Des Modèles Statistiques

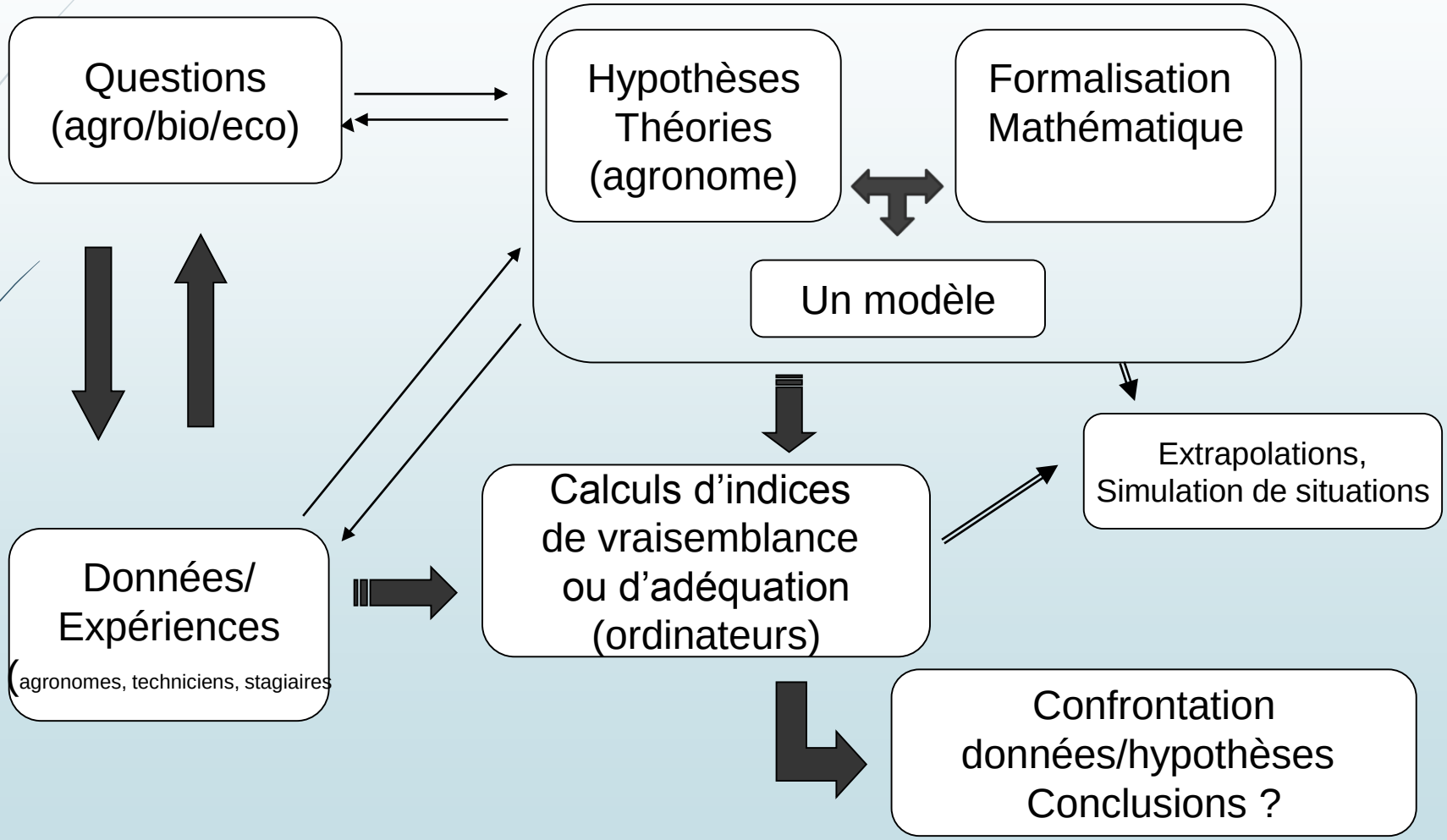
Rachid Senoussi

Unité Biosp, Inra Avignon

1. Préliminaires sur les modèles en général

- Un modèle est une représentation « humaine » de ce que pourrait être le réel.
- Un modèle ne peut décrire la « réalité » des choses mais seulement les relations/rapports entre ces choses.
- Un modèle n'est ni vrai ni faux. Son but est d'être utile.
- On peut même dire qu'un modèle est toujours faux, mais qu'il peut être très efficace pour résoudre/faire avancer des problèmes appliqués ou théoriques.
- Un modèle est constitué d'hypothèses de travail sur des questions appliquées en physique, biologie, économie ou autre.

