Estimation de paramètres, Incertitude des estimations, Sélection et validation de modèles

Samuel Soubeyrand
Unité Biostatistique et Processus Spatiaux
INRA PACA

Formation Modélisation ECCOREV
25–26 septembre 2014
CEREGE, Aix-en-Provence

Modélisation et statistique : introduction informelle

Un modèle

- est construit sur la base d'un ensemble d'hypothèses
- pour représenter le fonctionnement d'un système
  $\rightarrow$ continuum de représentations : de mécaniste à empirique, de stochastique à déterministe...

Usages d'un modèle :

- étude des propriétés du système modélisé
- simulation / prédiction
- estimation des **composantes inconnues** du modèle en "l'ajustant" à des données récoltées lors d'un suivi du phénomène étudié
  $\rightarrow$ problèmes inverses, **inférence statistique**...

Les inconnues

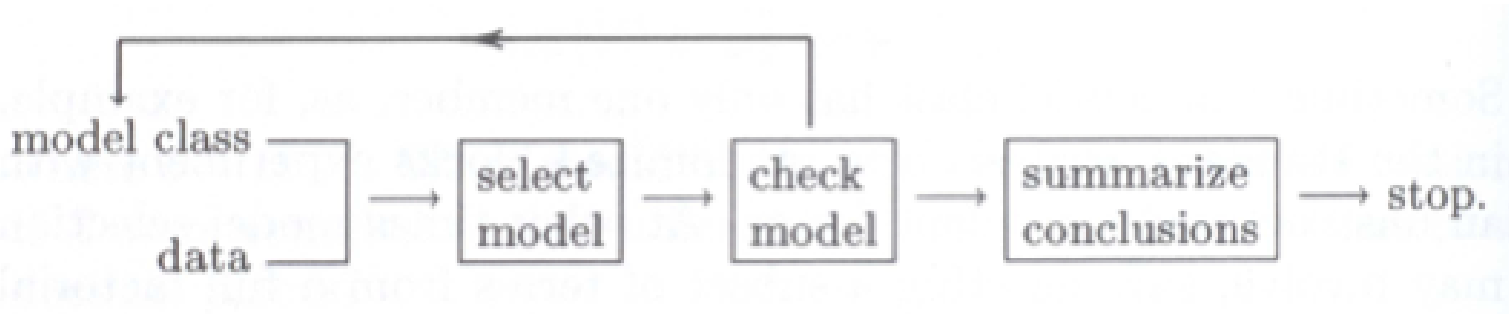
Nature des inconnues θ intervenant dans l'écriture d'un modèle $\mathcal{M}_\theta$

- Des paramètres
  ex. : coefficient d'une régression, taux de reproduction
- Des fonctions (paramètres de dimension infinie)
  ex. : paramètre de dispersion en fonction de la position dans l'espace
- Des variables latentes (ou cachées)
  ex. : temps d'infection d'individus observés infectés seulement après la période d'incubation

Estimer les inconnues d'un modèle

- θ appartient à un espace Θ non réduit à un singleton
- A disposition : des données supposées avoir été générées par un modèle contenu dans la classe de modèles $\{\mathcal{M}_\theta : \theta \in \Theta\}$
- Estimer les paramètres : évaluer sous quel(s) $\theta_0 \in \Theta$ ont été obtenues les données

Processus de l'analyse statistique basée sur un modèle



- *select model* : estimation et sélection
- Les éclairages donnés par cette présentation :
  - Quelques méthodes pour l'estimation de paramètres
  - Mesure de l'incertitude des estimations
  - L'importance des estimations jointes de paramètres
  - Evaluation d'une méthode d'estimation
  - Sélection de modèles
  - Validation d'un modèle

Modélisation stochastique

$$y = \mathcal{M}_\theta(x, \epsilon)$$

où

- y : variable(s) réponse(s)
- x : variable(s) explicative(s)
- ϵ : aléa (car $x \mapsto \mathcal{M}_\theta(x, 0)$ ne capture pas toute la variabilité du processus étudié)
- par convention, 0 est vu comme la valeur référence de ϵ

Objectif : estimer θ à partir d'observations multiples de (x, y), i.e. un échantillon $\{(x_i, y_i) : i = 1, \ldots, I\}$

- ! ! Les ϵ_i ne sont pas observés ! !
- L'estimation est facilitée lorsque les variables conditionnelles $y_i \mid x_i$ sont indépendantes (i.e. les ϵ_i sont indépendants)

Exemples de modèles

- Modèle linéaire : $y = x'\theta + \epsilon$

Exemples de modèles

- Modèle linéaire : $y = x'\theta + \epsilon$
- Modèle linéaire généralisé : $y \sim Poisson(e^{x'\theta})$

Exemples de modèles

- Modèle linéaire : $y = x'\theta + \epsilon$
- Modèle linéaire généralisé : $y \sim \mathit{Poisson}(e^{x'\theta})$
- Modèle non-linéaire généralisé mixte :

$$\pi_t = \text{logit}^{-1}\{x'\theta + \epsilon\} \qquad \epsilon \sim \mathit{Normale}(0, \sigma^2)$$

$$y \sim \mathit{Binomiale}\left(n, \prod_{t=1}^{T} \pi_t\right)$$

Exemples de modèles

- Modèle linéaire : $y = x'\theta + \epsilon$
- Modèle linéaire généralisé : $y \sim \mathit{Poisson}(e^{x'\theta})$
- Modèle non-linéaire généralisé mixte :

$$\pi_t = \text{logit}^{-1}\{x'\theta + \epsilon\} \qquad \epsilon \sim \mathit{Normale}(0, \sigma^2)$$

$$y \sim \mathit{Binomiale}\left(n, \prod_{t=1}^{T} \pi_t\right)$$

- Modèle mécanistico-statistique :

