_2 deux mesures de probabilité sur l'espace mesurable $(\Omega, \mathcal{F})$.

On dira que P_1 est absolument continue par rapport à P_2 et on notera $P_1 \ll P_2$ si pour tout $A \in \mathcal{F}$, $P_2(A) = 0$ implique que $P_1(A) = 0$.

P_1 et P_2 sont dites équivalentes (on note $P_1 \equiv P_2$) si $P_1 \ll P_2$ et $P_2 \ll P_1$.

Enfin P_1 et P_2 sont dites orthogonales (on note $P_1 \perp P_2$) si il existe $A \in \mathcal{F}$ tel que $P_1(A) = 1$ et $P_2(A) = 0$.

Considérons maintenant un processus stochastique quelconque $\{Y(s), s \in T\}$ pour un ensemble T, et soit $\mathcal{F} = \sigma(\{Y(s), s \in T\})$ la tribu engendrée par ce processus. Deux mesures P_1 et P_2 sont dites équivalentes sur $\{Y(s), s \in T\}$ si elles le sont sur $\mathcal{F}$.

En statistiques, si P_1 et P_2 sont les mesures associées à 2 modèles probabilistes, si $P_1 \equiv P_2$, alors on ne pourra pas correctement distinguer P_1 de P_2 avec P_1-probabilité 1 au regard des observations.

Plus précisemment, si $\{P_\phi, \phi \in \Theta\}$ est une famille de mesures équivalentes et $\hat{\phi}_n, n \geq 1$ est une séquence d'estimateurs, alors $\hat{\phi}_n$ ne peut pas être un estimateur faiblement convergent pour tout $\phi \in \Theta$. En effet, dans ce cas, pour tout ϕ de Θ, il existerait alors une sous-suite $\{\hat{\phi}_{n_k}, k \geq 1\}$ qui convergerait fortement (Zhang (2004)), c'est-à-dire $P_\phi(\hat{\phi}_{n_k} \to \phi, k \to \infty) = 1$; et par l'équivalence des mesures, pour tout ϕ', $P_{\phi'}(\hat{\phi}_{n_k} \to \phi, k \to \infty) = 1$).

D'autre part, la faible convergence de la sous-suite $\{\hat{\phi}_{n_k}, k \geq 1\}$ sous la mesure $P_{\phi'}$ implique de même l'existence d'une sous sous-suite qui converge vers ϕ' avec $P_{\phi'}$-probabilité 1 . Cette sous sous-suite converge donc vers 2 valeurs différentes sous la même mesure $P_{\phi'}$. Cette contradiction montre que $\hat{\phi}_n$ ne peut pas être faiblement convergent.

Pour conclure les généralités sur les mesures équivalentes ou orthogonales, considérons l'exemple suivant (Zhang, 2004). Soit $Y_1, \ldots, Y_n, \ldots$, une famille de variables aléatoires i.i.d de loi normales $\mathcal{N}(0, \sigma_i^2)$ sous $P_i, i = 1, 2$ avec $\sigma_1^2 \neq \sigma_2^2$. Alors, sur la tribu engendrée par les $Y_i, i = 1, \ldots, \infty$, les deux mesures sont orthogonales car si on définit l'ensemble A par

$$A = \left\{ \omega \in \Omega; \frac{1}{n} \sum_{i=1}^{n} Y_i(\omega)^2 \to \sigma_1^2 \text{ quand } n \to \infty \right\},$$

alors $P_1(A) = 1$ et $P_2(A) = 0$ par la loi des grands nombres.

Notons que 2 mesures peuvent n'être ni équivalentes, ni orthogonales.

Applications aux champs gaussiens

Supposons que m_1 et m_2 sont deux fonctions continues sur $\mathbb{R}^d$ et C_0 et C_1 sont deux fonctions continues et définies-positives sur $\mathbb{R}^d$. Soit $Y(.) = \{Y(s), s \in \mathbb{R}^d\}$ un champ aléatoire sur $\mathbb{R}^d$. Soit $\mathcal{D}$ un sous-ensemble fermé de $\mathbb{R}^d$ et soient $P_i = G_{\mathcal{D}}(m_i, C_i), i = 1, 2$, deux mesures gaussiennes sur la tribu engendrée par $Y(.)$ et définies par : sous P_i, $Y(.)$ est un champ gaussien stationnaire sur $\mathcal{D}$ dont l'espérance est donnée par la fonction m_i et dont la fonction de covariance est C_i.

Nous donnons quelques théorèmes généraux qui nous semblent intéressants pour l'étude de l'asymptotique sur un domaine borné. Ces théorèmes peuvent être trouvés dans Ibragimov et Rozanov (1978) et Stein (1999).

Théorème 8 *Les deux mesures P_1 et P_2 sont ou bien équivalentes ou bien orthogonales.*

Ce théorème permet de montrer l'équivalence entre deux mesures gaussiennes stationnaires sur un domaine borné en montrant qu'elles ne sont pas orthogonales.

Théorème 9 *$G_{\mathcal{D}}(0, C_0)) \equiv G_{\mathcal{D}}(m_1, C_1)$ si et seulement si $G_{\mathcal{D}}(0, C_0) \equiv G_{\mathcal{D}}(m_1, C_0)$ et $G_{\mathcal{D}}(0, C_0) \equiv G_{\mathcal{D}}(0, C_1)$.*

Ce théorème permet de séparer l'espérance et la covariance dans l'étude de l'équivalence entre deux mesures gaussiennes sur un domaine borné.

Pour l'espérance on a :

Théorème 10 *$G_{\mathcal{D}}(0, C_1) \equiv G_{\mathcal{D}}(m_1, C_1)$ si et seulement si m_1 peut être étendue à une fonction carrée intégrable sur tout $\mathbb{R}^d$ dont la transformation de Fourier $\tilde{m}_1$ vérifie*

$$\int_{\mathbb{R}^d} \frac{|\tilde{m}_1(\omega)|^2}{f(\omega)} d\omega < \infty$$

A partir d'un nombre infini d'observations d'une réalisation de $Y(.)$ sur un domaine borné, on peut donc confondre l'espérance nulle avec toute une fonction sous réserve de conditions de régularité de la fonction espérance mise en jeu.

Dans le cas $d = 1$, si $\mathcal{D} = [0, T], T > 0$, si il existe un entier p tel que $f(\omega)\omega^{2p} \approx 1$ quand $n \to \infty$, alors

$$G_{[0,T]}(0, C_1) = G_{[0,T]}(m_0, C_1)$$

si et seulement si m_1 est $p-1$ fois dérivable et si $m_1^{(p-1)}$ est absolument continue sur $[0, T]$ et si $m_1^{(p)}$ vérifie :

$$\int_0^T (m_1^{(p)}(t))^2 < \infty.$$

Pour les fonctions de covariance on a :

Théorème 11 *Soient f_0 et f_1 les densités spectrales respectives de C_0 et C_1. Supposons qu'il existe $\beta > d$ tel que $a < f_0(\omega)|\omega|^\beta < b$ quand $|v| \to \infty$, où $0 < a < b$.*

Si il existe $C \in \mathbb{R}^+$ tel que

$$\int_{|\omega|>C} \left(\frac{f_1(\omega) - f_0(\omega)}{f_0(\omega)} \right)^2 d\omega < \infty,$$

alors $G_{\mathcal{D}}(0, C_0) \equiv G_{\mathcal{D}}(0, C_1)$.

Ce théorème indique que si le comportement des deux densités spectrales associées à 2 modèles de covariance sont similaires aux hautes fréquences, autrement dit, si les deux fonctions de covariance ont le même comportement à l'origine, alors on ne pourra pas distinguer les deux modèles à partir d'observations dans un domaine borné, même en densifiant les points à l'intérieur du domaine. Si on considère que f_0 est la vraie densité spectrale que l'on cherche à estimer par f_1, l'hypothèse qu'il existe $\beta > d$ tel que $f_0(\omega)|\omega|^\beta$ soit borné quand $|v| \to \infty$ assure que la fonction de de covariance C_0, n'est pas trop régulière à l'origine. On retrouve ici un fait que nous avions déjà évoqué au chapitre 2. Matheron (1970) avait en effet démontré ce type de résultat, dans un cadre beaucoup plus général. Le formalisme des mesures gaussiennes équivalentes permet néanmoins de le faire de manière plus rigoureuse.

En terme d'estimation, Zhang (2004) utilise ce résultat pour démontrer que si $d = 1, 2$ ou 3, si C_0 et C_1 sont deux modèles de Matérn sur $\mathbb{R}^d$ paramètrés comme suit :

$$C_i(h) = \frac{\sigma_i^2(\alpha_i h)^\nu}{\Gamma(\nu)2^{\nu-1}} \mathcal{K}_\nu(\alpha_i h),$$

où $\mathcal{K}_\nu$ est la fonction de Bessel modifiée d'ordre 2, et ν est supposé connu, alors $G_{\mathcal{D}}(0, C_0) \equiv G_{\mathcal{D}}(0, C_1)$ si et seulement si $\sigma_0^2\alpha_0^{2\nu} = \sigma_1^2\alpha_1^{2\nu}$. Notons que contrairement au théorème 11, Zhang (2004) démontre l'équivalence entre les deux propositions.

Un cas particulier du théorème de Zhang est le modèle exponentiel, c'est à dire quand $\nu = 1/2$. Dans ce cas, seul le rapport variance sur portée est estimable de manière consistante (ce rapport est égal à la pente de la fonction de covariance à l'origine).

Pour conclure cette partie, notons que l'équivalence des mesures n'est pas un problème quand l'objectif de l'estimation des paramètres est l'interpolation aux sites ou l'on ne dispose pas d'information. En effet, deux krigeages menés sous deux mesures gaussiennes équivalentes sont asymptotiquement égaux (en valeur d'interpolation et en variance de prédiction). Nous référons une fois de plus à Stein (1999) et Zhang (2004) pour plus de détails.

5.3 Propriétés de l'estimateur du maximum de vraisemblance du coefficient de corrélation

5.3.1 Introduction

Dans cette section, nous allons étudier le comportement de l'estimateur par maximum de vraisemblance du coefficient de corrélation dans un modèle de corrélation intrinsèque et dans le cas de données hétérotopes. On rappelle que ce modèle s'écrit :

pour tout ensemble de sites $\{s_1, \ldots, s_n\} \in \mathcal{D}$, si $\mathbf{X} = (X(s_1), \ldots, X(s_n))'$ et $\mathbf{Y} = (Y(s_1), \ldots, Y(s_n))'$, alors :

$$\begin{pmatrix} \mathbf{X} \\ \mathbf{Y} \end{pmatrix} \sim \mathcal{N}(\mu \otimes \mathbb{1}_n, T \otimes H(\phi)) \tag{5.14}$$

avec $\mu = \begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}$, $H(\phi)$ une matrice de corrélation spatiale définie en appliquant une fonction de corrélation paramétrée par ϕ à la matrice de distance entre les sites et

$$T = \begin{pmatrix} \sigma_X^2 & \sigma_{XY} \\ \sigma_{XY} & \sigma_Y^2 \end{pmatrix} \text{ est une matrice définie positive.}$$

Ce modèle est une simple généralisation au contexte spatial du modèle utilisé pour l'inférence sur le coefficient de corrélation pour un échantillon bivarié. On a remplacé la matrice identité (conséquence de l'indépendance des observations) par $H(\phi)$ afin de prendre en compte les corrélations spatiales entre les observations.

Pascual et Zhang (2005) montrent que pour des données alignées si ϕ est connu, l'estimateur par maximum de vraisemblance du coefficient de corrélation a la même distribution que le coefficient de corrélation usuel

$$\tilde{r} = \frac{\sum_{i,j=1}^{n}(X_i - \bar{X})(Y_j - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

sous l'hypothèse d'indépendance des observations. Pour étudier comment l'hétérotopie dégrade le comportement de l'estimateur du coefficient de corrélation, le cadre i.i.d peut donc fournir une base de comparaison intéressante. Rappelons quelques résultats de ce contexte.

A taille finie

Sous l'hypothèse que les observations sont une réalisation du modèle (5.14) avec $H(\phi) = I_n$, l'estimateur $\tilde{r}$ du coefficient de corrélation a la densité de probabilité suivante :

- Pour $n = 2$, la distribution est un mélange de 2 dirac (une en +1 et l'autre en -1) dont les probabilités associées dépendent de la vraie valeur du coefficient de corrélation.

– Pour tout $n \geq 3$:

$$f_{\tilde{r}}(x) = \frac{(1-r)^{(n-1)/2}}{\pi\Gamma(n-2)}(1-x^2)^{(n-4)/2}\frac{d^{n-2}}{d(xr)^{n-2}}\left\{\frac{\cos^{-1}(-rx)}{\sqrt{1-r^2x^2}}\right\}. \tag{5.15}$$

Cette forme est difficilement utilisable pour de grandes tailles d'échantillon (des formes alternatives de cette formule, toutes aussi pittoresques, sont données dans Kendall et Stuart, 1979). Mais notons que pour $n = 3$, la densité a une forme de U pour $|r|$ proche de 0 et une forme de J pour $|r|$ plus grand et est unimodale pour $n \geq 4$ et d'autant plus disymètrique que $|r|$ est grand (voir figure 5.2).

Asymptotique

On peut montrer que

$$\sqrt{n}(\tilde{r} - r) \xrightarrow{\mathcal{L}} \mathcal{N}(0, (1-r^2)^2).$$

La convergence en loi est d'autant plus rapide que $|r|$ est proche de 0.

Enfin, notons que la variance asymptotique dépend de r. Pour construire des intervalles de confiance, on peut utiliser la fonction arctanh qui stabilise la variance (Fisher, 1921). En effet,

$$\sqrt{n}(\tanh^{-1}(\tilde{r}) - \tanh^{-1}(r)) \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1).$$

De plus la convergence en loi de $\tanh^{-1}(\tilde{r})$ vers la distribution gaussienne est beaucoup plus rapide que celle de $\tilde{r}$ et la variance de $\tilde{r}$ à taille finie est presque indépendante de r.

5.3.2 Modèle spatial

Observations isotopes

Paramètres de strucutre spatiale connus

Comme nous l'avons rappelé, pour un modèle de corrélation intrinsèque, dans le cas d'observations isotopes, quand le vecteur ϕ des paramètres de structure spatiale est connu, l'estimateur par maximum de vraisemblance du coefficient de corrélation a la même distribution (à taille finie) que le coefficient de corrélation usuel (Pascual et Zhang, 2005). Il est donc inutile de considérer plus en détail sa distribution asymptotique et les conditions pour que cette distribution soit atteinte : quel que soit le type d'asymptotique spatial considéré, l'estimateur du coefficient de corrélation $\hat{r}$ vérifie :

$$\sqrt{n}(\hat{r} - r) \xrightarrow{\mathcal{L}} \mathcal{N}(0, (1-r^2)^2).$$

Paramètres de structure spatiale inconnus

Dans le cas où ϕ est inconnu Pascual et Zhang (2005) montrent sur simulation que $\hat{r}(\hat{\phi})$, l'estimateur par maximum de vraisemblance du coefficient de corrélation obtenu pour ϕ inconnu est très proche de $\hat{r}(\hat{\phi})$, l'estimateur obtenu avec ϕ fixé à sa vraie valeur. D'un point de vue théorique, pour démontrer la convergence asymptotique de $\hat{r}$ et plus généralement de $\hat{\psi}_n$, on se retrouve confronté au même problème que pour un champ aléatoire univarié. On doit définir le type d'asymptotique spatial. On a le théorème suivant :

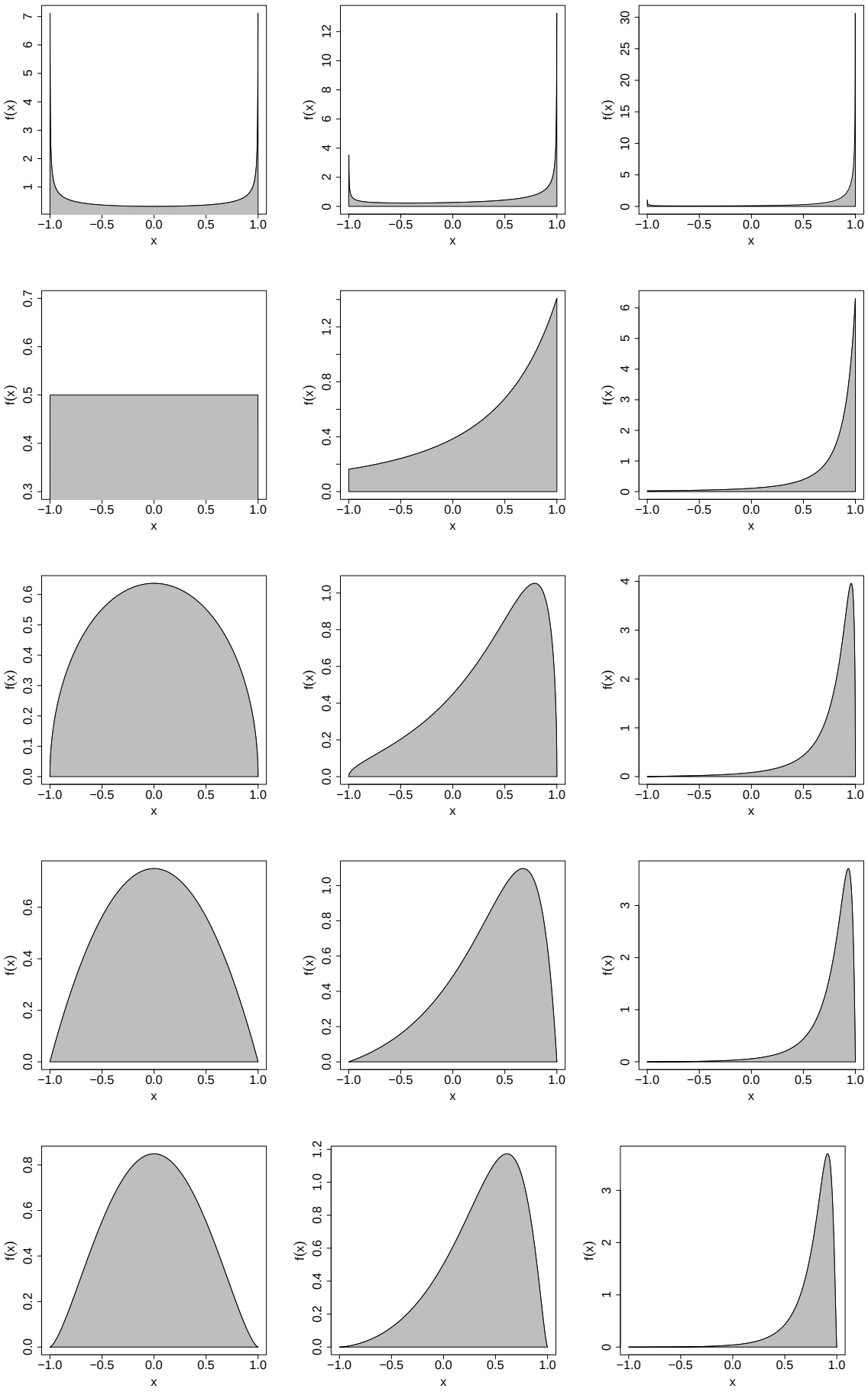

Fig. 5.2. Densités de l'estimateur du coefficient de corrélation pour $n = 3, 4, 5, 6$ et 7 (de haut en bas) et $r = 0, 0.4$ et 0.8 (de gauche à droite).

Théorème 12 *Soit*

$$W(.) = \left\{ \begin{pmatrix} X(s) \\ Y(s) \end{pmatrix}, s \in \mathcal{D} \right\}$$

un champ aléatoire bivarié stationnaire d'ordre 2 en corrélation intrinsèque. On considère pour tout n, $S_n = \{s_1, \ldots, s_n\}$, *un ensemble de sites d'observation.*

On suppose que quand n *tend vers l'infini,* S_n *évolue de telle sorte que les hypothèses du théorème 5 soient vérifiées pour le champ* $X(.)$ *et les paramètres* (σ_X^2, ϕ)[2].

Alors sur tout compact, $\hat{\psi}_n$, *l'estimateur par maximum de vraisemblance de* $\psi = (\sigma_X^2, r, \sigma_Y^2, \phi, \mu_X, \mu_Y)$ *vérifie :*

$$E[\mathcal{F}_n]^{1/2}(\hat{\psi}_n - \hat{\psi}) \Rightarrow_u \mathcal{N}_6(\boldsymbol{0}, I_6).$$

Démonstration. Voir annexe.

En asymptotique par densification du domaine avec un modèle de corrélation spatiale exponentiel $\rho(h) = e^{-h/\phi}$, Pascual et Zhang (2006) font la conjecture suivante :

puisque $\hat{\sigma}_X^2(\hat{\phi})/\hat{\phi}$ (respectivement $\hat{\sigma}_Y^2(\hat{\phi})/\hat{\phi}$) est un estimateur convergent pour σ_X^2/ϕ (respectivement $\sigma_Y^2(\hat{\phi})/\hat{\phi}$) à partir des observations de la variable X (respectivement Y) sous un schéma asymptotique par densification, pour les mêmes raisons $\hat{\sigma}_{XY}(\hat{\phi})/\hat{\phi}$ est sans doute un estimateur convergent pour σ_{XY}/ϕ. Dans ce cas,

$$\hat{r} = \frac{\hat{\sigma}_{XY}(\hat{\phi})/\hat{\phi}}{\hat{\sigma}_X(\hat{\phi})\hat{\sigma}_Y(\hat{\phi})/\hat{\phi}}$$

est un estimateur convergent pour r. Un théorème du même type que le théorème 11 pour la fonction de covariance croisée permettrait de conclure. Si l'on s'en tient aux simulations de Pascual et Zhang (2006), cette conjecture semble vraie.

Observations hétérotopes

Revenons au modèle spatial et introduisons quelques notations.

On note $\mathcal{S}_X = \{s_{x_1}, \ldots, s_{x_{n_X}}\}$ et $\mathcal{S}_Y = \{s_{y_1}, \ldots, s_{y_{n_Y}}\}$ les ensembles respectifs des n_X et n_Y sites d'observation des variables X et Y.

Soient $\mathbf{X} = (X(s_{x_1}), \ldots, X(s_{x_{n_X}}))'$ et $\mathbf{Y} = (Y(s_{y_1}), \ldots, Y(s_{y_{n_Y}}))'$.

On suppose que le vecteur $(\mathbf{X}', \mathbf{Y}')'$ est distribué selon le modèle de corrélation intrinsèque avec

$$\Sigma(\phi) = \begin{pmatrix} \sigma_X^2 H_{XX}(\phi) & \sigma_{XY} H_{XY}(\phi) \\ \sigma_{XY} H_{YX}(\phi) & \sigma_Y^2 H_{YY}(\phi) \end{pmatrix}$$

où $H_{XX}(\phi)$, $H_{XY}(\phi) = H_{YX}(\phi)'$ et $H_{YY}(\phi)$ sont les matrices de corrélations spatiales entre les éléments de $\mathbf{X}$, entre ceux de $\mathbf{X}$ et $\mathbf{Y}$, et entre ceux de $\mathbf{Y}$, respectivement.

On veut trouver un ensemble de conditions suffisantes à la convergence du maximum de vraisemblance de $\psi = (\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, r, \phi)$ sous l'hypothèse que quand n tend vers l'infini, les paramètres des deux champs $X(.)$ et $Y(.)$ puissent être estimés de manière consistante (au sens des théorèmes précédents).

Ces conditions concernent la proximité entre les ensembles $\mathcal{S}_X$ et $\mathcal{S}_Y$ quand n tend vers l'infini.

Nous allons distinguer le cas où les paramètres de structure spatiale sont connus et le cas où ils sont inconnus et doivent être estimés.

[2] Les hypothèses sont donc vérifiées pour $Y(.)$ et les paramètres (σ_Y^2, ϕ).

Structure de variance connue

On note $\lambda_k(M), k = 1, \ldots, n$ la fonction qui a toute matrice carrée de taille n associe la $k^{\text{ème}}$ valeur propre de M, les valeurs propres étant ordonnées par valeurs absolues croissantes.

Le théorème suivant donne la condition de proximité entre les ensembles de sites d'observations des deux variables, suffisantes pour la convergence.

Théorème 13 *On suppose que l'hypothèse (i) du théorème 5 est vérifiée.*

On suppose que quand n tend vers l'infini, on a

(i) n_X tend vers l'infini et $\mathcal{S}_X = \{s_{x_1}, \ldots, s_{x_{n_X}}\}$ croît de telle sorte que la matrice de variance du vecteur $\boldsymbol{X} = (X(s_{x_1}), \ldots, X(s_{x_{n_X}}))'$ vérifie les hypothèses du théorème 5 pour le paramètre σ_X^2.

(ii) n_Y tend vers l'infini et $\mathcal{S}_Y = \{s_{y_1}, \ldots, s_{y_{n_Y}}\}$ croît de telle sorte que la matrice de variance du vecteur $\boldsymbol{Y} = (Y(s_{y_1}), \ldots, Y(s_{y_{n_Y}}))'$ vérifie les hypothèses du théorème 5 pour le paramètre σ_Y^2.

Si pour tout compact $K \subset \Theta^q$, il existe une suite $(r_n(K))_{n\geq 1}$ telle que :

$$r_n(K) = o(n^{-1/2}) \tag{5.16}$$

et

$$tr(H_{XY}H_{YX})^{-1} \leq r_n(K) \tag{5.17}$$

alors

$$E[\mathcal{F}_n]^{1/2}(\psi)'(\hat{\psi}_n - \psi) \Rightarrow_u \mathcal{N}_5(\boldsymbol{0}, I_5).$$

De plus, on a

$$\frac{1}{v(r)}(\hat{r} - r) \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$$

avec

$$v(r) = \frac{1}{2}\frac{r^2 nA + 2n_X n_Y}{-r^4A^3 + r^2 nA(B-A) + n_X n_Y(2B - A)}$$

où

$$A = trace(Q_X M_X) \text{ et } B = trace(Q_X M_X Q_X)$$

avec $M_X = H_{XY}H_{YY}^{-1}H_{YX}H_{XX}^{-1}$ *et* $Q_X = (I_{nx} - r^2 M_X)^{-1}$

La condition donnée par l'équation (5.17) est une condition de proximité entre les deux nuages de sites d'observation. Elle assure essentiellement que les ensembles de sites d'observation des deux variables ne s'éloignent pas infiniment.

On appellera $v(r)$ la variance asymptotique de $\hat{r}$.

Sous l'hypothèse nulle de non corrélation entre les deux variables ($r = 0$), on a

$$v(0) = \frac{1}{\text{trace}(M_X)}.$$

On peut comparer $v(0)$ avec $\frac{1}{n}$, la variance asymptotique sous H_0 dans le cas i.i.d.

Definition 5.1. *On appellera nombre de données isotopes et indépendantes équivalent sous l'hypothèse de non corrélation ($r = 0$) la quantité suivante :*

$$N_{eq} = tr(H_{XX}^{-1}H_{XY}H_{YY}^{-1}H_{YX}).$$

On constate que l'augmentation des corrélations spatiales (par l'augmentation d'un paramètre de portée par exemple) a deux effets contradictoires sur la quantité N_{eq} traduisant le fait que

1) d'une part l'hétérotopie se compense (les termes de H_{XY} augmentent)

2) d'autre part en se "rapprochant" les uns des autres (en terme de corrélations spatiales), les $X(s_i)$ (de même les $Y(s_j)$) amènent de l'information redondante.

On peut borner N_{eq} par $\min(n_X, n_Y)$ car si

$$M = \begin{pmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{pmatrix}$$

est une matrice définie-positive, alors (voir par exemple Horn et Johnson (1985)) le rayon spectral de la matrice $M_{11}^{-1}M_{12}M_{22}^{-1}M_{21}$ est inférieur à 1 et donc sa trace est inférieure à $\dim(M_{11})$ et également à $\dim(M_{22})$.

Malgré de nombreuses tentatives, nous ne sommes pas parvenu à montrer que si $\rho_\phi(h)$ est une fonction croissante de ϕ, alors $N_{eq}(\phi)$ est également une fonction croissante en ϕ bien que les expérimentations numériques suggèrent ce résultat. On peut cependant montrer que quand ϕ tend vers l'infini, $N_{eq}(\phi)$ tend vers une limite qui dépend uniquement de la configuration spatiale et du degré de régularité à l'origine de la fonction de ϕ, $\rho_\phi(h)$.

Quand $r \neq 0$, il est plus délicat de comprendre comment l'hétérotopie dégrade le comportement asymptotique du coefficient de corrélation.

Il existe là aussi une notion de nombre équivalent : il suffit de poser

$$N_{eq}(r) = \frac{(1-r^2)^2}{v(r)}.$$

Mais comme r est inconnu, cette notion est moins opérante que dans le cas où $r = 0$. En effet ce dernier cas est souvent l'hypothèse nulle que l'on cherchera à tester et il peut être utile d'avoir une idée de la quantité d'information apportée par les données sur le coefficient de corrélation sous l'hypothèse nulle, en se basant sur les habitudes du contexte i.i.d hétérotope.

L'exemple suivant permet d'expliciter comment l'hétérotopie dégrade le comportement de $\hat{r}$ dans un cas simple.

Un cas d'école

Supposons que l'on observe une des deux variables sur un ensemble de sites $\{s_1, \ldots, s_n\}$ infiniment éloignés deux à deux (ou éloignés d'une distance supérieure à la portée pour un modèle de corrélation spatiale à portée finie). Et supposons que sur un cercle de rayon h autour de chacun de ces sites, on a une observation de la seconde variable.

Si ρ désigne la fonction de corrélation spatiale et $\tilde{\rho} = \rho(h)$, alors dans ce cas, $H_{XX} = H_{YY} = I_n$ et $H_{XY} = H_{YX} = \tilde{\rho} I_n$.

On peut déduire l'expression de $v(\hat{r})$:

$$v(\hat{r}) = \frac{(1-r^2\tilde{\rho}^2)^2}{\tilde{\rho}^2 n}.$$

On peut écrire $v(\hat{r})$ sous la forme

$$v(\hat{r}) = \frac{(1-r_{eq}^2)^2}{n_{eq}}.$$

Quand h grandit, l'hétérotopie dégrade le comportement de $\hat{r}$ à deux niveaux :

- r_{eq} : on estime d'autant mieux un coefficient de corrélation que celui-ci est grand en valeur absolue. Ici le coefficient de corrélation entre les variables $X(s_i)$ et $Y(s_i + h)$ est égal à $r\tilde{\rho}$.

