Le calcul bayésien approché

Notions de base et ouvertures

Samuel Soubeyrand

19 mars 2010

1. Introduction

Qu'est-ce que le calcul bayésien approché (ABC) ?

- C'est un ensemble de méthodes d'estimation de paramètres
- L'estimation des paramètres est faite dans une perspective bayésienne : on fournit des lois a posteriori
- Des simulations du modèle utilisé permettent l'inférence

L'ABC, en tant que méthode d'inférence bayésienne, est soumis au débat entre Bayésiens et Classiques (et le nourrit)

Les modèles stochastiques implicites

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

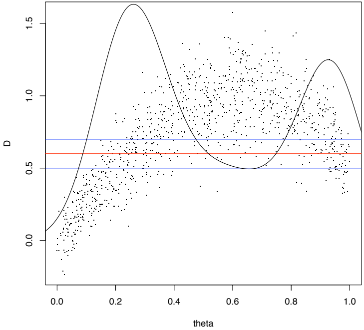
Principe de l'ABC

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



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

Pb 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 pour comparer les jeux de données $\mathcal{D}_i$ et $\mathcal{D}$

Notations :

- $S = s(\mathcal{D})$: statistiques résumées observées
- $S_i = s(\mathcal{D}_i)$: statistiques résumées pour les paramètres simulés θ_i

Exemple 1 : Estimer les paramètres d'un modèle de coalescence

- Coalescence : réunion des lignées phylogénétiques pour arriver à un plus proche ancêtre commun
- Données pour l'inférence : séquences génétiques ou marqueurs microsatellites pour les populations contemporaines
- Objectifs de l'inférence : tailles efficaces des populations, temps de divergence, taux de mutation, migration et recombinaison, topologie de l'arbre...
- Structure de dépendance complexe

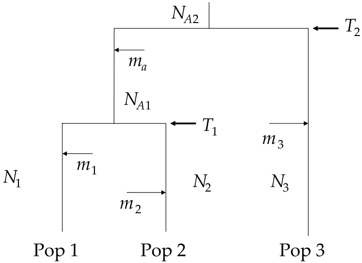
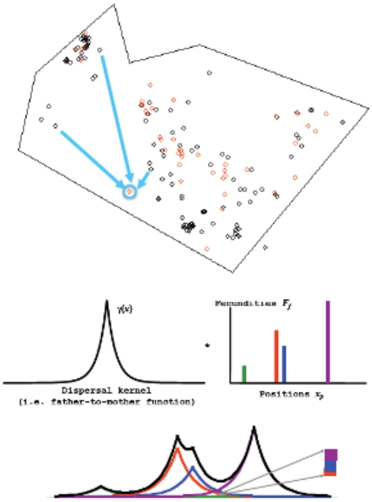


Figure 14.2. *Illustration of the three-population model. See text for description of parameters.*

Extrait de Beaumont et al (2008)

Exemple 2 : Estimer la fonction de dispersion du pollen

- Dispersion du pollen : composante majeure des flux de gènes chez les végétaux
- Données pour l'inférence : Positions et génotypes de certaines mères, génotypes de leurs graines, positions de certains pères, variation de fertilités
- Objectifs de l'inférence : paramètres de la fonction de dispersion...
- Structure de dépendance complexe



Extraits de Carpentier et al (2009)

Plan

1. Introduction
2. **Quelques algorithmes de calcul bayésien approché**
3. L'émergence de l'ABC
4. Travaux menés par BioSP sur l'ABC

Quelques algorithmes ABC

- Acceptation/rejet pour des stats discrètes
- Acceptation/rejet pour stats continues
- ABC avec régression linéaire locale
- ABC MCMC

Algorithme 1 : Acceptation/rejet pour des stats discrètes

(Fu and Lin, 1997)

Algorithme :

- Tirer indépendamment θ_i ($i = 1, \ldots, I$) dans π
- Simuler $\mathcal{D}_i$ sous $\mathcal{M}_{\theta_i}$ et calculer S_i
- Accepter θ_i si $S_i = S$