modus operandi



1.1. Sur les modèles mathématiques

- Une grande partie des mathématiques appliquées consiste à faire de la modélisation dans diverses disciplines
- De façon schématique, on peut distinguer
  - la modélisation déterministe où on ne prend pas en compte des variations aléatoires par le biais d'outils (EDO, EDP, filtrage,…)
  - et la modélisation stochastique (qui prend en compte ces variations aléatoires en essayant de leur associer des lois de probabilité
- Qualités d'un modèle: (parfois antagonistes)
  - Bon ajustement aux données et économe en paramètres
  - Robustesse (aux erreurs de mesures, à de petits écarts aux hypothèses de travail)
  - Facilité de manipulation (calculs numériques et stabilité des résultats)
  - Caractère prédictif
  - Interprétation des paramètres

Les modèles statistiques n'existent pas ! (Ou vraiment pas beaucoup)

- Ce sont les modèles stochastiques (probabilistes) qui importent le plus.
- Les statistiques complètent la modélisation stochastique en détaillant par des paramètres θ (vecteurs, fonctions numériques) les choix précis de lois déjà adoptées.
- Les « modèles statistiques » relèvent surtout de techniques permettant de préciser parmi les modèles en compétition ceux qui seraient les plus « proches », plus « vraisemblables » au vu des données récoltées.
- La modélisation statistique peut commencer déjà « sans modèle » par l'utilisation des « Statistiques Descriptives des données »: calcul de moyenne, de covariance, de distributions empiriques,…

1.2. La modélisation stochastique

- La modélisation stochastique a pour but essentiel de préciser des classes de lois de probabilité rendant compte des variations aléatoires des phénomènes d'intérêt, variations dues à des causes soit inconnues, soit impossible à mesurer (cachées, trop nombreuses,…)
- Pour cela, elle se donne un cadre formel permettant
  - de décrire les variations aléatoires mentionnés,
  - Et d'étudier les propriétés générales des phénomènes qu'ils engendrent.

- la modélisation statistique, plus appliquée, consiste essentiellement à choisir les outils appropriés pour confronter les données au modèle stochastique.
- Noter que le terme de modélisation statistique est très général et que, à la limite, toute démarche statistique en relève.

la modélisation statistique en constante évolution

- Les méthodes de modélisation statistique sont très nombreuses et en constante évolution/amélioration. Il s'agit plutôt de démarches /méthodes que de modèles figés pour:
  - Traiter des masses de données de plus en plus volumineuses (internet, biologie à haut débit, climat, imagerie, marketing...)
  - Et utiliser les nouveaux moyens de calcul tout aussi considérables

- Mais la question de base reste globalement l'«explication ou la mise en relation stochastique d'une variable privilégiée Y de nature parfois complexe , appelée variable à expliquer ou réponse, avec des variables dites explicatives X, expliquant "au mieux" Y.

2.1. Modèles (?) préliminaires

- Nettoyage des données (data management):
  - Statut des données manquantes, erronées, variables redondantes
- L'exploration des données :
  - Toutes les statistiques descriptives et caractéristiques empiriques : moyennes, covariances, histogrammes,
  - Estimation non paramétrique des densités, des intensités, …
  - ACP (analyse en Composantes Principales) pour les liaison des variables, ACM (Analyse des Correspondances Multiples) entre variables qualitative. Analyse discriminante, classification, méthodes CART (découpage de populations en fonction des variables explicatives,…
- Transformer des variables? Regrouper des variables? des modalités de variables?