- n_{eq} : la dégradation est ici due à la perte d'information induite par l'éloignement des points et à la relative ignorance de la vraie valeur de $Y(s_i)$ (ou de $X(s_i + h)$).

Notons que nous aurions pu démontrer ce résultat sans passer par le contexte spatial puisqu'on se retrouve dans un cas où les couples $(X(s_i), Y(s_i + h)), i = 1, \ldots, n$ sont identiquement distribués et mutuellement indépendants.

5.3.3 Une étude sur simulations

Nous avons mené une large étude sur simulation, pour tenter d'appréhender les caractéristiques distributionnelles à taille finie, de l'estimateur par maximum de vraisemblance du coefficient de corrélation.

Schéma expérimental :

Pour gérer l'hétérotopie, nous avons procédé de la manière suivante :

- nous avons placé les sites d'observation d'une des deux variables (disons X) sur une grille de $\mathbb{R}^2$ à mailles régulières (unité) de taille $n \times n$, pour différentes tailles de $n = 6, 7, 8, 9$, et 10.

- puis autour de chaque site d'observation s_i, $i = 1, \ldots, n^2$ de cette première variable, nous avons simulé un site d'observation pour Y, uniformément dans un carré dont les côtés, de longueur S sont parallèles aux mailles de la grille et dont s_i est le centre. Nous avons travaillé avec $S = j/4, j = 0, 1, 2, 3, 4$ où $j = 0$ correspond à l'isotopie.

En croisant ces différentes caractéristiques, on dispose donc de 25 configurations spatiales. Pour chacune d'entre elles, et pour $\phi = 0.5, 1$ et 2, et pour $r = 0, 0.4$ et 0.8 nous avons simulé 1000 jeux de données selon le modèle de corrélation intrinsèque avec une fonction de corrélation exponentielle : $\rho_\phi(h) = e^{-\phi h}$. Sans perte de généralité, nous avons toujours travaillé avec $\mu_X = \mu_Y = 0$ et $\sigma_X = \sigma_Y = 1$.

En résumé, pour chacune des 25 configurations spatiales, et pour chaque valeur de ϕ et de r, soit 225 configurations, nous avons simulé 1000 jeux de données. Notons que pour chacune des 225 configurations, on conserve le même échantillonnage spatial pour les 1000 jeux de données simulés.

Puis nous avons estimé les paramètres par maximum de vraisemblance.

Avant de présenter l'ensemble des résultats, nous montrons (fig. (5.3)) , pour trois configurations, les distributions empiriques de l'estimateur et la densité asymptotique (pour des nombres équivalents importants).

Dans ces cas, les distributions asymptotiques semblent être de bonnes approximations des distributions empiriques de $\hat{r}$.

Pour présenter l'ensemble des résultats, nous avons dans chaque configuration (pour r, n, S, ϕ et r fixés) :

- résumé la configuration spatiale par le nombre équivalent de données :

$$N_{eq} = \frac{(1 - r^2)^2}{v(r)}.$$

- résumé l'écart entre la distribution empirique de $\hat{r}$ et la distribution asymptotique donnée par le théorème 13 en effectuant un test d'adéquation de Kolmogorov.

A titre de comparaison, pour chaque valeur r, nous avons appliqué la même procédure, mais cette fois sur le coefficient de corrélation usuel, calculé à partir de données bivariées i.i.d simulées.

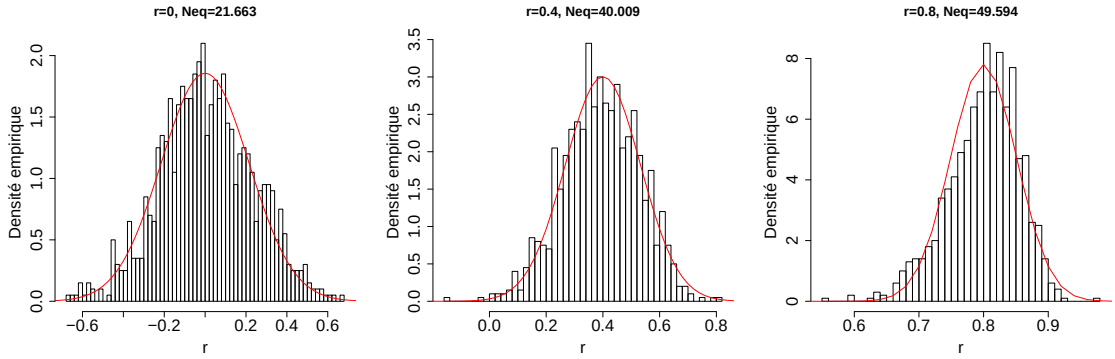

Fig. 5.3. Histogrammes des distributions empiriquess de $\hat{r}$ pour $r = 0, 0.4$ et 0.8 (de gauche à droite) pour une configuration spatiale donnée dans chacun des trois cas. En rouge, la densité de la distribution asymptotique correspondant.

Afin d'effectuer la comparaison dans les meilleures conditions, à chaque nombre équivalent pour r fixé, correspondait un échantillon i.i.d de taille N égale à l'arrondi de ce nombre équivalent.

Les figures (5.4) et (5.5) montrent ces résultats.

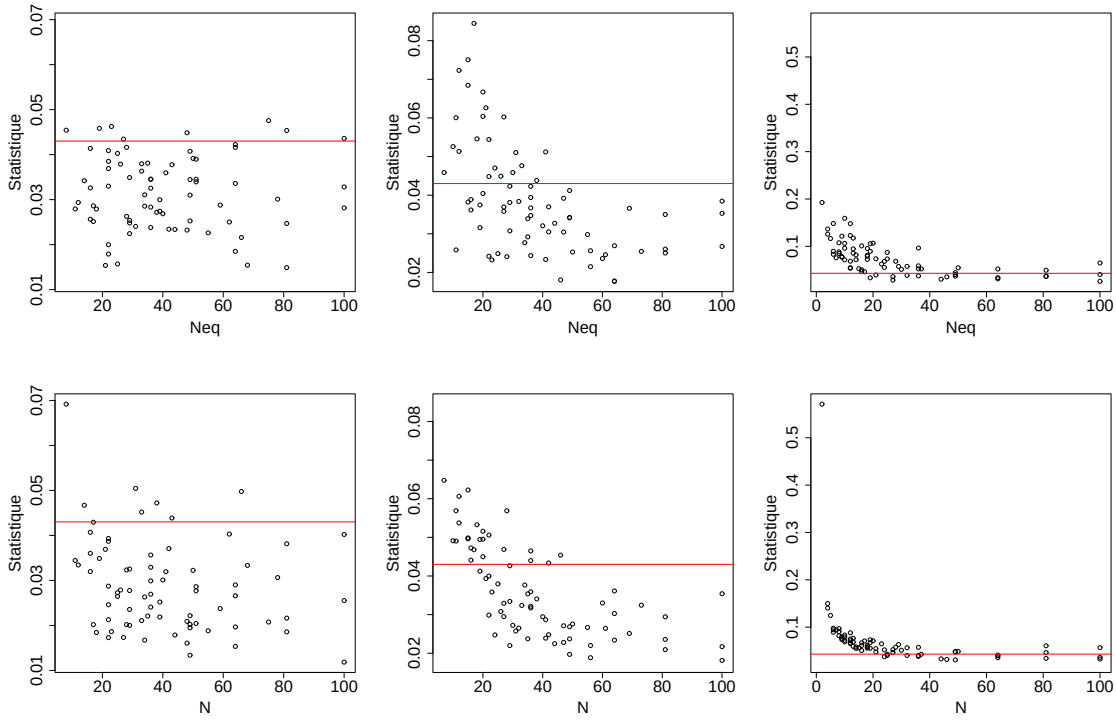

Fig. 5.4. Haut : Statistiques du test de Kolmogorov-Smirnov d'adéquation à la distribution asymptotique pour $\hat{r}$, en fonction du nombre équivalent de données i.i.d pour $r = 0, 0.4$ et 0.8 (de gauche à droite). Bas : Statistiques du test de Kolmogorov-Smirnov d'adéquation à la distribution asymptotique pour $\tilde{r}$ dans le cas i.i.d en fonction de la taille de l'échantillon pour $r = 0, 0.4$ et 0.8 (de gauche à droite). Ligne horizontale (rouge) : valeur du quantile d'ordre 0.05 pour la statistique de test. Voir détails dans le texte.

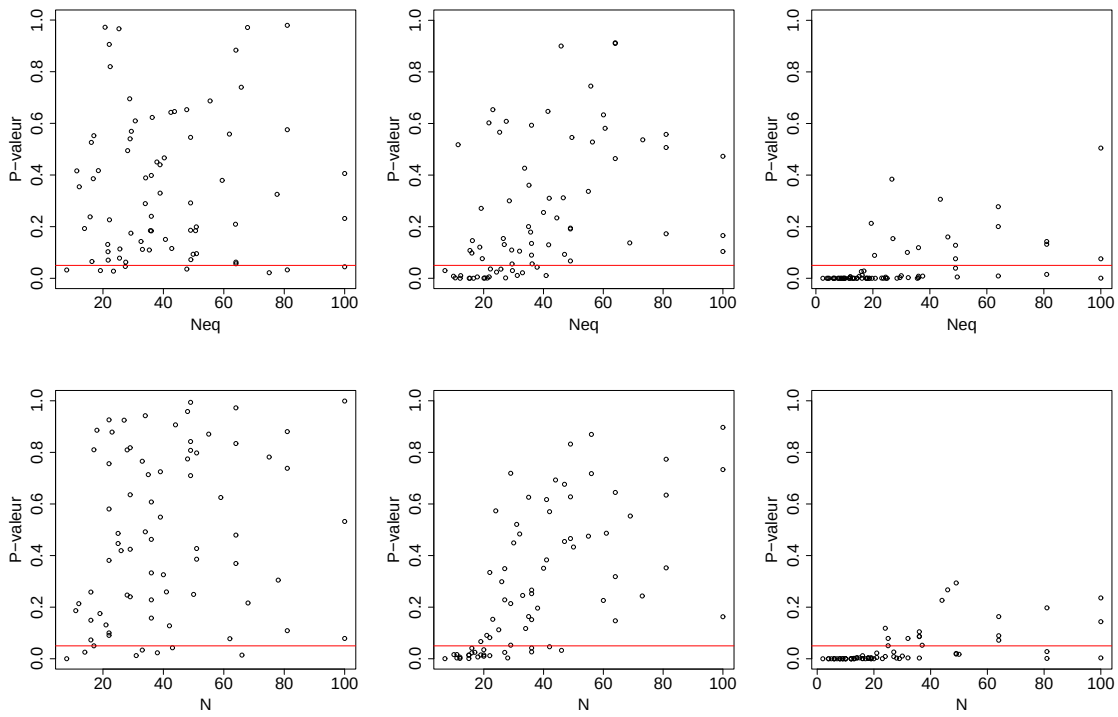

Fig. 5.5. Haut : P-valeurs du test de Kolmogorov-Smirnov d'adéquation à la distribution asymptotique pour $\hat{r}$, en fonction du nombre équivalent de données i.i.d pour $r = 0, 0.4$ et 0.8 (de gauche à droite). Bas : P-valeurs du test de Kolmogorov-Smirnov d'adéquation à la distribution asymptotique pour $\tilde{r}$ dans le cas i.i.d en fonction de la taille de l'échantillon pour $r = 0, 0.4$ et 0.8 (de gauche à droite). Ligne horizontale (rouge) : valeur du seuil 0.05. Voir détails dans le texte.

La distance de Kolmogorov semble diminuer en fonction du nombre équivalent, montrant ainsi que ce nombre est un bon indicateur de la quantité d'information apportée par les données sur le coefficient de corrélation. Notons que pour $r = 0$, la distribution asymptotique est rarement rejetée, même pour les petits nombres équivalents.

Si la comparaison des figures entre le cas spatial et le cas i.i.d montre quelques différences, il nous semble que les similitudes entre les deux sont tout de même très fortes.

Ces résultats nous permettent d'établir une règle empirique pour l'acceptation de la distribution asymptotique : si, en pratique dans le cas i.i.d, on accepte l'approximation asymptotique pour n donné, alors on acceptera cette approximation pour $N_{eq} \geq n$ dans le cas spatial hétérotope.

Qu'en est-il dans le cas où l'approximation asymptotique est rejetée ?

Nous représentons sur la figure (5.6) quelques distributions empiriques de $\hat{r}$ pour des configurations où le nombre équivalent est très petit.

La distribution théorique de $\tilde{r}$ à taille finie explique les grandes tendances de la distribution empirique. Cependant, dans tous les cas, un certain nombre de valeurs estimées sont très proches de 1 (comme dans le cas où $n < 4$) et ce mode n'apparaît pas dans les distributions théoriques (sauf pour $n \leq 3$). Plus r est grand et plus ce nombre de 1 est important ce qui accentue l'impression d'écart à la distribution théorique pour les grandes valeurs de r.

D'un point de vue pratique, il est rare que l'on utilise le coefficient de corrélation pour des tailles d'échantillon inférieures à 7 et nous déconseillons donc l'utilisation de notre méthode quand le nombre équivalent est très petit sous l'hypothèse que $r = 0$. Pour r général, quand le nombre

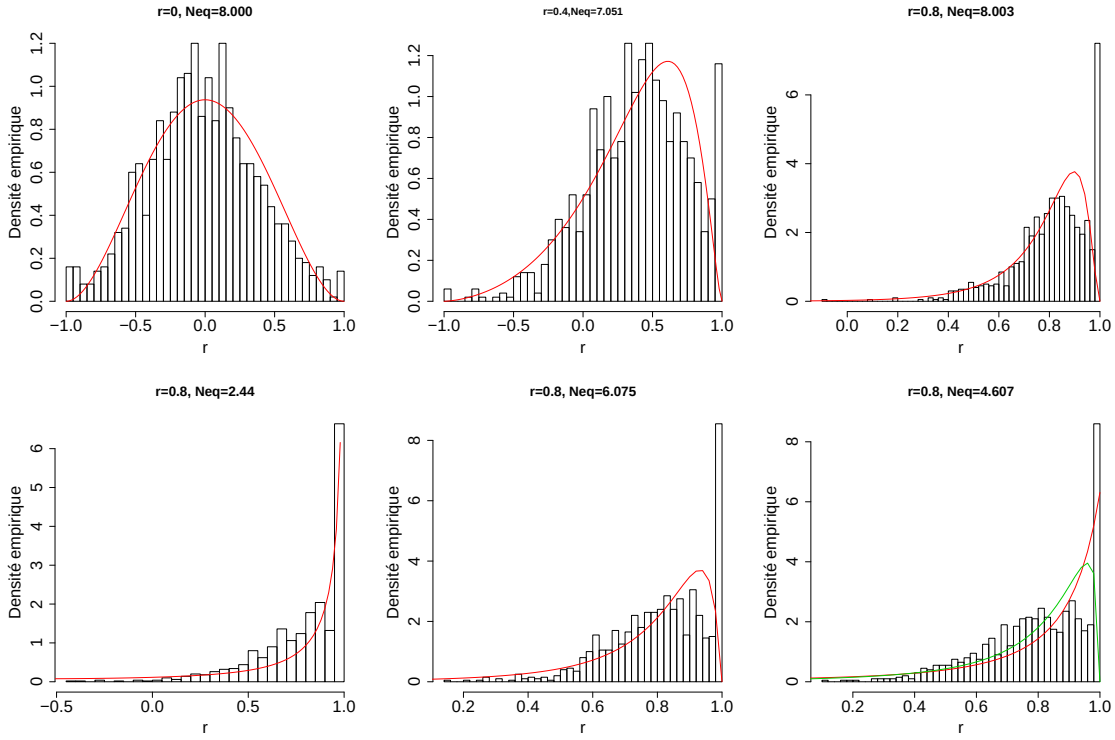

Fig. 5.6. Distributions empiriques de $\hat{r}$ dans différents cas. En rouge, la distribution théorique de $\tilde{r}$ à taille finie pour la taille d'échantillon la plus proche du nombre équivalent.Dans le cas où $N_{eq} = 2.44$, on représente la densité de $\tilde{r}$ pour $n = 3$, le cas $n = 2$ correspondant à un mélange de deux masses de Dirac. Pour la dernière figure, le nombre équivalent est environ égal à 4.607, nous représentons donc les courbes des distributions théoriques de $\tilde{r}$ pour $n = 4$ (rouge) et $n = 5$ (vert).

équivalent est petit, la seule indication fiable que l'on pourra éventuellement obtenir concerne le signe de la corrélation : si la valeur estimée de $\hat{r}$ est très proche de 1 (respectivement -1), la corrélation entre les deux variables est positive (respectivement négative).

5.3.4 Structure de covariance inconnue

Résultats asymptotiques

Pour établir les conditions de convergence asymptotique de l'estimateur par maximum de vraisemblance de $(\mu_X, \mu_Y, \sigma_X^2, r, \sigma_Y^2, \phi)$, supposons que la fonction de corrélations spatiales ρ_ϕ, n'a qu'un seul paramètre inconnu ϕ, et supposons que ρ_ϕ est deux fois dérivables en ϕ et que ses dérivées (premières et secondes) par rapport à ϕ sont des fonctions continues en h sur $[0, \infty[$. Enfin, supposons qu'il existe h_0 tel que

$$\left|\frac{\partial \rho_\phi}{\partial \phi}\right| \quad \text{et} \quad \left|\frac{\partial^2 \rho_\phi}{\partial \phi^2}\right|$$

sont décroissantes sur $[h_0, +\infty[$

Alors on a le théorème suivant :

Théorème 14 *On suppose que l'hypothèse (i) du théorème 5 est vérifiée.*

On suppose que quand n tend vers l'infini, on a

(i) n_X tend vers l'infini et $\mathcal{S}_X = \{s_{x_1}, \ldots, s_{x_{n_X}}\}$ croît de telle sorte que la matrice de variance du vecteur $\boldsymbol{X} = (X(s_{x_1}), \ldots, X(s_{x_{n_X}}))'$ vérifie les hypothèses du théorème 6 pour les paramètres (ϕ, σ_X^2).

(ii) n_Y tend vers l'infini et $\mathcal{S}_Y = \{s_{y_1}, \ldots, s_{y_{n_Y}}\}$ croît de telle sorte que la matrice de variance du vecteur $\boldsymbol{Y} = (Y(s_{y_1}), \ldots, Y(s_{y_{n_Y}}))'$ vérifie les hypothèses du théorème 6 pour le paramètre (ϕ, σ_Y^2).

Si il existe une constante $R > 0$ telle que

$$\forall n \in \mathbb{N}, \max_{i=1,\ldots,n_X} \min_{j=1,\ldots,n_Y} ||s_{x_i} - s_{y_j}|| < R,$$

et

$$\forall n \in \mathbb{N}, \max_{j=1,\ldots,n_Y} \min_{i=1,\ldots,n_X} ||s_{x_i} - s_{y_j}|| < R,$$

alors

$$E[\mathcal{F}_n]^{1/2}(\psi)'(\hat{\psi}_n - \psi) \Rightarrow_u \mathcal{N}_6(\boldsymbol{0}, I_6),$$

Quelques remarques :

-Notons que les conditions analytiques pour la fonction de covariance sont vérifiées par le modèle exponentiel.

- Les conditions géométriques asymptotiques imposées aux sites d'observations de chacune des deux variables sont plus fortes que dans le cas où ϕ est connu. Il est naturel que des hypothèses supplémentaires soient imposées dans ce cas puisque ϕ est inconnu mais il faut noter que la démonstration dans ce cas requiert de plus, que les hypothèses asymptotiques sur chacun des deux champs soient plus fortes, puisqu'on impose que les conditions du théorème 6 soient vérifiées au lieu de celles du théorème 5 dans le cas où ϕ est connu.

- La condition de proximité entre les ensembles de sites d'observation pour chaque variable est également plus forte que dans le cas où ϕ est connu, mais peut être est elle aussi plus évocatrice : elle assure que les sites d'observation des deux variables ne s'éloignent pas infiniment quand n tend vers l'infini.

La formule de la variance asymptotique de r (non donnée) est ici beaucoup plus complexe que dans le cas où ϕ est connu ; cependant, sous l'hypothèse nulle de non corrélation entre les deux variables ($r = 0$), on a également

$$v(0) = \frac{1}{\text{trace}(M_X)},$$

ce qui renforce le caractère génératif de la notion de nombre équivalent sous H_0.

A taille finie

Engager une étude par simulations sur le comportement de $\hat{r}$ dans le cas où ϕ est inconnu, de l'envergure de celle conduite dans le cas où ϕ est connu (225000 simulations et estimations) aurait un coût en temps de calcul inutile par rapport aux conclusions spécifiques à ce cas de figure.

En effet, le rajout de ϕ dans l'ensemble des paramètres à estimer rendrait cette tâche plus longue car l'ajout d'une dimension à l'espace des paramètres augmente les difficultés numériques et le nombre d'itérations nécessaires à la convergence de l'algorithme.

De plus, en comparant sur quelques configurations "bien choisies", les résultats obtenus dans les deux cas (ϕ connu vs ϕ inconnu), on peut tirer des conclusions suffisantes pour la compréhension globale du comportement de $\hat{r}$ dans le cas où les paramètres de structure spatiale sont inconnus.

Nous avons simulé pour deux configurations spatiales décrites dans le cas ϕ connu ($\phi = 0.5$ et 2 et $n = 7$ et $S = 0.5$ restent fixes), 49 sites d'observations. Dans la première configuration ($\phi = 0.5$), le nombre équivalent sous l'hypothèse que $r = 0$ est égal à 23.32, et à 12.05 dans la seconde configuration ($\phi = 2$) puis sur ces sites, 1000 jeux de données selon le modèle **(LM4a)** avec dans un premier temps $r = 0$. Nous avons estimé l'ensemble des paramètres avec ϕ connu et fixé à sa vraie valeur et ensuite avec ϕ inconnu.

La figure (5.7) montre les deux nuages de points.

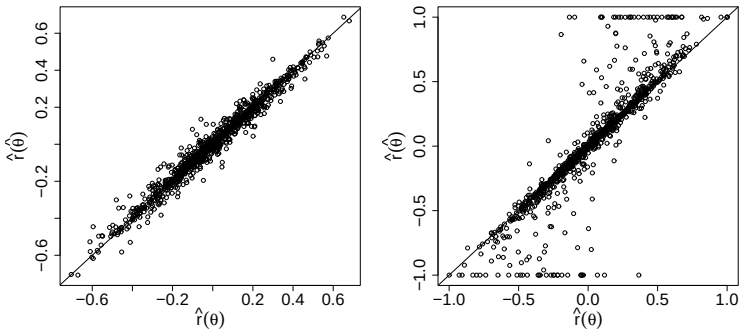

Fig. 5.7. Nuage de points des estimateurs de r dans le cas où ϕ est inconnu contre le cas où ϕ est connu avec un nombre équivalent égal à 23.32 (gauche) et 12.05 (droite).

On constate que pour N_{eq} relativement grand, les estimations de r dans les deux cas sont très proches (le coefficient de corrélation entre les deux est égal à 0.97), ce qui indique que les erreurs d'estimation faites sur le paramètre de portée influent peu sur le comportement de $\hat{r}$ dans ce cas. A l'inverse, dans le cas où N_{eq} est plus petit, si une part des couples $\hat{r}$ dans les cas ϕ connu vs ϕ inconnu sont sur la première bissectrice, un certain nombre d'entre eux ne s'y trouvent pas et parmi eux, la plupart prennent les valeurs -1 ou 1 quand ϕ est inconnu.

Si on observe le nuage de points des estimations $\hat{r}$ en fonction de $\hat{\phi}$ celles de ϕ (fig 5.8), on constate que $\hat{r}$ prend les valeurs 1 ou -1 quand le paramètre de portée est surestimé, c'est-à-dire quand les corrélations spatiales sont sous-estimées. Dans ce cas, tout se passe comme si la quantité d'information apportée par les données sur r était très faible, et donc comme si le nombre équivalent était faible lui aussi, ce qui explique le fait que, dans ces cas, on observe des valeurs de $\hat{r}$ à 1 ou -1 qui sont typiques de la très faible information.

De même, de faibles corrélations spatiales, surestimées (par le biais de la portée) pourraient conduire à un nombre équivalent lui aussi surestimé ; ce qui pourrait nous tromper quant à la qualité de l'information disponible sur r. En particulier, dans le cas d'un petit nombre équivalent estimé, on risque de croire que l'on dispose d'une quantité moyenne d'information sur r suffisante pour pouvoir conclure alors qu'elle est en réalité extrêmement faible.

C'est pourquoi dans les cas concrets, quand le nombre équivalent estimé est petit (< 15), une analyse variographique plus poussée de chacune des variables est nécessaire. On pourra par exemple quantifier par bootsrap paramètrique, la variabilité de $\hat{\phi}$, ou juger de la fiabilité de l'estimation des structures spatiales à la lumière des connaissances théoriques concernant l'estimation des paramètres de ces structures et rappellées dans la première partie de ce chapitre. Cette analyse devra s'accompagner d'une analyse de sensibilité de la fonction $N_{eq}(\phi)$ en fonction de ϕ.

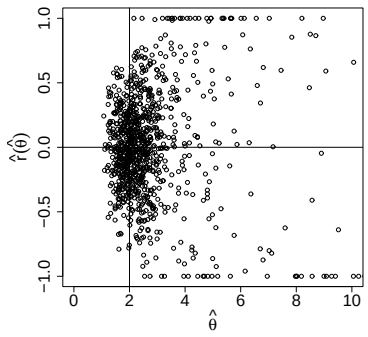

Fig. 5.8. Nuage de points des estimations de r dans le cas où ϕ est inconnu en fonction des estimations de ϕ.

On pouvait s'attendre à ce que sous H_0 les distributions de $\hat{r}$ dans les cas ϕ connu et ϕ inconnu soient proches si l'information apportée sur ϕ par les données est suffisante, puisque l'on a la même distribution asymptotique dans les deux cas d'après le théorème 14.

Pour $r \neq 0$, nous avons signalé que les formules de la variance asymptotique de r, $v(r)$, étaient différentes selon que ϕ soit connu ou inconnu. Cependant, en appliquant la même expérience que dans le cas $r = 0$ et pour une configuration spatiale résultante de bonne qualité ($N_{eq} = 29.952$ sous H_0), nous obtenons un coefficient de corrélation de 0.99 entre les deux cas (voir figure 5.9). L'effet dû à l'estimation de ϕ sur l'estimation de r semble donc "asymptotiquement négligeable".

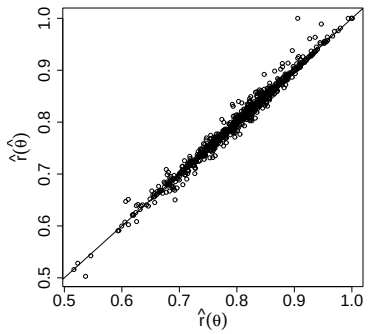

Fig. 5.9. Nuage de points des estimateurs de r dans le cas où ϕ est inconnu contre le cas où ϕ est connu avec un nombre équivalent égal à 29.952

6

L'alternative bayésienne

6.1 Introduction

Les paradigmes bayésiens et fréquentistes, basés sur deux philosophies différentes, se disputent encore une certaine primauté dans l'utilisation des statistiques (Efron, 2005). Les débats portent sur plusieurs points.

Tout d'abord, les fondements théoriques des deux écoles sont très différents. L'approche fréquentiste suppose que les observations sont une réalisation d'un processus stochastique décrit par un modèle (ici paramètrique) dont les paramètres sont supposés être des valeurs fixes. Une fois les paramètres estimés, l'objectif du statisticien est de quantifier la variabilité de l'estimateur, c'est à dire le comportement qu'aurait cet estimateur si l'on répétait "l'expérience" qui a généré les observations, dans les mêmes conditions. La variabilité de l'estimateur peut être calculée, en général par un résultat asymptotique (il s'agit alors d'une approximation), ou alors, elle peut être obtenue à valeurs de paramètres fixées, en simulant un grand nombre de fois les données sous le modèle choisi et en estimant à chaque fois les paramètres à partir des données simulées (bootstrap paramètrique).

L'approche bayésienne suppose également que les données sont issues d'un modèle paramétré. Mais le paramètre est vu comme une variable aléatoire au même titre que les observations. Une étape de la modélisation consiste alors à modéliser la connaissance *a priori* sur ce paramètre et l'incertitude associée à cette connaissance. L'inférence s'effectue ensuite sur la loi *a posteriori*, la version actualisée par la prise en compte des observations de la loi *a priori* .

Pour les tenants de l'école bayésienne, le fait de travailler conditionnellement aux observations est en cohérence avec la démarche statistique qui vise à remonter des effets aux causes (Robert, 1992).

Le point sur lequel se concentrent principalement les objections des tenants de l'école fréquentiste, est la nécessité dans le paradigme bayésien de choisir une loi *a priori* pour les paramètres. Dans l'absolu, les fréquentistes n'ont pas tort de soulever cette question puisque en pratique, on peut obtenir le résultat que l'on souhaite en utilisant une loi *a priori* "judicieusement" choisie. Cette critique, si elle concerne le manque d'objectivité potentiel de l'analyste ne nous semble pas pertinente ici. L'objection sur la difficulté de modéliser sa connaissance *a priori* et l'incertitude que l'on y associe nous semble un argument plus recevable. Les bayésiens proposent d'utiliser des priors non-informatifs lorsqu'aucune information n'est disponible *a priori* . Dans ce cas, on laisse les données conduire l'inférence et les résultats théoriques montrent que l'inférence dans les deux paradigmes est alors équivalente. Au regard de ces résultats, pour les tenants de l'école bayésienne, leur paradigme est plus général, il englobe le cadre fréquentiste dans le cas non-informatif ; et si de l'information *a priori* est disponible, il est le seul qui permette de l'intégrer.