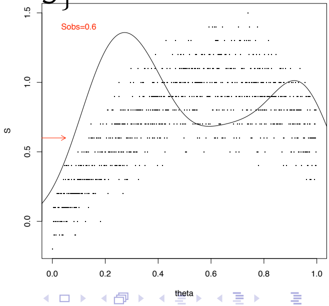
Ensemble des points acceptés pour construire la loi a posteriori :

$$\{\theta_i : i = 1, \ldots, I \text{ et } S_i = S\}$$

Estimation de la loi a posteriori :

$$\hat{p}(\theta \mid S) = \frac{\sum_{i=1}^{I} \delta_S(S_i)\delta_\theta(\theta_i)}{\sum_{i=1}^{I} \delta_S(S_i)}$$

converge vers $p(\theta \mid S) = \frac{p(S|\theta)\pi(\theta)}{p(S)}$



Algorithme 2 : Acceptation/rejet pour des stats continues

(Weiss and Von Haeseler, 1998)

Algorithme :

- Tirer indépendamment θ_i $(i = 1, \ldots, I)$ dans π
- Simuler $\mathcal{D}_i$ sous $\mathcal{M}_{\theta_i}$ et calculer S_i
- Accepter θ_i si $d(S_i, S) \leq \epsilon$ avec $\epsilon > 0$

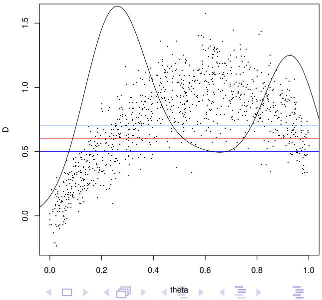
Ensemble des points acceptés pour construire la loi a posteriori :

$$\{\theta_i : i = 1, \ldots, I \text{ et } d(S_i, S) \leq \epsilon\}$$

Estimation de la loi a posteriori :

$$\hat{p}(\theta \mid S) = \frac{\sum_{i=1}^{I} \delta_{B(S,\epsilon)}(S_i)\delta_\theta(\theta_i)}{\sum_{i=1}^{I} \delta_{B(S,\epsilon)}(S_i)}$$

converge vers $p(\theta \mid d(S_i, S) \leq \epsilon) = \frac{p\{d(S_i,S)\leq\epsilon|\theta\}\pi(\theta)}{p\{d(S_i,S)\leq\epsilon\}}$



Algorithme 3 : ABC avec régression linéaire locale (Beaumont et al, 2003)

- Tirer $\theta_i \sim_{\text{iid}} \pi$, $\mathcal{D}_i \sim_{\text{indép.}} \mathcal{M}_{\theta_i}$, calculer $S_i = s(\mathcal{D}_i)$
- Si $\theta_i = \alpha + (S_i - S)\beta + \varepsilon_i$, alors on peut construire la loi a posteriori avec toutes les simulations en utilisant l'ensemble suivant de points : $\{\theta_i^* = \theta_i - (S_i - S)\hat{\beta} : i = 1, \dots, I\}$
- Si localement (autour de S) $\theta_i = \alpha + (S_i - S)\beta + \varepsilon_i$, alors on peut construire la loi a posteriori avec l'ensemble suivant de points : $\{\theta_i^* = \theta_i - (S_i - S)\hat{\beta} : i = 1, \dots, I\}$ où $\hat{\beta}$ est obtenu par régression linéaire locale en accordant plus de poids aux points tels que $d(S_i, S)$ est petit
  Est alors proposée l'estimation suivante de la loi a posteriori :

$$\hat{p}(\theta \mid S) = \frac{\sum_{i=1}^{I} K_{h_1}(d(S_i, S)) K_{h_2}(||\theta_i^* - \theta||)}{\sum_{i=1}^{I} K_{h_1}(d(S_i, S))}$$

Algorithme 4 : ABC-MCMC (Marjoram et al, 2003)

Iteration $i+1$ de l'algorithme :

