Estimateur de pseudo-vraisemblance pour les champs markoviens. Application dans le cas du PC.

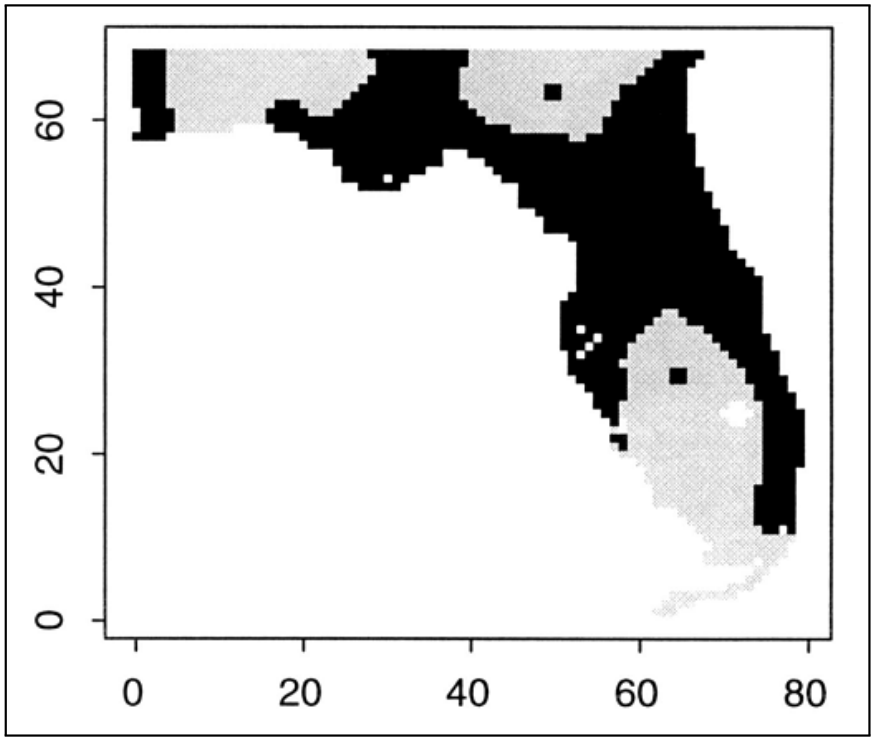
Besnik PUMO

INH - UMR SAGAH - ANGERS

23 novembre 2006

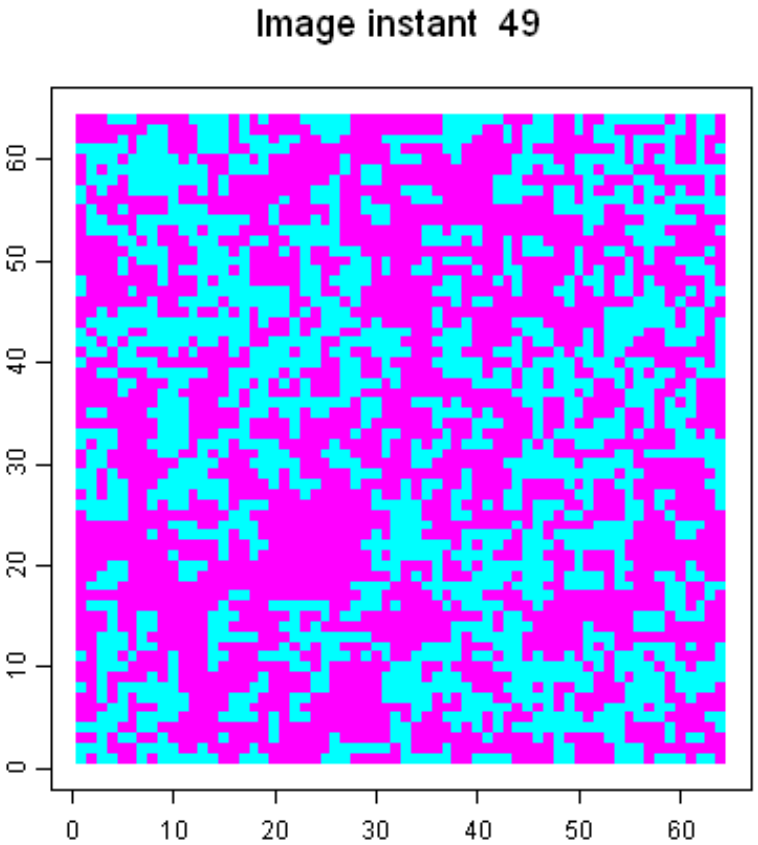
Modélisation de la distribution de *Zanthoxylum clava-herculis* : données binaires sur 180 espèces dans l'état de Californie (données recueillies Crumpacker)

