Estimation de Modèles Graphiques

Nicolas Verzelen

14 octobre 2005

1 Introduction

Lorsqu'on commence la modélisation statistique d'un système, on suppose souvent que les différentes variables aléatoires en jeu sont indépendantes et identiquement distribuées, ce dans un souci de simplicité. S'il est relativement facile de relaxer l'hypothèse de distributions identiques tout en gardant des propriétés statistiques intéressantes, il s'avére beaucoup plus difficile d'étudier des données dépendantes. Pourtant dans bon nombre de situations, il est absolument irréaliste de considérer les données comme indépendantes. Prenons une série temporelle représentant, par exemple, les cours du baril de pétrole. On ne peut sérieusement supposer les cours moyens d'une année indépendants du cours des années précédentes. Pour l'étude de telles données, on utilise souvent des tehniques de chaînes de Markov ou de martingale. Les chaînes de Markov et les martingales sont des outils de modélisation puissants lorsque les données présentent une orientation naturelle comme le temps dans le cas des séries temporelles. Cependant, si on ne peut pas définir (de manière trop arbitraire) une notion de passé et de futur pour les données, il s'avére maladroit d'introduire ce type de modèles.

Considérons par exemple l'analyse des taux de mortalité par cancer du poumon en Allemagne (Held *et al.*, 2005). Les données observées sont le nombre de cas de cancer du poumon enregistrés en 1986 et 1990 dans chacune des 544 provinces allemandes. L'analyse de données a pour but de mettre en évidence l'existence ou non d'un effet région sur le risque d'un cancer du poumon. Notons X_i la variable aléatoire qui représente le taux de mortalité par cancer du poumon dans la région i. On désire donc vérifier si l'espérance de X_i dépend de i. Il semble naturel (et cela se voit à la lecture des données) que les variables X_i non seulement ne sont pas indépendantes entre elles mais aussi que les taux de mortalité de deux régions sont d'autant plus corrélés que les régions sont proches l'une de l'autre. On pourrait modéliser directement la loi jointe des variables (X_i). Cependant, en dehors du cas gaussien cela se révéle complexe et surtout la forme de la loi jointe est souvent difficilement interprétable. Ainsi, choisir une forme de covariance exponentielle pourra par exemple très bien s'adapter aux données mais n'aura pas une signification particulière pour un épidémiologiste. Pour des modèles aussi complexes, il est important de pouvoir prendre en compte (ou tester) les connaissances des experts et donc de former des modèles facilement interprétables. Dans ce contexte, une méthode

de modélisation alternative consiste au choix des lois conditionnelles des variables X_i. Typiquement, on supposera que X_i, conditionnellement à un petit nombre de variable aléatoires $(X_j)_{j\in\mathcal{N}(i)}$, est indépendante de toutes les autres variables. Dans le cadre de notre exemple, cela signifierait par exemple que le taux de mortalité dans une région donnée, conditionnellement à la connaissance des taux de mortalité des région "proches", est indépendante des taux de mortalité dans toutes les autres régions. L'idée sous-jacente à ce modèle est de former un analogue sur un graphe des chaînes de Markov.

Ces modèles rencontrent un grand succès et sont largement utilisés non seulement en statistique mais aussi en ingénierie ou en informatique. Leur popularité s'explique par les possibilités de dialogues entre le statisticien (ingénieur ou informaticien) d'une part et le commanditaire de l'analyse de données d'autre part. De plus, ces modèles se marrient extrèmement bien avec une approche bayésienne et des méthodes MCMC[1]. Leur utilisation dans différents domaines (et le manque de communication entre ces domaines) rend toutefois les notations souvent peu cohérentes. Dans la suite, j'utiliserai indifférement les termes de modèles graphiques ou de champs markoviens pour de tels modèles.

2 Modèles graphiques

Dans cette section, nous définissons la notion de modèles graphiques et étudions plus précisément deux cas particuliers. Les champs markoviens ont été d'abord introduit en physique théorique pour modéliser des systèmes d'intéraction de particule comme le modèle d'Ising. Il existe une large littérature sur l'étude probabiliste de tels systèmes, en particulier sur l'existence de mesure d'équilibre et de transition de phase pour des systèmes infinis (Georgii, 1998).