$$\frac{\partial u}{\partial t} = D\,\Delta u + u(r - \gamma u) \qquad + \text{ cond. initiales et aux bords}$$

$$y \sim \mathit{Poisson}\left(\alpha \int_{\omega} u(\tau, x)\, dx\right)$$

Exemples de modèles

- Modèle linéaire : $y = x'\theta + \epsilon$
- Modèle linéaire généralisé : $y \sim \mathit{Poisson}(e^{x'\theta})$
- Modèle non-linéaire généralisé mixte :

$$\pi_t = \text{logit}^{-1}\{x'\theta + \epsilon\} \qquad \epsilon \sim \mathit{Normale}(0, \sigma^2)$$

$$y \sim \mathit{Binomiale}\left(n, \prod_{t=1}^{T} \pi_t\right)$$

- Modèle mécanistico-statistique :

$$\frac{\partial u}{\partial t} = D\,\Delta u + u(r - \gamma u) \qquad + \text{ cond. initiales et aux bords}$$

$$y \sim \mathit{Poisson}\left(\alpha \int_{\omega} u(\tau, x)\, dx\right)$$

- Modèle hiérarchique spatial multi-type :

$$\lambda(x) = \exp\{x'\theta + \epsilon(x)\} \qquad y^{(1)} \sim \mathit{Bernoulli}\left(1 - e^{-\int_{\omega^{(1)}} \lambda(x) dx}\right)$$

$$(\epsilon(x_1), \ldots, \epsilon(x_n)) \sim \mathit{Normale}_n(\mathbf{0}, \Sigma) \qquad y^{(2)} \sim \mathit{Poisson}\left(\int_{\omega^{(2)}} \lambda(x) dx\right)$$

Estimation via les moindres carrés ordinaires

- La plus intuitive des méthodes d'estimation : estimer θ en minimisant la somme des carrés des écarts entre les y_i et les $\hat{y}_{i,\theta} = \mathcal{M}_\theta(x_i, 0)$:

$$\hat{\theta} = \text{argmin}_{\theta \in \Theta} \sum_{i=1}^{I} \{y_i - \mathcal{M}_\theta(x_i, 0)\}^2$$

- Si $\mathcal{M}_\theta$ est un modèle linéaire ($y = x'\theta + \epsilon$), $\hat{\theta} = (\mathcal{X}'\mathcal{X})^{-1}\mathcal{X}'\mathbf{Y}$
- Dans le cas général, minimisation via un algorithme numérique itératif (Gauss-Newton, Nelder-Mead, recuit simulé...)
- On obtient une estimation ponctuelle... mais on peut dire plus

Estimation via le maximum de vraisemblance

- Nécessité d'ajouter une hypothèse distributionnelle sur l'aléa ϵ
  ex. : $\epsilon \sim \mathit{Normale}(0, \sigma^2)$

- On peut alors écrire la distribution conditionnelle jointe de $(y_1, \ldots, y_I)$ sachant $(x_1, \ldots, x_I)$
  - Cette distribution dépend de θ
  - Elle coincide avec la vraisemblance

- Si les variables conditionnelles $y_i \mid x_i$ sont indépendantes :

$$L\{\theta; (y_i, x_i) : i = 1, \ldots, I\} = \prod_{i=1}^{I} f_{Y|X}(y_i|x_i; \theta)$$

où $f_{Y|X}$ est la (densité de) probabilité de $y \mid x$ sous $\mathcal{M}_\theta$

Estimation via le maximum de vraisemblance

- Estimateur du maximum de vraisemblance :

$$\hat{\theta} = \text{argmax}_{\theta \in \Theta} L\{\theta; (y_i, x_i) : i = 1, \ldots, I\}$$

- Expression analytique de $\hat{\theta}$ pour certains modèles

- Dans le cas général, maximisation via un algorithme numérique itératif

  - Gauss-Newton, Nelder-Mead, recuit simulé
  - EM : *Expectation-Maximization*
  - MCEM : *Monte Carlo Expectation-Maximization*
  - ...

- On obtient une estimation ponctuelle... mais on peut dire plus

Généralisation : estimation par minimisation d'un contraste

- Supposons que les données $\{(x_i, y_i) : i = 1, \ldots, I\}$ ont été générées sous le modèle $\mathcal{M}_{\theta_0}$
- Un contraste est une fonction aléatoire $\theta \mapsto U_I(\theta)$
  - définie sur Θ
  - convergeant en probabilité, quand $I \to \infty$, vers $\theta \mapsto K(\theta, \theta_0)$ qui atteint un minimum strict quand $\theta = \theta_0$
- Estimateur du minimum de contraste :

$$\hat{\theta} = \text{argmin}_{\theta \in \Theta} U_I(\theta)$$

- Exemples : moindres carrés ordinaires et généralisés, maximum de vraisemblance pondérée ou non, maximum de pseudo-vraisemblance, méthode des moments...