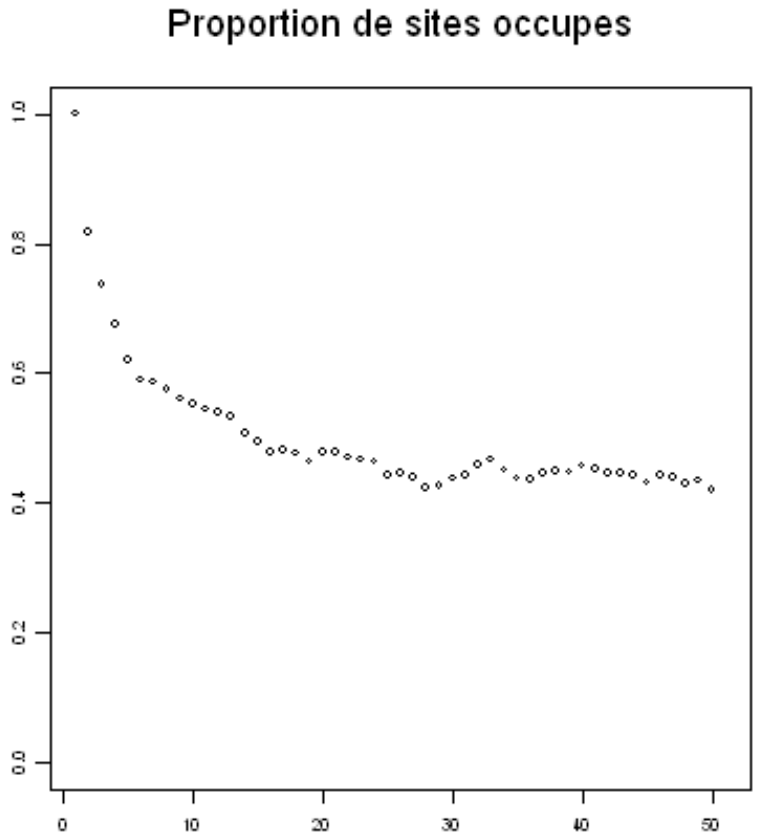
On étudie le lien entre $z_s = 0/1$ (abs./prés.) d'une espèce en fonction de 9 variables climatiques ($\mathbf{x}$).



Wu & Huffer (1997) : Le modèle $P(z_s = 1) = \frac{exp(\eta_s)}{1+exp(\eta_s)}$ où $\eta_s = \beta_0 + \mathbf{x}_s\beta_1$ est insuffisant et montre la nécessité de l'introduction de l'information spatiale.

Processus de Contact (PC)





- Un premier exemple
- Un deuxième exemple (simulations)
- Plan de l'exposé

1. La technique de pseudo-vraisemblance
   - Les auto-modèles
   - Théorème de Hammersley-Clifford
   - Auto-modèles binaires et Gaussien
   - Les estimateurs de EPVC et EPVC associé au codage K
2. Consistence, TCL et Tests
   - Consistence
   - TCL et Tests
3. Cas des Processus de Contact à temps discret
   - Les PC à temps discret
   - La PVC pour le PC
   - Résultats asymptotiques et tests
4. Comparaison des estimateurs
   - Exemple 1 - Régression auto-logistique
   - Exemple 2 - Processus de Contact

Notations

- $\mathbf{Z} = \{Z_s, s \in S\}$: le vecteur de $\sharp(S)$ v.a. dépendantes associé aux sites du réseau S - les observations $\mathbf{z} = \{z_s, s \in S\}$