Loin de ces querelles "idéologiques", le statisticien appliqué a une approche plus pragmatique. L'introduction croissante des statistiques dans de nombreux domaines des sciences de la vie ou de la terre, le type de données et les questions auxquelles il est confronté ont conduit au développement de modèles de plus en plus complexes, constitués de différentes couches hiérarchisées, et dont l'estimation dans un cadre fréquentiste devient extrêmement complexe, lorsqu'elle ne s'avère pas illusoire. Les algorithmes MCEM que nous avons présentés dans la section 4 sont des exemples pour lesquels, si elle est reste faisable, l'inférence est numériquement coûteuse et fournit uniquement une estimation du maximum de vraisemblance. A temps de calcul équivalent, un algorithme MCMC (*Monte Carlo Markov Chain*) dans le cadre bayésien aurait donné une approximation de la loi *a posteriori* et permis une inférence plus complète qu'une simple estimation. La théorie asymptotique qui permettrait d'apprécier la significativité du résultat nécessite des développements complexes dans le contexte spatial par exemple, et la qualité des résultats doit être validée par des simulations (cf chapitre 5). On atteint ici une limite de l'utilisation des statistiques fréquentistes.

"L'alternative bayésienne", le titre de ce chapitre, résume notre position à l'égard des querelles qui pourraient encore subsister entre les tenants des deux écoles. Nous présentons ici un résumé de la théorie bayésienne qui doit permettre de comprendre le paradigme et nous montrons comment utiliser l'outil MCMC en présentant deux algorithmes de la littérature, utilisés pour des modèles du chapitre 3 ; puis nous montrons comment les "combiner" pour résoudre l'inférence pour des modèles plus complexes de ce même chapitre.

6.2 Généralités

6.2.1 Introduction

Le contexte bayésien diffère du cadre fréquentiste dès l'étape de la modélisation puisque dans ce paradigme, les paramètres du modèle sont considérés comme aléatoires et leur distribution *a priori* doit être précisée comme un élément du modèle. Pour les utilisateurs des statistiques bayésiennes, le choix de ces priors est un moyen d'incorporer au modèle les connaissances disponibles sur le phénomène étudié, externes au jeu de données, ainsi que l'incertitude sur ces connaissances. Mais on peut également utiliser des priors non-informatifs si l'on veut laisser les données conduire l'inférence. L'inférence sur les paramètres est alors basée sur la loi *a posteriori*, la version actualisée par les observations de la loi *a priori* et obtenue par la formule de Bayes. Pour Banerjee *et al.* (2004), les statistiques bayésiennes fournissent une méthodologie inférentielle exacte, dans la mesure où elle n'est pas basée sur une distribution asymptotique des estimateurs, contrairement au cas fréquentiste.

6.2.2 Formule de Bayes

Plaçons nous dans le cadre général de l'inférence sur un vecteur de paramètres θ qui varie dans un ouvert Θ, à partir d'observations y dont la distribution s'écrit $\pi(y|\theta)$. On choisit la distribution *a priori* des paramètres $\pi(\theta)$. L'intérêt porte sur la loi *a posteriori* $\pi(\theta|y)$, la distribution des paramètres conditionnellement aux observations.

Cette distribution s'obtient à partir de la distribution des observations et de la loi *a priori* par la formule de Bayes :

$$\pi(\theta|y) = \frac{\pi(y|\theta)\pi(\theta)}{m(y)} \tag{6.1}$$

où

$$m(y) = \int_{\Theta} \pi(y|\theta)\pi(\theta)d\theta \tag{6.2}$$

est la distribution marginale de y. $m(y)$ étant une quantité qui ne dépend pas de θ, elle est souvent considérée comme une constante de normalisation de la loi *a posteriori*. Cette constante pouvant souvent être négligée dans l'analyse, l'équation (6.1) sera souvent réécrite :

$$\pi(\theta|y) \propto \pi(y|\theta)\pi(\theta), \tag{6.3}$$

qui signifie que la densité *a posteriori* en tant que fonction de θ est proportionnelle à la fonction de θ définie par $\theta \rightarrow \pi(y|\theta)\pi(\theta)$.

6.2.3 Choix des priors

Le choix des priors est un élément déterminant de l'analyse bayésienne. En effet, selon les priors choisis, on peut obtenir des résultats différents pour l'inférence d'un ensemble de paramètres. On peut par exemple montrer que le support de la loi *a posteriori* est inclus dans celui de la loi *a priori* (le support d'une loi π est l'ensemble $\{\theta \in \Theta, \pi(\theta) \neq 0\}$)[1]

Nous donnons ici quelques considérations sur les priors non informatifs et les priors conjugués.

Priors non informatifs

Lorsque l'espace d'état d'une variable aléatoire est discret et fini et que les paramètres sont les probabilités de chacun des évènements élémentaires, un choix naturel est d'affecter la même probabilité *a priori* à chaque évènement élémentaire. Cela revient à supposer que les différents évènements élémentaires sont équiprobables et cela semble correspondre à l'idée intuitive de l'absence d'information.

Il est des cas où l'intuition est d'un moindre secours pour choisir une loi *a priori* non informative. Par exemple si

$$x \sim \mathcal{B}\text{in}(n, p),$$

Villegas (1977) critique le choix de la loi uniforme sur $[0, 1]$ comme loi *a priori* non informative pour p car trop biaisée contre les extrêmes, 0 et 1.

Lorsque l'espace du paramètre est infini, il n'existe pas de lois de probabilité uniformes sur cet espace. On est alors conduit à choisir des priors impropres. Il ne s'agit plus de loi de probabilités mais de mesures σ-finie, de masse infinie sur l'espace des paramètres Θ :

$$\int_{\theta \in \Theta} \pi(\theta)d\theta = \infty.$$

Ces lois, dites également généralisées, peuvent conduire à des lois *a posteriori* bien définies.

Une critique fondamentale des lois *a priori* non informative concerne le problème d'invariance par reparamètrisation. Si on passe du paramètre $\theta \in \Theta$ au paramètre $\eta = g(\theta) \in g(\Theta)$ par une

[1] Si l'on est presque sûr que θ est égal à 0, un bon choix pour la loi *a priori* est de mettre toute la masse de probabilité en 0 ($\pi(\theta) = \delta_{\theta=0}$ est la masse de Dirac en 0). La loi *a posteriori* sera alors égale à la loi *a priori* . Dans cet exemple extrême, l'analyse statistique n'a aucune justification puisque l'on connaît par avance le résultat. Mais on comprend bien sur cet exemple que si l'information apportée par les données est faible par rapport à l'information *a priori*, cette dernière va devenir prépondérante et sa modélisation déterminante sur le résultat.

transformation bijective g, l'information *a priori* étant inexistante, on doit aussi utiliser une loi *a priori* non informative pour η.

Une possibilité est de considérer des invariances correspondant aux structures d'invariance du modèle. Pour expliquer ce que cela signifie, traitons les exemples d'invariance par translation et d'invariance par changement d'échelle :

Invariance par translation :

θ est un paramètre de position si la distribution des observations $\pi(x|\theta)$ s'écrit $f(x-\theta)$.

La famille des lois $f(x-\theta)$ est invariante par translation. C'est à dire que la distribution de $y = x - x_0$ s'écrit également $f(y-\theta^\star)$ avec $\theta^\star = \theta - x_0$.

Choisir une loi *a priori* pour θ qui respecte cette structure d'invariance revient à choisir une loi invariante par translation, i.e $\pi(\theta) = \pi(\theta - \theta_0)$, ce qui conduit à

$$\pi(\theta) = c \tag{6.4}$$

où c est une constante arbitraire.

Choisir une distribution invariante par translation pour un paramètre de position est conforme à l'intuition de la non information dans ce cas.

Le deuxième exemple est celui des paramètres d'échelle :

Invariance par changement d'échelle :

σ est un paramètre d'échelle si la distribution des observations $\pi(x|\sigma)$ s'écrit $1/\sigma f(x/\sigma)$ $(\sigma > 0)$. Cette famille est invariante par changement d'échelle. La loi *a priori* de σ devrait donc être invariante par changement d'échelle :

$$\pi(\sigma) = \frac{1}{c}\pi(\frac{\sigma}{c})$$

pour tout c, ce qui implique que

$$\pi(\sigma) \propto 1/\sigma. \tag{6.5}$$

Dans ce cas, la loi non-informative n'est plus constante.

Priors de Jeffreys :

La solution de considérer les structures d'invariance n'est que partiellement satisfaisante, celles-ci pouvant ne pas exister ou pouvant être choisies de différentes manières (Robert, 1992). Une approche plus globale qui ne prend pas en compte les structures d'invariance est l'utilisation des priors de Jeffreys basés sur l'information de Fisher. Si $I(\theta)$ désigne la matrice d'information de Fisher, (voir par exemple le chapitre 4), alors le prior de Jeffreys est donné par :

$$\pi^\star(\theta) = \sqrt{\det(I(\theta))},$$

modulo un coefficient de renormalisation si $\pi^\star$ est propre (c'est à dire si $\int \pi^\star(\theta)d\theta < \infty$).

Ces priors sont invariants par reparamétrisation. De plus, ils redonnent les lois obtenues dans les deux exemples (6.4) et (6.5).

Une justification peut être donnée dans le cas unidimensionnel. $I(\theta)$ est une mesure du pouvoir discriminant du modèle entre θ et $\theta + d\theta$. Favoriser *a priori* les valeurs de θ pour lesquelles ce pouvoir est le plus fort est le choix qui donne le moins de poids possible à la loi *a priori*. Les lois de Jeffrey

sont souvent impropres et généralement performantes. De plus elles permettent souvent de retrouver les estimateurs des statistiques fréquentistes.

Cependant, le calcul de la matrice d'information de Fisher peut être couteux (par exemple dans le cas des modèles hiérarchiques spatiaux) voire infaisable pour des modèles complexes. De plus Jeffreys déconseillait l'usage de ses priors dans les cas multidimensionnels (Robert, 1992). Enfin, la mise en place d'algorithmes MCMC est considérablement simplifiée par le choix de priors indépendants.

Cette dernière constatation est sans doute la raison qui fait, que dans la pratique, lorsqu'aucune information *a priori* n'est disponible, l'usage semble être, le plus souvent, de choisir des priors propres mais vagues (par exemple pour un paramètre d'espérance, la loi uniforme sur $[-10^n, 10^n]$ avec n grand relativement au problème considéré). Dans ce cas, une analyse de la robustesse au choix du prior peut être nécessaire. Elle est cependant peu souvent évoquée.

Priors conjugués

Pour certaines distributions des observations, le choix de la distribution *a priori* peut être fait de telle sorte que la loi *a posteriori*, ait la même forme que cette loi *a priori*. On dira que le prior est conjugué (relativement à la distribution des observations). Le choix de la distribution *a priori* est ici d'ordre pragmatique, il est fait pour des considérations mathématiques.

L'exemple le plus simple est le cas gaussien.

Supposons que l'on observe un échantillon $\mathbf{y} = (y_1, \ldots, y_n)'$ où

$$y_i \sim \mathcal{N}(\theta, \sigma^2),$$

avec σ^2 connu et supposons que la loi *a priori* de θ soit également gaussienne d'espérance μ et de variance τ^2. μ est l'espérance *a priori* de θ et permet d'intégrer la connaissance préalable sur l'espérance. La variance *a posteriori* τ^2 permet d'intégrer l'incertitude sur la connaissance de μ.

On peut alors montrer que la loi *a posteriori* de θ est encore une loi normale dont l'espérance est égale à

$$\frac{\sigma^2/n}{\sigma^2/n + \tau^2}\mu + \frac{\tau^2}{\sigma^2/n + \tau^2}\bar{y},$$

et de variance

$$\frac{1}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}}.$$

Plusieurs constatations peuvent être faites :

Tout d'abord l'espérance *a posteriori* est une moyenne pondérée de l'espérance *a priori* et de la moyenne arithmétique des observations. Le poids sur la moyenne des observations est d'autant plus grand que le nombre d'observations est grand et d'autant plus petit que la variance *a priori* est petite.

On note également que quand le nombre d'observations ou la variance *a priori* tendent vers l'infini (dans ce cas le prior devient non-informatif), la loi *a posteriori* de θ "tend" vers la distribution de $\bar{y}$, l'estimateur classique de θ. Quand le nombre d'observations grandit ou que la loi *a priori* est non informative, l'inférence bayésienne est équivalente à l'inférence classique (voir Robert, 1992, pour plus de détails).

6.3 Algorithmes MCMC

6.3.1 Principe

Dans la plupart des cas, le calcul de la distribution *a posteriori* sous forme analytique n'est pas envisageable. On cherche alors à simuler un échantillon (dont on choisit la taille) selon cette distribution. La méthode la plus utilisée est l'utilisation d'algorithmes MCMC (*Monte Carlo Markov Chain*).

L'idée est de mettre en place une chaine de Markov $(\psi^{(n)})_n$ ergodique dont la distribution stationnaire (encore dite invariante) est la loi *a posteriori*. En simulant une réalisation suffisamment grande de cette chaîne, on pourra supposer qu'à partir d'un certain n_0, les éléments simulés sont distribués selon la loi *a posteriori* que l'on appelle encore la loi objectif (*target density*). En supposant que pour $n > n_0$, on échantillonne exactement selon la loi objectif, on dispose alors d'une distribution approchée de la précision désirée (au temps de calcul près). L'inférence sur les paramètres (estimation bayésienne, intervalles de crédibilité,...) est alors basée sur cet échantillon.

Notons que les éléments de l'échantillon produit sont corrélés puisque provenant d'une chaîne de Markov.

On appelle phase de chauffe (*burn in*) les itérations effectuées pour $n < n_0$. De nombreuses méthodes empiriques permettent de choisir n_0 à partir de la chaîne produite (Robert, 1995).

Les algorithmes MCMC sont des outils très simples à mettre en œuvre, y compris pour des modèles complexes, ce qui explique certainement leur popularité et l'ascendant que prennent les statistiques bayésiennes sur la théorie classique.

6.3.2 Construction d'un algorithme MCMC

Nous avons choisi de présenter en détail quelques outils qui devraient permettre au lecteur non familier de ces algorithmes de pouvoir écrire un MCMC pour la plupart des modèles. Puis nous détaillons les choix qui permettent de programmer les algorithmes dans le cas des modèles spatiaux. Notons que nous ne présentons que les bases nécessaires et que des améliorations existent qui permettent, selon les modèles, d'augmenter les performances de l'algorithme de base.

L'échantillonneur de Gibbs

L'échantillonneur de Gibbs est un principe général d'algorithme qui permet de ramener la simulation d'un vecteur de dimension p selon la distribution multidimensionnelle de densité f à p simulations unidimensionnelles (ou plus généralement à k simulations, $k \leq p$).

Notons $f_i(x_i|x_1,\ldots,x_{i-1},x_{i+1},\ldots,x_p)$ la distribution conditionnelle de x_i connaissant $x_1,\ldots,x_{i-1},x_{i+1},\ldots,x_p$, $i=1,\ldots,p$, obtenue à partir de f.

Algorithme :

A l'étape t, on dispose d'un échantillon $\mathbf{x}^{(t)} = (x_1^{(t)},\ldots,x_p^{(t)})$.

Pour $i=1,\ldots,p$, on simule $x_i^{(t+1)}$ selon la distribution $f_i(.|x_1^{(t+1)},\ldots,x_{i-1}^{(t+1)},x_{i+1}^{(t)},\ldots,x_p^{(t)})$.

Les densités conditionnelles f_i sont appelées conditionnelles complètes.

Les propriétés de convergence de la chaine construite à partir de l'algorithme ci-dessus sont détaillées par exemple dans Robert et Casella (1999).

Lorsque toutes les conditionnelles complètes sont disponibles analytiquement et que l'on dispose de générateurs aléatoires pour simuler selon ces distributions, le problème de la simulation du vecteur

selon la distribution multidimensionnelle est résolu. A titre d'exemple, on peut citer la simulation d'un vecteur multigaussien (bien que dans ce cas l'échantillonneur de Gibbs ne soit pas forcément l'algorithme le plus efficace).

Lorsque toutes les distributions conditionnelles ne sont pas disponibles analytiquement, on se heurte au nouveau problème de simuler selon ces distributions.

L'algorithme de Metropolis-Hasting permet de simuler une observation selon une distribution (univariée ou multivariée) à partir d'une distribution selon laquelle on peut facilement simuler.

L'algorithme de Metropolis-Hasting

L'objectif de l'algorithme de Metropolis-Hasting est de simuler une observation (ou un échantillon) selon une distribution objectif f, à partir d'une distribution conditionnelle $q(y|x)$.

Pour la mise en œuvre de l'algorithme, on doit pouvoir simuler facilement à partir de la distribution $q(.|x)$, et sa densité $q(y|x)$ doit être disponible analytiquement (au moins à une constante multiplicative près) ou, à défaut, le rapport $f(y)/q(y|x)$ doit l'être.

Algorithme

A l'étape t, on dispose de $x^{(t)}$.

1. On génère $y^{(t)}$ selon $q(.|x^{(t)})$.
2. Puis, on prend

$$x^{(t+1)} = \begin{cases} y^{(t)} & \text{avec probabilité} \quad \rho(x^{(t)}, y^{(t)}) \\ x^{(t)} & \text{avec probabilité} \quad 1 - \rho(x^{(t)}, y^{(t)}) \end{cases},$$

où

$$\rho(x, y) = \min\left\{ \frac{f(y)}{f(x)} \frac{q(x|y)}{q(y|x)}, 1 \right\}.$$

La distribution $q(.|x)$ est appelée distribution instrumentale ou encore noyau de transition, $y^{(t)}$ est le candidat de l'étape t, et la quantité $\rho(x^{(t)}, y^{(t)})$ probabilité d'acceptation du candidat. Notons qu'il y a deux grand types de noyaux : ceux qui dépendent de l'ancienne valeur (ex : on simule $y^{(t)}$ selon une loi gaussienne d'espérance $x^{(t)}$) et ceux qui n'en dépendent pas.

L'algorithme de Metropolis-Hasting pourrait théoriquement être utilisé pour simuler un vecteur aléatoire de dimension p en utilisant une densité instrumentale multidimensionnelle. Mais quand p est grand, ce choix est rarement fait en pratique car la convergence d'un tel algorithme serait extrêmement lente. En effet, plus la dimension de l'espace des paramètres est grande, et plus la proportion de candidats rejetés est importante. On préférera utiliser l'algorithme de Gibbs comme cadre général, et lorsque pour un des paramètres, on ne pourra pas simuler dans une des conditionnelles complètes, on utilisera pour celui-ci l'algorithme de Metropolis-Hasting. Il en résultera ainsi un algorithme hybride, dit de Metropolis sous Gibbs.

Assemblage

Dans le cas des modèles hiérarchiques, 2 représentations graphiques du modèle aident à la construction de l'algorithme. Il s'agit du DAG (*Directed Acyclic Graph*) et du graphique d'indépendances conditionnelles.

Le DAG se construit en dirigeant une flèche d'origine A pointant sur B si $\pi(B|A)$ est donnée par le modèle.

A partir du DAG, on peut construire le graphe d'indépendances conditionnelles en reliant dans le DAG les co-parents d'un même enfant et en remplaçant les flèches par des segments. Le

graphe d'indépendances conditionnelles schématise les structures de dépendances du modèle entre les différents paramètres ou variables latentes.

Prenons l'exemple d'un modèle hiérarchique défini par la donnée de $\pi(B|A)$ et $\pi(D|B,C)$, où D représente les données, B une structure latente et A et C sont des paramètres supposés *a priori* indépendants. Le DAG et le graphe d'indépendance de ce modèle sont donnés respectivement aux figures (6.1) et (6.2).

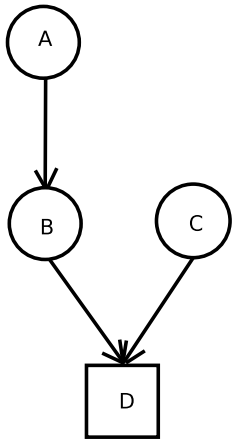

Fig. 6.1. DAG du modèle

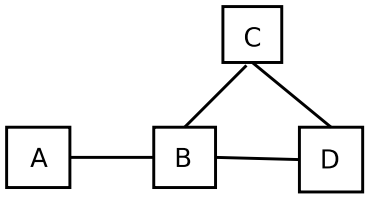

Fig. 6.2. Graphe d'indépendance conditionnelle du modèle

Pour lire le graphe d'indépendances conditionnelles, on note que les trois noeuds A, B, et C apparaissent dans le graphe de telle sorte que pour passer de A à C en suivant les segments, on est obligé de passer par B, alors A et C sont indépendants connaissant B ; de même, A et D sont indépendants connaissant B.

L'objectif de l'algorithme MCMC est d'échantillonner selon la distribution $\pi(A,B,C|D)$.

Si on utilise l'échantillonneur de Gibbs, il suffira de savoir échantillonner dans les distributions $\pi(A|B,C,D)$, $\pi(B|A,C,D)$ et $\pi(C|A,B,D)$.

En utilisant la formule de Bayes et les structures d'indépendance du modèle schématisées à l'aide des deux graphes, on obtient :

$\pi(A|B,C,D) \propto \pi(B|A)\pi(A)$,

$\pi(B|C,D,A) \propto \pi(D|B,C)\pi(B)$,

$\pi(C|A,B,D) \propto \pi(D|B,C)\pi(C)$.

On peut utiliser une étape de Metropolis-Hasting pour échantillonner dans une des distributions conditionnelles. Supposons par exemple que l'on veuille échantillonner A selon $\pi(A|B,C,D)$ en utilisant le noyau $q(.|.)$.

On simule un candidat $A^\star$ selon $q(.|A)$, et on accepte le candidat avec la probabilité

$$\rho(A^\star, A) = \min\left\{1, \frac{\pi(A^\star|B,C,D)q(A|A^\star)}{\pi(A|B,C,D)q(A^\star|A)}\right\}.$$

De nombreux auteurs choisissent la loi *a priori* comme noyau de transition, de telle sorte que la probabilité d'acceptation du candidat se simplifie ; elle est alors égale au minimum entre 1 et le rapport des vraisemblances conditionnelles (aux structures latentes) calculé entre le candidat et l'ancienne valeur :

$$\rho(A^\star, A) = \min\left\{1, \frac{\pi(A^\star|B,C,D)}{\pi(A|B,C,D)}\right\}.$$

6.3.3 Mise en place d'algorithmes dans le contexte spatial

Dans le cas gaussien

L'algorithme que nous présentons ici a été proposé par Banerjee et Gelfand (2002) pour résoudre un problème de régression entre deux variables spatiales observées en des sites qui ne coïncident pas dans le cas du modèle de corrélation intrinsèque **(LM4a)**. Nous avons choisi de le présenter pour montrer comment utiliser les priors conjugués. Nous avons calculé les expressions des distributions conditionnelles nécessaires à la mise en œuvre de l'algorithme, et qui n'étaient pas données dans l'article original.

On rappelle le modèle et on fixe quelques notations :

Soient $\{s_1, \ldots, s_n\}$ les sites pour lesquels au moins une des deux variables est observée, alors $\mathbf{W} = (X(s_1), \ldots, X(s_n), Y(s_1), \ldots, Y(s_n))'$ est un vecteur multigaussien d'espérance

$$\begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix} \otimes \mathbb{1}_n$$

et de matrice de variance $T \otimes H(\theta)$ où

$$T = \begin{pmatrix} \sigma_X^2 & \sigma_{XY} \\ \sigma_{XY} & \sigma_Y^2 \end{pmatrix}$$

et le $i^{\text{ème}}$ terme de $H(\theta)$ est donné par $\rho_\theta(||s_i - s_j||)$ pour une fonction de corrélation paramétrée par θ.

$\mathbf{x}_o$ et $\mathbf{y}_o$ sont les vecteurs d'observation respectifs des variables X et Y, et $\mathbf{w}_o = (\mathbf{x}'_o, \mathbf{y}'_o)'$ désigne le vecteur des observations ; $\mathbf{x}_m$ est le vecteur (aléatoire) des données manquantes, c'est-à-dire le vecteur de la variable X aux sites pour lesquels seule la variable Y est observée. De même $\mathbf{y}_m$ désigne le vecteur aléatoire des données manquantes pour la variable Y et $\mathbf{w}_m = (\mathbf{x}'_m, \mathbf{y}'_m)'$ est le vecteur des données manquantes.

Dans le contexte bayésien, on doit définir les priors pour les paramètres pour spécifier entièrement le modèle. Suivant Banerjee et Gelfand (2002), on prend comme distribution *a priori* pour

$$\mu = \begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}$$

une loi bigaussienne centrée et de matrice de variance :

$$\Sigma_\mu = \begin{pmatrix} 10^5 & 0 \\ 0 & 10^5 \end{pmatrix}.$$

Pour T, on prend une distribution inverse Wishart (voir annexe 3) centrée et de paramètres

$$\Sigma_T = \begin{pmatrix} 10^{-3} & 0 \\ 0 & 10^{-3} \end{pmatrix}$$

et $n_T = 2$.

On choisit un modèle exponentiel pour les corrélations spatiales :

$$\rho_\theta(h) = e^{-\theta h},$$

et on suppose que θ est *a priori* distribué selon une loi inverse gamma dont les paramètres sont choisis de telle sorte que l'espérance pour la portée pratique $(3/\theta)$ soit égale environ à la moitié de la distance maximale séparant deux points de l'échantillon (notée $\tilde{d}$) et la variance soit infinie : $\mathcal{IG}(2, 3/\tilde{d})$.

On met en place un échantillonneur de Gibbs pour échantillonner les paramètres et les valeurs manquantes selon la distribution

$$\pi(\theta, T, \mu_X, \mu_Y, \mathbf{x}_m, \mathbf{y}_m | \mathbf{x}_o, \mathbf{y}_o).$$

On initialise le vecteur des paramètres, par exemple en prenant la moyenne empirique des $\mathbf{x}_o$ pour $\mu_X^{(0)}$, et la variance empirique pour ${\sigma_X^2}^{(0)}$ (on procède de même avec $\mathbf{y}_o$ pour ${\mu_Y}^{(0)}$), 0 pour le coefficient de corrélation initial et une valeur arbitraire ou estimée à partir du variogramme expérimental pour $\theta^{(0)}$.

On peut simuler le vecteur des données manquantes initiales $\mathbf{w}_m^{(0)} = (\mathbf{x}_m^{(0)}, \mathbf{y}_m^{(0)})$ selon la distribution

$$\pi(\mathbf{w}_m | \theta^{(0)}, T^{(0)}, \mu_X^{(0)}, \mu_Y^{(0)}, \mathbf{w}_o),$$

qui est une loi gaussienne multivariée dont les paramètres se calculent par (2.6) et (2.7).

A l'étape $t+1$, on dispose des valeurs courantes $\theta^{(t)}, T^{(t)}, \mu_X^{(t)}, \mu_Y^{(t)}, \mathbf{w}_m^{(t)}$.

On simule simultanément les composantes du vecteur des valeurs manquantes $\mathbf{w}_m^{(t+1)}$ selon la distribution :

$$\pi(\mathbf{w}_m | \theta^{(t)}, T^{(t)}, \mu_X^{(t)}, \mu_Y^{(t)}, \theta^{(t)}, \mathbf{w}_o).$$

Pour la mise à jour de μ_X et μ_Y, on utilise une généralisation au cas multidimensionnel de la propriété de conjugaison présentée plus haut.

En effet, on peut montrer que si la distribution *a priori* du vecteur μ est une loi gaussienne d'espérance $\mathbf{0}$ et de matrice de variance Σ_μ, alors la distribution *a posteriori* de μ, c'est-à-dire, la distribution de μ connaissant $\mathbf{w}_0, \mathbf{w}_m, \theta, T$ est encore une loi gaussienne multivariée d'espérance :

$$\left(\mathbb{1}_n' H(\theta)^{-1} \mathbb{1}_n T^{-1} + \Sigma_\mu^{-1}\right)^{-1} T^{-1} \begin{pmatrix} \mathbb{1}_n' H(\theta)^{-1} \mathbf{x} \\ \mathbb{1}_n' H(\theta)^{-1} \mathbf{y} \end{pmatrix},$$

et de matrice de variance :

$$\left(\mathbb{1}_n' H(\theta)^{-1} \mathbb{1}_n T^{-1} + \Sigma_\mu^{-1}\right)^{-1}.$$

Donc la mise à jour du paramètre μ s'effectue en simulant $\mu^{(t+1)}$ selon une loi bigaussienne d'espérance

$$\left(\mathbb{1}_n' H(\theta^{(t)})^{-1}\mathbb{1}_n {T^{(t)}}^{-1} + \Sigma_\mu^{-1}\right)^{-1} T^{-1} \begin{pmatrix} \mathbb{1}_n' H(\theta^{(t)})^{-1}\mathbf{x} \\ \mathbb{1}_n' H(\theta^{(t)})^{-1}\mathbf{y} \end{pmatrix}$$

et de matrice de variance :

$$\left(\mathbb{1}_n' H(\theta^{(t)})^{-1}\mathbb{1}_n {T^{(t)}}^{-1} + \Sigma_\mu^{-1}\right)^{-1}.$$

Pour la mise à jour de T, on utilise également un résultat de conjugaison. En effet, si la distribution *a priori* de T est une loi inverse Wishart centrée de paramètres Σ_T et n_T, alors ((Banerjee *et al.* , 2000)) :

la loi *a posteriori* de T, c'est-à-dire la distribution de T sachant $\mathbf{w}, \mu_X, \mu_Y$ et θ est encore une distribution inverse Wishart centrée de paramètres

$$\Sigma_T + \begin{pmatrix} (\mathbf{x}-\mu_X\mathbb{1}_n)' H(\theta)^{-1}(\mathbf{x}-\mu_X\mathbb{1}_n) & (\mathbf{x}-\mu_X\mathbb{1}_n)' H(\theta)^{-1}(\mathbf{y}-\mu_Y\mathbb{1}_n) \\ (\mathbf{x}-\mu_X\mathbb{1}_n)' H(\theta)^{-1}(\mathbf{y}-\mu_Y\mathbb{1}_n) & (\mathbf{y}-\mu_Y\mathbb{1}_n)' H(\theta)^{-1}(\mathbf{y}-\mu_Y\mathbb{1}_n) \end{pmatrix}$$

et $n_T + n$

On met donc donc à jour T en simulant $T^{(t+1)}$ selon une loi inverse Wishart de paramètres

$$\Sigma_T + \begin{pmatrix} (\mathbf{x}-\mu_X^{(t+1)}\mathbb{1}_n)' H(\theta^{(t)})^{-1}(\mathbf{x}-\mu_X^{(t+1)}\mathbb{1}_n) & (\mathbf{x}-\mu_X^{(t+1)}\mathbb{1}_n)' H(\theta^{(t)})^{-1}(\mathbf{y}-\mu_Y^{(t+1)}\mathbb{1}_n) \\ (\mathbf{x}-\mu_X^{(t+1)}\mathbb{1}_n)' H(\theta^{(t)})^{-1}(\mathbf{y}-\mu_Y^{(t+1)}\mathbb{1}_n) & (\mathbf{y}-\mu_Y^{(t+1)}\mathbb{1}_n)' H(\theta^{(t)})^{-1}(\mathbf{y}-\mu_Y^{(t+1)}\mathbb{1}_n) \end{pmatrix}$$

et $n_T + n$.