2.2. Méthode statistiques non paramétriques

- On a seulement besoin d'hypothèses de travail assez générales (données iid, ou stationnaires, ou invariants par permutation,…):
- Des méthodes d'estimation par noyaux de densité, d'intensité de processus, de fonctions de covariance d'un champs aléatoire, de la fonction de survie, etc…
- Des tests non paramétriques : égalité des moyennes de 2 échantillons, test des rangs, …
- Test d'égalité de distributions Kolmogorov-Smirnov, etc…
- Toute la famille des tests par permutations en statistique spatiale sur la détection d'agrégats spatiaux, sur l'homogénéité directionnelle, …
- Extension a de la statistique semi-paramétrique: une partie paramétrique à inférer et une partie non paramétrisée (modèle de Cox,…)

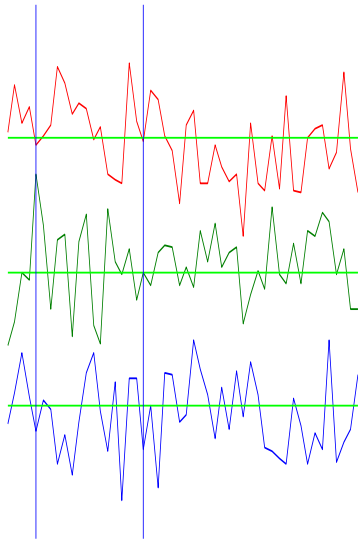
2.3. Modèles Paramétriques classiques(échantillon iid ou presque)

- Le modèle linéaire (gaussien) de base: le plus simple, le plus ancien
  - Régression linéaire, analyse de *variance/covariance*, les variables explicatives sont déterministes (effets fixes).
  - Cadre gaussien très efficace au niveau du formalisme
- Le modèle linéaire généralisé: extension au non gaussien et description des paramètres par des fonctions de liens très générales et plus seulement linéaires: régression logistique, de Poisson, loglinéaire, durée de survie,…
- Modèles linéaires généralisés: On explique Y de façon non linéaire à partir de fonctions inconnues des X (on fait alors de la statistique non paramétrique:
  - Exemples: régression non paramétrique, GAM (Generalized Additive Models), Réseaux de neurones
- Les modèles mixtes: On étend les modèles précédents au cas où les variables explicatives sont elles aussi aléatoires avec spécification de leurs lois (effets aléatoires): Ceci permet d'expliquer une plus grande variabilité des données

2.4. Modèles paramétriques spatio-temporels et autres

- Les modèles probabilistes plus complexes se dénomment selon
  - le caractère continu/discret de l'espace et du temps
  - S'il y a une flèche (du temps ou d'ordre) pour les processus temporels et pour certains processus sur graphes (causalité directionnelle)
  - La structure de voisinage des points supports: temporel/ spatial/ spatio temporel/ sur réseaux
  - Le types d'objets à étudier sur ces espaces
- Modèle Statistique: On le décrit par la donnée d'une famille de variables aléatoires $\{ X(t), t \in T\}$ (observations) générées selon un ensemble donné de lois conjointes $P_\theta(dx(t), t \in T)$ où θ est un paramètre vectoriel de dimension finie ou infinie inconnu.
- Le but inférer sur θ: l'estimer et estimer la qualité de cette estimation, …

Processus à temps discret



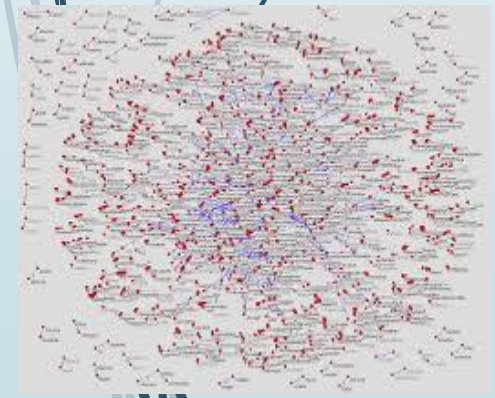
Des développements à l'infini sur ce type de données