  $\mathbf{Z}_A := \{Z_s, s \in A\}$ pour $A \subset S$

  Pour un processus spatio-temporel, $\mathbf{Z}(t)$: la configuration à l'instant t

- $\partial_s \subset S$: le voisinage de s ; ∂_A : voisinage de $A \subset S$

Un modèle spatial (possible) sur réseau peut être défini à partir de S, d'un système de voisinage $\{N_s, s \in S\}$ associé et d'une famille de distributions conditionnelles appropriées - Th. Hammersley-Clifford

Théorème de Hammersley-Clifford

Besag (1974) : Soit $Q(\mathbf{z}) = log\{P(\mathbf{z})/P(\mathbf{0}\}$. Alors $P(\mathbf{z}) = \frac{exp(Q(\mathbf{z}))}{\sum_{\mathbf{z}} exp(Q(\mathbf{z}))}$. Le modèle est bien défini dans S si :

$$\begin{aligned} Q(\mathbf{z}) &= \sum_{s} z_s G_i(z_s) + \sum_{s,s'} \sum_{\text{voisins}} z_s z_{s'} G_{s,s'}(z_s, z_{s'}) \\ &+ \sum \sum \sum_{s,s',s'' \ \text{voisins}} z_s z_{s'} z_{s''} G_{s,s',s''}(z_s, z_{s'}, z_{s''}) + \cdots \end{aligned}$$

$$\Rightarrow \frac{P(Z_s = z_s | \mathbf{z}_{S \setminus s})}{P(Z_s = 0 | \mathbf{z}_{S \setminus s})} = exp\{Q(\mathbf{z}) - Q(\mathbf{z}_{z_s=0})\}$$

Les auto-modèles - famille des distributions exponentielles (au plus deux sites sont voisins)

$$Q(\mathbf{x}) = \sum_{s} z_s G_s(z_s) + \sum_{s,s'} \sum_{\text{voisins}} \beta_{s,s'} z_s z_{s'}$$

Modèle binaire : $\Rightarrow \; P\left(z_s|\mathbf{z}_{S\setminus s}\right) = \frac{exp\{z_s(\alpha_s+\sum_{s'\in\partial_s}\beta_{s,s'}z_{s'})\}}{1+exp\{\alpha_s+\sum_{s'\in\partial_s}\beta_{s,s'}z_{s'}\}}$

$$\Downarrow$$

$$P(\mathbf{z}) = \frac{exp\{\sum_s z_s(\alpha_s+\sum_{s'\in\partial_s}\beta_{s,s'}z_{s'})\}}{\sum\sum_{\mathbf{z}\in\{0,1\}^{\sharp S}} exp\{\sum_s z_s(\alpha_s+\sum_{s'\in\partial_s}\beta_{s,s'}z_{s'})\}}$$

Modèle gaussien : $Z_s \mid z_{S\setminus s} \sim G\left(\mu_s + \sum_{s'\in\partial_s}\beta_{s,s'}(z_{s'}-\mu_{s'}), \sigma^2\right)$

$$\Downarrow$$

$$P(\mathbf{z}) = (2\pi\sigma^2)^{-n/2}|B|^{1/2} exp\left\{-\frac{1}{2\sigma^2}\,(\mathbf{z}-\boldsymbol{\mu})'\,B\,(\mathbf{z}-\boldsymbol{\mu})\right\}$$

$B > 0 : diag(B) = 1,\; B_{s,s'} = B_{s',s} = -\beta_{s,s'}$! ! !