Pour le (ou les) paramètre(s) du modèle de corrélations spatiales, on doit utiliser une étape de Metropolis-Hasting. On procède de la manière suivante :

on propose un candidat $\theta^\star$ pour $\theta^{(t+1)}$ selon le noyau de densité $q(.|\theta^{(t)})$. On accepte le candidat avec probabilité :

$$\rho(\theta^\star, \theta^{(t)}) = \min\left\{1, \frac{\pi(\mathbf{w}_m^{(t+1)}, \mathbf{w}_0|\theta^\star, T^{(t+1)}, \mu_X^{(t+1)}, \mu_Y^{(t+1)})\pi(\theta^\star)q(\theta^{(t)}|\theta^\star)}{\pi(\mathbf{w}_m^{(t+1)}, \mathbf{w}_0|\theta^{(t)}, T^{(t+1)}, \mu_X^{(t+1)}, \mu_Y^{(t+1)})\pi(\theta^{(t)})q(\theta^\star|\theta^{(t)})}\right\}.$$

Dans le cas d'un modèle (GLMM2)

L'algorithme suivant est dû à Diggle *et al.* (1998) pour estimer les paramètres dans le cas d'un modèle **(GLMM2)**. Nous reprenons les notations de la deuxième formulation du modèle **(GLMM2)** du chapitre 3.

Initialisation :

On choisit les paramètres initiaux pour β_0 et β_1 en supposant les effets aléatoires nuls, c'est-à-dire, en utilisant la procédure d'estimation des GLM classiques. A l'étape $t+1$,

On met à jour successivement β_0 et β_1 de la manière suivante :

- pour β_0, on propose un candidat $\beta_0^\star$ pour $\beta_0^{(t+1)}$ selon un noyau à densité $q_{\beta_0}(.|\beta_0^{(t)})$

et on accepte la candidat avec probabilité

$$\rho(\beta_0^\star, \beta_0^{(t)}) = \min\left\{1, \frac{\pi(\mathbf{z}|\varepsilon^{(t)}, \beta_0^\star, \beta_1^{(t)})\pi(\beta_0^\star)q_{\beta_0}(\beta_0^{(t)}|\beta_0^\star)}{\pi(\mathbf{z}|\varepsilon^{(t)}, \beta_0^{(t)}, \beta_1^{(t)})\pi(\beta_0^{(t)})q_{\beta_0}(\beta_0^\star|\beta_0^{(t)})}\right\}.$$

- pour β_1, on propose un candidat $\beta_1^\star$ pour $\beta_1^{(t+1)}$ selon un noyau à densité $q_{\beta_1}(.|\beta_1^{(t)})$

et on accepte la candidat avec probabilité

$$\rho(\beta_1^\star, \beta_1^{(t)}) = \min\left\{1, \frac{\pi(\mathbf{z}|\varepsilon^{(t)}, \beta_0^{(t+1)}, \beta_1^\star)\pi(\beta_1^\star)q_{\beta_1}(\beta_1^{(t)}|\beta_1^\star)}{\pi(\mathbf{z}|\varepsilon^{(t)}, \beta_0^{(t+1)}, \beta_1^{(t)})\pi(\beta_1^{(t)})q_{\beta_1}(\beta_1^\star|\beta_1^{(t)})}\right\}.$$

Puis, on met à jour le vecteur d'effets aléatoires ε composante par composante (selon le principe de l'échantillonneur de Gibbs) en utilisant une étape de Metropolis-Hasting :

pour $i = 1, \ldots, n$, on propose un candidat $\varepsilon_i^\star$ pour $\varepsilon_i^{(t+1)}$ en simulant selon la distribution (qui fait office de noyau de transition) :

$$\pi(\varepsilon_i|\varepsilon_1^{(t+1)}, \ldots, \varepsilon_{i-1}^{(t+1)}, \varepsilon_{i+1}^{(t)}, \ldots, \varepsilon_n^{(t)}, \beta_0^{(t+1)}, \beta_1^{(t+1)}, \theta^{(t)}, \sigma_\varepsilon^{2(t)}),$$

qui est une loi gaussienne univariée dont l'espérance et la variance sont obtenues par (2.6) et (2.7)[2]. De cette manière, la probabilité d'acceptation du candidat se simplifie : elle est donnée par

$$\rho(\varepsilon_i^\star, \varepsilon_i^{(t)}) = \min\left\{1, \frac{\pi(z_i|\varepsilon_i^\star, \beta_0^{(t+1)}, \beta_1^{(t+1)})}{\pi(z_i|\varepsilon_i^{(t)}, \beta_0^{(t+1)}, \beta_1^{(t+1)})}\right\}.$$

Puis on met à jour θ en proposant un candidat $\theta^\star$ pour $\theta^{(t+1)}$ selon un noyau à densité $q_\theta(.|\theta^{(t)})$ et on accepte le candidat avec probabilité

$$\rho(\theta^\star, \theta^{(t)}) = \min\left\{1, \frac{\pi(\varepsilon|\theta^\star, \sigma_\varepsilon^{2(t)})\pi(\theta^\star)q_\theta(\theta^{(t)}|\theta^\star)}{\pi(\varepsilon|\theta^{(t)}, \sigma_\varepsilon^{2(t)})\pi(\theta^{(t)})q_\theta(\theta^\star|\theta^{(t)})}\right\}.$$

Enfin, on met à jour σ_ε^2 en proposant un candidat $\sigma_\varepsilon^{2\star}$ pour $\sigma_\varepsilon^{2(t+1)}$ selon un noyau à densité $q_{\sigma_\varepsilon^2}(.|\sigma_\varepsilon^{2(t)})$

et on accepte le candidat avec probabilité

$$\rho(\sigma_\varepsilon^{2\star}, \sigma_\varepsilon^{2(t)}) = \min\left\{1, \frac{\pi(\varepsilon|\sigma_\varepsilon^{2\star}, \theta^{(t+1)})\pi(\sigma_\varepsilon^{2\star})q_{\varepsilon^2}(\sigma_\varepsilon^{2(t)}|\sigma_\varepsilon^{2\star})}{\pi(\varepsilon|\sigma_\varepsilon^{2(t)}, \theta^{(t+1)})\pi(\sigma_\varepsilon^{2(t)})q_{\varepsilon^2}(\sigma_\varepsilon^{2\star}|\sigma_\varepsilon^{2(t)})}\right\}.$$

Cas du modèle (GLMM4a)

Pour le modèle **(GLMM4a)**, on peut combiner les deux algorithmes précédents de la manière suivante :

A l'étape $t+1$, à partir des paramètre courants, on simule les effets aléatoires :

Pour $i = 1, \ldots, n_Z$, on propose un candidat $y_i^\star$ pour $y_i^{(t+1)}$ en simulant selon la distribution :

$$\pi(y_i|y_1^{(t+1)}, \ldots, y_{i-1}^{(t+1)}, y_{i+1}^{(t)}, \ldots, y_{n_Z}^{(t)}, \mathbf{y}_m^{(t)}, \mathbf{x}_o, \mathbf{x}_m^{(t)}, T^{(t)}, \theta^{(t)}, \mu_X^{(t)}, \mu_Y^{(t)})$$

qui est une loi gaussienne univariée dont l'espérance et la variance sont obtenues par (2.6) et (2.7) et on accepte le candidat avec une probabilité donnée par

$$\rho(y_i^\star, y_i^{(t)}) = \min\left\{1, \frac{\pi(z_i|y_i^\star)}{\pi(z_i|y_i^{(t)})}\right\}.$$

On met ensuite à jour simultanément $\mathbf{y}_m$ et $\mathbf{x}_m$ en les simulant dans la distribution

$$\pi(\mathbf{y}_m, \mathbf{x}_m|\mathbf{x}_o, y_1^{(t+1)}, \ldots, y_{n_Z}^{(t+1)}, T^{(t)}, \theta^{(t)}, \mu_X^{(t)}, \mu_Y^{(t)})$$

[2] Pour limiter le nombre de calculs, notons que la variance conditionnelle pour chaque composante ε_i est donnée, d'après l'équation (2.13), par l'inverse du $i^{\text{ème}}$ terme diagonal de la matrice inverse de $\Sigma^{(t)}$, la matrice de variance globale de ε à l'étape t.

qui est une loi gaussienne multivariée dont l'espérance et la variance sont obtenues par (2.6) et (2.7)

Puis on met à jour les paramètres T, θ, μ_X et μ_Y comme dans l'algorithme présenté dans la sous-section 6.3.3.

Modèle hiérarchique bivarié

De même, on peut facilement construire un algorithme MCMC pour le modèle bivarié hiérarchique (dont la structure bivariée latente est modélisée par un modèle **(LM4a)**). Nous donnons quelques éléments pour mettre en place l'algorithme dans le cas où l'on a n_1 observations de la variable Z_1 et n_2 pour la variable Z_2.

A chaque itération, on simule les effets aléatoires, puis on applique une itération de l'algorithme présenté dans la sous-section 6.3.3 pour mettre à jour les paramètres.

Pour simuler les effets aléatoires, on procède de manière similaire à l'algorithme de Diggle *et al.* (1998)

Notons pour $j = 1, 2$ et $i = 1, \ldots, n_j$, $z_{i,j}$ et $y_{i,j}$ l'observation de la variable j au site i et l'effet aléatoire associé.

A l'étape $t+1$, pour $i = 1, \ldots, n_1$, on propose un candidat $y_{1i}^{\star}$ pour $y_{1i}^{(t+1)}$ selon la distribution

$$\pi(y_{1,i}|y_{1,1}^{(t+1)}, \ldots, y_{1,i-1}^{(t+1)}, y_{1,i+1}^{(t)}, \ldots, y_{1,n_1}^{(t)}, y_{2,1}^{(t)}, \ldots, y_{2,n_2}^{(t)}, T^{(t)}, \theta^{(t)}, \mu_1^{(t)}, \mu_2^{(t)}),$$

qui est une loi gaussienne univariée dont l'espérance et la variance sont obtenues par (2.6) et (2.7) et on accepte le candidat avec une probabilité donnée par

$$\rho(y_{1,i}^{\star}, y_{1,i}^{(t)}) = \min\left\{1, \frac{\pi(z_{1,i}|y_{1,i}^{\star})}{\pi(z_{1,i}|y_{1,i}^{(t)})}\right\}.$$

Puis pour $i = 1, \ldots, n_2$, on propose un candidat $y_{2i}^{\star}$ pour $y_{2i}^{(t+1)}$ selon la distribution

$$\pi(y_{2,i}|y_{1,1}^{(t+1)}, \ldots, y_{1,n_1}^{(t)}, y_{2,1}^{(t+1)}, \ldots, y_{2,i-1}^{(t+1)}, y_{2,i+1}^{(t)}, \ldots, y_{2,n_2}^{(t)}, T^{(t)}, \theta^{(t)}, \mu_1^{(t)}, \mu_2^{(t)}).$$

et on accepte le candidat avec une probabilité donnée par

$$\rho(y_{2,i}^{\star}, y_{2,i}^{(t)}) = \min\left\{1, \frac{\pi(z_{2,i}|y_{2,i}^{\star})}{\pi(z_{2,i}|y_{2,i}^{(t)})}\right\}.$$

6.3.4 Comparaison bayésien/fréquentiste dans le cas d'un modèle de corrélation intrinsèque (LM4a)

Pour conclure cette partie, nous comparons sur un jeu de données simulé selon le modèle de corrélation intrinsèque, les résultats de l'inférence menée dans le cadre bayésien et dans le cadre fréquentiste (en utilisant l'algorithme hybride Fisher scoring/Newton-Raphson). Cet exemple nous permet de mettre en évidence un des problèmes que l'on peut rencontrer avec les algorithmes MCMC quand la "distribution objectif" est une variable aléatoire de grande dimension.

On place les sites d'observation de la variable X sur une grille rectangulaire et régulière dans $\mathbb{R}^2$ de pas unité. Pour chacun de ces sites, on simule un site d'observation pour la variable Y, uniformément dans un carré dont les côtés sont de taille unité et sont parallèles aux mailles de la grille et dont le site d'observation de X est le centre (voir figure 6.3).

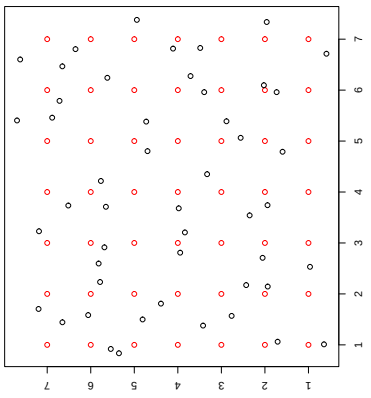

Fig. 6.3. Sites d'observations du jeu de données simulé

Puis on génère les observations selon le modèle de corrélation intrinsèque avec $\sigma_X^2 = \sigma_Y^2 = \theta = 1$ et $\mu_X = \mu_Y = 0$ et $r = 0.5$.

On applique ensuite l'algorithme présenté dans la sous-section 6.3.3. Comme rien n'avait été précisé à ce propos dans l'article original, nous avons choisi comme noyau de transition pour θ, sa loi *a priori* $\pi(\theta)$. On effectue 50 000 itérations.

La figure (6.4) montre, à titre d'exemple, les séries produites par la chaîne de Markov pour trois paramètres en fonction de l'itération courante.

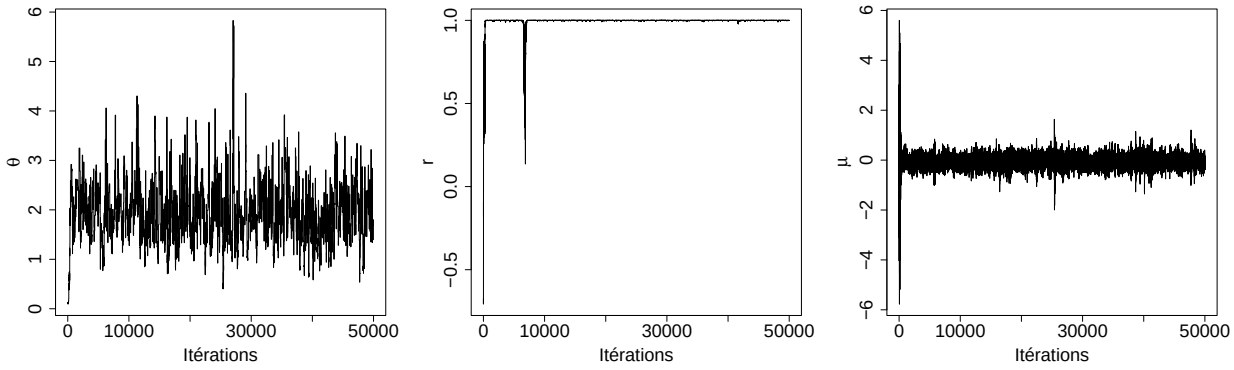

Fig. 6.4. Chaînes MCMC pour trois paramètres. De gauche à droite : θ, r et μ_X.

Les chaînes semblent donc avoir convergé rapidement vers leur distribution stationnaire. On remarque que pour le coefficient de corrélation r, la chaîne semble indiquer que la loi *a posteriori* a une grande densité autour d'une valeur très proche de 1 (voir figure (6.5).

Pour comparer les résultats des deux paradigmes, nous représentons pour μ_X, σ_X^2, r et θ les histogrammes de l'échantillon MCMC auxquels nous ajoutons la valeur de l'estimation par maximum de vraisemblance correspondante (fig. (6.6), (6.7), (6.8) et (6.9)). Pour avoir une idée plus précise de l'information apportée par les données, nous avons également représenté sur le même graphique, les priors et les posteriors empiriques.

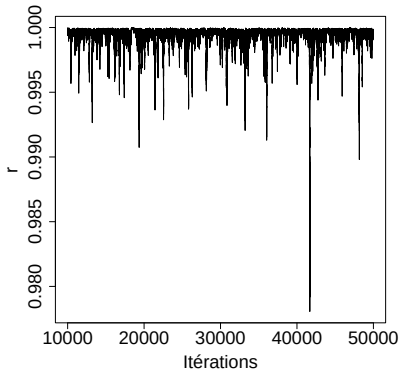

Fig. 6.5. Chaîne MCMC pour le paramètre r à partir de la $10000^{\text{ème}}$ itération.

Pour les paramètres μ_X^2 et σ_X^2, l'information apportée par les observations est importante relativement à la prior et les modes de ces lois *a priori* coïncident avec le maximum de vraisemblance (figures (6.6) et (6.7)). Notons qu'il en est de même pour μ_Y et σ_Y^2 (non représentés).

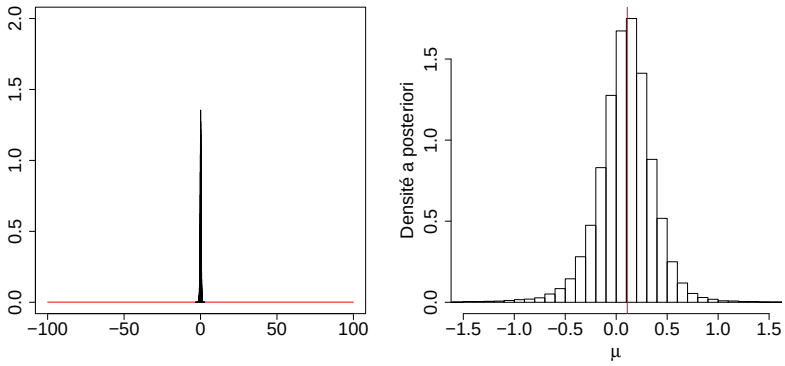

Fig. 6.6. Pour μ_X. Gauche : loi *a priori* théorique (rouge) et *a posteriori* (noire) empirique. Droite : Loi *a posteriori* empirique. En rouge, la valeur du maximum de vraisemblance.

La loi *a priori* pour r obtenue à partir de la distribution inverse Wishart pour T semble être une loi uniforme. La loi *a posteriori* indique que r est très proche de 1 alors que le maximum de vraisemblance est autour de 0.6 pour ce paramètre et le nombre équivalent de données isotopes i.i.d est important (autour de 20). Il ne s'agit donc pas d'un problème lié à la qualité de l'information apportée par les observations, d'autant que le paramètre de portée n'est que légèrement surestimé (voir figure 6.9).

Le problème semble donc provenir de la chaîne de Markov elle-même. Plusieurs répétitions de l'expérience ont toujours conduit à ce résultat. Cependant, il suffit d'avoir trois sites d'observation en commun pour les deux variables (hétérotopie partielle) pour que cette anomalie soit levée. Une tentative d'explication peut être donnée : à chaque itération, les valeurs manquantes sont simulées pour les deux variables ; au moment où le coefficient de corrélation s'approche très près de un (ce qui, compte tenu des conditions de l'expérience advient presque toujours) chaque valeur manquante, pour la variable X par exemple, sera simulée à partir de ses voisins (X du fait de la dépendance

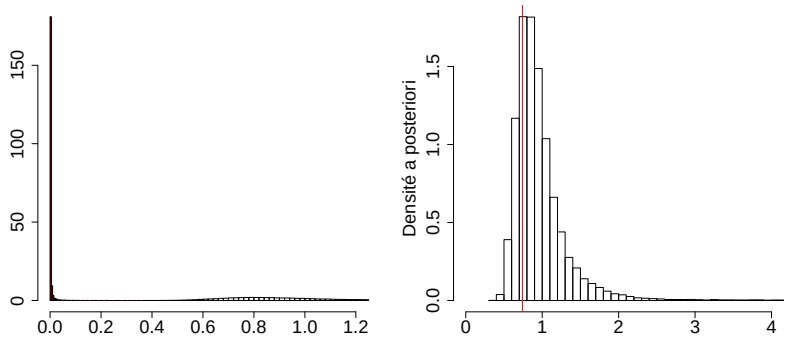

Fig. 6.7. Pour σ_X^2. Gauche : loi *a priori* théorique (rouge) et *a posteriori* (noire) empirique. Droite : Loi *a posteriori* empirique. En rouge, la valeur du maximum de vraisemblance.

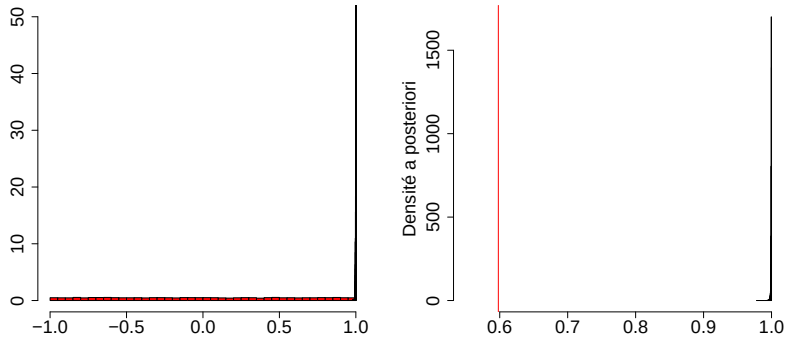

Fig. 6.8. Pour r. Gauche : loi *a priori* simulée (rouge) et *a posteriori* (noire) empirique. Droite : Loi *a posteriori* empirique. En rouge, la valeur du maximum de vraisemblance.

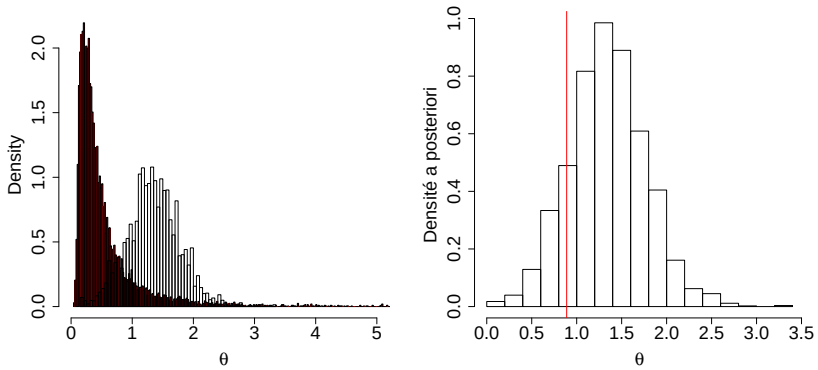

Fig. 6.9. Pour θ. Gauche : loi *a priori* empirique (rouge) et *a posteriori* (noire) empirique. Droite : Loi *a posteriori* empirique. En rouge, la valeur du maximum de vraisemblance.

ponctuelle) et de la variable Y au même site, variable qui aura le poids le plus important du fait de la valeur du coefficient de corrélation, nourrissant ainsi "l'illusion" que les deux variables sont liées linéairement, et ainsi de suite. Trois couples observés aux mêmes sites empêchent donc cette particularité de se produire, ce qui pourrait également être le cas avec une hétérotopie totale mais très faible (en terme de configuration géométrique de l'échantillonnage ou de structures spatiales).

Pour être plus général, le problème est dû au fait que la "distribution objectif" est de grande dimension et que ses composantes sont très corrélées, empêchant la chaîne de Markov d'être suffisamment mélangeante.

Pour pallier cette difficulté, sans doute eut-il mieux valu sacrifier l'élégance de l'algorithme et sa rapidité, en mettant à jour chaque paramètre par une étape de Metropolis-Hasting, sans simuler les valeurs manquantes. La méthode proposée par Banerjee et Gelfand (2002) ne semble en effet pas adaptée à l'hétérotopie totale.

Concernant le paramètre de portée, il semble que l'algorithme MCMC le sous-estime et considère donc les corrélations spatiales plus fortes que ce que nous observons dans le cadre fréquentiste. Le coefficient de corrélation étant lui même surestimé, il est possible qu'il y ait un effet de compensation. Pour le vérifier, nous avons relancé l'algorithme en supposant les paramètres de la matrice T connus et fixés au maximum de vraisemblance. La posterior pour le paramètre de portée indique alors que les corrélations spatiales sont plus fortes (voir figure 6.10), confirmant ainsi l'effet de compensation.

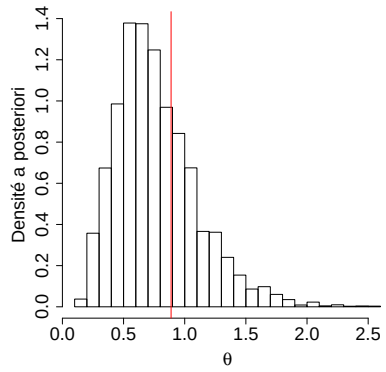

Fig. 6.10. Pour θ quand les autres paramètres de la matrice de covariance sont fixés au maximum de vraisemblance. Loi *a posteriori* empirique. En rouge, la valeur du maximum de vraisemblance.

7

Etudes de cas

Dans cette section, nous présentons deux cas d'étude.

La première étude concerne le dépérissement de la vigne et la dégradation des sols. Elle nous permet de mettre en évidence un type d'hétérotopie qui résulte de l'impossibilité, en raison du coût des analyses et de la nature destructive du protocole d'obervation, de saisir les deux variables aux mêmes sites. Cette étude fut à l'origine des orientations prises par cette thèse notamment en ce qui concerne la modélisation. Elle met également en évidence certaines perspectives et limites de notre travail.

Le second cas d'étude, qui porte sur les liens entre l'histoire climatique et la biodiversité intraspécifique actuelle d'une espèce d'arbre, est plus étroitement lié au cadre méthodologique développé durant la thèse. Il permet de montrer l'intérêt pratique des développements théoriques du chapitre 6, notamment en ce qui concerne la qualité d'un jeu de données relativement à la configuration spatiale de l'échantillonnage.

7.1 Dépérissement de la vigne et dégradation des sols

7.1.1 Présentation de l'étude

Pour expliquer le dépérissement de la vigne dans la basse vallée de la Peyne, un projet de recherche GESSOL fut engagé par le ministère de l'environnement (2003) et l'INRA a été chargé de cette étude. Les responsables de l'étude ont avancé l'hypothèse que sous l'effet des pratiques culturales (en particulier le travail du sol et le passage des engins), la structure de certains sols fragiles pouvait se dégrader au point de limiter de façon importante les paramètres de porosité et donc de modifier considérablement l'ensemble des transferts hydriques ou gazeux (Legros *et al.* , 1998). L'hypothèse corollaire conduisait à penser que le dépérissement d'un nombre important de pieds de vigne pouvait être lié à cette dégradation de structure.

Cette liaison fonctionnelle type de sol/dépérissement a tout d'abord été abordée à grande échelle à l'aide de cartes pédologiques pour le sol, et d'une note de dépérissement pour plus de 500 parcelles à partir de photographies aériennes (Lagacherie *et al.* , 2001).

Pour affiner les résultats de l'analyse statistique effectuée à l'échelle de toute la zone d'étude, des prélèvements du sol (192 carottages) ont été effectués à l'intérieur de certaines parcelles et la description exhaustive de l'état de la vigne a été faite dans une zone entourant ces sondages. Nous nous sommes concentrés exclusivement sur ce second jeu de données.

Sur 4 communes de la région concernée, 96 placettes (portion de parcelles) de 20 mètres sur 15 mètres ont été choisies, 3 par parcelle. Sur chaque placette, deux sondages pédologiques ont été

effectués et deux propriétés ont été retenues pour caractériser les signes de dégradation des sols étudiés : l'hydromorphie et le type textural, basé sur le taux d'argile. Il s'agit de deux variables à 3 modalité ordonnées. L'état de chaque pied de vigne de la placette est résumé sous forme binaire selon que le pied est sain (0) ou dépérissant (1). Les sondages étant effectués entre les pieds de vigne, il en résulte un jeu de données totalement hétérotope.

7.1.2 Nos choix méthodologiques

Devant la nature catégorielle des variables explicatives, nous avons choisi de ne pas modéliser ces variables comme aléatoire, et pour traiter l'hétérotopie, nous avons simplement considéré que les caractéristiques du sol, observées au niveau d'un sondage, définissaient le sol dans une couronne autour de ce sondage. Nous avons obtenu 3 jeux de données en choisissant 3 rayons de couronnes : 1m (4 pieds), 1,5m (entre 4 et 8 pieds) et 2m (entre 4 et 12 pieds) (voir figure 7.1). Le nombre de pieds autour d'un sondage peut varier en fonction de l'écartement des rangs (gobelet ou palissade) et de la localisation du sondage dans la placette (près d'un bord ou non).

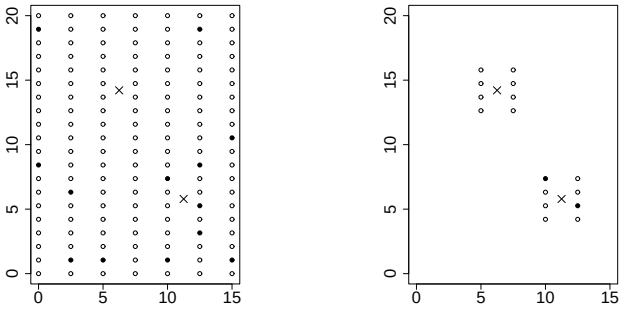

Fig. 7.1. Gauche : une placette schématisée, les croix désignent les sites de sondage du sol et les cercles désignent les pieds de vignes (sains=cercles vides, dépérissant=cercles pleins). Droite : pieds séléctionnés dans une couronne de 1,5 mètres autour du sondage.

Il résulte de cette procédure, 192 individus statistiques géoréférencés. La variable d'intérêt peut être modélisée comme une loi binômiale (le nombre de pieds dépérissants dans la couronne) à condition de considérer que les pieds de vignes à l'intérieur d'une couronne sont mutuellement indépendants (aux effets du sol observé près). Une procédure classique basée sur un GLM (lien logit, effets des deux variables du sol et de leur interaction) montre que les effets sol n'expliquent qu'une très faible part de la variabilité du dépérissement (voir Desassis *et al.* , 2005). Pour modéliser la variabilité non prise-en-compte, on utilise un modèle GLMM à effets aléatoires spatialement corrélés **(GLMM2)**.