- Tirer un θ^* candidat selon la loi de proposition $q(\theta^* \mid \theta_i)$
- Simuler $\mathcal{D}^*$ sous $\mathcal{M}_{\theta^*}$ et calculer $S^* = s(\mathcal{D}^*)$
- Si $d(S^*, S) \leq \epsilon$ et $U(0,1) \leq \min\left(1, \frac{\pi(\theta^*)q(\theta_i|\theta^*)}{\pi(\theta_i)q(\theta^*|\theta_i)}\right)$ alors $\theta_{i+1} = \theta^*$
- Sinon $\theta_{i+1} = \theta_i$

A l'état stationnaire on échantillonne dans la loi

$$p(\theta \mid d(S_i, S) \leq \epsilon) = \frac{p\{d(S_i, S) \leq \epsilon \mid \theta\}\pi(\theta)}{p\{d(S_i, S) \leq \epsilon\}}$$

Améliorations et autres algorithmes ABC (non exhaustif !)

Une littérature foisonante :

- ABC avec régression non-linéaire hétéroscédastique (Blum and François, 2009)
- ABC MCMC avec relaxation de la tolérence dans le MCMC (Wegmann et al., 2009)
- ABC PMC utilisant l'échantillonnage d'importance (Beaumont et al., 2009)
- ABC et sélection de modèles (Leuenberger and Wegmann, 2010 ; Toni et al., 2010)
- ABC et sélection des statistiques en utilisant l'ε-exhaustivité relative (Joyce and Marjoram, 2008)

Mais attention à ne pas aller trop vite :

- Ex : l'algorithme ABC PRC de Sisson et al (2007) est biaisé (cf. Beaumont et al, 2008)

Limites des algorithmes actuels

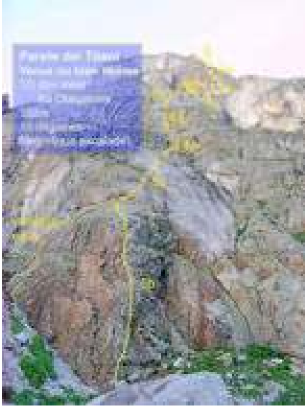
- Choix de la distance entre statistiques $d(S_i, S)$ pour comparer les jeux de données
- Trouver des statistiques informatives (tendre vers l'exhaustivité)
- Sélection des statistiques $S = s(\mathcal{D})$ car fléau de la dimension et problème de redondance d'information
- Sélection du seuil de tolérence $(d(S_i, S) \leq \epsilon)$ ou d'autres paramètres de tuning
- On estime $p(\theta \mid S)$ et pas $p(\theta \mid \mathcal{D})$
- Modèles simulés lentement qui limitent le nombre d'itérations réalisables

Plan

1. Introduction
2. Quelques algorithmes de calcul bayésien approché
3. **L'émergence de l'ABC**
4. Travaux menés par BioSP sur l'ABC

L'émergence de l'ABC

- La voie évolutioniste
- La voie statisticienne
- La jonction des voies



La voie évolutioniste

- Pourquoi cette voie ? Parce que les scientifiques travaillant en théorie de l'évolution ont construit des modèles stochastiques implicites dont ils voulaient estimer les inconnues
- Les tenants évolutionistes de l'ABC font remonter cet ensemble de méthodes à 1997

La voie évolutioniste

- Les tenants évolutionistes de l'ABC font remonter cet ensemble de méthodes à 1997
  - 1997 - Tavaré et al. - Utilise la vraisemblance calculable de statistiques résumées plutôt que la vraisemblance incalculable du jeu de données pour inférer un temps de coalescence

La voie évolutioniste

- Les tenants évolutionistes de l'ABC font remonter cet ensemble de méthodes à 1997
  - 1997 - Tavaré et al. - Utilise la vraisemblance calculable de statistiques résumées plutôt que la vraisemblance incalculable du jeu de données pour inférer un temps de coalescence
  - 1997 - Fu and Li - Approxime par Monte-Carlo la vraisemblance d'une statistique discrète sous un paramètre discret pour inférer l'âge d'un plus proche ancêtre commun