- On obtient une estimation ponctuelle... mais on peut dire plus

Mesurer l'incertitude de $\hat{\theta}$

... et plus généralement décrire les propriétés de $\hat{\theta}$

- Une estimation ponctuelle ne suffit pas
- Le $\hat{\theta}$ est-il éloigné du θ_0 "vrai" ?
- **Il faut** fournir des mesures de l'incertitude de $\hat{\theta}$
  ex. : variances d'estimation, intervalles de confiance, distributions...
- Les mesures d'incertitude doivent prendre en compte :
  - l'aléa lié à l'échantillonnage
  - la structure probabiliste de l'aléa ϵ

Mesurer l'incertitude de $\hat{\theta}$

- A l'aide d'un argument exact :
  ex. : $y = x'\theta + \epsilon, \quad \epsilon \sim_{indep.} Normale(0, \sigma^2), \quad \sigma^2$ connu
  pour ce modèle linéaire : $\hat{\theta} \sim Normale(\theta_0, \sigma^2(\mathcal{X}'\mathcal{X})^{-1})$

Mesurer l'incertitude de $\hat{\theta}$

- A l'aide d'un argument exact :
  ex. : $y = x'\theta + \epsilon, \quad \epsilon \sim_{indep.} Normale(0, \sigma^2), \quad \sigma^2$ connu
  pour ce modèle linéaire : $\hat{\theta} \sim Normale(\theta_0, \sigma^2(\mathcal{X}'\mathcal{X})^{-1})$

- A l'aide de l'asymptotique (quand $I \to \infty$) :
  Sous des hypothèses de régularité, l'estimateur $\hat{\theta}$ du minimum de contraste est consistent est asymptotiquement normal :
  - $\hat{\theta}$ converge en probabilité vers θ_0
  - $\sqrt{I}(\hat{\theta} - \theta_0)$ converge en loi vers la loi $Normale(0, \mathcal{I}_{\theta_0}^{-1}\Gamma_{\theta_0}\mathcal{I}_{\theta_0}^{-1})$

Mesurer l'incertitude de $\hat{\theta}$

- A l'aide d'un argument exact :
  ex. : $y = x'\theta + \epsilon, \quad \epsilon \sim_{indep.} Normale(0, \sigma^2), \quad \sigma^2$ connu
  pour ce modèle linéaire : $\hat{\theta} \sim Normale(\theta_0, \sigma^2(\mathcal{X}'\mathcal{X})^{-1})$

- A l'aide de l'asymptotique (quand $I \to \infty$) :
  Sous des hypothèses de régularité, l'estimateur $\hat{\theta}$ du minimum de contraste est consistent est asymptotiquement normal :
  - $\hat{\theta}$ converge en probabilité vers θ_0
  - $\sqrt{I}(\hat{\theta} - \theta_0)$ converge en loi vers la loi $Normale(0, \mathcal{I}_{\theta_0}^{-1}\Gamma_{\theta_0}\mathcal{I}_{\theta_0}^{-1})$

- A l'aide de simulations (bootstrap (non-)paramétrique) :
  1. Simuler B jeux de données
     - soit sous le modèle $\mathcal{M}_{\hat{\theta}}$
     - soit par ré-échantillonnage des données
  2. Estimer θ pour les B jeux de données
  3. Utiliser les B estimations de θ pour caractériser la loi de $\hat{\theta}$

Estimation bayésienne

- Les paramètres θ sont vus comme des variables aléatoires ayant une distribution de probabilité *a priori* π
- Ce que l'on cherche à quantifier en bayésien :
  - la distribution *a posteriori*

$$f(\theta \mid Y) = \frac{f(Y \mid \theta)\pi(\theta)}{\int_\Theta f(Y \mid \alpha)\pi(\alpha)d\alpha}$$

  - et ses caractéristiques : moments a posteriori, maximum a posteriori, quantiles a posteriori, intervalles de crédibilité...
- D'un point de vue fréquentiste : on obtient à la fois des estimations ponctuelles et des mesures d'incertitude

Estimation bayésienne

- Exemple : Nombre de succès sur n essais indépendants
  - modèle binomial : $Y \mid \theta \sim \text{Binomiale}(n, \theta)$
  - prior beta : $\theta \sim \text{Beta}(a, b)$
  - posterior [1] : $\theta \mid Y \sim \text{Beta}(a + Y, b - n + Y)$

- Mais la loi a posteriori peut être difficilement calculable
  - Modèle avec nombreux paramètres et variables latentes :

$$f(\theta_1, \ldots, \theta_K \mid Y) = \frac{f(Y \mid \theta_1, \ldots, \theta_K)\pi(\theta_1, \ldots, \theta_K)}{\int_{\Theta_1} \cdots \int_{\Theta_K} f(Y \mid \theta_1, \ldots, \theta_K)\pi(\alpha_1, \ldots, \alpha_K)d\alpha_1 \cdots d\alpha_K}$$

  - Les intégrales multiples (grande dimension) rendent difficile le calcul de la posterior jointe $f(\theta_1, \ldots, \theta_K \mid Y)$, des posteriors marginales $f(\theta_k \mid Y)$, des moments a posteriori $E(\theta_i^q \mid Y)$...

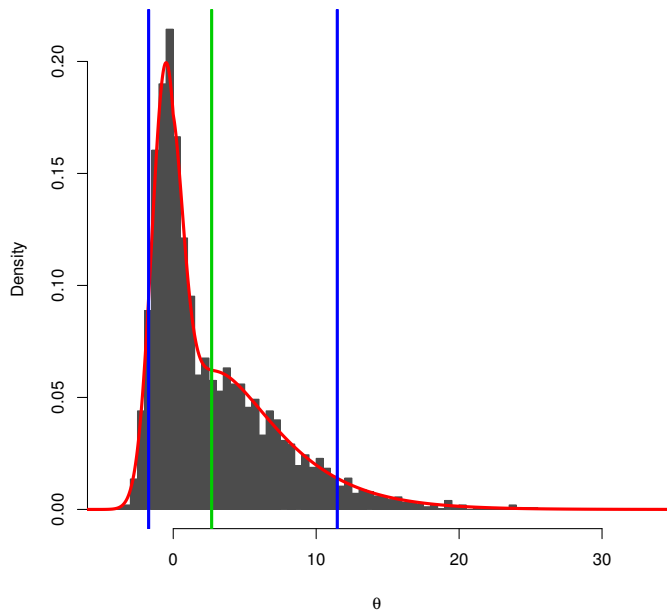
1. Détail du calcul :

$$f(\theta \mid Y) = \frac{C_n^Y \theta^Y (1-\theta)^{n-Y} \frac{\theta^{a-1}(1-\theta)^{b-1}}{B(a,b)}}{\int_0^1 C_n^Y \alpha^Y (1-\alpha)^{n-Y} \frac{\alpha^{a-1}(1-\alpha)^{b-1}}{B(a,b)} d\alpha} = \frac{\theta^{(a+Y)-1}(1-\theta)^{(b+n-Y)-1}}{B(a+Y, b+n-Y)}$$