- Systèmes dynamiques à temps discret
- Autorégressifs (AR, ARMA, ARMAX,…)
  - $X(n)= f(X(n-1), \ldots, X(n-p), \varepsilon(n),\ldots, \varepsilon(n-q), U(n))$, n entier
  - f: inconnue à estimer ou paramétrée par un vecteur θ à inférer
  - X observation, ε innovation (non observée), U variable de contrôle du système
  - Donner loi initiale $\pi_\theta (dx_0,..,dx_p)$ et loi $Q_\theta (d\varepsilon)$ de l'innovation (souvent bruit blanc)
- Chaînes de Markov homogènes et non homogènes sur un espace E
  - Donner une loi initiale $\pi_\theta (dx_0)$ (souvent inutile si ergodicité)
  - Donner une probabilité de transition $P_\theta (x, dy)$
  - Exemple des système de dynamiques populationnelles,…
- Calcul de vraisemblance souvent assez simple

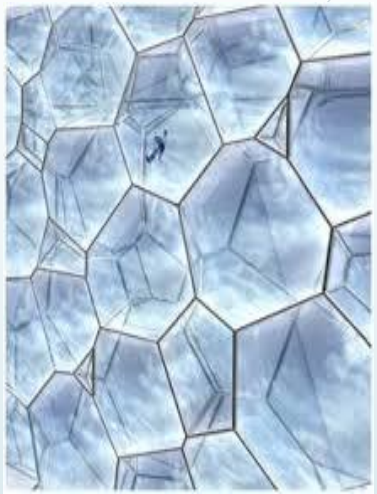
Champs aléatoires à temps discret

Idem, des développements à l'infini sur ce type de données

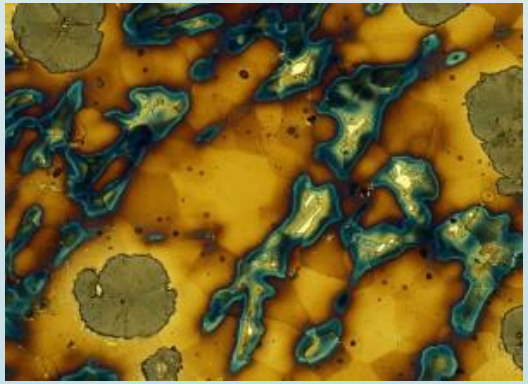
- Systèmes sur réseaux spatiaux discrets ou sur graphes orientés
- Autorégressifs Spatiaux (AR, ARMA, ARMAX,…)
  - $X(n)= f(X(n-1), \ldots, X(n-p), \varepsilon(n),\ldots, \varepsilon(n-q), U(n))$, $n =(n_1,\ldots, n_k)$ vecteur d'entiers
  - f: inconnue à estimer ou paramétrée par un vecteur θ à inférer
  - X observation, ε innovation (non observée), U variable de contrôle du système
  - Donner un ensemble de lois conditionnelles loi $Q_\theta (d\varepsilon' \mid)$ de l'innovation (souvent bruit blanc)
- Champs de Markov homogènes et non homogènes sur un espace E
  - Donner une loi pour l'origine $\pi_\theta (dx_0)$
  - Donner un ensemble de lois de probabilités conditionnelles cohérentes $P_\theta (dx \mid X(\text{voisins de } x))$ pour assurer l'existence d'une loi de proba pour le champ
- Calcul de vraisemblance n'est plus simple mais que l'on peut approximer par simulations et algorithmes stochastiques



Champs spatiaux à indice continu, suite



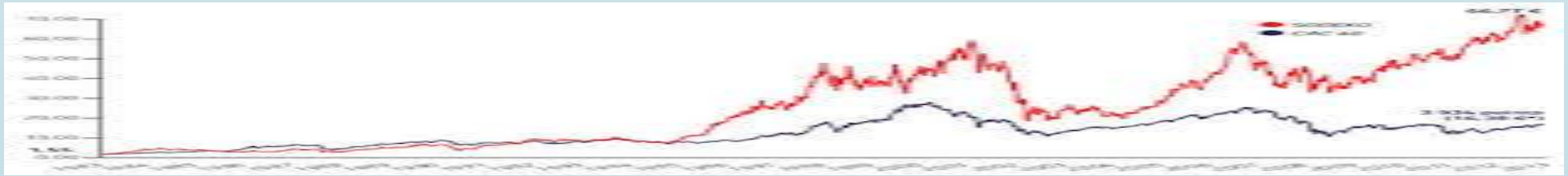
- Généralisation à des processus de points marqués (position et différentes caractéristiques d'un arbre, ...)
- Généralisation à des objets géométriques (modèles booléens, shot noises, filaments, fissures, alvéoles,…
- Généralisation à des mesures aléatoires dynamiques : répartitions des populations interactives et dynamiques,
- Toutefois, sauf cas simples il est rare de pouvoir calculer des vraisemblance mais seulement définir des contrastes, des approximation de vraisemblance à partir d'EDP ou par simulations