Modèle

Rappelons le modèle :

Soit $S(.)$ un champ aléatoire, d'espérance nulle et de fonction de covariance exponentielle

$$C(h) = \sigma^2 \rho_\theta(h)$$

avec

$$\rho_\theta(h) = \alpha \mathbb{1}_{[h=0]} + (1-\alpha)e^{-\phi h},$$

où α est la part d'effet de pépite dans la variance totale et ϕ est le paramètre de portée.

On suppose que conditionnellement à une réalisation $\mathbf{S} = (S_1, \ldots, S_{192})$ de $S(.)$, le nombre de pieds N_i dépérissants autour du $i^{\text{ème}}$ sondage est distribué selon une loi binômiale $\mathcal{B}(n_i, p_i)$ où n_i est le nombre de pieds de vigne dans la couronne, et p_i est la probabilité qu'un de ces pieds soit dépérissant. On suppose que les N_i sont mutuellement indépendants connaissant $\mathbf{S}$. On suppose que

$$p_i = E[N_i|S_i] = \frac{e^{\mu_i + S_i}}{1 + e^{\mu_i + S_i}},$$

où μ_i est la somme d'une constante et des effets du sol à estimer ainsi que leur interaction, soit au total 9 paramètres à estimer pour μ_i (d'après le système des contraintes).

Algorithme

Pour estimer les 12 paramètres (9 pour l'espérance et 3 pour la structure spatiale de $S(.)$), on se place dans le cadre bayésien et on met en place un algorithme MCMC.

Pour les lois *a priori* des paramètres, on choisit :

- des lois uniformes sur $[-10, 10]$ pour les 9 paramètres d'espérance, ce qui permet de couvrir un large spectre de probabilités dans l'échelle logit.

- pour $\sigma_s^2 = \sigma^2(1-\alpha)$ et $\sigma_p^2 = \sigma^2\alpha$ les paramètres de variance de $S(.)$ reparamétrisés, on choisit des priors impropres :

$$\pi(x) \propto \frac{1}{x}.$$

- pour ϕ on choisit une distribution inverse Gamma $\mathcal{IG}(2, 6/M)$ où M est la distance maximale séparant deux sites du domaine (comme proposé par Banerjee et Gelfand, 2002).

Pour l'algorithme, on utilise celui correspondant au modèle **(GLMM2)** et présenté dans le chapitre 6. Il ne reste qu'à définir les noyaux de transition :

- pour les paramètres de μ_i que l'on peut noter τ_{ij}, $j = 1, \ldots, 9$, on choisit un noyau gaussien dont l'espérance est la valeur du paramètre à l'itération précédente, et la variance 0.01.

- pour ϕ, on utilise la loi *a priori* de ϕ, $\pi(\phi)$,

- pour σ_p^2 et σ_s^2, on utilise le produit entre l'ancienne valeur et une variable log normale, c'est-à-dire l'exponentielle d'une variable gaussienne. Pour les paramètres de la loi gaussienne, on prend 0 pour l'espérance et 0.01 pour la variance. Avec ce choix, la probabilité d'acceptation, dans l'algorithme de Metropolis-Hasting pour la mise à jour des paramètres de variance, se simplifie. Par exemple pour $\sigma_s^{2\star}$ un candidat pour la nouvelle valeur de σ_s^2, cette probabilité s'écrit (voir Klein *et al.* , 2007) :

$$\rho(\sigma_s^{2\star}, \sigma_s^2) = \min\left\{1, \frac{\pi(\mathbf{S}|\sigma_s^{2\star}, \sigma_p^2, \phi)}{\pi(\mathbf{S}|\sigma_s^2, \sigma_p^2, \phi)}\right\}.$$

On simule une chaîne de 100 000 itérations dont on conserve les 90 000 dernières.

7.1.3 Résultats

Nous présentons brièvement les résultats obtenus.

Effets du sol

Les chaînes de Markov pour les effets du sol semblent converger rapidement (voir fig 7.2). Au niveau des effets individuels, lorsque l'on considère les couronnes de rayon 1,50 mètre, la texture moyenne apparaît comme une condition moins propice au dépérissement, bien que 0 appartienne à l'intervalle de crédibilité à 95 % ([-1.51,0.01]).

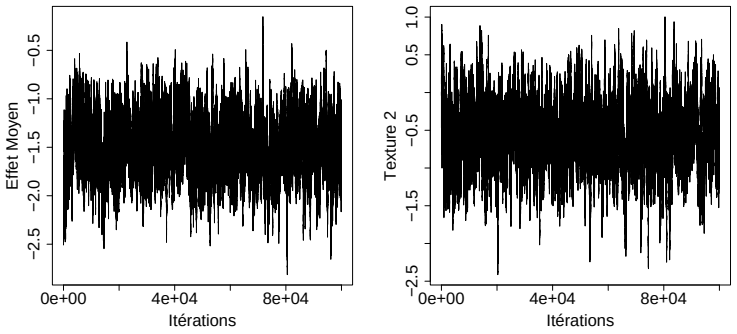

Fig. 7.2. Evolution des chaînes de Markov pour 2 paramètres, l'effet moyen (gauche) et l'effet texture 2 (droite) pour les couronnes de 1 mètre.

Pour les couronnes de rayon 2 mètres, aucun effet du sol n'est détecté par la méthode, ce qui pourrait être compatible avec l'échelle spatiale de la variation des caractéristiques du sol décrites par les 2 variables sélectionnées.

Ces résultats confirment les analyses effectuées à grande échelle. En effet, la mortalité de la vigne, estimée grossièrement à partir des photographies aériennes, était significativement moins importante sur le type de sol 651 (de la carte pédologique), qui correspond, dans la classification pédologique utilisée, à un sol de texture moyenne.

Structures spatiales du champ latent

Les résultats concernant les paramètres de structure du champ latent indiquent la quasi-absence de structure spatiale. En effet, comme le montre la figure (7.3), dans les trois cas, toute la variance du champ de résidus est expliquée par l'effet de pépite et la partie structurée spatialement n'a pratiquement pas de poids. La loi *a posteriori* du paramètre de portée ϕ semble identique à sa loi *a priori* (fig 7.4), ce qui est une conséquence attendue de la très faible valeur de σ_s^2 estimée tend à rendre ϕ non identifiable.

L'observation de structures en damier du dépérissement (corrélations spatiales à courtes distances) et la recherche par les expérimentateurs de la plus grande hétérogénéité possible entre les deux sondages d'une même placette (et ce afin de détecter au mieux les effets du sol responsables de ce dépérissement) peuvent expliquer, en l'absence d'effets du sol importants, l'importance de l'effet de pépite.

7.1.4 Discussion

Les résultats obtenus par notre analyse semblent confirmer l'exisstence d'un effet du sol. Mais une forte variabilité du dépérissement observé entre les sondages n'est pas explicable en dehors du

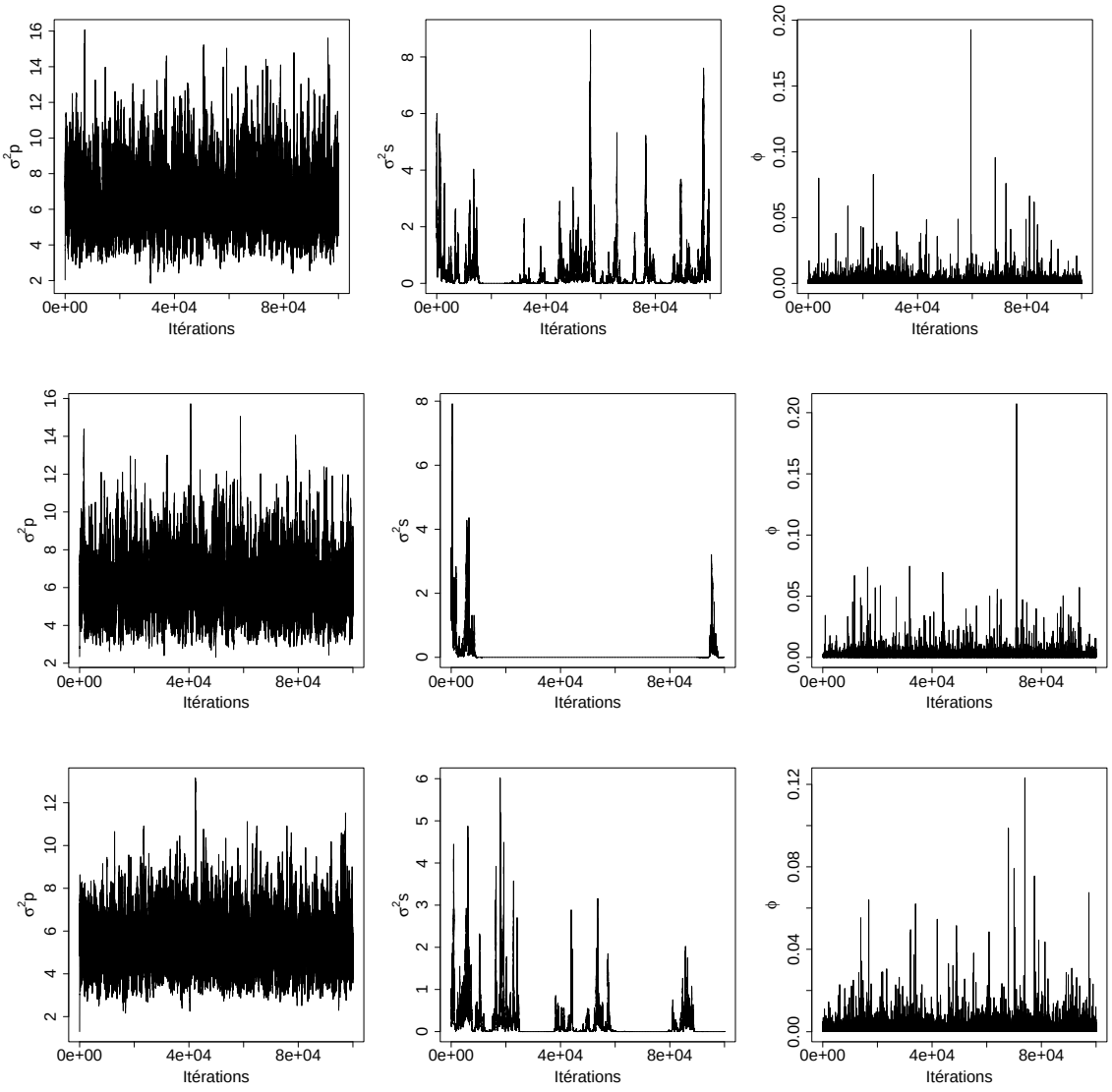

Fig. 7.3. Evolution des chaînes de Markov pour 3 paramètres, l'effet de pépite (gauche), la variance de la part structurée spatialement (centre) et du paramètre de portée (droite) pour les couronnes de 1 mètre pour des couronnes de 1 m (haut), 1,50 m (milieu) et 2 m (bas).

protocole de collecte des données. La recherche de la plus grande hétérogénéité à l'intérieur d'une parcelle se justifie pour identifier au mieux les causes responsables du dépérissement ; mais cette méthodologie n'a pas été suivie lors du choix des placettes échantillonnées parmi l'ensemble des parcelles de la zone d'étude. En effet, la mortalité des pieds à l'intérieur des placettes était très supérieure à celle observée sur l'ensemble des parcelles, ce qui empêche une vision d'ensemble du phénomène et rend difficile la mise en évidence des causes de la mortalité. En particulier, aucun des paramètres correspondant aux effets du sol (et qui mesure les écarts à la proportion moyenne de pieds dépérissants) n'a été évalué significativement positif.

Ces choix méthodologiques peuvent également expliquer l'absence de structures spatiales du champ latent pour les couronnes de plus petites tailles. Mais nos propres choix méthodologiques (priors et noyaux de transition) auraient pu avoir une influence déterminante sur les résultats. En

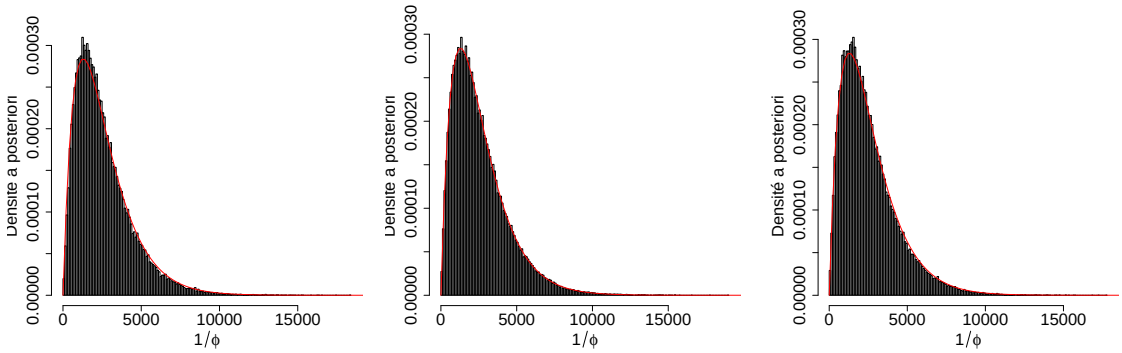

Fig. 7.4. Lois *a posteriori* empirique (histogramme) et densité *a priori* (courbe rouge) pour l'inverse du paramètre de portée ($1/\phi$). A gauche : couronnes de 1 mètre. Au centre : couronnes de 1,50 mètre. A droite : couronnes de 2 mètres.

effet nous avons choisi des priors impropres pour les paramètres de variance et une espérance *a priori* de la portée très grande par rapport aux distances intervenant dans le jeu de données[1]. Si la portée *a priori* est trop grande par rapport à ce qu'indiquent les observations, cela pourrait conduire l'algorithme à éliminer la composante structurée spatialement de la variabilité résiduelle totale ($\sigma_s^2 \sim 0$).

Pour s'assurer que nos choix permettent d'estimer une portée plus faible, nous avons mené une petite étude par simulation. Pour chaque taille de couronne, nous avons simulé une réalisation, aux sites d'observation, d'un champ gaussien stationnaire avec $\sigma_p^2 = \sigma_s^2 = 2$ et $\phi = 1/200$. Nous avons fixé tous les effets sols à zéro et l'espérance constante à -2. Dans les trois cas, nous avons ensuite simulé le nombre de pieds dépérissants en chaque site selon des lois binômiales dont les tailles étaient données par chacun des trois jeux de données réelles, et le paramètre de probabilité calculé à partir des réalisations du champ selon le modèle. Nous avons ensuite appliqué l'algorithme avec les mêmes choix de priors et de noyaux que précédemment (mais sur 50 000 itérations). Conformément aux vraies valeurs, aucun effet du sol n'est détecté comme significatif. En ce qui concerne les paramètres de structure, les figures (7.5) et (7.6) détaillent ces résultats. Il semble que l'on puisse estimer ces paramètres pour la plus grande taille de couronne (2m), malgré un effet de pépite qui représente 50% de la variabilité totale de la résiduelle. En revanche, pour les deux plus petites, l'algorithme a un comportement proche du cas réel, bien que l'on détecte un faible apport d'information sur la portée dans le cas des couronnes de taille 1,5 mètre.

Nos choix de loi *a priori* peuvent être critiqués, mais leur caractère vague (variance *a priori* de ϕ infinie) doit permettre aux données de ne pas leur donner trop de poids. Nos simulations montrent donc que peu d'information est apportée par les données sur les structures spatiales du champ sous-jacent, quand les paramètres de taille des variables binômiales (ici, nombre de pieds considérés autour de chaque sondage) sont petits. Quand la taille des couronnes est plus grande, les structures spatiales du champ latent sont néanmoins détectables en dépit d'un fort effet de pépite dans le cadre de nos simulations. Si des simulations supplémentaires, avec différents effets de pépite, peuvent permettre de mieux comprendre l'information sur les structures des champs sous-

[1] La portée pratique a une espérance *a priori* égale à la moitié de la distance maximale entre deux sondages. Or les sites d'observations sont répartis sur 4 communes distantes, ce qui implique une grande portée *a priori* par rapport aux distances dans une commune.

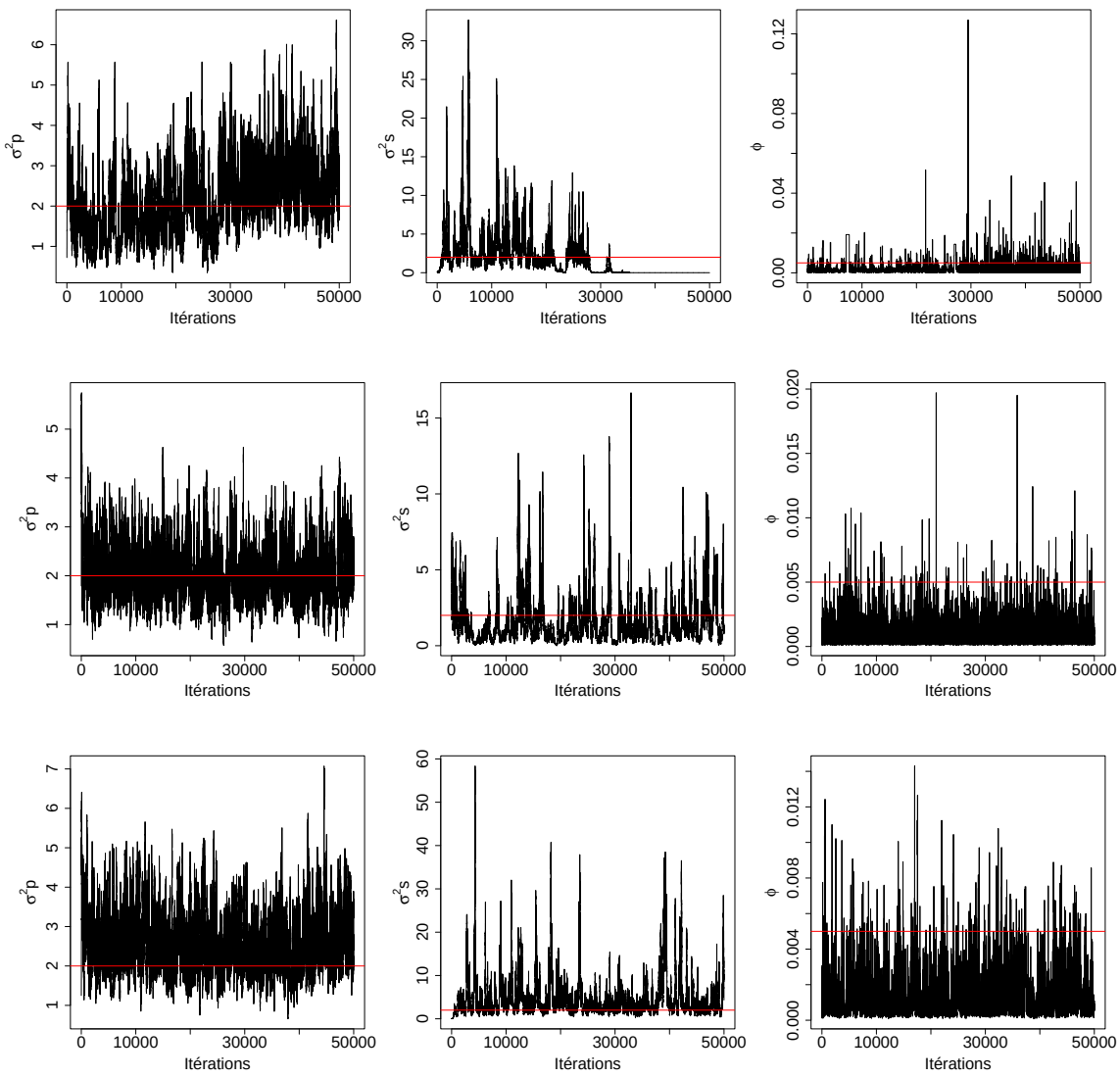

Fig. 7.5. Evolution des chaînes de Markov pour 3 paramètres, l'effet de pépite (gauche), la variance de la part structurée spatialement (centre) et du paramètre de portée (droite) pour les couronnes de 1 mètre pour des couronnes de 1 m (haut), 1,50 m (milieu) et 2 m (bas). Axe rouge : vraie valeur du paramètre.

jacents qu'amènent des données binômiales, la présente étude semble montrer l'absence de structures spatiales dans le cas du dépérissement.

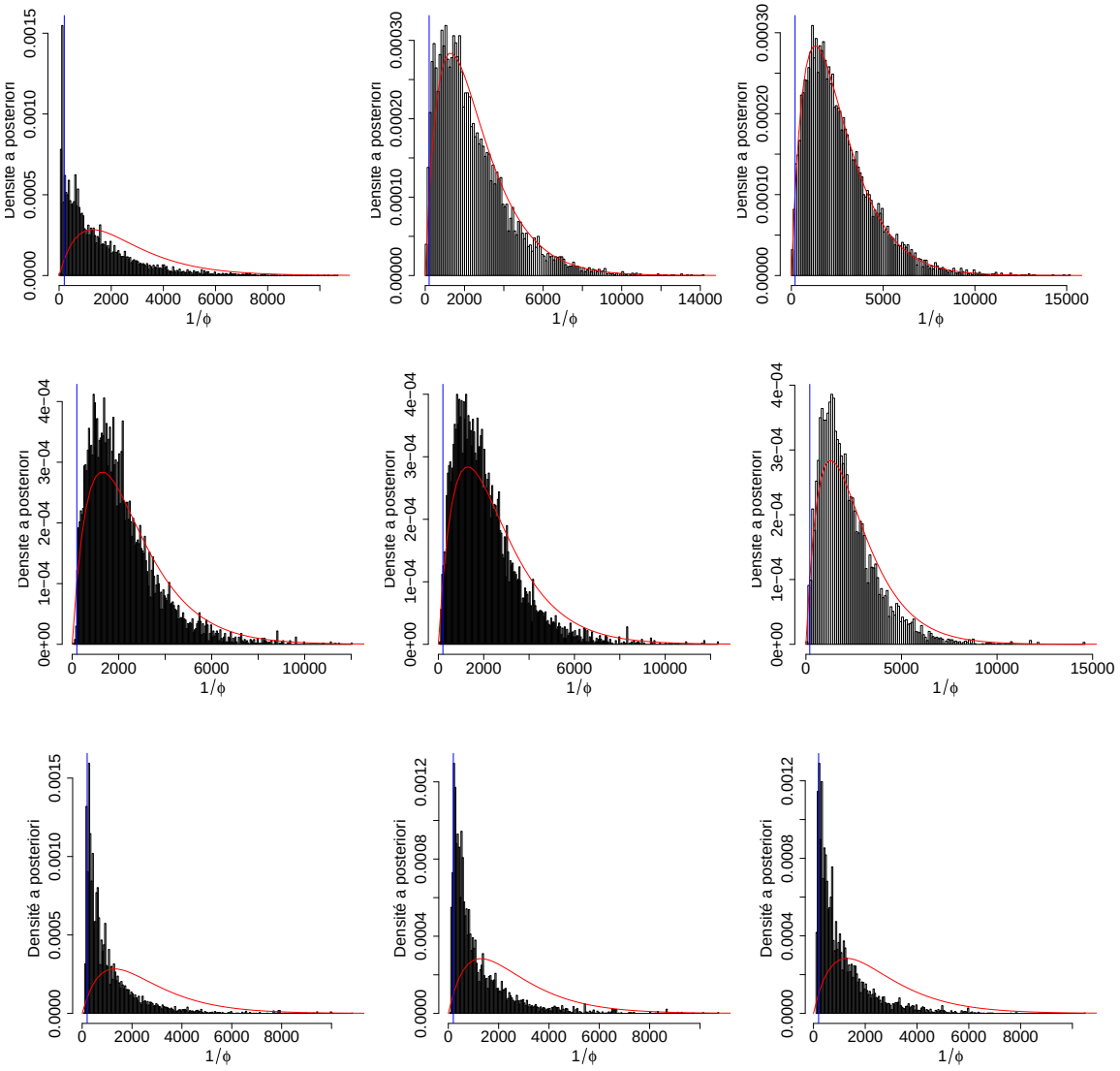

Fig. 7.6. Evolution des lois *a posteriori* de $1/\phi$ entre les itérations [20 000,30 000] (gauche) , [30 000,40 000] (centre) ,[40 000,50 000] (droite) pour des couronnes de 1 m (haut), 1,50 m (milieu) et 2 m (bas). Courbe rouge : distribution *a priori*. Axe bleu : vraie valeur de $1/\phi$.

7.2 Histoire climatique et biodiversité

Comme nous l'avons annoncé en introduction de ce chapitre, l'étude que nous présentons ici à une vocation illustrative et ne prétend en aucun cas apporter des réponses décisives concernant le problème étudié, mais plutôt de montrer l'intérêt de nos outils et la manière de les utiliser.

7.2.1 Problèmatique et données

Avant d'entrer dans les détails sur le jeu de données et sur la problématique, il nous semble utile de préciser les circonstances qui nous ont amenés à conduire cette étude. Lors de mon premier comité de thèse que j'avais axé essentiellement sur le problème de l'hétérotopie, Avner Bar Hen (Professeur

de biomathématiques, Paris XIII) membre de ce comité, nous a proposé un jeu de données qu'il avait à sa disposition et qui semblait de nature à enrichir nos travaux.

Le jeu de données provient de deux études menées indépendamment sur la même zone géographique, l'Europe centrale (voir figure 7.7).

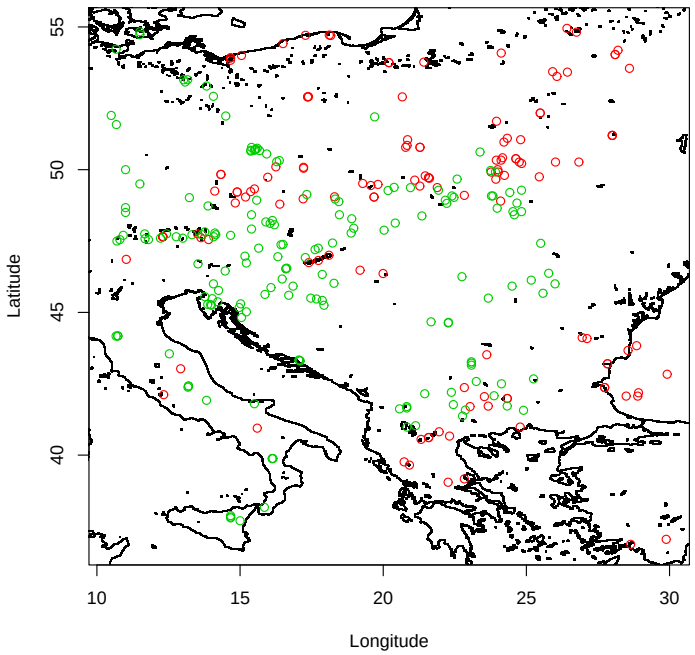

Fig. 7.7. Carte des sites d'observation. En rouge : les sites de reconstitution du climat. En vert : les sites des fôrets sur lesquelles la biodiversité est mesurée.

Le premier jeu de données est issu d'une étude paléoclimatique. A partir de carottages du sol en 133 sites, le climat passé a été reconstitué sur une période allant de -14000 ans av JC à nos jours. En chaque site, on dispose donc de deux séries temporelles, une pour la température et l'autre pour les précipitations. Notons que le climat pour être reconstitué nécessite la présence de pollen fossilisé dans les carottes, ce qui explique que l'on ne dispose pas d'une donnée à chaque pas de temps pour tous les sites.

Le second jeu de données contient, pour 126 forêts, deux mesures caractéristiques de la biodiversité actuelle pour une espèce d'arbre : l'hétérozygotie et le taux allélique.

Toutes ces variables sont des variables continues et par conséquent peuvent être modélisées par des lois gaussiennes.

En l'absence d'une concertation possible avec des interlocuteurs spécialisés dans les domaines d'application (l'idéal eut été d'être en contact avec les initiateurs des deux études), nous n'avons pas cherché à explorer toutes les possibilités de modélisation adaptées à ce problème. Nous avons plutôt "forcé" les données à rentrer dans le cadre méthodologique vers lequels nous nous étions préalablement orientés.

Nous avons donc sélectionné une variable par jeu de données : la température pour les données paléoclimatiques et l'hétérozygotie pour la biodiversité intra-spécifique. Le choix de ces deux variables

a été fait sur la base d'analyses exploratoires qui nous permettaient de penser qu'elles étaient les plus à même de mettre en valeur notre méthodologie (notamment par la présence de structures spatiales apparentes). Nous avons ensuite simplifié les observations temporelles en moyennant en chaque site les séries climatiques sur 9 périodes de 1000 ans et une de 6000 [-16000,-1000]) du fait de la pauvreté des observations sur cette période (voir fig. 7.8).

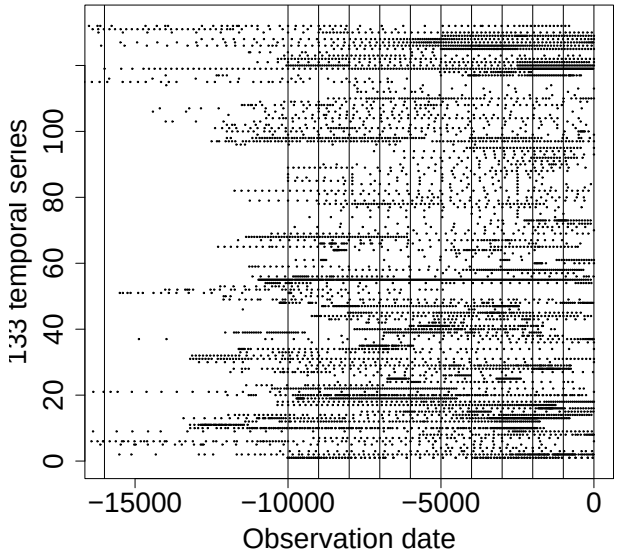

Fig. 7.8. Dates d'observation (absisse) pour les 133 séries temporelles (ordonnée). Les lignes verticales indiquent les limites des périodes sélectionnées.

On dispose donc après ce traitement de 10 jeux de données pour l'histoire climatique et notre objectif était alors de mettre en évidence les périodes climatiques anciennes où le climat est lié significativement à la biodiversité intraspécifique actuelle. Cette thématique rejoint les préoccupations du moment associées à la compréhension des changements climatiques globaux et de l'adaptation (Petit *et al.* , 2003).