La $PVC_S(\theta)$, noté aussi $PVC(\theta)$ est :

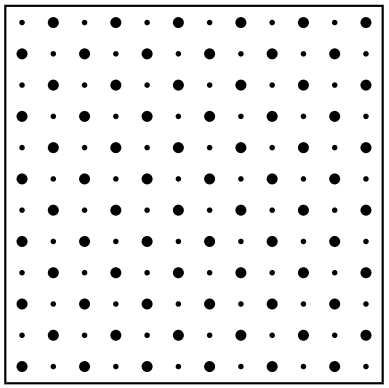
$$PVC_S(\theta) = \prod_{s \in S^o} P(z_s | z_{\partial_s}; \theta))$$

et l'estimateur de PVC ou EPMC :

$$\hat{\theta} = \arg\max_\theta PVC(\theta)$$

Un sous-ensemble $K \subset S$ de sites est un codage si les sites de K sont indépendants conditionnellement à $\mathbf{z}_{S \setminus K}$; l'EPVC associé à K est

$$\hat{\theta}_K = \arg\max_\theta PVC_K(\theta)$$



De nombreux résultats existent sur la consistence des EPVC (Guyon 1995, chapitre 5) - On suppose $\sharp(S(n)) \to \infty$: Le "contraste" de PVC est :

$$U(n;\theta) = -\frac{1}{\sharp(S(n))} \sum_{s \in S^o(n)} log(P(z_s | z_{\partial_s};\theta))$$

- Z est un champ appartenant à la famille exponentielle (ϕ), stationnaire et ergodique - Les estimateurs par Codage et PVC sont consistent (p.s.) sous certaines conditions de régularité de ϕ, $S(n)$ et $U(n)$ (Guyon 1995, §**5.2.1**).

- Théorème sur les estimateurs de minimum de contraste (**Théorème 3.4.1**, Guyon 1995) ; voir aussi Comets (1992 ; Th. 2.1 et 3.1)

- Z est un champ appartenant à la famille exponentielle (ϕ), stationnaire et ergodique - Les estimateurs par Codage et PVC vérifient le TCL sous certaines conditions de régularité de ϕ, $S(n)$ et $U(n)$.

- TCL pour les estimateurs de minimum de contraste (**Théorème 3.4.5**, Guyon 1995) ; voir aussi Comets et Janžura (1998 ; Th. 4.1)

- Test des différences des contrastes pour les paramètres (**Théorème 3.4.6**, Guyon 1995)

PC : Harris (1974)

- Un modèle spatio-temporel à valeurs binaires

$$(\mathbf{Z}(t), t \geq 0),\ Z_s(t) \in \{0, 1\}$$

  - $Z_s(t) = 1$ (0) : présence (absence) d'un plant au site s à l'instant t,
  - $\mathbf{Z}(t)$: configuration du processus à l'instant t

- PC à temps continu ou à temps discret

Définition

- ∂_0 le voisinage de 0 ; $\partial_s = s + \partial_0$ le voisinage de s

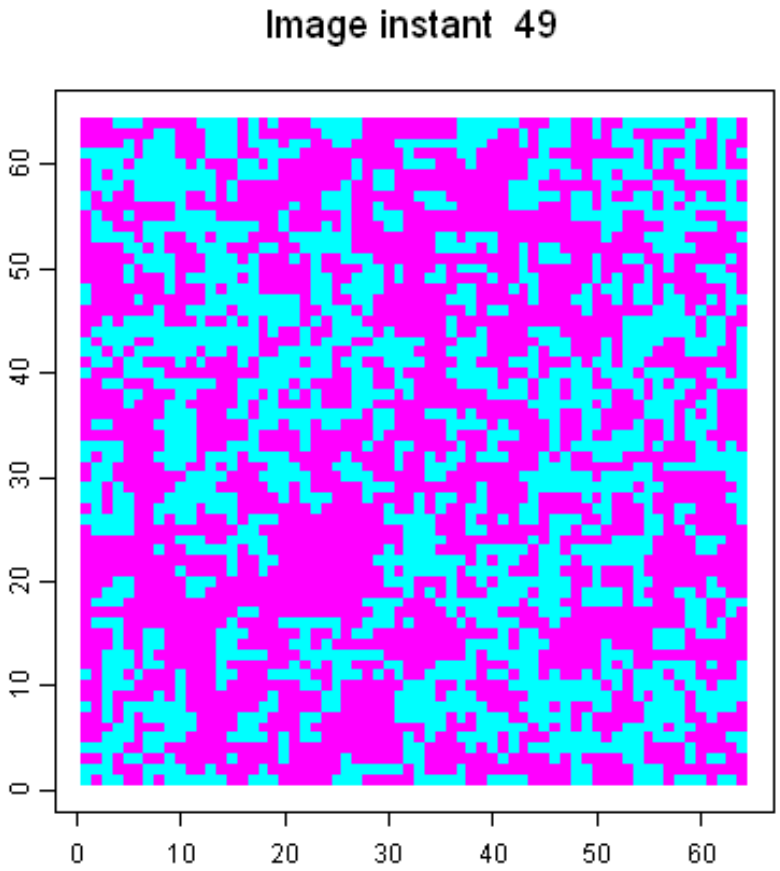
- La probabilité de transition $P(z_s(t+1)|\mathbf{z}(t))$ est invariante dans l'espace-temps