Processus à temps continu

Des développements à l'infini autour des:

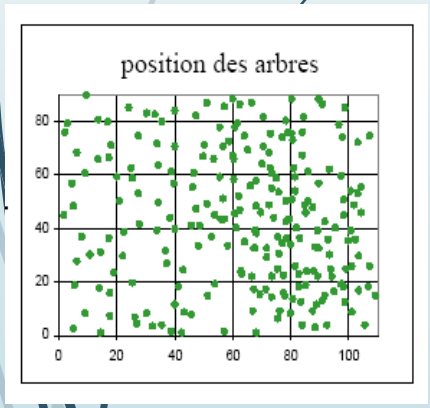
- Processus de Diffusion (Equation Différentielle Stochastique) à base du mouvement brownien :
  - $dX(t) = a_\theta(t,X(t))\,dt + \sigma_\theta(t;X(t))dW(t)$, $t>0$ (le temps)
  - $A_\theta(.)$ drift; $\sigma_\theta(.)$ coefficient de diffusion, W : brownien
- Généralisation à des Equation Différentielles Stochastiques à base de processus de Levy
  - Incluant à la fois des processus directeur s browniens, des processus de sauts de poisson et autres « monstruosités » d'irrégularité
  - Ajout éventuel d'un processus de contrôle . Paramétrisation des coefficients
  - Redonne des exemples de processus autorégressifs en temps continu, des processus invariants par changement d'échelle, …
- En général, calcul des vraisemblances pour des observation $X(t_1),\ldots..X(t_n)$ impossible. Les lois sont décrites formellement par des EDP diffusions et intégro-différentielle en fontion des conditions initiales et des coefficients.
- On peut courageusement faire de l'approximation numérique, proposer des méthode par quasi-vraisemblance ou par simulations.



Champs spatiaux à indice continu

Des développements à l'infini autour des:

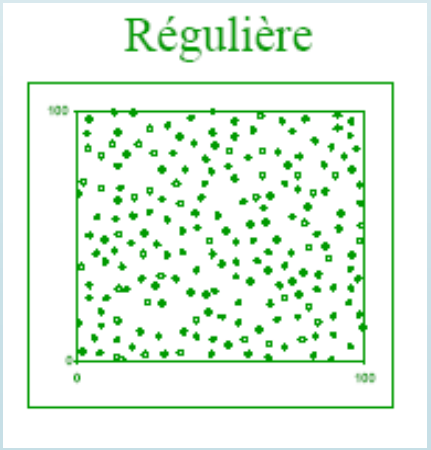
- Processus des champs aléatoires Z(x) (Géostatistique):
  - Principalement champs gaussiens (stationnaires, isotropes)
  - Paramétrisation de la tendance $m_\theta(x)=E_\theta(Z(x))$ et de la structure de covariance $c_\theta(x,y)=Cov_\theta(Z(x),Z(y))$ (resp $=c_\theta(x-y)$; $=c_\theta(|x-y|)$)
  - Estimation des paramètres puis prédiction des valeurs (krigeage)
  - Calcul de la vraisemblance possible
- Extension aux cas non gaussiens, non isotropes, non stationnaire,…
  - Vraisemblance non calculable en général
- Existe une théorie des EDP stochastiques (difficile) que l'on commence seulement à manipuler/appliquer en modélisation des champs aléatoires