2.1 Notations

On se donne $\mathcal{G} = (V,E)$ un graphe non orienté fini. Soit $(X_\alpha)_{\alpha\in V}$ une collection de variable aléatoires indexées par les sommets de $\mathcal{G}$ et prenant leur valeur dans l'espace $\left(\prod_{\alpha\in V}\chi_\alpha\right)$. Les ensembles χ_α sont soit un espace d'état discret soit un $\mathbb{R}$-espace vectoriel de dimension finie. Pour A une partie de V, on note $\chi_A = \left(\prod_{\alpha\in A}\chi_\alpha\right)$ et $\chi = \chi_V$. Pour x un élément de χ et A une partie de V, x_A désigne la projection canonique de x sur χ_A.

Définition 2.1 *On appelle* ***clique*** *un ensemble de sommets deux à deux reliés. On dit aussi que cet ensemble est complet.*

Définition 2.2 *Soient A, B, C trois ensembles disjoints de sommets. On dit que C sépare A et B, si tout chemin qui part d'un élément de A et arrive en un élément de B passe par au moins un point de C.*

Définition 2.3 *Si $\mathcal{G}$ est un graphe, le système de voisinage N sur $\mathcal{G}$ est la famille $N = \{\mathcal{N}_s\}_{s\in V}$ des voisins de s pour s un sommet de V.*

1. Pour plus de détails sur les MCMC pour des champs markoviens, on pourra consulter Knorr-Held & Rue (2002)

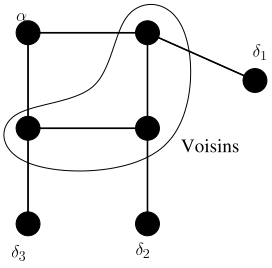

FIG. 1 – *Propriété de Markov locale: X_α conditionnellement à $X_{Voisins}$ est indépendant des X_δ*

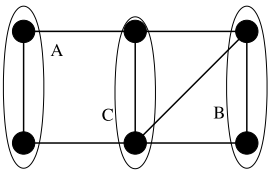

FIG. 2 – *Propriété de Markov globale: X_A conditionnellement à X_C est indépendant des X_B*

2.2 Description générale

2.2.1 Propriétés de Markov

Dans un premier temps, nous désirons formaliser la notion de propriété de Markov sur un graphe. Nous allons définir quatre propriétés différentes qui se révélent équivalentes sous des conditions classiques.

Définition 2.4 *Une mesure de probabilité $\mathbb{P}$ sur χ obéit à (P) la propriété de Markov par paire par rapport à $\mathcal{G}$ si pour tout couple (α,β) de sommets non-adjacents les variables X_α et X_β sont indépendantes conditionnellement à toutes les autres variables*

Définition 2.5 *Une mesure de probabilité $\mathbb{P}$ sur χ obéit à (L) la propriété de Markov locale par rapport à $\mathcal{G}$ si pour tout sommet $\alpha \in V$,*

$$(X_\alpha | X_\beta : \beta \sim \alpha) \perp \{X_\delta : \delta \nsim \alpha)$$

Définition 2.6 *Une mesure de probabilité $\mathbb{P}$ sur χ obéit à (G) la propriété de Markov globale si pour tout triplet (A,B,C) de sous-ensembles disjoints de V tel que C sépare A de B dans $\mathcal{G}$,*

$$\{X_\alpha : \alpha \in A | X_\beta : \beta \in C\} \perp \{X_\delta : \delta \in C\}$$

Définition 2.7 *Une mesure de probabilité $\mathbb{P}$ sur χ se factorise par rapport à $\mathcal{G}$ si pour toutes cliques $a \subset V$, il existe des fonctions positives ψ_a qui ne dépendent de x qu'à travers x_a seulement et il existe une mesure produit $\mu = \otimes_{\alpha \in V} \mu_\alpha$ sur χ tels que $\mathbb{P}$ a pour densité f par rapport à μ et f est de la forme:*

$$f(x) = \prod_{a \ clique} \psi_a(x) \tag{1}$$

Dans ce cas, on dit que $\mathbb{P}$ suit la propriété de factorisation (F).