Estimation bayésienne

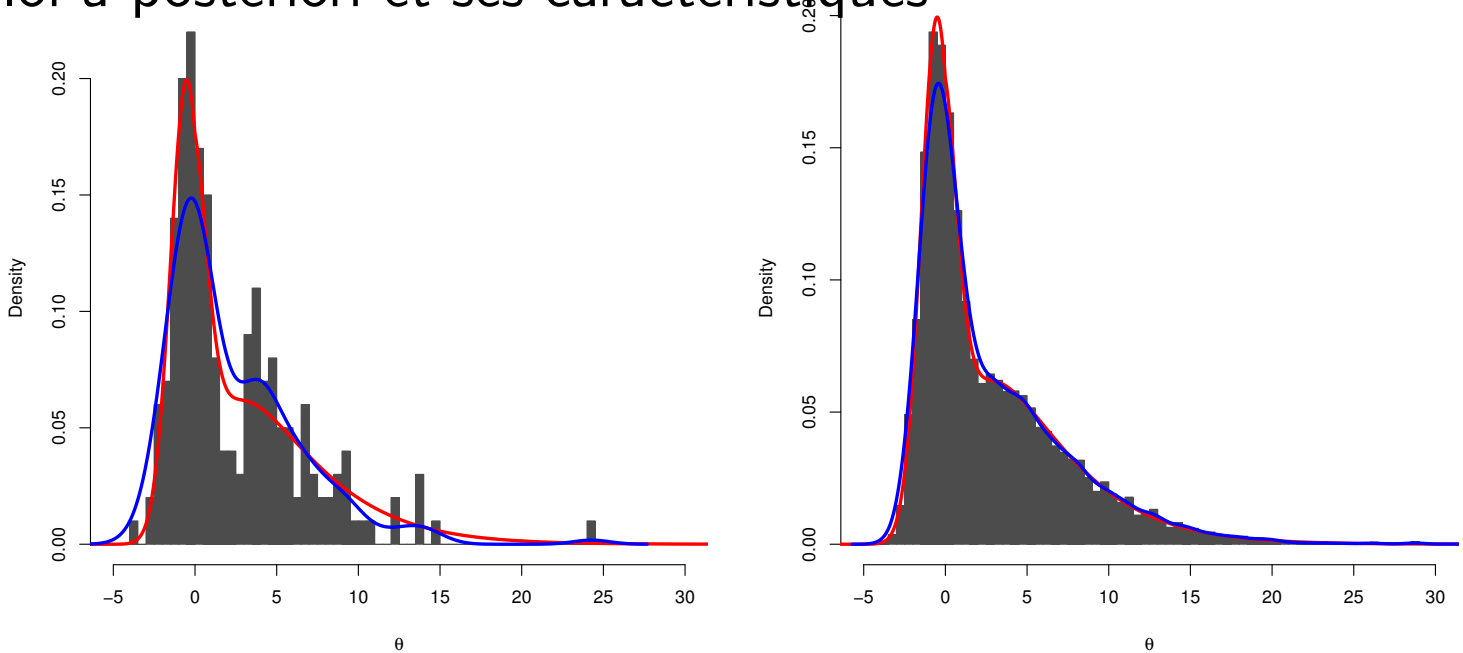
- Ce que peuvent les méthodes numériques (ré-échantillonnage d'importance, MCMC, *Population Monte Carlo*...) :
  - générer un échantillon issu de la loi a posteriori
  - sans passer par le calcul d'intégrales multiples

- A quoi sert cet échantillon ?
  - connaître intimement la loi a posteriori
  - estimer la densité a posteriori
  - estimer les moments a posteriori
    ex : $E(\theta \mid Y) \approx \frac{1}{I} \sum_{i=1}^{I} \theta^{(i)}$
  - estimer des intervalles de crédibilité
  - ...



Estimation bayésienne

- Générer un grand échantillon pour correctement approcher la loi a posteriori et ses caractéristiques



Echantillons de taille 200 (histo gauche) et 10000 (histo droite)

Densités a posteriori vraies (rouge) et estimées (bleu)

ABC – *Approximate Bayesian Computation*

Dans des disciplines comme la biologie évolutive, la dynamique de populations, l'épidémiologie

- On construit des modèles décrivant des scénarios complexes (structures de dépendance spatio-temporelle complexes)
- Dont les lois de probabilité ne peuvent être écrites de manière concise (vraisemblance incalculable en pratique)
- Mais qui peuvent être utilisés pour simuler des données