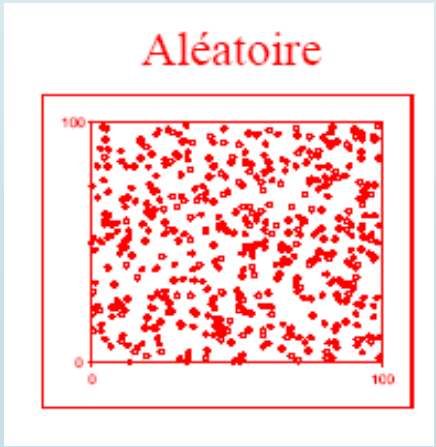


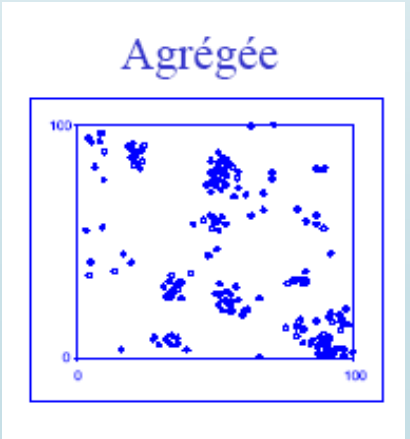
Champs spatiaux à indice continu

Des développements à l'infini autour des:

- Processus de points dans l'espace corrélés entre eux
  - Définition d'une fonction d'intensité (ou de densité de pt par unité de volume)
  - Définition de différentes notions de corrélations spatiales: covariance des effectifs par zone, fonction d'interaction de paires, fonction de Ripley, etc.. mesure de Palm, mesure s de Janossy (« lois conditionnelle » infinitésimales)
  - Quelques modèles utiles et paramétrés (Modèles deStrauss, Neymann-Scott, Markoviens, de Cox, ..) pour les modèles fréquents
  - Paramétrisation des coefficients







3.1. Listing des méthodes d'inférence sur le paramètre

- Méthode du maximum de vraisemblance: trouver θ qui maximise la vraisemblance (ou son logarithme) quand cela est possible
  - Méthode heuristiquement justifiée et aussi théoriquement pour beaucoup des cadres présentés précédemment (en s'appuyant sur la loi des grands nombres et du théorème central limite) et l'entropie des lois (information de Kullback)
  - Elle est souvent la plus efficace des méthodes (mais pas toujours)
  - Elle permet de réaliser des tests sur différentes hypothèses de travail dès que le calcul est possible et sans effort supplémentaire de calcul
  - C'est un peu la reine des méthodes (si applicable)
- On peut généraliser cette approche MV par des fonctions similaires qui jouent le même rôle: Méthodes du minimum de Contraste
  - Définir des contrastes entre 2 paramètres (moindres carrés, méthode des moments, quasi-vraisemblance, vraisemblance partielle de Cox, etc…: mais à justifier impérativement au niveau théorique (martingales positives, etc…)

méthodes d'inférence sur le paramètre, suite

- L'approche Bayésienne: On suppose que le paramètre θ n'est en fait pas une valeur fixe et inconnu mais est issu d'une réalisation d'une variable aléatoire de loi a priori $\pi(d\,\theta)$ contenant les informations que l'on détient sur ce paramètre. Ce qui importe alors est de mieux connaître la loi du paramètre au su des observations récoltées.

  $\Rightarrow$ Trouver la loi a posteriori $\pi(d\,\theta \mid X)$.

- On peut alors calculer toutes sorte de choses sur θ : sa moyenne, sa covariance, ses quantiles, etc … en principe !

$\Rightarrow$ Problèmes de Calculs dans les 3 cas précédents soit pour le calcul de densités soit pour le calcul de la minimisation lui-même soit les 2 !

3.2. Approximation de vraisemblance / loi a posteriori

- Dès que le modèle est un peu sophistiqué/complexe, se posent des problèmes ardus de calcul (intégration sur espace à grande dimension, …).
- Impératif de développer des méthodes numériques pour inférer :
  - Ces méthodes reposent elles même sur des modèles stochastiques en général mais sont en principe indépendantes du modèle étudié.
  - Il est important de séparer formellement ces différentes étapes dans la modélisation (ne pas confondre outil technique secondaire de calcul/approximation et hypothèses de travail (modèle de base).
- En exemples de calcul approché:
  - la méthode ABC (Approximate Bayesian Computation) pour le cadre bayésien: approcher la loi par simulations de données et retenir celles proches de l'observation.
  - Approximation numérique de densités pour des processus de diffusions.
  - Approximation de vraisemblance de processus ponctuels par simulations dites exactes.