7.2.2 Analyses et résultats

Analyses exploratoires

En présence de jeux de données hétérotopes, tracer un nuage de points entre les deux variables n'est pas réalisable. On doit donc envisager d'autres outils exploratoires. Tout d'abord, des corrélations spatiales de chaque variable sont indispensables pour espérer capturer un lien éventuel entre les variables. Les variogrammes expérimentaux (fig (7.9)) mettent en évidence, outre un effet de pépite important, l'existence de correlations spatiales.

Pour le lien entre les variables, nous avons construit une fonction empirique de corrélation croisée ; c'est-à-dire que pour différentes distances, nous avons calculé le coefficient de corrélation entre les couples de $X(s_i)$ et de $Y(s_j)$ séparés de ces distances. La figure (7.10) montre, pour deux périodes, des comportements différents à l'origine de cette fonction : dans un cas (à gauche), il semble qu'il y ait une liaison, alors que dans l'autre, si elle existe, la liaison est moins apparente.

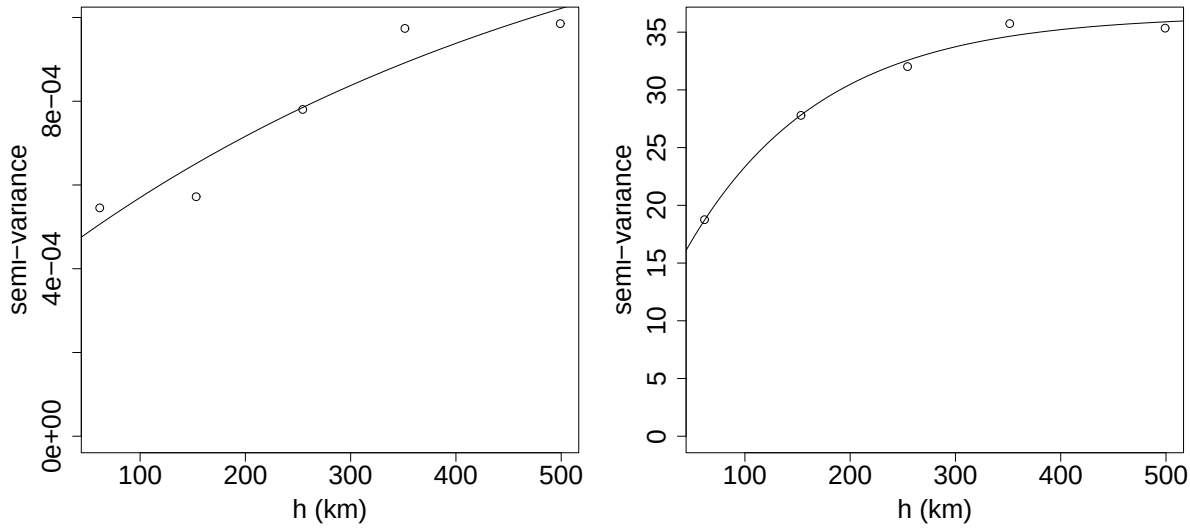

Fig. 7.9. Variogrammes expérimentaux de l'hétérozygotie (gauche) et des températures moyennes entre -6999 et -6000 (droite) et ajustement par un modèle exponentiel et un effet de pépite.

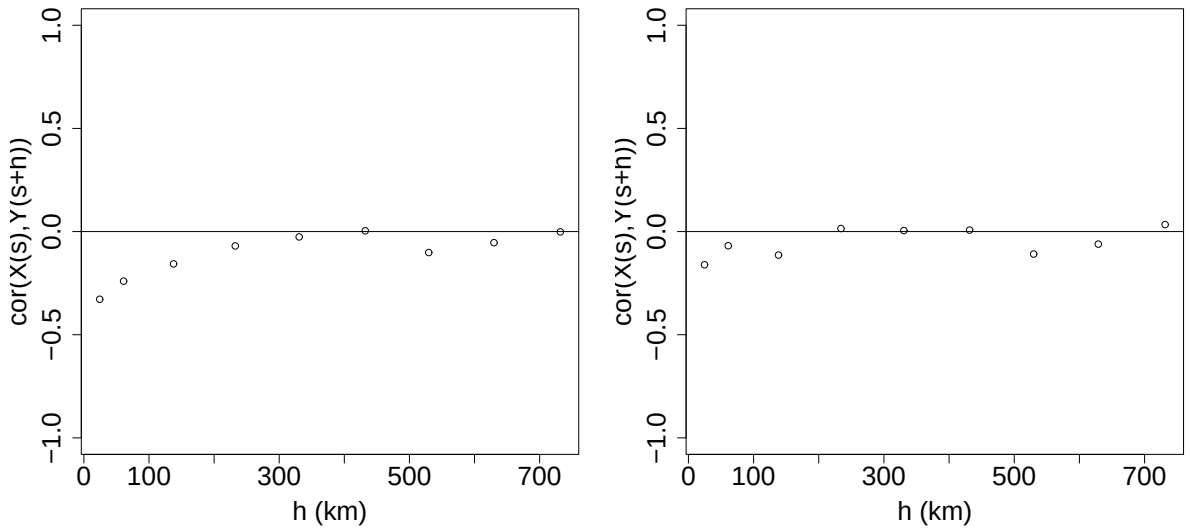

Fig. 7.10. Fonctions de corrélation croisées empiriques entre l'hétérozygotie et les températures moyennes entre -6999 et -6000 (gauche) et -5999 et -5000 (droite).

Sur la base de cette analyse exploratoire, on voudrait avoir une approche plus quantitative en utilisant les outils développés dans les chapitres précédents.

Approche inférentielle paramétrique

On utilise un modèle de corrélation intrinsèque pour modéliser la relation entre les deux variables.

Comme modèle de corrélations spatiales, les variogrammes expérimentaux suggèrent d'utiliser un modèle exponentiel en conjonction avec un effet de pépite.

On estime les paramètres du modèle pour chaque période de temps en utilisant l'algorithme hybride Fisher scoring/Newton-Raphson en reparamétrant comme suit :

$$\alpha^\star = \frac{1}{1+e^\alpha}.$$

Pour évaluer si les liaisons, mises en évidence par le coefficient de corrélation, sont significatives, on utilise un test du rapport de vraisemblance (*Likelihood Ratio Test*). C'est-à-dire que l'on calcule L, le log du rapport entre la vraisemblance maximisée en fonction de tous les paramètres et la vraisemblance maximisée en fixant le coeffcient de corrélation r à 0. On compare ensuite $2L$ à sa loi asymptotique sous H_0 : la distribution du Chi-2 à 1 degré de liberté (nombre de paramètres fixés sous H_0). Le nombre équivalent de données isotopes i.i.d étant pour toutes les périodes relativement faible, nous avons complété notre approche en effectuant un bootstrap de la statistique du test sous H_0. Les résultats sont consignés dans le tableau (7.1).

Tab. 7.1. Tableau des résultats

Périodes		N_{eq}	Paramètres estimés					P-valeurs	
De	à		$\hat{\mu}_X$	$\hat{\sigma}_X$	$\hat{\alpha}\times 100$	$1/\hat{\phi}$	$\hat{r}$	asymptotique	bootstrap
-16000	-10000	5.70	-7.3	6.33	48.54	420.35	-0.84	0.051	0.100
-9999	-9000	6.01	-5.6	6.39	47.44	409.39	-0.95	0.014	0.026
-8999	-8000	7.05	-4.1	6.03	42.69	302.29	-0.81	0.019	0.037
-7999	-7000	6.43	-3.5	6.17	47.21	313.71	-0.76	0.045	0.064
-6999	-6000	8.11	-3.4	5.53	35.47	235.61	-0.78	0.019	0.033
-5999	-5000	7.21	-3.7	5.18	43.77	303.34	-0.46	0.232	0.278
-4999	-4000	7.44	-2.8	5.13	44.55	230.99	-0.49	0.221	0.257
-3999	-3000	7.06	-3.3	5.70	49.13	240.61	-0.53	0.206	0.242
-2999	-2000	6.15	-3.2	5.03	53.69	336.82	-0.81	0.108	0.136
-1999	-1000	7.63	-3.8	5.46	41.10	241.92	-0.92	0.013	0.023
-999	0	7.66	-4.7	5.78	37.98	195.03	-0.46	0.319	0.363

A l'exception des trois périodes comprises entre -5900 et -3000, les coefficients de corrélation calculés sont très élevés. Cependant, ces valeurs doivent être relativisées du fait de la pauvreté de l'information apportée par les données sur le coefficient de corrélation. En effet, le nombre équivalent (< 10) suggère la plus grande précaution pour interpréter ces résultats. Toutefois, les test basés sur la statistique LRT semblent suggérer que, pour les périodes comprises entre -9999 et -6000 d'une part et entre -1999 et -1000 d'autre part, il y a une corrélation négative entre les deux variables.

La forte variabilité de la portée entre les périodes suggère d'accorder une plus grande liberté de structure spatiale au modèle, par exemple en laissant les effets de pépite libres pour chaque variable.

Ces considérations sur les modèles et l'ajustement ou sur la faible information apportée par les données sur le coefficient de corrélation, n'excluent pas de s'interroger sur les hypothèses implicites faites par notre famille de modèles et notamment la dépendance ponctuelle. L'état actuel des connaissances concernant le problème de la biodiversité intra-spécifique et de l'histoire climatique (Petit *et al.* , 2003) nous conduirait plutôt à envisager des effets directionnels. En effet, les variations climatiques sont essentiellement dues au retrait des zones glaciaires au cours des périodes étudiées, libérant peu à peu du sud au nord un espace propice à la recolonisation par les forêts, permettant ainsi aux refuges anciens de la Méditérannée de s'étendre vers le nord. Il est admis que les zones de croisement des différents flux de recolonisation ont une grande richesse génétique.

8

Conclusions

Notre objectif initial était d'adapter certaines procédures des statistiques classiques telles que la régression linéaire, au contexte spatial et plus précisément aux données géoréférencées et hétérotopes. Ce travail devait permettre le traitement de données continues que l'on peut modéliser à partir de champs gaussiens, mais nous voulions aussi pouvoir traiter des données pour lesquelles l'hypothèse gaussienne est inappropriée, par exemple des données binaires ou de comptages, en se plaçant dans le contexte des modèles linéaires généralisés.

En voulant comprendre comment s'organisaient les différents modèles disponibles dans la littérature (Mardia et Marshall (1984), Diggle *et al.* (1998), Banerjee et Gelfand (2002)), nous avons explicité le schéma particularisation/généralisation qui permet de passer de l'un à l'autre. En effectuant cette caractérisation, nous avons peu à peu construit une généalogie de ces modèles qui nous a permis d'identifier une classe qui les généralisait. Dans le cas gaussien, nous avons de plus positionné cette nouvelle classe de modèles par rapport aux modèles linéaires de corrégionalisation de la géostatistique multivariée. Il s'avère que nos modèles en sont des sous-modèles et nous avons montré qu'ils se caractérisaient par une propriété porteuse de sens quant à la nature des liaisons spatiales entre les différentes variables : la dépendance ponctuelle. Ces modèles, riches en termes de stuctures spatiales que l'on peut modéliser (contrairement à Banerjee et Gelfand (2002), nos modèles n'imposent pas une structure spatiale commune à toutes les variables), restent plus parcimonieux que les modèles LMC traditionnellement utilisés par les géostatisticiens, et certains de leurs paramètres portent sans doute plus de sens pour le statisticien classique ou l'utilisateur : il s'agit par exemple de paramètres de régression et de variance résiduelle.

La construction du schéma de liaisons entre ces modèles apparentés nous a également permis de proposer de nouveaux modèles pour lesquels aucune des variables observées n'est supposée gaussienne. Ces modèles pourraient être utiles en écologie des populations pour modéliser les interactions entre deux espèces à partir de comptages, que ces derniers soient observés aux mêmes sites ou non.

Concernant l'inférence des modèles présentés, nous avons choisi de nous placer essentiellement dans le cadre fréquentiste, suivant en cela des auteurs comme Mardia et Marshall (1984), Zhang (2002) ou encore Pascual et Zhang (2006). Dans ce cadre, les dépendances spatiales imposent d'estimer les paramètres en utilisant des algorithmes numériques, à défaut de pouvoir les calculer analytiquement. Nous avons exploité les propriétés spécifiques de convergence et de robustesse des algorithmes de Newton-Raphson et des scores de Fisher pour proposer un algorithme hybride qui s'est montré très efficace (bien qu'encore rudimentaire) pour associer la robustesse des scores de Fisher et la rapidité de l'algorithme de Newton-Raphson pour converger. Ce nouvel algorithme nous a permis d'explorer efficacement le comportement des estimateurs à taille finie par simulations (très

nombreuses), ce qui aurait été difficile et plus long avec les algorithmes classiques, l'information apportée par les données sur le(s) paramètre(s) de lien entre les deux variables pouvant être très faible, quand les données sont hétérotopes.

Dans le cas des modèles hiérarchiques, c'est-à-dire en présence de données qui ne peuvent être modélisées directement comme des variables gaussiennes, la vraisemblance s'écrit comme une intégrale de grande dimension et ne peut donc être utilisée sous sa forme analytique. Comme souvent dans le cas des GLMM non spatiaux, l'utilisation d'algorithmes stochastiques semblait s'imposer. Nous avons adapté au contexte spatial les outils d'estimation (essentiellement des algorithmes MCEM) développés par Booth et Hobert (1999) pour estimer les GLMM. L'utilisation de l'échantillonnage d'importance (*importance sampling*) au lieu de l'algorithme de Metropolis-Hasting utilisé par Zhang (2002) s'avère par exemple plus efficace pour mener à bien l'étape E ; la taille des échantillons simulés est mise à jour à chaque itération de l'EM afin d'optimiser le compromis temps de calcul/précision au fur et à mesure que l'on s'approche du maximum de vraisemblance. Le temps de calcul, bien que considérablement réduit par rapport aux approches plus anciennes, reste encore rédhibitoire pour envisager sérieusement de quantifier la variabilité des estimateurs par bootstrap paramétrique.

Nous nous sommes ensuite intéressés au comportement asymptotique des estimateurs dans le contexte gaussien. Après un rappel des deux grands types d'asymptotique utilisés dans le contexte spatial (par accroissement du domaine et par densification du domaine), nous nous sommes concentrés essentiellement sur l'asymptotique par accroissement du domaine. En nous basant sur les résultats de Sweeting (1980), nous avons précisé certains éléments des démonstrations effectuées par Mardia et Marshall (1984) ou encore Cressie et Lahiri (1996) pour la normalité asymptotique dans le cas d'un modèle de régression pour résidus spatialement corrélés ; nous avons notamment montré un de leur postulat et nous avons étendu aux plus grandes dimensions un résultat connu en dimension 1.

Puis dans une deuxième partie, nous nous sommes concentrés sur le problème de l'hétérotopie. Nous avons montré que sous certaines conditions asymptotiques (géométriques), l'estimateur du coefficient de corrélation dans un modèle spatial de corrélation intrinsèque est asymptotiquement normal. Nous avons explicité la variance de cet estimateur en fonction des caractéristiques spatiales de l'échantillon et de la structure de corrélation. Sous l'hypothèse que le coefficient de corrélation est nul, cette expression prend une forme particulièrement simple que l'on peut comparer à la variance asymptotique du coefficient de corrélation usuel calculé dans le cas classique (observations isotopes et i.i.d). Cette comparaison nous permet d'introduire pour notre problème, la notion de **nombre équivalent de couples isotopes et indépendants** ; c'est-à-dire qu'à partir d'un jeu de données hétérotopes, connaissant les structures spatiales des variables analysées et la géométrie de l'échantillonnage, nous sommes capables d'apprécier, au regard de la précision de l'estimateur du coefficient de corrélation, la qualité d'un jeu de données. Autrement dit, cette présentation des résultats nous permet d'appréhender la qualité géométrique de l'échantillonnage à partir de l'utilisation du coefficient de corrélation dans le contexte usuel. Quand les paramètres de structure spatiale sont inconnus, l'expression de la variance asymptotique de l'estimateur du coefficient de corrélation est beaucoup plus complexe que dans le cas où ces paramètres sont connus, sauf sous l'hypothèse où $r = 0$; dans ce cas, l'expression de la variance asymptotique est inchangée. Les résultats obtenus par simulation montrent que l'approximation asymptotique concernant le coefficient de corrélation peut être acceptée (au sens du test d'adéquation de Kolmogorov-Smirnov) dès lors que le nombre

équivalent est du même ordre de grandeur que le nombre de données i.i.d nécessaire pour que l'approximation asymptotique pour le coefficient de corrélation usuel le soit. Ainsi, pour un nombre de données équivalent supérieur à 20, il semble que l'on puisse utiliser l'approximation asymptotique sous l'hypothèse nulle de non corrélation. Pour des nombres équivalents plus petits, obtenir la distribution théorique de notre estimateur semble être un défi difficile car, dans le cas i.i.d, l'expression de la densité fait intervenir une dérivée d'ordre $n-1$, où n est la taille de l'échantillon. Or le nombre équivalent n'est en général pas un entier.

Pourquoi un chapitre sur le bayésien ?

Au regard du succès des méthodes bayésiennes pour effectuer l'inférence des modèles spatiaux, notamment hiérarchiques, nous n'avons pas fait l'économie d'explorer certains aspects de ce paradigme. Sans vouloir se placer dans le champ comparatif des deux paradigmes, notre expérience nous enseigne que :

- la mise en œuvre des algorithmes MCMC est plus aisée que celle des algorithmes de maximisation de la vraisemblance, tant du point de vue de l'écriture des algorithmes que de la programmation ;

- les algorithmes MCMC fournissent (approximativement) toute la distribution *a posteriori* des paramètres ; ce qui permet de discuter la qualité du résultat (relativement à la distribution *a priori*) et d'apprécier l'information apportée par les données, contrairement au cadre fréquentiste où l'estimation d'une valeur unique de chaque paramètre est donnée comme résultat ;

- dans le cas gaussien, le temps de calcul numérique nécessaire pour simuler un échantillon suffisamment grand dans la posterior (dans le cadre bayésien) est très largement supérieur à celui requis pour obtenir le maximum de vraisemblance (dans le cadre fréquentiste) ; or, disposant maintenant de résultats asymptotiques (théoriques) et à taille finie (empiriques), nous sommes à même, à partir de la valeur estimée du coefficient de corrélation et du nombre équivalent, d'apprécier l'information amenée par les données et ce faisant, la qualité de l'estimation ;

- néanmoins, dans le cas hiérarchique, en l'absence de résultats sur la variabilité des estimateurs, et compte tenu du temps de calcul nécessaire pour l'apprécier par simulations, l'approche bayésienne semble mieux à même de fournir des résultats propres à la discussion (posterior) ;

- cependant, dans l'état actuel de notre travail, sans ignorer que notre investigation de l'approche bayésienne n'a pas été aussi approfondie que l'approche fréquentiste, il nous semble que la convergence vers les objectifs (convergence des chaînes MCMC dans le cadre bayésien et convergence vers un maximum dans le cadre fréquentiste) est plus facilement garantie dans le cadre fréquentiste (voir chapitre 6).

A propos des cas d'étude.

L'étude sur le dépérissement de la vigne nous a conduit à étudier l'hétérotopie dans le cadre d'observations non gaussiennes (binaires). C'est aussi avec ce jeu de données que nous avons formulé les hypothèses nécessaires à la construction d'un modèle hiérarchique. Dans ce cas précis, le traitement de l'hétérotopie a été limité, du fait de la nature particulière des variables explicatives (des facteurs à modalité) qui rend la modélisation des structures spatiales difficile. C'est néamoins à partir de ce jeu de données que nous avons élaboré les questions autours de l'hétérotopie.

Une fois ces questions posées, le jeu de données sur l'histoire climatique et la biodiversité semblait propice à l'illustration des outils développés autours de l'hétérotopie (modèles, algorithmes et nombre équivalent). La mise en œuvre de nos méthodes apparaît comme une approche exploratoire efficace pour mettre en évidence la corrélation ponctuelle entre deux variables hétérotopes avant toute autre modélisation. Cependant, pour apprécier le potentiel de cette méthode, nous avons parfaitement

conscience qu'il faudrait l'utiliser sur de nombreux cas d'étude pour lesquels la dépendance ponctuelle est une hypothèse plus appropriée.

Perspectives.

Au terme de cette étude et de l'analyse du potentiel des méthodes développées, il est possible d'énoncer quelques pistes pour des développements ultérieurs.

Une première extension de ce travail pourraît être la généralisation à des modèles faisant intervenir plus de deux variables. Il s'agirait d'obtenir des sous modèles du LMC, plus parcimonieux que ce dernier, afin de mettre en place facilement des procédures automatiques pour estimer l'ensemble des paramètres. En effet la procédure classique pour estimer les LMC est d'abord de choisir les paramètres de corrélations spatiales (la portée par exemple) à partir des variogrammes expérimentaux, puis d'estimer les matrices de corrélations (voir Zhang, 2007). Avec des modèles plus parcimonieux, on peut imaginer que l'ensemble des paramètres soit estimable directement et automatiquement. La généralisation dans le cas hiérarchique suivrait immédiatement.

Dans le cadre de ce travail, nous avons cherché à adapter les outils d'estimation des GLMM les plus récents, ce qui nous a conduit à développer des algorithmes efficaces. Nul doute que dans ce domaine, verront le jour des développements qui, associés à l'amélioration de la puissance des calculateurs, permettront d'explorer le comportement des estimateurs dans le cadre des modèles hiérarchiques. En attendant, des développements asymptotiques dans le cadre de ces modèles, pourraient permettre de mieux appréhender l'information qu'amènent les données sur les champs latents et la variabilité des estimateurs de liaison entre variables qui en résulte. L'idée du nombre équivalent pourrait par exemple être étendue à ce problème.

A

La dépendance ponctuelle

Nous démontrons ici que si $\left\{W(s) = \begin{pmatrix} X(s) \\ Y(s) \end{pmatrix}, s \in \mathcal{D}\right\}$ est un champ aléatoire bivarié gaussien, stationnaire en espérance et en covariance, alors :

Y est autokrigeable par rapport à X si et seulement si :

$$\forall (s_1, s_2) \in \mathcal{D}^2 Y(s_1) \perp\!\!\!\perp X(s_2) | X(s_1). \tag{A.1}$$

On utilise la propriété suivante (voir par ex. Whittaker (1990)) :

Théorème 15 *Si* $(X_1, X_2, X_3) \sim \mathcal{N}\left(\boldsymbol{0}, \begin{pmatrix} 1 & \rho_{12} & \rho_{13} \\ \rho_{21} & 1 & \rho_{23} \\ \rho_{31} & \rho_{32} & 1 \end{pmatrix}\right)$,

alors

$$X_2 \perp\!\!\!\perp X_3 | X_1 \Leftrightarrow \rho_{23} - \rho_{21}\rho_{13} = 0$$

Plus généralement, si

$$(X_1, X_2, X_3) \sim \mathcal{N}\left(\begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho_{12} & \rho_{13} \\ \rho_{21} & \sigma_2^2 & \rho_{23} \\ \rho_{31} & \rho_{32} & \sigma_3^2 \end{pmatrix}\right),$$

on applique le théorème 15 au vecteur centré et réduit

$$\begin{pmatrix} \frac{X_1 - \mu_1}{\sigma_1} \\ \frac{X_2 - \mu_2}{\sigma_2} \\ \frac{X_3 - \mu_3}{\sigma_3} \end{pmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{pmatrix} 1 & r_{12} & r_{13} \\ r_{21} & 1 & r_{23} \\ r_{31} & r_{32} & 1 \end{pmatrix}\right)$$

où

$$r_{ij} = \frac{\rho_{ij}}{\sigma_i \sigma_j}, i, j = 1, \ldots, 3$$

et on obtient l'équivalence suivante :

$$\begin{aligned} X_2 \perp\!\!\!\perp X_3 | X_1 &\Leftrightarrow r_{23} - r_{21} r_{13} = 0 \\ &\Leftrightarrow \sigma_1^2 \rho_{23} - \rho_{21}\rho_{13} = 0. \end{aligned} \tag{A.2}$$

Partons de l'équation (A.1). Elle est équivalente par (A.2) à

$$\sigma_Y^2 \mathrm{Cov}(Y(s_1), X(s_2)) = \mathrm{Cov}(Y(s_1), X(s_1)) \mathrm{Cov}(X(s_1), X(s_2)).$$

D'autre part, d'après la stationnarité en covariance, il existe ρ_X et ρ_{XY} deux fonctions vérifiant :

$$\forall (s_1, s_2) \in \mathcal{D}^2, \mathrm{Cov}(X(s_1), X(s_2)) = \rho_X(\overrightarrow{s_1 s_2})$$
$$\mathrm{Cov}(Y(s_1), X(s_2)) = \rho_{XY}(\overrightarrow{s_1 s_2}).$$

On a donc indépendance entre (A.1) et la proposition suivante : $\exists \beta \in \mathbb{R}$ tel que

$$\forall (s_1, s_2) \in \mathcal{D}^2, \rho_{XY}(\overrightarrow{s_1 s_2}) = \beta \rho_X(\overrightarrow{s_1 s_2}),$$

ce qui achève la démonstration.

B

Démonstrations du chapitre 5

Cette annexe contient les démonstrations du chapitre 5 (Comportement des estimateurs).

A plusieurs reprises dans cette partie, nous utiliserons l'ordre des matrices (semi) définie-positive. Introduisons quelques notations et propriété :

Dans la suite, si A et B sont deux matrices carrées de même dimension, on notera $A \leq B$ (resp. $A < B$) si la matrice $B - A$ est semi définie-positive (resp. définie-positive).

On a :

Lemme 1 *(Horn et Johnson, 1985)*

Soeint A et B deux matrices carrées symétriques de taille n et T une matrice de dimension $n \times m$:

Si $A \leq B$, alors

$$T'AT \leq T'BT,$$

$$B^{-1} \leq A^{-1}$$

Si $\lambda_k(A)$ et $\lambda_k(B)$, $k = 1, \ldots, n$ sont les $k^{ème}$ valeurs propres respectives de A et B rangées dans l'ordre croissant, alors

$$\lambda_k(A) \leq \lambda_k(B).$$

Démonstration du théorème 4 :

$$\begin{aligned} E_\psi[||A_{\phi\phi}^{-1} L_{\phi\phi} A_{\phi\phi}^{-1} - I_q||^2] &= E_\psi[\mathrm{tr}(A_{\phi\phi}^{-2} L_{\phi\phi} A_{\phi\phi}^{-2} L_{\phi\phi} - 2A_{\phi\phi}^{-2} L_{\phi\phi} + I_q)] \\ &= E_\psi[\mathrm{tr}(A_{\phi\phi}^{-2} L_{\phi\phi} A_{\phi\phi}^{-2} L_{\phi\phi})] - \mathrm{tr}(I_q) \end{aligned}$$

car

$$E_\psi[A_{\phi\phi}^{-2} L_{\phi\phi}] = I_q. \tag{B.1}$$

En utilisant à nouveau (B.1), on obtient :

$$E_\psi[||A_{\phi\phi}^{-1} L_{\phi\phi} A_{\phi\phi}^{-1} - I_q||^2] = \sum_{i,j=1}^{q} \mathrm{cov}_\psi((A_{\phi\phi}^{-2} L_{\phi\phi})_{ij}, (A_{\phi\phi}^{-2} L_{\phi\phi})_{ji})$$

$$\mathrm{cov}_\psi((A_{\phi\phi}^{-2} L_{\phi\phi})_{ij}, (A_{\phi\phi}^{-2} L_{\phi\phi})_{ji}) = \sum_{r,s=1}^{q} b_{ir}^{\psi,\phi} b_{js}^{\psi,\phi} \mathrm{cov}_\psi((L_{\phi\phi})_{rj}, (L_{\phi\phi})_{si})$$

Par soucis de clarté, on pose $\Sigma(\phi)^\star = \Sigma^\star$ où $\star$ désigne les éventuels indices ou exposants.

Rappelons que

$$(L_{\phi\phi})_{ij} = -\frac{1}{2}\text{tr}(\Sigma^{-1}\ddot{\Sigma}_{ij} + \Sigma^{(i)}\Sigma_j) - \frac{1}{2}\mathbf{Z}'\Sigma^{(ij)}\mathbf{Z}$$

pour des matrices $\ddot{\Sigma}_{ij}, \Sigma^{(i)}, \Sigma_j$ et $\Sigma^{(ij)}$ définies au chapitre 3.

On a donc

$$\text{cov}_\psi((L_{\phi\phi})_{rj}, (L_{\phi\phi})_{si}) = \frac{1}{4}\text{cov}_\psi(\mathbf{Z}'\Sigma^{(rj)}\mathbf{Z}, \mathbf{Z}'\Sigma^{(si)}\mathbf{Z})$$

$$\text{cov}_\psi((L_{\phi\phi})_{rj}, (L_{\phi\phi})_{si}) = \frac{1}{4}\left(E_\psi[\mathbf{Z}'\Sigma^{(rj)}\mathbf{Z}\mathbf{Z}'\Sigma^{(si)}\mathbf{Z}] - E_\psi[\mathbf{Z}'\Sigma^{(rj)}\mathbf{Z}]E_\psi[\mathbf{Z}'\Sigma^{(si)}\mathbf{Z}']\right)$$

Notons v_{ab}^{ij} le $(a,b)^{\text{ème}}$ élément de $\Sigma^{(ij)}$ pour $i,j = 1,\ldots,q$ et $a,b = 1,\ldots,n$.

$$E_\psi[\mathbf{Z}'\Sigma^{(rj)}\mathbf{Z}\mathbf{Z}'\Sigma^{(si)}\mathbf{Z}] = \sum_{a,b,c,d=1}^{n} v_{ab}^{rj}v_{cd}^{si}E_\psi[z_a z_b z_c z_d] \tag{B.2}$$

$\mathbf{Z}$ étant un vecteur multigaussien centré et de matrice de covariance Σ, on a (voir par exemple Stein, 1999) :

$$E_\psi[z_a z_b z_c z_d] = \sigma_{ab}\sigma_{cd} + \sigma_{ac}\sigma_{bd} + \sigma_{ad}\sigma_{bc}$$

où σ_{ij} est le $(i,j)^{\text{ème}}$ terme de Σ.