L'ABC permet d'estimer les paramètres de ces *modèles stochastiques implicites*

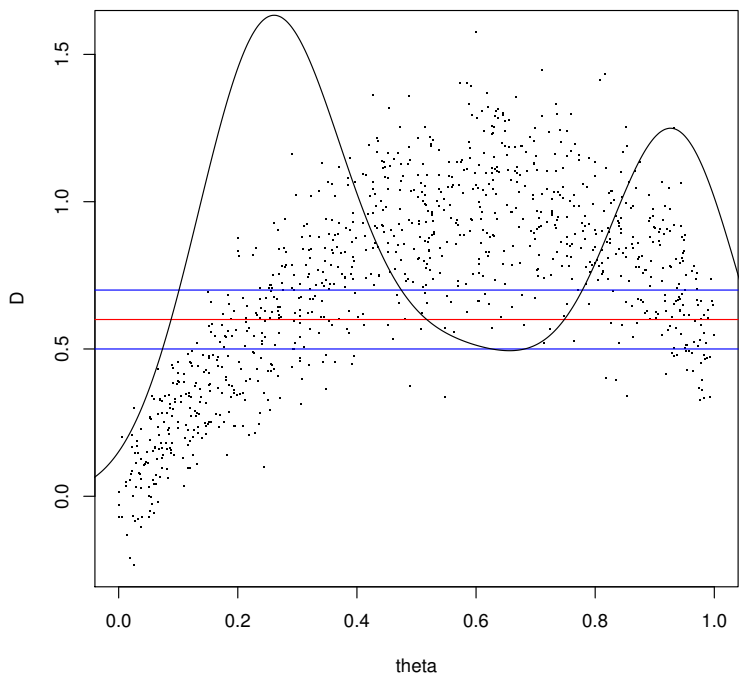
ABC – Principe

Notations :

- $\mathcal{D}$: jeu de données observé
- θ : paramètres du modèle
- π : loi a priori des paramètres
- $\mathcal{M}$: Modèle paramétré par θ sensé générer les données $\mathcal{D}$

Procédure :

- Simuler des jeux de paramètres θ_i, $i = 1, \ldots, I$, sous la loi a priori π
- Pour chaque jeu de paramètres θ_i, simuler un jeu de données $\mathcal{D}_i$
- Retenir les $\mathcal{D}_i$ "ressemblant" à $\mathcal{D}$ pour construire la loi a posteriori de θ



ABC – Le rôle des statistiques résumées

- Dernière étape de la procédure ABC :
  - Retenir les $\mathcal{D}_i$ "ressemblant" à $\mathcal{D}$ pour construire la loi a posteriori de θ

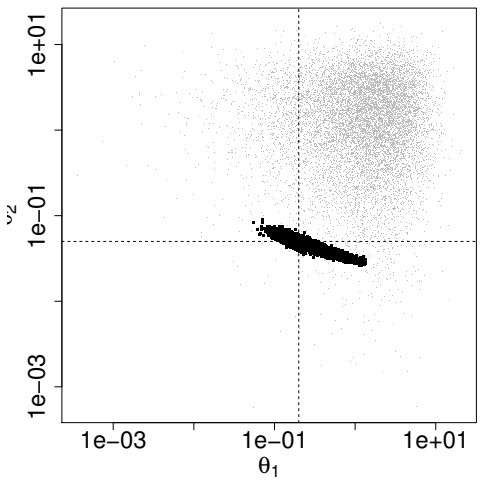
- Problème quand les $\mathcal{D}_i$ et $\mathcal{D}$ sont de grande dimension et dans des espaces continus

- Dans l'ABC, on se sert de statistiques résumées $S = s(\mathcal{D})$ et $S_i = s(\mathcal{D}_i)$ pour comparer les jeux de données $\mathcal{D}_i$ et $\mathcal{D}$

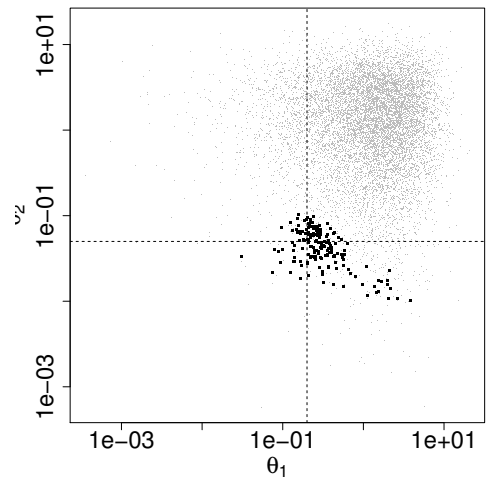
- On obtient un échantillon tiré approximativement dans la loi *a posteriori* $f(\theta \mid S)$

ABC – Attention aux interprétations

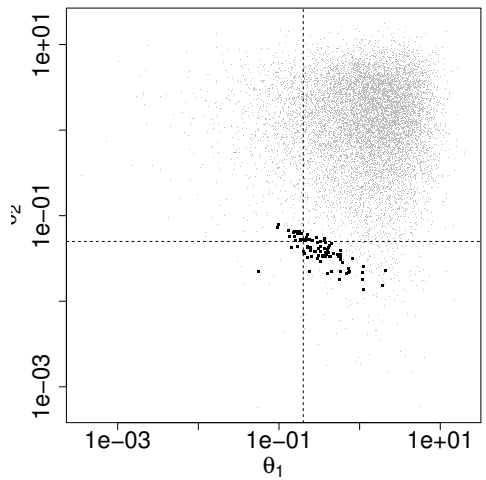
- L'ABC fournit un échantillon tiré approximativement dans la loi *a posteriori* $f(\theta \mid S)$
- Dans le cas général, cet échantillon ne coincide pas avec un échantillon tiré dans la loi *a posteriori* classique
- L'utilisation de statistiques résumées a transformé l'information contenue dans les données



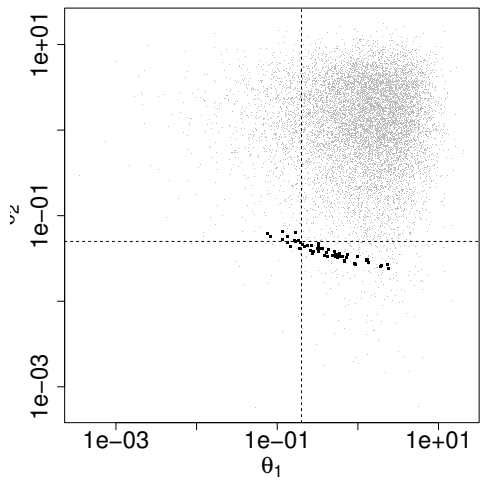
MCMC



ABC(S_1)