Ces quatre propriétés de Markov ne sont pas équivalentes dans le cas général. Cependant, on a les implications suivantes.

Proposition 2.8 *Quelque soit le graphe fini non orienté $\mathcal{G}$ et la probabilité $\mathbb{P}$ sur χ, on a les relations:*

$$(F) \Rightarrow (G) \Rightarrow (L) \Rightarrow (P)$$

Par ailleurs, sous les conditions dites d'Hammersley-Clifford, on a l'équivalence entre les différentes propriétés de Markov.

Théorème 2.9 (Hammersley-Clifford) *Une mesure de probabilité $\mathbb{P}$ qui possède une densité f strictement positive par rapport à une mesure produit μ satisfait la propriété de Markov par paire par rapport au graphe non orienté $\mathcal{G}$ si et seulement si f de factorise par rapport à $\mathcal{G}$*

Les hypothèses du théorème 2.9 ne sont pas minimales. Kaiser & Cressie (2000) démontrent un résultat analogue sous des conditions plus générales. Ce théorème fondamental permet de se ramener à la propriété de Markov par paire pour utiliser la propriété globale et la factorisation de la densité. On peut maintenant définir les champs markoviens.

Définition 2.10 *Soit X un processus indéxé par V et à valeur dans χ. X est un champ markovien par rapport au graphe $\mathcal{G}$ (ou par rapport au système de voisinage N) s'il vérifie la propriété (L) de Markov locale par rapport à $\mathcal{G}$ et les hypothèses d'Hammersley-Clifford (2.9).*

Il est immédiat que tout processus à valeur dans χ et vérifiant les hypothèses d'Hammersley-Clifford (2.9) est un champ markovien par rapport au graphe complet (tous le sommets sont voisins les uns des autres). Inversement, un champ markovien par rapport à un graphe sans arête correspond à un ensemble de variables indépendantes.

2.2.2 Spécification locale

On a évoqué précédemment que l'intérêt d'un champ markovien est souvent dans la connaissance des lois conditionelles. Les modèles statistiques utilisés se basent sur les lois conditionelles. Il est donc important de pouvoir relier les lois "locales" à la loi jointe du processus.

Définition 2.11 *On définit la fonction π^s pour s un site par:*

$$\pi^s(x) = P\left(X_s = x_s | X_{\mathcal{N}_s} = x_{\mathcal{N}_s}\right) \tag{2}$$

*ici P désigne la densité de la loi de X_s conditionnellement à $X_{\mathcal{N}_s}$ par rapport à la mesure μ_s. La famille $\{\pi^s\}_{s\in V}$ est appelé la **spécification locale** du champ markovien.*

La loi jointe $\mathbb{P}$ est entiérement caractérisée par la donnée du voisinage et du système de spécification locale. Cependant, si on se donne un système de spécification locale, on n'est nullement assuré que celui-ci définit bien une loi jointe. Il est donc fondamental de pouvoir rapidement vérifier si un système de

spécification est admissible (définit bien une loi jointe) et de pouvoir exhiber la loi jointe.

Résumons très rapidement l'approche par potentiels des champs markoviens; pour plus de détails, on pourra consulter Cressie (1991) (ch. 6.4). A partir du système de spécification locale, on peut définir les fonctions $G_S(x)$ pour S une clique. Si ces fonctions sont invariantes par permutation des cliques S, la fonction negpotentiel Q se définit de manière unique par:

$$Q(x) = \sum_{S \ clique} G_S(x(s_1) \dots x(s_p)) \tag{3}$$

Enfin si la fonction $exp(Q)$ est intégrable par rapport à μ et de somme Z, on obtient la densité du processus par rapport à la loi jointe μ:

$$f(x) = exp(Q(x))/Z$$

La mesure $\mathbb{P}$ définie par sa densité f par rapport à μ vérifie alors les spécifications locales initiales.