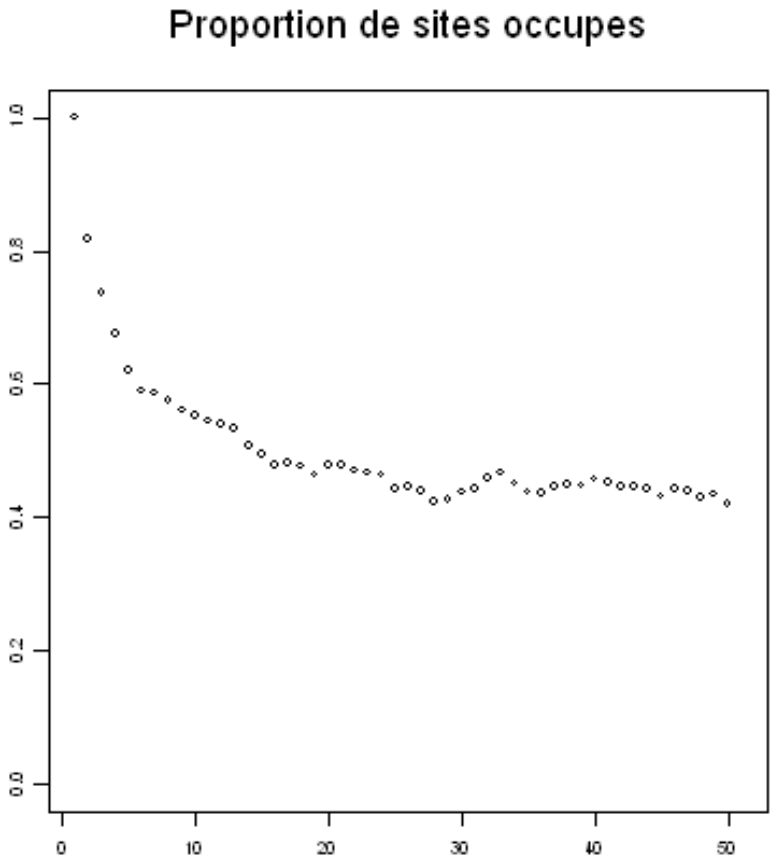
  Règles de l'évolution (de t à $t+1$ - synchrone) :

  a. Chaque plant disparaît avec une probabilité γ,
  b. Un plant en vie produit un nouveau plant dans $s' \in \partial_s$ avec probabilité $f(s'-s)$,
  c. Au plus un plant survit en s.

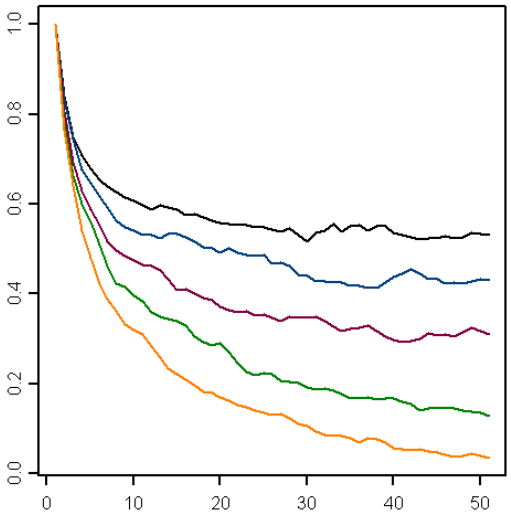
  Tous les événements sont indépendants (dans l'espace-temps).