En injectant cette formule dans l'équation (B.2), en notant que les matrices $\Sigma^{(ij)}, i,j = 1,\ldots,q$ et Σ sont symétriques et en faisant les arrangements adéquats, on déduit :

$$E_\psi[\mathbf{Z}'\Sigma^{(rj)}\mathbf{Z}\mathbf{Z}'\Sigma^{(si)}\mathbf{Z}] = 2\text{tr}(\Sigma\Sigma^{(rj)}\Sigma\Sigma^{(si)}) + \text{tr}(\Sigma\Sigma^{(rj)})\text{tr}(\Sigma\Sigma^{(si)}).$$

d'où

$$\text{cov}_\psi((L_{\phi\phi})_{rj}, (L_{\phi\phi})_{si}) = \frac{1}{2}\text{tr}(\Sigma\Sigma^{(rj)}\Sigma\Sigma^{(si)}).$$

L'équation (5.5) est donc équivalente à :

$$\sum_{r,s,i,j=1}^{q} b_{ir}^{\psi,\phi}b_{js}^{\psi,\phi}\text{tr}(\Sigma\Sigma^{(rj)}\Sigma\Sigma^{(si)}) \to_u 0 \tag{B.3}$$

Pour $E_\psi[||A_{\phi\phi}^{-1}L_{\phi\beta}A_{\beta\beta}^{-1}||^2]$, on raisonne de manière similaire.

On rappelle que le $(r,j)^{\text{ème}}$ terme de $L_{\phi\beta}$ est donné par $x_j'\Sigma^{(r)}X\beta + x_j'\Sigma^{(r)}\mathbf{W}$.

Ce qui conduit à :

$$\begin{aligned}
E_\psi[||A_{\phi\phi}^{-1}L_{\phi\beta}A_{\beta\beta}^{-1}||^2] &= E_\psi[\text{tr}(A_{\phi\phi}^{-2}L_{\phi\beta}A_{\beta\beta}^{-2}L_{\beta\phi})] \\
&= \sum_{i=1}^{q}\sum_{j=1}^{p}\text{cov}_\psi((A_{\phi\phi}^{-2}L_{\phi\beta})_{i,j}, (A_{\beta\beta}^{-2}L_{\beta\phi})_{j,i}) \\
&= \sum_{r,i=1}^{q}\sum_{s,j=1}^{p} b_{ir}^{\psi,\phi}b_{js}^{\psi,\beta}\text{cov}((L_{\phi\beta})_{r,j}, (L_{\beta\phi})_{s,i}) \\
&= \sum_{r,i=1}^{q}\sum_{s,j=1}^{p} b_{ir}^{\psi,\phi}b_{js}^{\psi,\beta}x_j'\Sigma^{(r)}\Sigma\Sigma^{(i)}x_s
\end{aligned}$$

En remplaçant $\Sigma^{(r)}$ et $\Sigma^{(j)}$ par leurs expressions, on obtient l'équivalence entre l'équation (5.4) et

$$\sum_{r,i=1}^{q}\sum_{s,j=1}^{p} b_{ir}^{\psi,\phi}b_{js}^{\psi,\beta}x_j'\Sigma^{-1}\Sigma_r\Sigma^{-1}\Sigma_i\Sigma^{-1}x_s \rightarrow_u 0 \tag{B.4}$$

Démonstration du théorème 5 :

Pour montrer (5.6), on utilise les lemmes suivants :

Lemme 2 *Si λ est la plus petite valeur propre d'une matrice carrée M, alors $M - \lambda I \geq 0$*

Démonstration. Notons $N = M - \lambda I$ et λ_N une valeur propre de N, alors

$$\det(N - \lambda_N I) = \det(M - (\lambda + \lambda_N)I) = 0$$

donc $\lambda_N + \lambda$ est une valeur propre de M supérieure à λ par définition. D'où $\lambda_N \geq 0$ et $M - \lambda I$ est définie positive.

Lemme 3 *Si A est une matrice symétrique, λ_1 sa plus petite valeur propre et λ_n la plus grande et B est une matrice semi définie-positive, alors $\lambda_1 tr(B) \leq tr(AB) \leq \lambda_n tr(B)$*

Démonstration. Voir Horn et Johnson (1985).

D'après le lemme 2, on a $\Sigma^{-1} - \lambda_n^{-1}I_n \geq 0$ donc $X'\Sigma^{-1}X \geq \lambda_n^{-1}X'X$ et par inversion, $(X'\Sigma^{-1}X)^{-1} \leq \lambda_n(X'X)^{-1}$. Par les hypothèses (ii) et (iv), $\lambda_n(X'X)^{-1}$ tend uniformément en ϕ vers la matrice nulle et donc $(X'\Sigma^{-1}X)^{-1}$ qui est semi définie-positive, également.

Pour montrer la deuxième partie de (5.6), notons c_{ij} le $(i,j)^{\text{ème}}$ terme G_n^{-1}, et $D = \text{diag}(t_{11}^{1/2}, \ldots, t_{qq}^{1/2})$. On a $A_{\phi\phi}^{-2} = 2D^{-1}G_n^{-1}D^{-1}$. D'où $b_{ij}^{\psi,\phi} = 2t_{ii}^{-1/2}t_{jj}^{-1/2}c_{ij}$.

En appliquant 2 fois le lemme 3, on a $t_{ii} = \text{tr}(\Sigma^{-1}\Sigma_i\Sigma^{-1}\Sigma_i) \geq \lambda_n^{-2}||\Sigma_i||^2$

d'où

$$|b_{ij}^{\psi,\phi}| \leq \frac{2\lambda_n^2}{||\Sigma_i||||\Sigma_j||}|c_{ij}| \tag{B.5}$$

et donc par les hypothèses (i),(ii) et (iii),

$$b_{ij}^{\psi,\phi} = o(n^{-1/2}), \tag{B.6}$$

ce qui conduit à $A_{\phi\phi}^{-1} \rightarrow_u \mathbf{0}$.

Pour montrer (B.3), on note :

$$\mathcal{S}_n^{\phi} = \sum_{r,s,i,j=1}^{q} b_{ir}^{\psi,\phi}b_{js}^{\psi,\phi}\text{tr}(\Sigma\Sigma^{(rj)}\Sigma\Sigma^{(si)})$$

Par l'inégalité de Cauchy-Schwartz pour le produit scalaire matriciel $< A, B >= \text{tr}(AB')$ on a :

$$|\mathcal{S}_n^{\phi}| \leq \sum_{r,s,i,j=1}^{q} |b_{ir}^{\psi,\phi}||b_{js}^{\psi,\phi}|||\Sigma\Sigma^{(rj)}||||\Sigma\Sigma^{(si)}||$$

Soit $||.||_s$ la norme spectrale définie par $||A||_s = \sqrt{\lambda_n(AA')}$ où $\lambda_n(M)$ désigne la plus grande valeur propre de M. La norme spectral est sous-multiplicative, c'est à dire que $||AB||_s \leq ||A||_s||B||_s$ pour toute matrices A et B (voir par exemple Horn and Johnson, 1985). Notons enfin que $||A|| \leq \sqrt{n}||A||_s$. En remplaçant $\Sigma^{(rj)}$ par son expression, il suit :

$$
\begin{aligned}
||\Sigma\Sigma^{(rj)}|| &\leq \sqrt{n}||\Sigma_r\Sigma^{-1}\Sigma_j\Sigma^{-1} + \Sigma_j\Sigma^{-1}\Sigma_r\Sigma^{-1} - \ddot{\Sigma}_{rj}\Sigma^{-1}||_s \\
&\leq 2\sqrt{n}||\Sigma_r\Sigma^{-1}\Sigma_j\Sigma^{-1}||_s + \sqrt{n}||\ddot{\Sigma}_{rj}\Sigma^{-1}||_s \\
&\leq \left(2\frac{\lambda_n^r\lambda_n^j}{\lambda_1^2} + \frac{\lambda_n^{rj}}{\lambda_1}\right)\sqrt{n}
\end{aligned}
$$

et donc par (i), (ii) et (B.6), $\mathcal{S}_n^\phi \rightarrow_u \mathbf{0}$.

Enfin, pour montrer (B.4), on note :

$$
\mathcal{S}_n^\beta = \sum_{r,i=1}^{q}\sum_{s,j=1}^{p} b_{ir}^{\psi,\phi}b_{js}^{\psi,\beta}x_j'\Sigma^{(r)}\Sigma\Sigma^{(i)}x_s
$$

En remplaçant $\Sigma^{(r)}$ et $\Sigma^{(i)}$ par leur expression, on obtient :

$$
\begin{aligned}
\mathcal{S}_n^\beta &= \sum_{r,i=1}^{q}\sum_{s,j=1}^{p} b_{ir}^{\psi,\phi}b_{js}^{\psi,\beta}x_j'\Sigma^{-1}\Sigma_r\Sigma^{-1}\Sigma_i\Sigma^{-1}x_s \\
&= \sum_{s,j=1}^{p} b_{js}^{\psi,\beta}x_j'\Sigma^{-1/2}H\Sigma^{-1/2}x_s
\end{aligned}
$$

avec

$$
H = \sum_{r,i=1}^{q} b_{ir}^{\psi,\phi}\Sigma^{-1/2}\Sigma_r\Sigma^{-1}\Sigma_i\Sigma^{-1/2}
$$

On rappelle que $b_{js}^{\psi,\beta}$ est le $(j,s)^{\text{ème}}$ terme de la matrice $-(X'\Sigma^{-1}X)^{-1}$, ce qui conduit à :

$$
\begin{aligned}
\mathcal{S}_n^\beta &= -\text{tr}((X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1/2}H\Sigma^{-1/2}X) \\
&= \text{tr}(MH)
\end{aligned}
$$

avec $M = \Sigma^{-1/2}X(X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1/2}$

On applique à nouveau l'inégalité de Cauchy-Schwartz et on note que $tr(MM') = p$ ce qui conduit à

$$
\begin{aligned}
|\mathcal{S}_n^\beta| &\leq \sqrt{p}||H|| \\
&\leq \sqrt{np}||H||_s \\
&\leq \sqrt{np}\sum_{r,i=1}^{q}|b_{ir}^{\psi,\phi}|||\Sigma^{-1/2}||_s^4||\Sigma_r||_s||\Sigma_i||_s \\
&\leq \sqrt{np}\sum_{r,i=1}^{q}|b_{ir}^{\psi,\phi}|\frac{\lambda_n^r\lambda_n^i}{\lambda_1^4}
\end{aligned}
$$

On conclut en utilisant les hypothèses (i) et (ii) et (B.6).

Démonstration du théorème 6 :

Lemme 4 *Soit A une matrice carrée de taille n, symétrique et de valeurs propres $\lambda_1 \leq \cdots \leq \lambda_n$. Pour tout $1 \leq p \leq n$, on note A_p n'importe quelle sous-matrice principale de A de taille p (obtenue à partir de A en supprimant $n-p$ lignes et les colonnes correspondantes), et $\lambda_1^p \leq \cdots \leq \lambda_p^p$ ses valeurs propres. Alors pour tout $1 \leq k \leq p$,*

$$
\lambda_k \leq \lambda_k^p \leq \lambda_{k+n-p}.
$$

Lemme 5 *Soient A une matrice $n \times n$ symétrique et inversible et B une matrice $n \times n$ telles que*

$$||A - B||_s ||A^{-1}||_s < 1,$$

alors B^{-1} est inversible et

$$||B^{-1}||_s \leq \frac{||A^{-1}||_s}{1 - ||I_n - A^{-1}B||_s}.$$

Démonstration. Une matrice B est inversible si et seulement si il existe une norme matricielle sous multiplicative telle que $||I - B|| < 1$ (voir par exemple Horn et Johnson, 1985). On écrit $B = A(I_n - A^{-1}(A - B))$. A étant inversible, B est inversible si et seulement si $I_n - A^{-1}(A - B)$ est inversible, c'est à dire si il existe une norme sous multiplicative telle que $||A^{-1}(A - B)|| < 1$. Or $||A^{-1}(A - B)||_s \leq ||A^{-1}||_s ||A - B||_s < 1$ par hypothèse, ce qui prouve que B est inversible. Puis il suffit de noter que

$$B^{-1} = A^{-1} + (I_n - A^{-1}B)B^{-1}$$

et d'appliquer la sous-multiplicativité de la norme spectrale et l'inégalité triangulaire.

Pour montrer le théorème, on va montrer que

- (5.12) $\Rightarrow$ (5.9)
- (5.13) $\Rightarrow$ (5.10)

Notons $\Lambda = \max\{|\lambda_n|, |\lambda_n^i|, |\lambda_n^{ij}|, i, j = 1, \ldots, q\}$.

Introduisons la norme matricielle du maximum des sommes en colonnes définie par :

$$||A||_c = \max_{1 \leq i \leq n} \sum_{j=1}^{n} |a_{ij}|.$$

Il s'agit d'une norme matricielle sous-multiplicative et par conséquent $||A||_s \leq ||A||_c$ pour toute matrice carrée, la norme spectrale étant la plus petite norme matricielle sous-multiplicative.

On a donc

$$\forall \sigma_{ij}^{(.)} \in \{\sigma_{ij}, \sigma_{ij}^k \text{ et } \sigma_{ij}^{kl}; k, l = 1, \ldots, q\}, \quad \Lambda \leq \max_{1 \leq i \leq n} \sum_{j=1}^{n} |\sigma_{ij}^{(.)}|.$$

Et donc (5.9) est vérifiée en utilisant (5.12).

Notons $\text{diag}((q_l)_{1 \leq l \leq n})$ la matrice dont le $(i,j)^{\text{ème}}$ terme est donné par $q_{ij} = q_i$ si $i = j$ et 0 sinon.

Posons $A = \text{diag}((\sigma_{ii})_{1 \leq i \leq n})$ et $B = \Sigma$.

On a

$$||A - B||_s \leq \max_i \sum_{\substack{j=1 \\ j \neq i}}^{n} |\sigma_{ij}| \leq \eta_1(K) \min_i \sigma_{ii}$$

par (5.13). En notant que

$$||A^{-1}||_s = \frac{1}{\min_i \sigma_{ii}},$$

on peut appliquer le résultat du lemme 5 ce qui conduit à :

$$||\Sigma^{-1}||_s \leq \frac{1}{\min_i \sigma_{ii}(1 - \max_i \sum_{\substack{j=1 \\ j \neq i}}^{n} |\frac{\sigma_{ij}}{\sigma_{ii}}|)}$$

d'où par (5.13)

$$\lambda_1 \geq \min_i \sigma_{ii}(1 - \eta_1(K)),$$

ce qui montre (5.10).

Démonstration du théorème 7

Nous allons montrer que les hypothèses du théorème 6 sont vérifiées.

Lemma B.1.

$$\sqrt{\sum_{i=1}^{d} u_i^2} \geq \frac{\sqrt{d}}{d} \sum_{i=1}^{d} |u_i|$$

Démonstration. On utilise la concavité de la fonction racine carrée.

Lemma B.2.

$$\sqrt{\sum_{i=1}^{d} u_i^2} \leq \sum_{i=1}^{d} |u_i|$$

Démonstration. Il suffit d'élever les deux termes au carré.

On note H_n la matrice des corrélations spatiales pour n observations, H_n^{ϕ} sa dérivée par rapport à ϕ.

De plus, si $\Sigma_n = \sigma_2 H(\phi)$, on note σ_{ij} ses éléments, $\sigma_{ij}^{\phi}, \sigma_{ij}^{\sigma^2}, \sigma_{ij}^{\phi\phi}, \sigma_{ij}^{\sigma^2\sigma^2}, \sigma_{ij}^{\sigma^2\phi}$ les éléments respectifs des matrices

$$\frac{\partial \Sigma_n}{\partial \phi}, \frac{\partial \Sigma_n}{\partial \sigma^2}, \frac{\partial^2 \Sigma_n}{\partial \phi^2}, \frac{\partial^2 \Sigma_n}{\partial {\sigma^2}^2}, \frac{\partial^2 \Sigma_n}{\partial \sigma^2 \partial \phi}.$$

$$\begin{aligned}
\forall n \in \mathbb{N}^{\star}, \forall j \in [\![1, n]\!], \sum_{i=1}^{n} |\sigma_{ij}| &\leq \sigma^2 \sum_{\mathbf{i} \in \mathbb{Z}^d} \exp(-\phi ||\mathbf{i} * h||) \\
&= \sigma^2 \sum_{i_1, \ldots, i_d \in \mathbb{Z}} \exp\left(-\phi \sqrt{\sum_{s=1}^{d} i_s^2 h_s^2}\right) \\
&\leq \sigma^2 \sum_{i_1, \ldots, i_d \in \mathbb{Z}} \exp\left(-\phi \frac{\sqrt{d}}{d} \sum_{s=1}^{d} |i_s| h_s\right) \\
&= \sigma^2 \sum_{i_1, \ldots, i_d \in \mathbb{Z}} \prod_{s=1}^{d} \exp\left(-\phi \frac{\sqrt{d}}{d} |i_s| h_s\right) \\
&= \sigma^2 \prod_{s=1}^{d} \left(\frac{1 + e^{-\phi \alpha_s}}{1 - e^{-\phi \alpha_s}}\right)
\end{aligned}$$

avec

$$\alpha_s = \frac{h_s \sqrt{d}}{d}.$$

$\sum_{i=1}^{n} |\sigma_{ij}|$ est donc bornée par une fonction des paramètres qui est continue. Donc pour tout compact K, il existe une constante $C(K)$ telle que

$$\sum_{i=1}^{n} |\sigma_{ij}| < C(K).$$

$$\begin{aligned}\forall n \in \mathbb{N}^\star, \forall j \in [\![1,n]\!], \sum_{i=1}^{n} |\sigma_{ij}^{\phi}| &\leq \sigma^2 \sum_{\mathbf{i}\in\mathbb{Z}^d} ||\mathbf{i}*\mathbf{h}|| \exp(-\phi||\mathbf{i}*\mathbf{h}||) \\ &\leq \sigma^2 \sum_{i_1,\dots,i_d\in\mathbb{Z}} \left(\sum_{s=1}^{d} |i_s|h_s\right) \exp\left(-\phi\frac{\sqrt{d}}{d}\sum_{s=1}^{d}|i_s|h_s\right) \\ &= -\sigma^2\sqrt{d}\frac{\partial}{\partial\phi}\left(\sum_{\mathbf{i}\in\mathbb{Z}^d}\exp\left(-\phi\frac{\sqrt{d}}{d}\sum_{s=1}^{d}|i_s|h_s\right)\right) \\ &= -\sigma^2\sqrt{d}\frac{\partial}{\partial\phi}\left(\prod_{s=1}^{d}\frac{1+e^{-\phi\alpha_s}}{1-e^{-\phi\alpha_s}}\right)\end{aligned}$$

Notons

$$p_s = e^{-\phi\alpha_s}.$$

On en déduit :

$$\forall n \in \mathbb{N}^\star, \forall j \in [\![1,n]\!], \sum_{i=1}^{n} |\sigma_{ij}^{\phi}| \leq 2\sigma^2 d \sum_{s=1}^{d}\left(h_s\frac{p_s}{(1-p_s)^2}\prod_{\substack{r=1\\r\neq s}}^{d}\left(\frac{1+p_r}{1-p_r}\right)\right)$$

On en déduit donc que la quantité $\sum_{i=1}^{n}|\sigma_{ij}^{\phi}|$ est bornée sur tout compact.
Pour $\sum_{i=1}^{n}|\sigma_{ij}^{\phi\phi}|$, notons

$$\tilde{h} = \min_{1\leq i\leq d} h_i,$$

$$\hat{h} = \max_{1\leq i\leq d} h_i,$$

et

$$\alpha = \frac{\sqrt{d}}{d}\tilde{h}.$$

On a

$$\begin{aligned}\forall n \in \mathbb{N}^\star, \forall j \in [\![1,n]\!], \sum_{i=1}^{n} |\sigma_{ij}^{\phi\phi}| &\leq \sigma^2 \sum_{\mathbf{i}\in\mathbb{Z}^d} ||\mathbf{i}*\mathbf{h}||^2 \exp(-\phi||\mathbf{i}*\mathbf{h}||) \\ &\leq \sigma^2\hat{h}^2 \sum_{i_1,\dots,i_d\in\mathbb{Z}}\left(\sum_{s=1}^{d} i_s^2\right)\prod_{s=1}^{d} e^{-\phi\alpha|i_s|} \\ &\leq \sigma^2\hat{h}^2 \sum_{i_1,\dots,i_d\in\mathbb{Z}}\left(\sum_{s=1}^{d} i_s^2 e^{-\phi\alpha|i_s|}\exp\left(-\phi\alpha\sum_{r\neq s}|i_r|\right)\right) \\ &\leq \sigma^2\hat{h}^2 \sum_{s=1}^{d}\sum_{i_s\in\mathbb{Z}}\left(i_s^2 e^{-\phi\alpha|i_s|}\sum_{i_1,\dots,i_{s-1},i_{s+1},\dots,i_d\in\mathbb{Z}}\exp\left(-\phi\alpha\sum_{r=1}^{d-1}|i_r|\right)\right) \\ &= \frac{1}{\alpha^2}\sigma^2\hat{h}^2\sum_{s=1}^{d}\sum_{i_s\in\mathbb{Z}}\left(\frac{\partial^2}{\partial\phi^2}e^{-\alpha\phi|i_s|}\right)\left(\frac{1+e^{-\alpha\phi}}{1-e^{-\alpha\phi}}\right)^{d-1} \\ &= 2\sigma^2\hat{h}^2 d\frac{e^{-\alpha\phi}}{(1-e^{-\alpha\phi})^2}\left(\frac{1+e^{-\alpha\phi}}{1-e^{-\alpha\phi}}\right)^{d}\end{aligned}$$

Il suit que $\sum_{i=1}^{n}|\sigma_{ij}^{\sigma^2}|$ et $\sum_{i=1}^{n}|\sigma_{ij}^{\phi\sigma^2}|$ sont également bornées par des fonctions continues en σ^2 et ϕ. Ainsi (5.12) est vérifiée.

Prouvons (5.13) :

$$\begin{aligned}\forall n \in \mathbb{N}^\star, \forall j \in [\![1,n]\!], \sum_{\substack{i=1\\ i\neq j}}^{n} |\sigma_{ij}| &\leq \sigma^2 \sum_{\mathbf{i}\in\mathbb{Z}^d\setminus\{\mathbf{0}\}} \exp(-\phi||\mathbf{i}*\mathbf{h}||)\\ &= \sigma^2\left(\frac{1+e^{-\alpha}}{1-e^{-\alpha}}-1\right)^d\\ &= \sigma^2\left(\frac{2e^{-\alpha}}{1-e^{-\alpha}}\right)^d\end{aligned}$$

Donc pour un compact K de Θ, il existe $\eta_1(K) \in]0,1[$ telle que

$$\forall n \in \mathbb{N}^\star, \forall j \in [\![1,n]\!], \sum_{\substack{i=1\\ i\neq j}}^{n} |\sigma_{ij}| \leq \eta_1(K)\sigma_{jj},$$

dès lors que

$$\tilde{h} > \frac{\sqrt{d}\log 3}{\phi_m}.$$

Montrer que l'hypothèse (iii) est vérifiée pour la matrice dérivée de Σ_n par rapport à σ^2 :
notons h_{ij} la distance séparant deux points de la grille indicés par i et j (dans $\{1,\dots,n\}$) ; on a

$$\mathrm{tr}({\Sigma_n}^{\sigma^2}{\Sigma_n}^{\sigma^2}) = \sum_{j=1}^{n}\sum_{i=1}^{n} e^{-2\phi h_{ij}}.$$

Or,

$$\sum_{j=1}^{n}\sum_{i=1}^{n} e^{-2\phi h_{ij}} \geq \sum_{i=1}^{n} e^{-2\phi\tilde{h}}$$

puisque quand n tend vers l'infini, la grille se remplit de manière homogène dans toutes les directions. On a donc

$$\mathrm{tr}({\Sigma_n}^{\sigma^2}{\Sigma_n}^{\sigma^2}) \geq n e^{-2\phi\tilde{h}}.$$

On montre de même que l'hypothèse (iii) est vérifiée pour la matrice dérivée de Σ_n par rapport à ϕ.

Pour que la démonstration soit complète, le théorème 6 requiert que $G_n(\phi)$ tende vers une matrice $G(\phi)$ qui n'est pas singulière.

Ici, on a

$$G_n(\phi) = \begin{pmatrix} 1 & a_n \\ a_n & 1 \end{pmatrix}$$

avec

$$a_n = \frac{\mathrm{tr}(H_n^{-1}H_n^{\phi})}{\sqrt{n}\sqrt{\mathrm{tr}(H_n^{-1}H_n^{\phi}H_n^{-1}H_n^{\phi})}}.$$

Or, le seul endroit où cette hypothèse a servi est l'équation B.5 dans la démonstration du théorème 5. La convergence de $G_n(\phi)$ vers une matrice $G(\phi)$ n'est donc pas nécessaire. Il suffit juste que les

éléments de la matrice $G_n^{-1}(\phi)$ soient bornés pour tout n. Dans le cas présent, il suffit qu'il existe $\delta \in]0,1[$ tel que $a_n^2 < 1 - \delta$[1].

Posons

$$b_n = n\mathrm{tr}(H_n^{-1} H_n^\phi H_n^{-1} H_n^\phi) - \mathrm{tr}(H_n^{-1} H_n^\phi)^2$$

Il suffit de montrer qu'il existe $\delta > 0$ tel que

$$\frac{b_n}{n\mathrm{tr}(H_n^{-1} H_n^\phi H_n^{-1} H_n^\phi)} > \delta$$

pour tout n.

Or $n\mathrm{tr}(H_n^{-1} H_n^\phi H_n^{-1} H_n^\phi) \leq n^2 C(K)$, donc il suffit de montrer qu'il existe $\delta > 0$ tel que

$$\frac{b_n}{n^2} > \delta.$$

Posons $c_n(\phi) = \mathrm{tr}(H_n^\phi H_n^{-1})/n$

$$\begin{aligned} ||H_n^\phi - c_n(\phi) H||^2 &= ||H_n^{1/2}(H_n^{-1/2} H_n^\phi H_n^{-1/2} - c_n(\phi) I_n) H^{1/2}||^2 \\ &\leq ||H_n||_s^2 ||H_n^{-1/2} H_n^\phi H_n^{-1/2} - c_n(\phi) I_n||^2 \end{aligned}$$

où la dernière inégalité s'obtient en appliquant 2 fois le lemme 3.

On remarque que

$$||H_n^{-1/2} H_n^\phi H_n^{-1/2} - c_n(\phi) I_n||^2 = \frac{b_n}{n}$$

et comme $||H_n||_s$ est borné, il suffit de montrer qu'il existe $\delta > 0$ tel que

$$\frac{||H_n^\phi - c_n(\phi) H||^2}{n} \geq \delta.$$

Considèrons

$$c_n^\star(\phi) = \mathrm{argmin}_{c \in \mathbb{R}} ||H_n^\phi - c H_n||^2$$

On a

$$c_n^\star(\phi) = \frac{\mathrm{tr}(H_n^\phi H_n)}{||H_n^2||^2}.$$

Pour tout n, on a donc

$$\begin{aligned} \frac{||H_n^\phi - c_n(\phi) H_n||^2}{n} &\geq \frac{||H_n^\phi - c_n^\star(\phi) H_n||^2}{n} \\ &= \frac{||H_n^\phi||^2 - c_n^\star(\phi)^2 ||H_n||^2}{n} \end{aligned}$$

où la dernière égalité vient du théorème de Pythagore.

Il ne reste donc plus qu'à montrer que

$$\frac{||H_n^\phi||^2 ||H_n||^2 - \mathrm{tr}(H_n^\phi H_n)^2}{n||H_n||^2} > \delta > 0$$

[1] Dans le cas général où la dimension de ϕ est $p \geq 2$, il suffit de montrer que la plus petite valeur propre de $G_n(\phi)$ est minorée pour tout n par une constante $\delta > 0$, ce qui assurera que les éléments de $G_n(\phi)^{-1}$ soient majorés en valeurs absolues. Mais comme la trace de $G_n(\phi)$, qui est égale à la somme de ses valeurs propres, est constante égale à p, il est équivalent de montrer que le déterminant de $G_n(\phi)$ (produit des valeurs propres) est minorée par une constante $\delta > 0$, ce qui est en général beaucoup plus aisé.

où encore

$$\frac{||H_n^\phi||^2||H_n||^2 - \mathrm{tr}(H_n^\phi H_n)^2}{n^2} > \delta > 0$$

$$\begin{aligned}
||H_n^\phi||^2||H_n||^2 - \mathrm{tr}(H_n^\phi H_n)^2 &= \mathrm{tr}({H_n^\phi}^2)\mathrm{tr}(H_n^2) - \mathrm{tr}(H_n^\phi H_n)^2 \\
&= \left(\sum_{i=1}^n\sum_{j=1}^n {\sigma_{ij}^\phi}^2\right)\left(\sum_{k=1}^n\sum_{l=1}^n \sigma_{kl}^2\right) - \left(\sum_{i=1}^n\sum_{j=1}^n \sigma_{ij}^\phi\sigma_{ji}\right)^2 \\
&= \sum_{i=1}^n\sum_{j=1}^n\sum_{k=1}^n\sum_{l=1}^n {\sigma_{ij}^\phi}^2\sigma_{kl}^2 - \sum_{i=1}^n\sum_{k=1}^n\left(\sum_{j=1}^n \sigma_{ij}^\phi\sigma_{ji}\right)\left(\sum_{l=1}^n \sigma_{kl}^\phi\sigma_{kl}\right) \\
&= \sum_{i=1}^n\sum_{j=1}^n\sum_{k=1}^n\sum_{l=1}^n {\sigma_{ij}^\phi}^2\sigma_{kl}^2 - \sum_{i=1}^n\sum_{k=1}^n\sum_{j=1}^n\sum_{l=1}^n \sigma_{ij}^\phi\sigma_{ji}\sigma_{kl}^\phi\sigma_{kl} \\
&= \sum_{i=1}^n\sum_{j=1}^n\sum_{k=1}^n\sum_{l=1}^n (h_{ij}^2 - h_{ij}h_{kl})e^{-2\phi(h_{kl}+h_{ij})} \\
&= \sum_{\substack{i,j,k,l=1,\dots,n \\ h_{kl}<h_{ij}}} (h_{ij}-h_{kl})^2 e^{-2\phi(h_{kl}+h_{ij})}
\end{aligned}$$

Soient $0 < \lambda_1 < \lambda_2 < \lambda_3$, trois constantes telles qu'il existe $n_0 \in \mathbb{N}$ tel que, quel que soit $n > n_0$:

$$\forall (i,k) \in \{1,\dots,n\}^2, \exists (i,k) \in \{1,\dots,n\}^2, h_{kl} < \lambda_1 < \lambda_2 < h_{ij} < \lambda_3.$$

On a

$$\begin{aligned}
\sum_{\substack{i,j,k,l=1,\dots,n \\ h_{kl}<h_{ij}}} (h_{ij}-h_{kl})^2 e^{-2\phi(h_{kl}+h_{ij})} &\geq \sum_{\substack{i,j,k,l=1,\dots,n \\ h_{kl}<\lambda_1<\lambda_2<h_{ij}<\lambda_3}} (h_{ij}-h_{kl})^2 e^{-2\phi(h_{kl}+h_{ij})} \\
&\geq n^2(\lambda_2-\lambda_1)^2 e^{-2\phi\lambda_3},
\end{aligned}$$

ce qui achève la démonstration.