On peut donc bien montrer l'existence de la loi jointe et la calculer à partir de la seule connaissance du système de lois conditionnelles. Cependant, le calcul de la somme Z s'avère complexe. Dès lors, l'inférence par maximum de vraisemblance est généralement très lente voire impossible. Néanmoins, à partir de la fonction Q on peut facilement obtenir la probabilité d'un événement (ou la densité) à une constante près et réaliser ainsi de l'inférence par pseudo-vraisemblance (Guyon, 1987).

Toutes ces définitions peuvent paraître un peu théorique et un peu obscures pour l'instant. Appliquons-les à un exemple classique de champ markovien.

Exemple: modèle d'Ising sur un réseau torique
On considère un processus (X_s) pour s un couple d'entiers dans le carré $[-A : A]^2$. Les voisins de $s = (u,v)$ sont $(u+1,v)$, $(u-1,v)$, $(u,v+1)$, $(u,v-1)$. Les sites situés au bord du carré sont reliés à l'autre bord (voir figure 3). Les variables aléatoires X_s prennent les valeurs $\{+1, -1\}$. On peut définir le modèles soit directement par la loi jointe soit avec les lois conditionnelles.

Définissons $\mathcal{E}$ la fonction d'énergie d'une réalisation par:

$$\mathcal{E}(x) = \sum_{v \sim w} x_v x_w$$

La loi du modèle d'Ising à la température T est alors définie par:

$$P(x) = \frac{\exp(\mathcal{E}(x)/T)}{Z}$$

où Z est une constante de normalisation.

On peut aussi donner les lois conditionnelles: soit s un site et $n+$ le nombre de sites voisins de s qui ont pour valeur $+1$.

$$\frac{P(X_s = 1|n+)}{P(X_s = -1|n+)} = \exp(4((n+) - 2)/T)$$

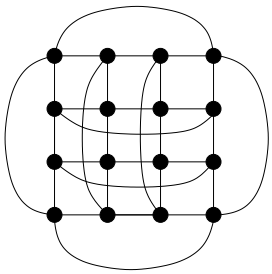

FIG. 3 – *Graphe du champ markovien correspondant au modèle d'Ising sur un carré de taille 4 avec conditions au bord toriques.*

On vérifie facilement que ces deux définitions correspondent bien au même processus.

Dans la suite on va se concentrer sur deux cas particuliers de champs markoviens:

1. l'espace χ est fini
2. le processus X suit une loi gaussienne

2.3 Espace d'état χ fini

Dans cette section, on se restreint au cas où chacune des variables aléatoires X_α prend ses valeurs dans un espace d'états fini χ_α. Considérons un $|n|$ échantillon de ce processus. Les données observées sont alors souvent des comptes de la forme $n = \{n(i)\}_{i\in\chi}$. Ici, $|n| = \sum_{i\in\chi} n(i)$ désigne le nombre total de données. Pour a une partie de V, i_a désigne un élément de χ_a. On trouvera une présentation générale de ce cas dans Lauritzen (1996).

Ces modèles s'appliquent naturellement à l'étude de systèmes experts probabilistes. Les systèmes experts sont des programmes de décision automatiques. Il s'agit à partir d'une connaissance partielle de l'état d'un système d'en reconstituer l'état complet. Considérons par exemple un programme de diagnostic automatique qui à partir des résultats d'un examen médical donne une "prédiction" de la maladie du patient. En ce sens, l'estimation des systèmes experts classiques ressemble dans sa formulation à un problème d'apprentissage. Dans le cas des systèmes experts probabilistes, on ne veut pas seulement reconstituer l'état le plus probable mais aussi évaluer la distribution des états possibles conditionnellement aux données. Dans le cadre de notre exemple, cela revient à donner la distribution des états possibles du patient conditionnellement à ses résultats d'analyse médicales.

2.3.1 Spécification des modèles

Les lois correspondant à de tels modèles sont multinomiales et caractérisées de manière unique par leur vecteur moyen $(m_i)_{i\in\chi}$.