Un PC simulé - voisinage 4 ppv





PC super critique



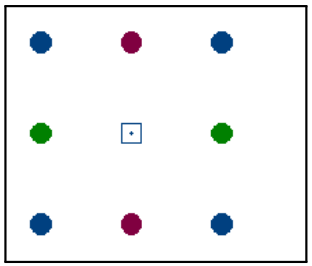
Deux situations sont possibles :

- PC supercritique : le processus survit avec une probabilité positive
- PC sous-critique : le processus disparaît avec probabilité 1

Théorème (Convergence 2.7, 2.15 et 3.4 (Durrett, 1995))

- $\mathbf{Z}^\infty$ *d'un PC supercritique ne dépend pas* $\mathbf{Z}(0)$ *(conditionnellement à la survie)*
- $\mathbf{Z}^\infty$ *est ergodique (par rapport aux translations spatiales)*
- $\mathbf{Z}(t) \Rightarrow P(\tau < \infty)\delta_0 + P(\tau = \infty)\mathbf{Z}^\infty$

- Le paramètre : $\theta = (\gamma, \lambda_1, \ldots, \lambda_p) \in (0,1)^{p+1}$



$p = 3$

- Le logarithme de la PVC normalisée : $n(T)$: Nbre total de sites informatifs

$$log\ PVC(T;\theta) = \frac{1}{n(T)} \sum_{t=0}^{T-1} \log \prod_{s\ informatif} P(Z_s(t+1)|\mathbf{z}(t))$$

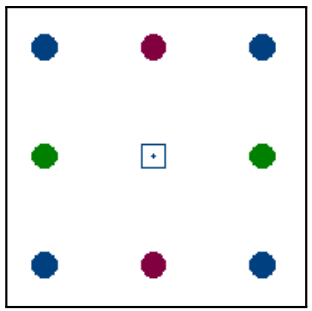
Condition **(I)** : $n(T) \to \infty$ *quand* $T \to \infty$.

Convergence et TCL (Guyon & Pumo 2005, 2006). Sous **(I)** $\hat{\theta}_T$ est consistent (p.s.) et asymptotiquement gaussien :

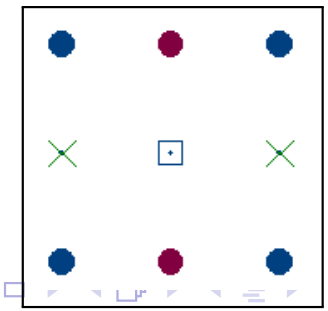
$$\sqrt{n(T)}\,[J_T(\theta_o)]^{-1/2}\,A_T(\theta_o)(\hat{\theta}_T - \theta_o) \xrightarrow{d} \mathbf{G}_{p+1}(0, I_{p+1}).$$

Tests sur les paramètres. Soit $H0$ définissant un PC d'ordre $p - q \geq 1$. Sous HO et la condition **(I)** : $2 \cdot \sqrt{n(T)}\,[-\ell_T(\hat{\alpha}_T) + \ell_T(\hat{\theta}_T)] \overset{d}{\sim} \sum_{i=1,(p-q)+1} \upsilon_{i,T}\chi^2_i(1)$

Voisinage d'ordre $p = 3$

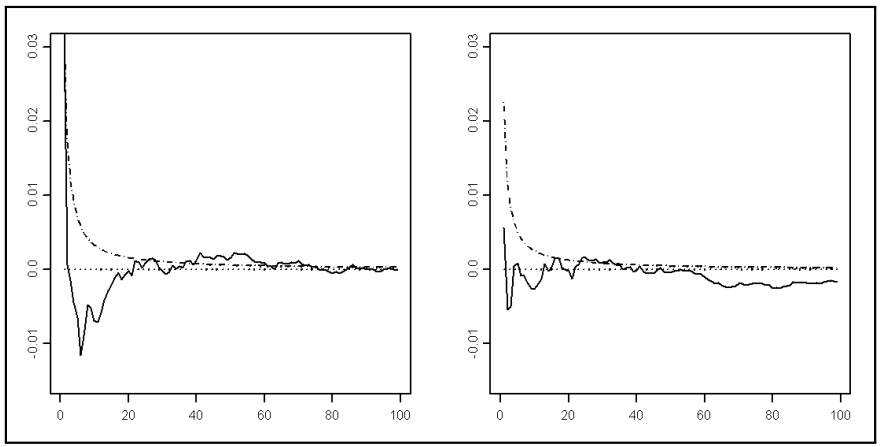


Voisinage sous $H0$ ($p' = 2$)