ABC(S_2)



ABC(S_3)

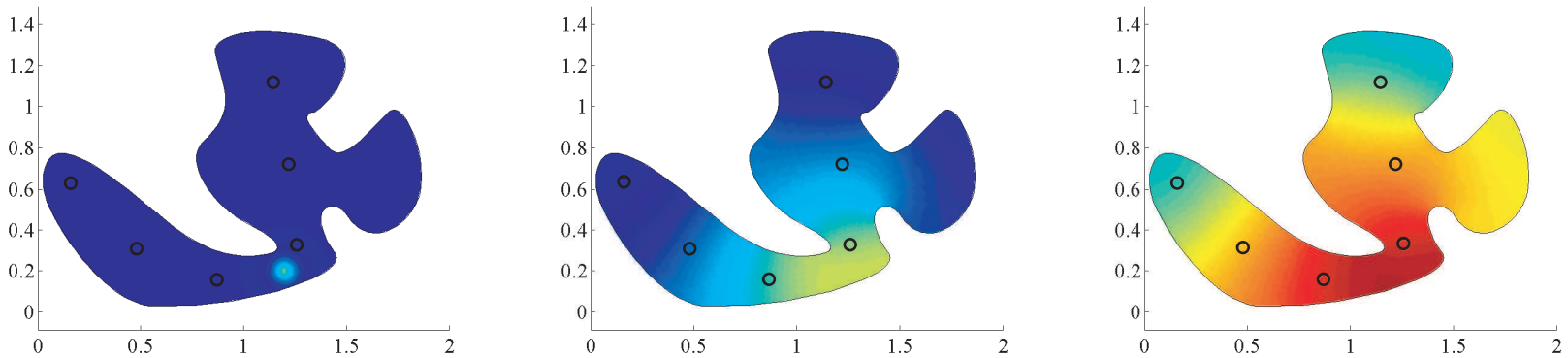
Révolution statistique liée aux algorithmes numériques

- L'ère de l'analytique :
  - Fournir une expression analytique d'un estimateur du maximum de vraisemblance ou d'une loi a posteriori contraint à utiliser des modèles relativement simples et en nombre limité
- L'ère du numérique :
  - La classe des modèles "ajustables" aux données est en hyper-expansion

L'approche mécanistico-statistique à l'ère du numérique

Cadre d'étude :

$t = -2$ $t = 0$ $t = 0.9$



- Exemples de phénomènes d'intérêt : dynamique d'invasion, dynamique épidémique...
- On dispose de données issues d'un suivi spatio-temporel de la population d'intérêt : données bruitées, dégradées (binarisation), agrégées, non-exhaustives, censurées...
  ex. : mesures de dégâts
- On dispose d'un modèle mécaniste représentant la dynamique
  ex. : EDP, EDS, IBM, modèle de dispersion stochastique...

L'approche mécanistico-statistique à l'ère du numérique

- Les données (bruitées, dégradées, agrégées, non-exhaustives, censurées...) sont des fonctions aléatoires de l'évolution spatio-temporelle de la densité réelle de la population

- Construction d'un modèle du processus d'observation
  - conditionnel au modèle mécaniste
  - décrivant le lien stochastique entre densité de population et données

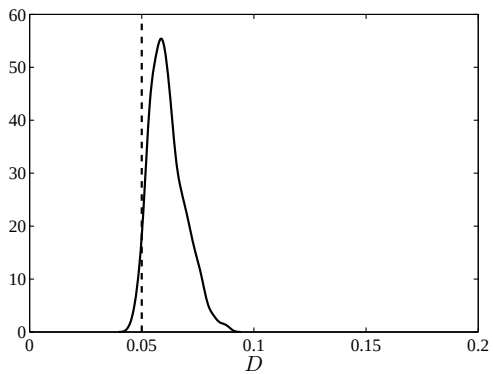
- Structure hiérarchique du modèle (state-space model) :
  - Modèle du processus d'observation
  - Modèle mécaniste de la dynamique
  - Modèle a priori pour les paramètres (si inférence bayésienne)

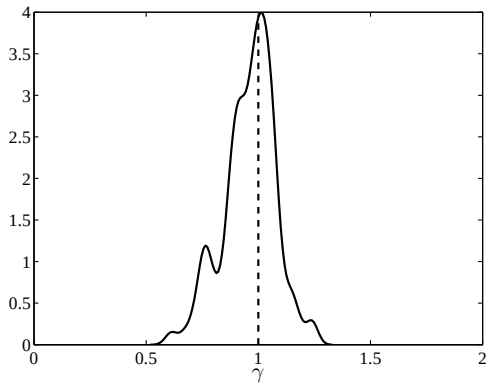
- Estimation des paramètres à l'aide des méthodes décrites précédemment

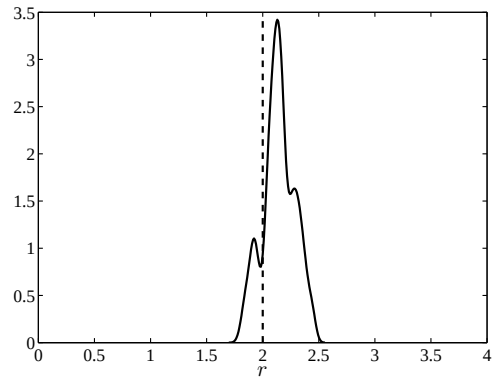
L'approche mécanistico-statistique à l'ère du numérique