Définition 2.12 (modèles log-linéaires) *Soit H un sous espace de $\mathbb{R}^\chi$. On désigne par $M(H)$ l'ensemble des vecteurs moyens de la forme* $\exp(v)$ *où $v \in H$ et qui vérifie $\sum_{i\in\chi} exp(v_i) = 1$. On appelle modèle linéaire étendu $\bar{M}(H)$ l'adhérence de $M(H)$.*

Un modèle hiérarchique $M(H_{\mathcal{G}})$ est caractérisé par:

$$H_{\mathcal{G}} = \sum_{a \ clique} F_a$$

où F_a désigne le sous espace des fonctions de χ dans $\mathbb{R}$ qui ne dépendent que de leur composante en a:

$$g \in F_a \Leftrightarrow g(i) = g(j) \text{ pour tout } i, j \text{ tel que } i_a = j_a$$

De même, on peut définir un modèle hiérarchique étendu.

Comme l'espace d'état est fini, il s'en suit que $M(H_{\mathcal{G}})$ correspond à l'ensemble des champs markoviens par rapport au graphe $\mathcal{G}$ et à valeurs dans χ. Par contre, son adhérence ne contient pas uniquement des champs markoviens, mais présente l'avantage d'admettre un maximum de vraisemblance avec probabilité 1.

2.3.2 Maximum de vraisemblance

Dans cette section, on s'intéresse à l'estimation du maximum de vraisemblance sur $\bar{M}(H_{\mathcal{G}})$.

Théorème 2.13 *L'estimé par maximum de vraisemblance $\widehat{m}$ du vecteur moyen dans le modèle hiérarchique étendu est l'unique élément de $\bar{M}(\mathcal{H}_{\mathcal{G}})$ qui satisfait l'équation:*

$$\widehat{m}_{i_a} = n(i_a), \quad i_a \in \chi_a, \ a \text{ clique pour } \mathcal{G} \tag{4}$$

Le théoréme 2.13 donne les équations à résoudre mais pas de méthodes pour les résoudre. Il existe des algorithmes itératifs de résolution approchée de ce système (Lauritzen, 1996).

Une fois obtenu le maximum de vraisemblance pour un modéle donné, on peut essayer de comparer différents modèles imbriqués en utilisant un test de déviance.

2.3.3 Déviance entre modèles log-affines

Considérons un modèle hiérarchique étendu $\bar{M}_0 = \bar{M}(H_{\mathcal{G}})$ et un sous modèle $\bar{M}_1$ décrit par $H_{\mathcal{K}}$, où $\mathcal{K}$ est un sous graphe de $\mathcal{G}$. On définit $\chi^+ = \{i \in \chi | n(i) > 0\}$. La statistique de déviance entre les modèles est donnée par:

$$d_{01} = -2 \log \frac{L(\widehat{m}_1)}{L(\widehat{m}_0)} = 2 \sum_{i \in \chi^+} log \frac{\widehat{m}_0(i)}{\widehat{m}_1(i)} \tag{5}$$

Les résultats classiques sur le maximum de vraisemblance pour des modèles exponentiels montrent que sous l'hypothèse $m \in M_1$, la déviance suit asymptotiquement une loi du χ^2 avec un degrée de liberté égal à:

$$f = dim H_{\mathcal{G}} - dim H_{\mathcal{K}}$$

Cependant, on ne connaît pas grand chose de la qualité d'approximation de la déviance par une variable χ^2 lorsque n est petit.

Finalement en statistique "classique", on ne connaît pour l'instant que des résultats asymptotiques d'estimation de modèles graphiques discrets. Cependant, l'importance des applications potentielles en font actuellement un sujet d'étude privilégié. Bien que très peu répandues en France, les recherches se concentrent actuellement vers une approche bayésienne. On parle dans ce cas de réseau bayésien.