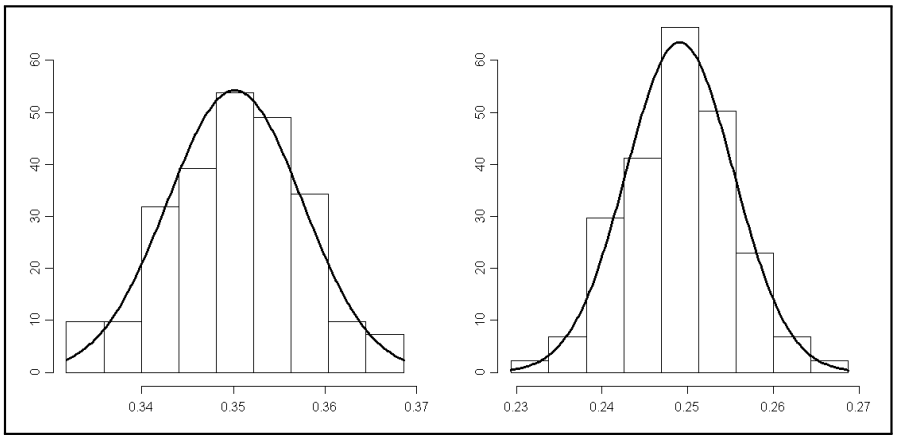


Validation numérique des résultats asymptotiques

PC supercritique ($\gamma = 0.35$, $\lambda = 0.25$) sur un réseau périodique 64×64



Le biais et écart-type pour T variant de 1 à 100



Histogrammes obtenus à partir de 100 estimations

- Guyon et Kunsch (1992) ont comparé les estimateurs MV, PVC et par Codage pour le modèle de Ising - voir aussi Guyon 1995 (§5.3.4)
- Etude comparative par simulation - basée sur 500 simulations (Wu et Huffer, 1997)

Paramètres	Méthode	$\hat{\beta}_0$ ($\hat{s}_{\hat{\beta}_0}$)	$\hat{\beta}_1$ ($\hat{s}_{\hat{\beta}_1}$)	$\hat{\gamma}$ ($\hat{s}_{\hat{\gamma}}$)
(-1, 2, 1)	COD	-0.98 (0.36)	2.04 (0.30)	1.00 (0.12)
	PVC	-0.96 (0.35)	2.04 (0.30)	0.99 (0.12)
	MCMC	-0.93 (0.32)	2.06 (0.28)	0.98 (0.10)

- Etude comparative sur l'exemple (Wu et Huffer, 1997) - Le modèle :

$$P(Z_s = z_s|\mathbf{z}_{S\setminus s}) = \frac{exp(\eta_s)}{1 - exp(\eta_s)} \text{ où } \eta_s = \beta_0 + \mathbf{x}'_s\boldsymbol{\beta}_1 + \gamma \sum_{s' \in \partial_s} z'_s$$

$$\mathbf{x}' = (TM, TMM, ELV)$$

TMM : Température minimale du mois le plus froid,

TM : Temp. Moyenne du mois le plus froid, *ELV* : Altitude

Méthode	β_0 $(\hat{s}_{\hat{\beta}_0})$	TM $(\approx \hat{\sigma}_{TM})$	TMM $(\approx \hat{\sigma}_{TMM})$	ELV $(\approx \hat{\sigma}_{ELV})$	γ $(\approx \hat{\sigma}_{\gamma})$
COD	-7.76 (0.88)	-4.28 (1.01)	3.99 (1.03)	-0.88 (0.29)	4.53 (0.48)
PVC	-7.05 (0.75)	-3.80 (0.94)	3.58 (0.97)	-0.74 (0.26)	4.04 (0.39)
MCMC	-4.78 (0.26)	-1.64 (0.17)	1.53 (0.16)	-0.32 (0.04)	2.55 (0.14)