Démonstration du théorème 12
Introduisons un lemme :

Lemme 6 *Soient M une matrice symétrique et N une matrice définie-positive, toutes deux de dimension n. On note $\lambda_1 \leq \cdots \leq \lambda_n$ et $\mu_1 \leq \cdots \leq \mu_n$ les valeurs propres respectives de M et de $M+N$, alors*

$$\forall k \in \{1,\dots,n\}, \quad \lambda_k \leq \mu_k.$$

Démonstration. Voir Magnus et Neudecker (1999)

Lemme 7 *Soient A et B deux matrices carrées de tailles respectives n et p dont les valeurs propres respectives sont données par $\lambda_1,\dots,\lambda_n$ et $\mu_1,\dots,\mu_p$, alors les np valeurs propres de la matrice $A \otimes B$ sont données par $\{\lambda_i\mu_j, i = 1\dots,n, j = 1,\dots,p\}$.*

Conséquence : le produit de kronecker de deux matrices est défini-positif si et seulement si elles sont toutes deux définie-positives ou définie-négatives.

Démonstration. Voir par exemple Magnus et Neudecker (1999)

La plus grande valeur propre de $\Sigma = T \otimes H(\phi)$ est égale au produit des plus grandes valeurs propres respectives de T et $H(\phi)$ d'après le lemme 7. La plus grande valeur propre de H est bornée sur tout compact par hypothèse et celle de T également car la fonction qui à toute matrice carrée de dimension n associe sa $k^{\text{ème}}$ valeur propre est une fonction continue sur l'espace des matrices de dimension n (voir par exemple Horn et Johnson, 1985). On montre de même que la plus petite valeur propre de Σ est minorée sur tout compact par une constante strictement positive.

Montrons que les valeurs propres des dérivées premières et secondes en les paramètres de la matrice de covariance sont bornées.

Pour les dérivées premières en fonction des paramètres σ_X^2 et σ_Y^2, on a :

$$\frac{\partial \Sigma}{\partial \sigma_X^2} = \begin{pmatrix} H(\phi) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \quad \text{et} \quad \frac{\partial \Sigma}{\partial \sigma_Y^2} = \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & H(\phi) \end{pmatrix}.$$

et le résultat est une conséquence directe des hypothèses.

Pour

$$\frac{\partial \Sigma}{\partial \sigma_{XY}} = \begin{pmatrix} \mathbf{0} & H(\phi) \\ H(\phi) & \mathbf{0} \end{pmatrix},$$

on applique le lemme 6 à $M = \frac{\partial \Sigma}{\partial \sigma_{XY}}$ et $N = \begin{pmatrix} H(\phi) & \mathbf{0} \\ \mathbf{0} & H(\phi) \end{pmatrix}$ en notant que

$$M + N = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \otimes H(\phi)$$

a comme plus grande valeur propre $2\lambda_n H(\phi)$, bornée sur tout compact, par hypothèse.

Pour

$$\Sigma_i = \frac{\partial \Sigma}{\partial \phi_i} = T \otimes H_i(\phi),$$

on applique le lemme 1 en posant :

$$M = \lambda I_2 \otimes H_i^2(\phi) \quad \text{et } N = T^2 \otimes H_i^2(\phi).$$

où

$$\lambda = \frac{1}{2}\left(\sqrt{(\sigma_X^4 - \sigma_Y^4)^2 + 4\sigma_{XY}^2(\sigma_X^2 + \sigma_Y^2)^2} + (\sigma_X^4 + \sigma_Y^4 + 2\sigma_{XY}^2)\right)$$

est la plus grande valeur propre de T^2. $H_i^2(\phi)$ étant définie positive et $\lambda I_2 - T^2$ également par construction et d'après le lemme 2, donc $M - N$ est définie positive d'après le lemme 7. On en déduit que la plus grande valeur propre de Σ_i est bornée sur tout compact quand n tend vers l'infini par le lemme 1 et les hypothèses.

Pour les dérivées secondes, on procède de manière similaire.

L'hypothèse (5.11) du théorème 5 se vérifie directement en calculant la trace du carré des matrices des dérivées et en appliquant l'hypothèse :

pour tout compact K tel que $K^\phi \subset \Theta^\phi$,

$$\operatorname{tr}(H(\phi)H(\phi))^{-1} \leq r_n(K),$$

où $r_n(K) = o(n^{-1/2})$.

Montrons que $G_n(\phi)$ tend vers une matrice $G(\phi)$ non singulière.

En utilisant le fait que $\mathrm{tr}(A \otimes B) = \mathrm{tr}(A)\mathrm{tr}(B)$ par le lemme 7, on obtient matrice d'information de Fisher :

$$E[\mathcal{F}_n] = \frac{1}{2\Delta^2} \begin{pmatrix} n\sigma_Y^4 & -2n\sigma_Y^2\sigma_{XY} & n\sigma_{XY}^2 & \sigma_Y^2\Delta\mathrm{tr}(H^{-1}H_i) \\ -2n\sigma_Y^2\sigma_{XY} & 2n(\sigma_{XY}^2 + \sigma_X^2\sigma_Y^2) & -2n\sigma_X^2\sigma_{XY} & -2\sigma_{XY}\Delta\mathrm{tr}(H^{-1}H_i) \\ n\sigma_{XY}^2 & -2n\sigma_X^2\sigma_{XY} & n\sigma_X^4 & \sigma_X^2\Delta\mathrm{tr}(H^{-1}H_i) \\ \sigma_Y^2\Delta\mathrm{tr}(H^{-1}H_i) & -2\sigma_{XY}\Delta\mathrm{tr}(H^{-1}H_i) & \sigma_X^2\Delta\mathrm{tr}(H^{-1}H_i) & 2\Delta^2\mathrm{tr}(H^{-1}H_iH^{-1}H_i) \end{pmatrix}.$$

où $\Delta = \det(T) = \sigma_X^2\sigma_Y^2 - \sigma_{XY}^2$.

On normalise la matrice $E[\mathcal{F}_n]$ pour obtenir $G_n(\phi)$ et on calcule le déterminant avec Maple. On obtient après simplification :

$$\det(G_n(\phi)) = \frac{(a_n^2 - 1)(r^2 - 1)^3}{r^2 + 1},$$

où

$$a_n = \frac{\mathrm{tr}(H^{-1}H_i)}{\sqrt{n\mathrm{tr}(H^{-1}H_iH^{-1}H_i)}}.$$

Or par hypothèse, il existe $\delta < 1$ telle que $a_n^2 < \delta$, donc pour tout compact $K \subset \Theta$, $\det(G_n(\phi))$ est minorée par une constante strictement positive, ce qui achève la démonstration.

Démonstration du théorème 13 :

Posons :

$$M = 2\begin{pmatrix} \sigma_X^2 H_{XX} & \mathbf{0} \\ \mathbf{0} & \sigma_Y^2 H_{YY} \end{pmatrix} \quad \text{et } N = \Sigma.$$

M et N sont définie-positives. De plus,

$$M - N = \begin{pmatrix} \sigma_X^2 H_{XX} & -\sigma_{XY}H_{XY} \\ -\sigma_{XY}H_{YX} & \sigma_Y^2 H_{YY} \end{pmatrix} \quad \text{est définie-positive}$$

On peut donc appliquer le lemme 1. On déduit que

$$\lambda_n(\Sigma) \leq 2\max\{\lambda_n(\sigma_X^2 H_{XX}), \lambda_n(\sigma_Y^2 H_{YY})\}, \tag{B.7}$$

et donc $\lambda_n(\Sigma)$ est bornée sur tout compact par hypothèse.

Montrons que la plus petite valeur propre de Σ est minorée pour tout compact $K \subset \Theta$:

Pour tout compact $K \subset \Theta$, il existe une constante $\delta(K) \in]0,1[$ telle que $\forall r \in K^r$,

$$r^2 \leq 1 - \delta(K),$$

où K^r est le compact projeté de K dans lequel varie r.

Posons $M = \Sigma$ et $N = \delta(K)\begin{pmatrix} \sigma_X^2 H_{XX} & \mathbf{0} \\ \mathbf{0} & \sigma_Y^2 H_{YY} \end{pmatrix}$.

$M - N$ est définie-positive ; donc par le lemme 1, on a :

$$\lambda_1(\Sigma) \geq \delta(K)\min\{\lambda_1(\sigma_X^2 H_{XX}), \lambda_1(\sigma_Y^2 H_{YY})\}. \tag{B.8}$$

et donc $\lambda_1(\Sigma)$ est minoré par une constante strictement positive sur tout compact.

Pour les autres hypothèses, on procède comme dans la démonstration du théorème 12.

Calcul de l'information de Fisher :

On ne s'intéresse qu'à la sous-matrice de la matrice d'information de Fisher correspondant aux paramètres de la matrice de variance-covariance et on ne s'occupe pas des paramètres d'espérance puisque la matrice d'information de Fisher totale est bloc-diagonale.

On rappelle que le $(i,j)^{\text{ème}}$ élément de cette sous-matrice s'écrit :

$$\frac{1}{2}\text{tr}(\Sigma^{-1}\Sigma_i\Sigma^{-1}\Sigma_j)$$

où Σ_i et Σ_j sont les matrices dérivées de Σ respectivement par rapport aux $i^{\text{ème}}$ et $j^{\text{ème}}$ paramètres.

En utilisant la formule d'inversion matricielle par bloc suivante,

$$\begin{pmatrix} A & B \\ B' & C \end{pmatrix} = \begin{pmatrix} E^{-1} & -A^{-1}BF^{-1} \\ -C^{-1}B'E^{-1} & F^{-1} \end{pmatrix},$$

avec $E = A - BC^{-1}B'$ et$F = C - B'A^{-1}B$, on obtient

$$\Sigma^{-1} = \begin{pmatrix} \frac{1}{\sigma_X^2}Q_X H_{XX}^{-1} & -\frac{r}{\sigma_X\sigma_Y}H_{XX}^{-1}H_{XY}Q_Y H_{YY}^{-1} \\ -\frac{r}{\sigma_X\sigma_Y}H_{YY}^{-1}H_{YX}Q_X H_{XX}^{-1} & \frac{1}{\sigma_Y^2}Q_Y H_{YY}^{-1} \end{pmatrix}$$

avec

$$Q_X = [I_{n_X} - r^2 M_X]^{-1}, \quad Q_Y = [I_{n_Y} - r^2 M_Y]^{-1},$$

$$M_X = H_{XX}^{-1}H_{XY}H_{YY}^{-1}H_{YX}, \quad \text{et} \quad M_Y = H_{YY}^{-1}H_{YX}H_{XX}^{-1}H_{XY}.$$

D'autre part

$$\frac{\partial\Sigma}{\partial\sigma_X} = \begin{pmatrix} 2\sigma_X H_{XX} & r\sigma_Y H_{XY} \\ r\sigma_Y H_{YX} & \mathbf{0} \end{pmatrix},$$

$$\frac{\partial\Sigma}{\partial\sigma_Y} = \begin{pmatrix} \mathbf{0} & r\sigma_Y H_{YY} \\ r\sigma_Y H_{YX} & 2\sigma_Y^2 H_{XY} \end{pmatrix},$$

et

$$\frac{\partial\Sigma}{\partial r} = \begin{pmatrix} \mathbf{0} & \sigma_X\sigma_Y H_{XY} \\ \sigma_X\sigma_Y H_{YX} & \mathbf{0} \end{pmatrix}.$$

En utilisant le fait que $Q_X H_{XX}^{-1} = H_{XX}^{-1}Q_X'$ et $Q_Y H_{YY}^{-1} = H_{YY}^{-1}Q_Y'$ (par simple calcul) et le fait que $H_{XY}Q_Y = Q_X' H_{XY}$ (par la symètrie de Σ^{-1}), on obtient la matrice d'information de Fisher associée au vecteur $(\sigma_X^2, r, \sigma_Y^2)$:

$$E[\mathcal{F}_n] = \begin{pmatrix} \frac{2n_X + r^2A}{\sigma_X^2} & -\frac{rA}{\sigma_X} & -\frac{r^2A}{\sigma_X\sigma_Y} \\ -\frac{rA}{\sigma_X} & 2B - A & -\frac{rA}{\sigma_Y} \\ -\frac{r^2A}{\sigma_X\sigma_Y} & -\frac{rA}{\sigma_Y} & \frac{2n_Y + r^2A}{\sigma_Y^2} \end{pmatrix}.$$

En utilisant le logiciel Maple, on inverse cette matrice et on effectue les simplifications qui amènent à la formule annoncée.

Démonstration du théorème 14 :

La plupart des hypothèses à montrer l'ont été dans la démonstration du théorème 13.

Pour les majorations spécifiques à ce théorème, notons :

$$h_{ij}^{XX} = ||s_{x_i} - s_{x_j}||, x_i, x_j = 1, \ldots, n_X;$$

$$h_{ij}^{YY} = ||s_{y_i} - s_{y_j}||, y_i, y_j = 1, \ldots, n_Y;$$

$$h_{ij}^{XY} = ||s_{x_i} - s_{y_j}||, x_i = 1, \ldots, n_X; y_j = 1, \ldots, n_Y.$$

Il faut par exemple montrer que pour tout compact $K \subset \Theta$, il existe une constante $C(K) > 0$ telle que :

$$\max_{j=1,\ldots,n} \sum_{i=1}^{n} |\sigma_{ij}^{\phi}| < C(K).$$

Par les hypothèses (i) et (ii), il suffit donc de montrer que pour tout compact $K \subset \Theta$, il existe une constante $C_1(K) > 0$ telles que

$$\max_{j=1,\ldots,n} \sum_{i=1}^{n} \left| \frac{\partial \rho_{\phi}(h_{ij}^{XY})}{\partial \phi} \right| < C1(K),$$

et une constante $C_2(K) > 0$ telle que

$$\max_{j=1,\ldots,n} \sum_{i=1}^{n} \left| \frac{\partial \rho_{\phi}(h_{ji}^{XY})}{\partial \phi} \right| < C2(K).$$

Notons f_{ϕ} la fonction de h définie par

$$f_{\phi}(h) = \left| \frac{\partial \rho_{\phi}(h)}{\partial \phi} \right|.$$

$$\sum_{i=1}^{n} f_{\phi}(h_{ij}^{XY}) = \sum_{\substack{i=1,\ldots,n \\ h_{ij}^{XY} \leq h_0}} f_{\phi}(h_{ij}^{XY}) + \sum_{\substack{i=1,\ldots,n \\ h_{ij}^{XY} > h_0}} f(h_{ij}^{XY}).$$

Il existe un entier N tel que

$$\forall n \in \mathbb{N}, \forall j \in \{1, \ldots, n\}, \sharp\{i, h_{ij}^{XY} < h_0\} < N.,$$

où $\sharp(E)$ désigne le cardinal de l'ensemble E.

En effet, sinon $\sum_{i=1}^{n} h_{ij}^{XX}$ ne pourrait pas être bornée pour tout n et pour tout $j \in \{1, \ldots, n\}$.

Donc

$$\sum_{\substack{i=1,\ldots,n \\ h_{ij} \leq h_0}} f_{\phi}(h_{ij}^{XY})$$

est bornée (car f_{ϕ} est continue par hypothèse).

Considérons

$$i_0 = \text{argmin}_{\tilde{i}=1,\ldots,n} h_{\tilde{i}j}^{XY}.$$

Par l'inégalité triangulaire, on a

$$h_{i_0 i}^{XX} \leq h_{i_0 j}^{XY} + h_{ij}^{XY}. \tag{B.9}$$

$$\sum_{\substack{i=1,\dots,n\\ h_{ij}>h_0}} f(h_{ij}^{XY}) = \sum_{\substack{i=1,\dots,n\\ h_{ij}^{XY}>h_0\\ h_{i_0 i}^{XX}-R\leq h_0}} f(h_{ij}^{XY}) + \sum_{\substack{i=1,\dots,n\\ h_{ij}^{XY}>h_0\\ h_{i_0 i}^{XX}-R>h_0}} f(h_{ij}^{XY})$$

Pour la même raison que précédemment, le cardinal de l'ensemble

$$\{i \in \{1,\dots,n\}, h_{i_0 i}^{XX} - R \leq h_0\}$$

est borné pour tout n et tout $i_0 \in \{1,\dots,n\}$. Il ne reste donc plus qu'à majorer

$$\sum_{\substack{i=1,\dots,n\\ h_{ij}^{XY}>h_0\\ h_{i_0 i}^{XX}-R>h_0}} f(h_{ij}^{XY})$$

Par hypothèse et par (B.9), on a

$$f_\phi(h_{ij}^{XY}) \leq f_\phi(h_{i_0 i}^{XX} - R)$$

Or,

$$\sum_{i=1}^{n} f_\phi(h_{i_0 i}^{XX})$$

est bornée par hypothèse, donc $\sum_{i=1}^{n} f_\phi(h_{i_0 i}^{XX} - R)$ l'est également (car f_ϕ est décroissante sur $]h_0, +\infty[$) et donc le résultat suit.

L'hypothèse (iii) du théorème 5 se montre comme elle se montrait dans le cas du théorème 7 en utilisant la condition de proximité.

C

Les distributions

Nous présentons ici quelques distributions utilisées dans la thèse :

C.1 Distribution de Wishart

La distribution de Wishart généralise la distribution de Chi-2 au cas multivariable :

si $X_1, \ldots, X_n$ sont des vecteurs de taille p aléatoires et indépendants de distribution gaussienne d'espérance μ et de matrice de variance Σ, alors

$$W = \sum_{i=1}^{n} X_i X_i'$$

est une matrice aléatoire (définie-positive).

Par définition, W est distribué selon une loi de Wishart de paramètres μ, Σ, n et p que nous notons $\mathcal{W}_p(\mu, \Sigma, n)$.

C.2 Distribution inverse Wishart

Par définition, W est distribué selon une loi inverse Wishart de paramètres μ, Σ, n et p que nous notons

$$\mathcal{IW}_p(\mu, \Sigma, n),$$

si W^{-1} est distribuée selon une loi de Wishart de paramètres μ Σ^{-1}, $n + p - 1$ et p

$$\mathcal{W}_p(\mu, \Sigma, n + p - 1).$$

C.3 Distribution de Student multivariée

Soit $\mathbf{y}$ un vecteur multigaussien de dimension p, centré et de matrice de variance Σ et u une variable de Chi-2 à n degrés de liberté ; alors, la variable

$$\mathbf{z} = \frac{\mathbf{y}}{\sqrt{n/u}} + \mu$$

est distribuée selon une loi de Student multivariée de paramètres Σ, μ et n (on note $\mathbf{z} \sim \mathcal{T}(\mu, \Sigma, n)$).

La densité de $\mathbf{z}$, s'écrit :

$$f(z) = \frac{\Gamma[(n+p)/2]}{\Gamma(n/2) n^{p/2} \pi^{p/2} |\Sigma|^{1/2} \left[1 + \frac{1}{n}(x-\mu)' \Sigma^{-1} (x-\mu)\right]^{(n+p)/2}}.$$

C.4 Distribution inverse Gamma

On dira qu'une variable aléatoire X est distribuée selon une loi inverse Gamma de paramètre de forme $\alpha > 0$ et de paramètre d'échelle $\beta > 0$, notée $\mathcal{IG}(\alpha, \beta)$ si $1/X$ est une variable aléatoire distribuée selon une loi Gamma de paramètres α et $1/\beta$.

Dans ce cas, on a

$$E[X] = \frac{\beta}{\alpha - 1} \text{ pour } \alpha > 1,$$

et

$$\text{Var}(X) = \frac{\beta^2}{(\alpha-1)^2(\alpha-2)}, \text{ pour } \alpha > 2.$$

Enfin, la densité de X s'écrit :

$$f(x; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{-\alpha-1} \exp\left(\frac{-\beta}{x}\right).$$

Références

Adler, R. (1981). *Geometry of Random Fields.* New-York : John Wiley & Sons.

Banerjee, S., B. P. Carlin, et A. E. Gelfand (2004). *Hierarchical Modeling and Analysis for Spatial Data.* Boca Raton : Chapman-Hall CRC Press.

Banerjee, S. et A. E. Gelfand (2002). Prediction, interpolation and regression for spatially misaligned data. *Sankhya : The Indian Journal of Statistics 64*, 227–245.

Banerjee, S., A. E. Gelfand, et W. Polasek (2000). Geostatistical modelling for spatial interaction data with application to postal service performance. *Journal of Statistical Planning and Inference 90*, 87–105.

Booth, J. G. et J. P. Hobert (1999). Maximizing generalized linear mixed model likelihoods with an automated monte carlo em algorithm. *Journal of the Royal Statistics Society, B 61*, 265–285.

Booth, J. G. et Hobert, J. P. (1998). Standards errors of prediction in generalized linear mixed models. *Journal of the American Statistical Association 93*, 262–272.

Breslow, N. E. et D. G. Clayton (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Associaton 88*, 9–25.

Breslow, N. E. et X. Lin (1995). Biais correction in generalized linear mixed models with a single component of dispersion. *Biometrika 82*, 81–91.

Chilès, J.-P. et P. Delfiner (1999). *Geostatistics : modeling spatial uncertainty.* New York : Wiley.

Christakos, G. (1992). *Random Field Models in Earth Sciences.* San Diego : Academcic Press, Inc.

Cressie, N. (1993). *Statistics for Spatial Data.* New York : Wiley.

Cressie, N. et S. N. Lahiri (1993). The asymptotic distribution of reml estimators. *Journal of Multivariate Analysis 45*, 217–233.

Cressie, N. et S. N. Lahiri (1996). Asymptotics for reml estimation of spatial covariance parameters. *Journal of Statistical Planning and Inference 50.*

Dean, C. B., M. D. Ugarte, et A. F. Militino (2004). Penalized quasi-likelihood with spatially correlated data. *Computational Statistics & Data Analysis 45*, 235–248.

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

Desassis, N., P. Monestiez, J. N. Bacro, P. Lagacherie, et J. M. Robbez-Masson (2005). Mapping unobserved factors on vine plant mortality. In P. Renard (Ed.), *Geostatistics for Environmental Applications*, GEOENV, pp. 125–136. Springer.

Diggle, P. J., J. A. Tawn, et R. A. Moyeed (1998). Model-based geostatistics. *Applied statistics 47*(3), 299–350.

Efron, B. (2005). Bayesians, frequentists, and scientists. *Journal of the American Statistical Association 100*, 1–5.

Evans, M. et T. Swartz (1998). *Bayesian Integration Using Multivariate Student Importance Sampling* (27 ed.)., Chapter Computing Science and Statistics, pp. 456–461. Fairfax Station : Interface Foundation of North America.

Gabriel, E. (2004). *Détection de zones de changement abrupt dans des données spatiales et applications à l'agriculture de précision*. Thèse de Doctorat, Université Montpellier II.

Goulard, M. et M. Voltz (1992). Linear coregionalization model : tools for estimation and choice of cross-variogram matrix. *Mathematical Geology 24*, 269–286.

Hart, J. (1953). Central tendency in areal distributions. *Economic Geography 30*, 48–59.

Horn, R. A. et C. R. Johnson (1985). *Matrix Analysis*. Cambridge : Cambridge University Press.

Hrafnkelsson, B. et N. Cressie (2003). Hierarchical modeling of count data with applications to nuclear fall-out. *Environmental and Ecological Statistics 10*, 179–200.

Ibragimov, I. A. et Y. A. Rozanov (1978). *Gaussian Random Processes*. New York : Springer-Verlag.

Journel, A. G. et C. J. Huigbregts (1978). *Mining Geostatistics*. New York : Academic Press.

Kaiser, M. S. et N. Cressie (1997). Modeling poisson variables with positive spatial dependence. *Statistics & Probability letters 35*, 423–432.

Klein, E. K., N. Desassis, et S. Oddou-Muratorio (2007). Pollen flow in the wildservice tree, *sorbus torminalis* (l.) crantz. iv. full inter-individual variance of male fertility estimated jointly with dispersal kernel. *Soumis.*.

Knight, K. (2000). *Mathematical Statistics*. New-York : Chapman and Hall.

Lagacherie, P., C. Collin-Bellier, et N. Goma-Fortin (2001). Evaluation et analyse de la variabilité spatiale de la mortalité des ceps dans un vignoble languedocien à partir de photographies aériennes à haute résolution. *J. Int. Sci. Vigne Vin 35*, 141–148.

Legros, J.-P., J.-P. Argillier, G. Callot, A. Carbonneau, et F. Champagnol (1998). Les sols viticoles du languedoc. Un état préoccupant. *Progrès Agric. Vitic 13-14*, 296–298.

Mardia, K. V. et R. J. Marshall (1984). Maximum likelihood estimation of models for residual covariance in spatial regression. *Biometrika 71*(1), 135–146.

Matheron, G. (1962). *Traité de Géostatistique appliquée. Tome 1*. Number 14. Technip Paris edition.

Matheron, G. (1963). Principles of geostatistics. *Economic Geology 58*, 1246–1266.

Matheron, G. (1970). La théorie des variables régionalisées et ses applications. *Cahier du Centre de Morphologie Mathématique de Fontainebleau Fasc.5, Ecole des Mines de Paris*.

Matheron, G. (1978). Estimer et choisir. *Cahier du Centre de Morphologie Mathématique de Fontainebleau Fasc.7, Ecole des Mines de Paris*.

McCullagh, P. et J. A. Nelder (1989). *Generalized Linear Models* (2$^{\text{de}}$ ed.). Cambridge : Chapman & Hall.

McCulloch, C. E. (1994). Maximum likelihood variance components estimation for binary data. *Journal of the American Statistical Association 89*, 330–335.

McCulloch, C. E. (1997). Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American Statistical Association 92*, 162–170.

McCulloch, C. E. et S. R. Searle (2001). *Generalized Linear Mixed Models*. Wiley & sons.

McLachlan, G. J. et T. Krishnan (1997). *The EM Algorithm and Extensions*. Wiley.

Monestiez, P., L. Dubroca, E. Bonnin, J.-P. Durbec, et C. Guinet (2006). Geostatistical modelling of spatial distribution of balaenoptera physalus in the northwestern mediterranean sea from sparse count data and heterogeneous observation efforts. *Ecological Modelling 193*, 615–628.

Monestiez, P., M. Goulard, et G. Charmet (1994). Geostatistics for spatial genetic structures : study of wild populations of perennial ryegrass. *Theoretical and Applied Genetics 88*, 33–41.

Oliver, M. J., C. Lajaunie, R. Webster, K. R. Muir, et J. R. Mann (1993). Estimating the risk of childhood cancer. In A. Soares (Ed.), *Geostatistics Troia 92*, pp. 899–910. Kluwer.

Pascual, J. et H. Zhang (2006). Estimation of linear correlation coefficient of two correlated spatial processes. *Sankhya : The Indian Journal of Statistics 68*, 307–325.

Petit, R. J., I. Aguinagalde, J.-L. de Beaulieu, C. Bittkau, S. Brewer, R. Cheddadi, R. Ennos, S. Fineschi, D. Grivet, M. Lascoux, A. Mohanty, G. Müller-Starck, B. Demesure-Musch, A. Palmé, J. P. Martin, S. Rendell, et G. G. Vendramin (2003). Glacial refugia : Hotspots but not melting pots of genetic diversity. *Science 300*.

Rao, C. R. et H. Toutenburg (1995). *Linear Models. Least squares and Alternatives*. Springer Series in Statistics. New-York : Springer-Verlag.

Robert, C. P. (1992). *L'analyse statistique bayésienne*. Paris : Economica.

Robert, C. P. (1995). Convergence control techniques for marov chain monte carlo algorithms. *Statistical Science 10*, 231–253.

Robert, C. P. et G. Casella (1999). *Monte Carlo Statistical Methods*. New-York : Springer.

Royle, J. A. et L. M. Berliner (1999). A hierarchical approach to multivariate spatial modeling and prediction. *Journal of Agricultural, Biological and Environmental Statistics 4*, 29–56.

Schabenberger, O. et C. A. Gotway (2005). *Statistical Methods for Spatial Data Analysis*. Chapman & Hall/CRC.

Stein, M. (1999). *Interpolation of Spatial Data, Some Theory for Kriging*. Springer Series in Statistics. New York : Springer.

Sweeting, T. (1980). Uniform asymptotic normality of the maximum likelihood estimator. *The Annals of Statistics 8*(6), 1375–1381.

Villegas, C. (1977). On the representation of ignorance. *Journal of the American Statistical Association 85*, 1159–1164.

Wackernagel, H. (2003). *Multivariate Geostatistics - An Introduction with Application, 3rd Edition*. New York : Springer-Verlag.

Wei, G. C. G. et M. A. Tanner (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.

Whittaker, J. (1990). *Graphical Models in Applied Multivariate Statistics*. New York : Wiley.

Zhang, H. (2002). On estimation and prediction for spatial generalized linear mixed models. *Biometrics 58*(1), 129–136.

Zhang, H. (2004). Inconsistent estimation and asymptotically equal interpolation in model based-geostatistics. *Journal of the American Statistical Association 99*, 250–261.

Zhang, H. (2007). Maximum-likelihood estimation for multivariate spatial linear coregionalisation models. *Environmetrics 18*, 125–139.

Zhang, H. et D. L. Zimmerman (2005). Toward reconciling two asymptotic frameworks in spatial statistics. *Biometrika 134*, 583–603.

Zhu, Z. et M. L. Stein (2005). Spatial sampling design for parameter estimation of the covariance function. *Journal of statistical planning and inference 134*, 583–603.