2.4 Champ de Markov gaussien

Dans cette partie, on s'intéresse à des champs markoviens pour lesquels la loi jointe est celle d'un vecteur gaussien. En effet, dans le cas gaussien, le graphe et les indépendances conditionnelles apparaîtront naturellement dans la matrice de précision. L'utilisation de champs gaussiens se justifie ici comme souvent par la simplicité des calculs et des résultats en comparaison du cas général. De fait, une importante communauté en statistique spatiale et en analyse d'image utilise ces modèles. Pour un panorama des applications des modèles gaussiens, on pourra consulter Rue & Held (2005).

2.4.1 spécification des modèles

Dans un premier temps, nous rappelons quelques résultats classiques sur les distributions de variables gaussiennes. En particulier, nous donnons une caractérisation des lois conditionnelles pour un vecteur gaussien.

Lemme 2.14 *Considérons Y un vecteur gaussien de matrice de covariance Σ inversible. Si on considère (A,B,C) une partition des index de ce vecteur, on obtient alors l'équivalence entre les deux proposition suivantes:*

1. *les variables aléatoires Y_A et Y_B sont indépendantes conditionnellement à Y_C*
2. $\Sigma^{-1}_{A,B} = 0$

Les résultats de Besag (1974) et un peu de calculs permettent de donner une caractérisation des champs markoviens gaussiens appelés aussi GMRF (Gaussian Markov Random Field) ou CAR (Conditionnal AutoRegressive Model). On peut trouver une démonstration dans Cressie (1991) (ch. 6.4).

Théorème 2.15 *Considèrons un champ markovien gaussien X par rapport au graphe $\mathcal{G}$, de covariance Σ et de moyenne α. On peut toujours décomposer de manière unique Σ^{-1} en $D^{-1}(I - C)$ où la matrice D est diagonale positive et C est de diagonale nulle. Dans ce cas, on a les résultats suivants:*

1. $C_{ij} = 0$ *si* $i \nsim j$ *dans* $\mathcal{G}$
2. $var(X_i|X_j : j \neq i) = D_{ii}$
3. $\mathbb{E}(X_i|X_j : j \neq i) = \alpha_i + \sum c_{ij}(X_j - \alpha_j)$

Réciproquement, supposons qu'il existe deux matrices C et D telles que D est diagonale, C est de diagonale nulle et que X est un vecteur gaussien de moyenne

α et de matrice de covariance $(I - C)^{-1}D$. Définissons le graphe $\mathcal{H}$ qui a pour sommets les index de X et dont les arêtes sont définies par:

$$i \sim j \text{ si } C_{ij} \neq 0$$

Alors X est un champ markovien gaussien par rapport au graphe $\mathcal{H}$ et il vérifie les propriétés précédentes. Il est donc beaucoup plus facile dans le cas gaussien de vérifier qu'un système de spécification local est bien admissible. Notons que la condition de positivité d'Hammersley-Clifford est équivalente à l'inversibilité de la matrice de covariance.

La connaissance du graphe du champ markovien permet donc de connaître les 0 de la matrice de précision du vecteur gaussien. De plus, la matrice de précision peut être explicitée à partir des espérances conditionnelles et des variances conditionnelles du vecteur gaussien.

2.4.2 Maximum de vraisemblance

Considérons un n-échantillon $(y^1, \ldots, y^n)$ d'un champ markovien gaussien par rapport au graphe $\mathcal{G}$ de loi:

$$Y \sim \mathcal{N}(\xi, \Sigma) \tag{6}$$

Soit A une matrice, on note $A(\mathcal{G})$ la matrice définie par:

$$A(\mathcal{G})_{\gamma\mu} = \begin{cases} 0 & \text{si} \quad \gamma \nsim \mu \\ A_{\gamma\mu} & \text{sinon} \end{cases}$$

Notons e le vecteur de taille $|V|$ et dont toutes les composantes sont égales à 1. y est la matrice de taille $n \times |V|$ dont les colonnes sont formées des données.