Comparaison des méthodes : Critère $SAE = \sum_{i=1}^{m} |y_i - \hat{p}_i|$ et $SSE = \sum_{i=1}^{m} (y_i - \hat{p}_i)^2$:

Critère	COD	PVC	MCMC
SAE	758.58	707.19	568.36
SSE	699.48	665.07	429.93

1. MCMC plus performant que PVC et PVC plus performant que COD
2. quand γ important ceci est plus évident ; sinon MCMC et PVC assez proches
3. PVC ou COD exigent moins de calculs et peuvent servirent de valeurs pour initier l'algorithme MCMC

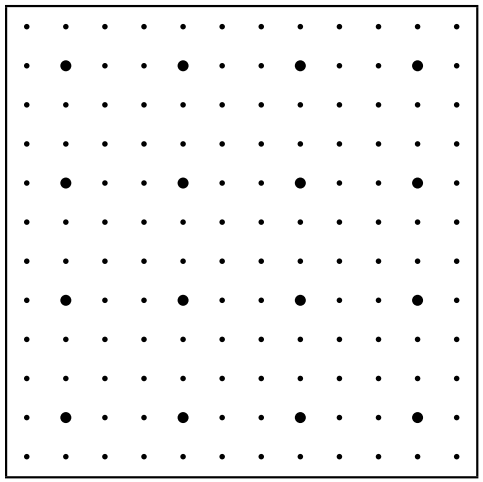
Simulations : $\gamma_o = 0.35$, $\lambda_o = 0.25$

	γ	$\sigma_{\hat{\gamma}}$	λ	$\sigma_{\hat{\lambda}}$
PVC	0.3572	0.0079	0.2434	0.0064
K-codage	0.3456	0.0218	0.2361	0.0181

Estimations PVC et K-codage

Estimation ($T = 4$) des paramètres et leurs écart-types

			$\lambda_o = 0.2$		
γ_o	$\hat{\gamma}_4$	$\hat{\sigma}_{\hat{\gamma}_4}$	$\hat{\lambda}_4$	$\hat{\sigma}_{\hat{\lambda}_4}$	n_4
0.2	0.189	0.6	0.193	0.4	15286
0.4	0.399	0.008	0.188	0.005	13565
0.6	0.607	0.009	0.206	0.008	10452



- Besag J. (1974) Spatial Interaction and the Statistical Analysis of Lattice Systems, *JRSS B*, 36, 192-225.
- F. Comets (1992) On consistency of a class of estimators for exponential families of markov random fields on the lattice, *The Ann. of Stat.*, 20, 455-468.
- F. Comets, M. Janžura (1998) A CLT for conditionally centred random fields with an application to markov fields, *J. Appl. Prob.*, 35, 608-621.
- Durrett R., Levin S.A. (1994) Stochastic spatial models : a user's guide to ecological applications, *Phil. Trans. R. Soc. Lond.*, B 343, 329-350.
- Guyon X. (1995) *Random fields on a network : modelling, statistics and applications*, Springer, Berlin.
- Guyon X., Künsch H.R. (1992) Asymptotic comparison of estimators in the Ising model, L.N.S. 74, 177-198.
- Guyon X., Pumo B. (en révision) Space-time estimation of a particle system model, en révision pour *Statistics*.
- Jensen J.L., Künsch H.R.(1994) On asymptotic normality of pseudo-likelihood estimate for pairwise interaction processes, Ann. Inst. Statist. Math., 46, 475-486.
- Pumo B. (2007) Parameter estimation of the CP of order d, accepté pour publication dans *Comm. in Statistics*.
- Wu H., Huffer F. (1997) Modelling the distribution of plant species using the autologistic regression model, Enviromental and Ecological Statistics, 4, 49-64.
- Voir aussi Livre de Stan Z. Li