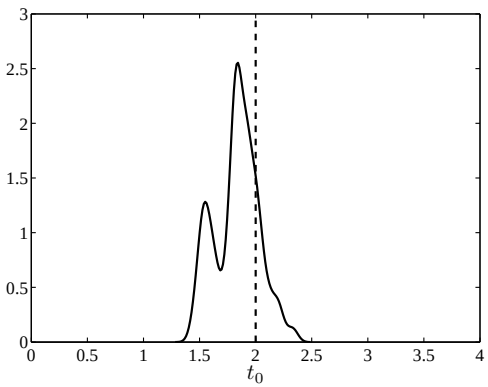
Estimation de la date et du lieu d'introduction d'une espèce invasive, et autres paramètres

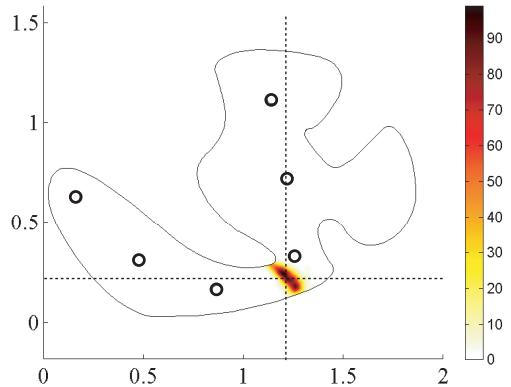
Distributions a posteriori marginales :











L'importance des estimations jointes de paramètres

- Etude du lien entre taille et survie des gambusies
- Lois a posteriori bivariées :

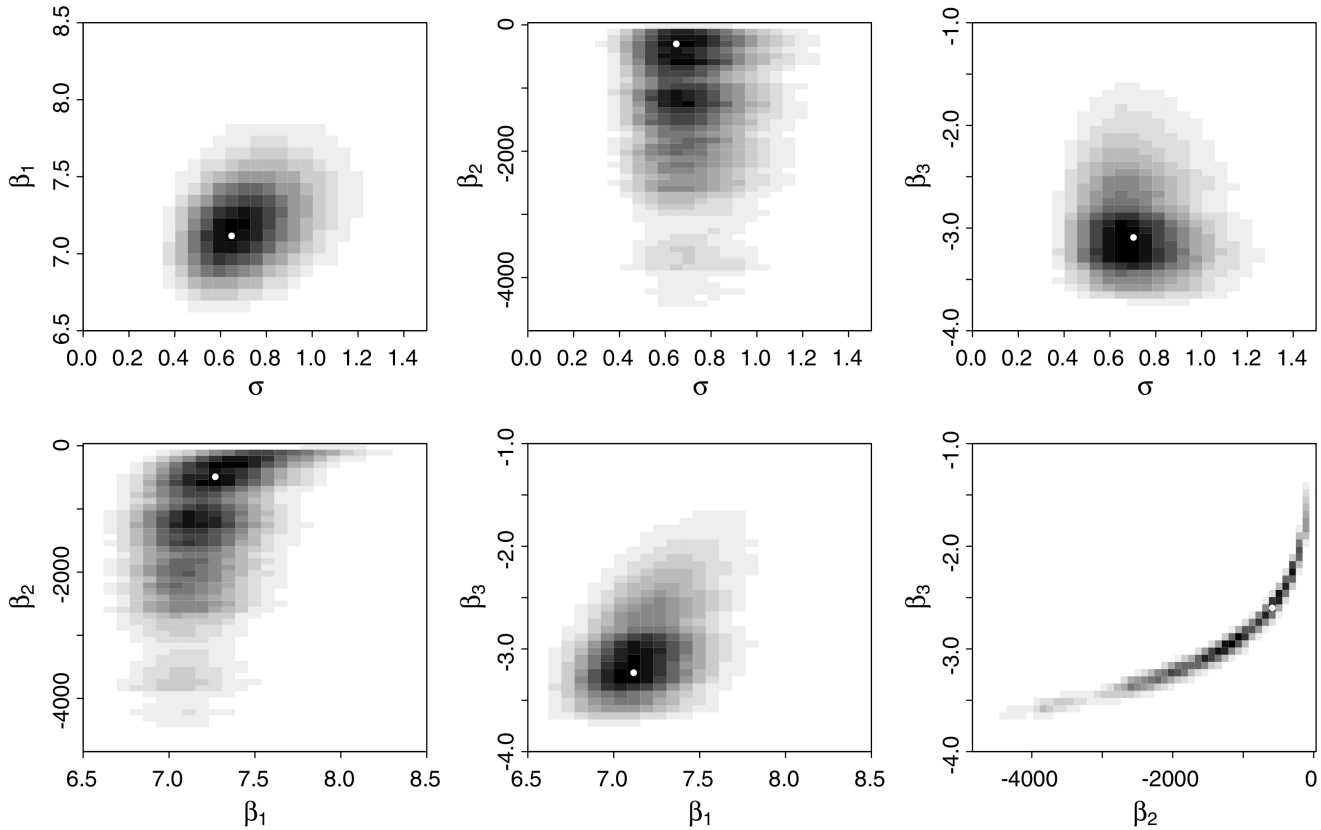


Fig. 2. Bivariate histograms representing the two-dimensional posterior distributions of the standard deviation σ and the regression parameters β_1, β_2 and β_3 (high densities are in black). White dots correspond to the posterior modes.