Théorème 2.16 *Pour un champ markovien gaussien par rapport à un graphe $\mathcal{G}$, le maximum de vraisemblance de la moyenne inconnue et de la matrice de covariance inconnue existe si:*

$$ssd = y^t y - y^t e e^t y / n$$

est définie positive. Si $n > |V|$, cela arrive avec probabilité 1. Quand le maximum de vraisemblance existe, il vérifie alors:

$$\begin{aligned} \widehat{\xi} &= \bar{y} = y^t e / n \\ n\widehat{\Sigma}(\mathcal{G}) &= ssd(\mathcal{G}) \end{aligned} \tag{7}$$

Comme dans le cas des variables discrétes, ce théorème ne donne pas explicitement le maximum de vraisemblance mais uniquement un système d'équation à résoudre. Cependant, dans le cas où $n > |V|$, on est assuré que le système admet une unique solution et il existe des algorithmes de résolution approchée du système (Lauritzen, 1996). Notons que la condition $n > |V|$ pour l'existence du maximum de vraisemblance est seulement suffisante mais non nécessaire. A partir des équations du maximum de vraisemblance, on observe qu'une condition nécessaire est $n > max_{C \text{ clique}}|C|$.

2.4.3 Déviance

Comme dans les modèles discrets, on peut appliquer les résultats classiques des modèles exponentiels pour former des tests de déviance.

Considérons le graphe $\mathcal{G}_0$ et $\mathcal{G}_1$ un sous-graphe de celui-ci. Plaçons-nous dans le cas où $n > |V|$, on peut calculer $(\widehat{K}_0,\widehat{\xi}_0)$ et $(\widehat{K}_1,\widehat{\xi}_1)$ les maximums de vraisemblance dans chacun de ses deux modèles. Dans ce cas, on peut calculer la déviance:

$$D = -2 \log \frac{L(\widehat{K}_1,\widehat{\xi}_1)}{L(\widehat{K}_0,\widehat{\xi}_0)} = n \left(\log det \widehat{K}_0 - \log det \widehat{K}_1 \right) \tag{8}$$

Proposition 2.17 *Sous l'hypothèse H_1, la déviance d converge en distribution vers une loi du χ^2 de degrés de liberté $|E_0| - |E_1|$.*

Finalement, on peut former des tests d'appartenance à un modèle aussi bien dans le cas gaussien que dans le cas d'espace d'états finis. Cependant, ces tests ne sont pas entiérement satisfaisants car il s'agit uniquement d'une procédure asymptotique. Or dans le cas où la taille du champ est importante et la quantité de données relativement petite, on ne peut plus faire confiance aux résultats asymptotiques.

Nous désirons donc trouver une procédure non asymptotique de sélection du graphe et d'estimation du champ markovien. Pour réaliser cet objectif, on va s'intéresser à des méthodes de sélection de modèles pénalisée.

3 Estimation et sélection de modèles pénalisée

Dans la partie précédente, on a revu quelques méthodes d'estimation par maximum de vraisemblance des paramêtres d'un champ markovien. Cependant, l'inférance du graphe des relations se révêle plus difficile. Il semble alors plus adroit de l'envisager sous l'angle de la sélection de modèles. En effet, si on "choisit" un graphe muni de trop peu d'arêtes, il risque d'exister un biais important entre la vraie loi jointe et l'estimateur obtenue. Inversement, si le graphe choisi a trop d'arêtes on se retrouve face au problème classique d'overfitting: trop de paramêtres à estimer pour trop peu de données ce qui engendre une grande variance de l'estimateur.

Soit $s \in \mathcal{S}$ une quantité à estimer. Considérons une famille $(S_m)_{m\in\mathcal{M}}$ de modèles et un estimateur $\widehat{s}_m$ associé à chacun de ces modèles. Soit l un contraste. L'objet de la sélection de modèle est de choisir $\widehat{m} \in \mathcal{M}$ à partir des données pour construire un estimateur $\tilde{s} = \widehat{s}_{\widehat{m}}$. Dans notre cas, un modèle S_m correspond typiquement à un ensemble de champs markoviens sur un graphe $\mathcal{G}_m$. Par exemple, S_m peut désigner l'ensemble des champs markoviens gaussiens de moyenne nulle par rapport à $\mathcal{G}_m$.