3.4.1. Outil des algorithmes stochastiques pour l'optimisation

- Soit f : E $\rightarrow R_+$ une fonction (dite objectif) sur un espace d'état E

Pb: trouver x^* tel que $f(x^*)=\inf(f(x), x\in E)$ avec f complexe, irrégulière, …

- Outil: chercher une Chaîne de Markov X_n ergodique dont la loi stationnaire sera concentrée sur l'ensemble des solutions x^*
- Développement d'algorithmes performants: Metropolis-Hasting, algorithmes populationnels (dits évolutionnaires avec des notions de séléction, mutations, …), algorithmes des fourmis,…
- En gros si E discret avec une structure de voisinage : On fixe une température T>0 et on définit une loi de Gibbs $\pi_T(x)= \exp(-f(x)/T)/Z_T$
  - Z_T est une constante de normalisation (de Boltzmann) qui pose problème à calculer. On construit alors une CM X_n avec transition

$$p_T(x,y)=\begin{cases} 1/|V(x)| & si\ y\in V(x)\ et\ f(x)\le f(y) \\ \exp(-(f(y)-f(x))/T) & si\ y\in V(x)\ et\ f(x)>f(y) \\ 0 & sinon \end{cases}$$

  - On a X_n cv en loi vers π_T et si $T=T(n)\rightarrow 0$ et si x^* unique alors $X_n \rightarrow x^*$

3.4.2. Outil MCMC: MonteCarlo Markov Chains pour la simulations de lois

- Soit une loi de proba $\pi(dx)$ sur un espace d'état E que l'on veut étudier (formule non explicite ou très compliquée [exemples modèles markoviens, modèle Ising, …]):
  - On se propose de construire une Chaîne de Markov ergodique dont la loi stationnaire serait $\pi(dx)$ et qui serait facile à simuler
- Démarches similaires au cas de l'optimisation précédent en trouvant une proba de transition adaptée qui n'utilise que le rapport $\pi(dx)/\ \pi(dy)$ et se débarrasse de facteurs communs intraitables
  - Methode MCMC, Méthode « Importance Sampling », et autres
- Application au calcul d'intégrales et de modèles à variables cachées, à l'analyse statistique des images,…

4. Sélection de Modèles

- Il n'y a pas de théorie générale pour décider du choix d'un modèle parmi d'autres
- Dans le cadre d'une famille de modèles paramétrés emboîtés (où certains paramètres seraient superflus) il y a des méthodes basées sur une pénalisation de la vraisemblance (ou du contraste considéré)
  - Si k nombre de paramètres et n le nb observations
  - Akaïké Information Criterion : $AIC= -2\log(Lv) +2k$)
  - On a aussi une formule de l'AIC corrigé si $n/k < 40$
  - Bayesian Information Criterion (Schwarz) : $BIC= -2\log(Lv) +k \log(n)$
- Ces méthodes se généralisent à quelques cas non standards (donnée chronologiques, etc..) et pour des modèles non emboîtés avec certains a priori sur les modèles

5. Validation d'un modèle

- Plusieurs façons de concevoir la validation d'un modèle:
- Cas classique : Regarder la distribution des résidus (d'une régression par exemple: R^2 statistique,…)
- Sinon, des considérations plus générales dépendantes du modèle lui même
  - On simule des données selon le modèle estimé et on compare quelques statistiques calculées sur les simulations à celles calculées sur l'échantillon.
  - On utilise une partie des données pour regarder la qualité de prédiction des données non utilisées (on peut répéter cela plusieurs fois : bootstrap)

6. En conclusion

- Pas de bibliographie sinon des centaines, voire des milliers de références sur ces généralités…
- Analyse de sensibilité du modèle laissée de coté
- Les logiciels : R, WinBUGS
  - un nombre incalculable de packages prêt à l'usage sous R.
  - Mais le plus souvent, il faut être capable de fabriquer son propre programme car votre modèle est forcément spécifique
  - Mais utiliser quelques modules généralistes de calcul (optimisation, simulation,…).
- Encore et toujours poser non seulement les bonnes questions mais aussi voir si on peut les traduire en terme de modèles proba/stat opérationnels.
- Ne pas hésiter à consulter.