- **Il faut** prendre en compte les dépendances entre les estimations des paramètres pour faire de la prédiction

Evaluation d'une méthode d'estimation

- Cas favorables : on dispose d'une théorie exacte pour mesurer l'efficacité d'une méthode d'estimation
- Autres cas (même lorsqu'il existe une théorie asymptotique) : il peut être judicieux de vérifier l'efficacité des méthodes d'estimation
- Problèmes possibles : biais d'estimation, maximisation numérique de la vraisemblance inefficace, MCMC non convergeant...
- L'approche simulatoire :
  - Estimation pour un grand nombre de jeux de données simulés
  - Définition un plan d'expérience pour les simulations
  - Simulations sous un modèle différent de $\mathcal{M}_\theta$ pour tester la robustesse de la méthode d'estimation
  - Critères : MSE-like, taux de couvertures des intervalles de confiances...

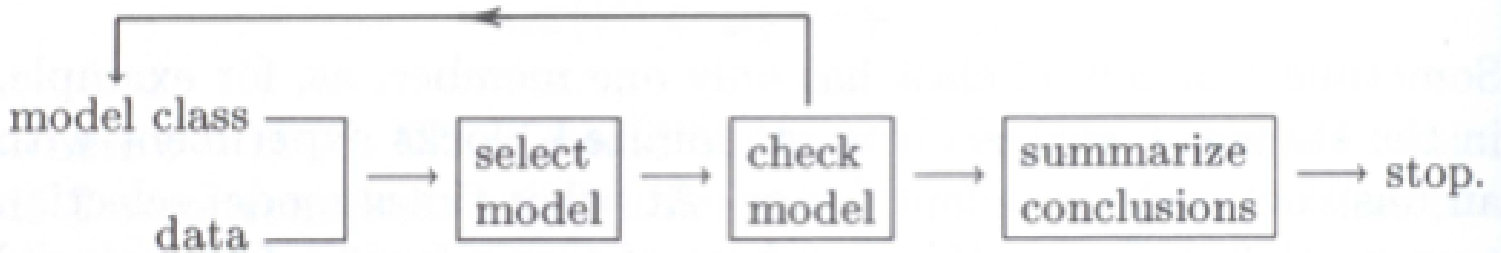
Sélection de modèles

- Question : quel est le modèle le plus approprié pour représenter les données ?
- Ici les structures des modèles considérés sont différentes
- Nombreuses approches :
  - AIC, QAIC, BIC, DIC, Facteurs de Bayes
  - Critères de prédiction calculés par validation croisée
  - ...

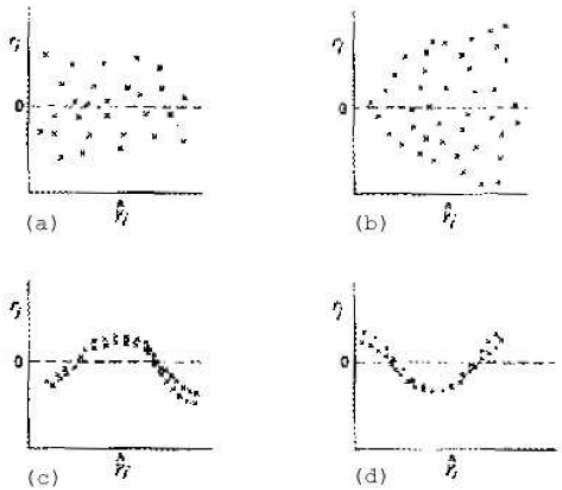
- Exemple d'alternative à la sélection : dans le cas de la prédiction, *(Bayesian) model averaging*
  $\rightarrow$ les modèles sont pondérés pour construire une prédiction

Validation de modèle

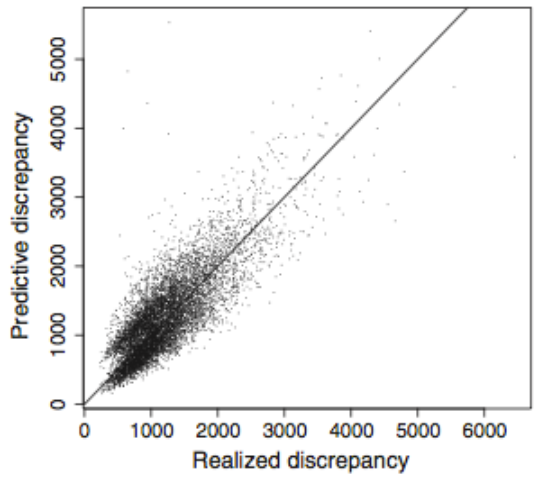
- Le modèle sélectionné est le “meilleur” des modèles... dans la classe des modèles qui a été considérée
- Mais représente-t-il bien les données ?



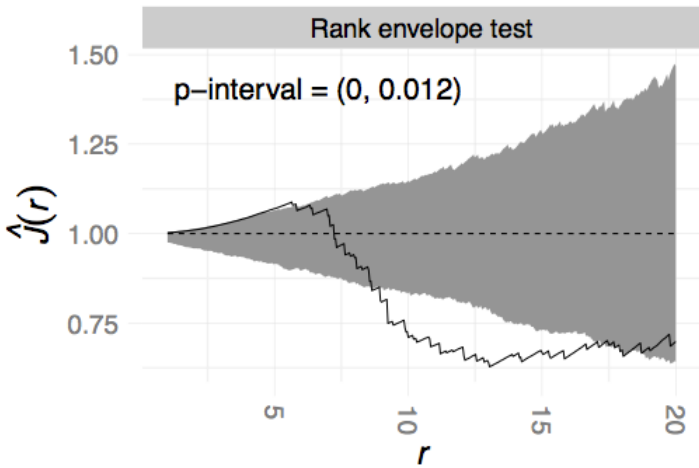
- Nombreuses approches :



analyse des résidus



posterior predictive check



envelope tests

Message final

- La statistique d'aujourd'hui permet d'ajuster aux données des modèles proches des mécanismes
  $\rightarrow$ approche mécanistico-statistique

- Mais il est judicieux d'évaluer la méthode d'estimation utilisée