L'objectif serait de choisir m^* l'oracle tel que:

$$m^* = arg \inf_{m\in\mathcal{M}} \mathbb{E}_P[l(s,\widehat{s}_m)]$$

Cela n'est cependant pas possible. On se contente souvent de démontrer des inégalités quasi-oracles de la forme:

$$\mathbb{E}_P[l(s,\tilde{s})] \leq C \inf_{m\in\mathcal{M}}(\mathbb{E}_P[l(s,\widehat{s}_m)] + \chi(m))$$

où $\chi(m)$ est un terme qui représente la variance de l'estimateur.

Une procédure de sélection est la pénalisation. Les estimateurs $\widehat{s}_m$ sont ici des minimiseurs du contraste empirique γ et X représente les observations. La sélection de modèle pénalisée revient à minimiser le critére suivant:

$$\widehat{m} = arg \min_{m\in\mathcal{M}}\{\gamma(X,\widehat{s}_m)) + pen(m)\}$$

La pénalité représente la "complexité" du modèle. L'objectif principal est de bien calibrer cette pénalité pour obtenir des inégalités de type oracle.

Il n'existe pas pour l'instant de résultat non asymptotique sur l'estimation pénalisée de modèles markoviens. Les statisticiens se contentent souvent d'heuristiques comme celle d'Akaike (1974). Néanmoins, certains travaux pour des modèles différents se rapprochent de notre problématique. Ainsi Baraud *et al.* (2001) ont étudié l'estimation adaptative de modèles autorégressifs. Il est intéressant de remarquer que les champs de Markov peuvent être considérés comme des modèles autorégressifs, sauf que les erreurs ne sont pas indépendantes entre elles. Massart (2003) a aussi démontré des inégalités de type oracle pour le maximum de vraisemblance pénalisé. Bien que son étude se situe dans un cadre très général et n'est pas directement exploitable dans notre cas, elle peut constituer un angle privilégiée d'approche.

4 Perspectives et enjeux

Pendant ma thèse, je m'attacherai à trouver des méthodes d'estimation de champ markovien qui permettent à la fois d'estimer le graphe et les lois conditionnelles. Pour réaliser cet objectif, je pense dans un premier temps essayer des procédures d'estimation pénalisée et en dériver des inégalités de type oracle pour des champs markoviens gaussiens. Une grande importance sera attachée à l'application de ces résultats à l'analyse de données longitudinales.

Références

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, **19**:716–723.

Baraud, Y., Comte, F. & Viennet, G. (2001). Adaptative estimation in autoregression or β-mixing regression via model selection. *The Annals of Statistics*, **29**:839–875.

Besag, J. E. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society, Series B*, **36**:192–225.

Cressie, N. A. C. (1991). *Statistics for Spatial Data, rev. edn.*. John Wiley and Sons: New York.

Georgii, H. (1998). *Gibbs Measures and Phase Transitions*. W. de Gruyter.

Guyon, X. (1987). Estimation d'un champ par pseudo-vraisemblance conditionelle: Etude asymptotique et application au cas markovien. *Spatial Processes and Spatial Time Series Analysis (Proceedings of the Sixth Franco-Belgian Meeting of Statisticians,1985)*, p. 15–62.

Held, L., Natario, I., Fenton, S., Rue, H. & Becker, N. (2005). Towards joint disease mapping. *Statistical Methods in Medical Research*, **14**:61–82.

Kaiser, M. & Cressie, N. (2000). The contruction of multivariate distributions from markov random models. *Journal of Multivariate Analysis*, **73**:199–220.

Knorr-Held, L. & Rue, H. (2002). On block updating in markov random field models for disease mapping. *Biometrics*, **58**:597–614.

Lauritzen, S. (1996). *Graphical Models*. Oxford Science Publications, Oxford, U.K.

Massart, P. (2003). *St-flour lectures notes. Empirical processes and Adaptative estimation*.

Rue, H. & Held, L. (2005). *Gaussian Markov Random Fields: Theory and Applications*, tome 104. Chapman & Hall, London (Monographs on Statistics and Applied Probability).