La voie évolutioniste

- Les tenants évolutionistes de l'ABC font remonter cet ensemble de méthodes à 1997
  - 1997 - Tavaré et al. - Utilise la vraisemblance calculable de statistiques résumées plutôt que la vraisemblance incalculable du jeu de données pour inférer un temps de coalescence
  - 1997 - Fu and Li - Approxime par Monte-Carlo la vraisemblance d'une statistique discrète sous un paramètre discret pour inférer l'âge d'un plus proche ancêtre commun
  - 1998 - Weiss and Von Haesler - Fait du Fu and Li pour des stats continues en intégrant un seuil de tolérence $d(S_i, S) \leq \varepsilon$

La voie évolutioniste

- Les tenants évolutionistes de l'ABC font remonter cet ensemble de méthodes à 1997
  - 1997 - Tavaré et al. - Utilise la vraisemblance calculable de statistiques résumées plutôt que la vraisemblance incalculable du jeu de données pour inférer un temps de coalescence
  - 1997 - Fu and Li - Approxime par Monte-Carlo la vraisemblance d'une statistique discrète sous un paramètre discret pour inférer l'âge d'un plus proche ancêtre commun
  - 1998 - Weiss and Von Haesler - Fait du Fu and Li pour des stats continues en intégrant un seuil de tolérence $d(S_i, S) \leq \varepsilon$
  - 1999 - Pritchard et al. - Adapte l'algorithme de Weiss and Von Haesler à la mode bayésienne : c'est l'ABC acceptation-rejet

La voie évolutioniste

- Les tenants évolutionistes de l'ABC font remonter cet ensemble de méthodes à 1997
  - 1997 - Tavaré et al. - Utilise la vraisemblance calculable de statistiques résumées plutôt que la vraisemblance incalculable du jeu de données pour inférer un temps de coalescence
  - 1997 - Fu and Li - Approxime par Monte-Carlo la vraisemblance d'une statistique discrète sous un paramètre discret pour inférer l'âge d'un plus proche ancêtre commun
  - 1998 - Weiss and Von Haesler - Fait du Fu and Li pour des stats continues en intégrant un seuil de tolérence $d(S_i, S) \leq \varepsilon$
  - 1999 - Pritchard et al. - Adapte l'algorithme de Weiss and Von Haesler à la mode bayésienne : c'est l'ABC acceptation-rejet
  - 2003 - Beaumont el al. - Introduit le terme "Approximate Bayesian Computation" ; ABC avec régression linéaire locale

La voie évolutioniste

- Les tenants évolutionistes de l'ABC font remonter cet ensemble de méthodes à 1997
  - 1997 - Tavaré et al. - Utilise la vraisemblance calculable de statistiques résumées plutôt que la vraisemblance incalculable du jeu de données pour inférer un temps de coalescence
  - 1997 - Fu and Li - Approxime par Monte-Carlo la vraisemblance d'une statistique discrète sous un paramètre discret pour inférer l'âge d'un plus proche ancêtre commun
  - 1998 - Weiss and Von Haesler - Fait du Fu and Li pour des stats continues en intégrant un seuil de tolérence $d(S_i, S) \leq \varepsilon$
  - 1999 - Pritchard et al. - Adapte l'algorithme de Weiss and Von Haesler à la mode bayésienne : c'est l'ABC acceptation-rejet
  - 2003 - Beaumont el al. - Introduit le terme "Approximate Bayesian Computation" ; ABC avec régression linéaire locale
  - 2003 - Marjoram et al. - Oriente l'exploration de l'espace des paramètres dans l'ABC ; MCMC sans vraisemblance

La voie évolutioniste

- Les tenants évolutionistes de l'ABC font remonter cet ensemble de méthodes à 1997
  - 1997 - Tavaré et al. - Utilise la vraisemblance calculable de statistiques résumées plutôt que la vraisemblance incalculable du jeu de données pour inférer un temps de coalescence
  - 1997 - Fu and Li - Approxime par Monte-Carlo la vraisemblance d'une statistique discrète sous un paramètre discret pour inférer l'âge d'un plus proche ancêtre commun
  - 1998 - Weiss and Von Haesler - Fait du Fu and Li pour des stats continues en intégrant un seuil de tolérence $d(S_i, S) \leq \varepsilon$
  - 1999 - Pritchard et al. - Adapte l'algorithme de Weiss and Von Haesler à la mode bayésienne : c'est l'ABC acceptation-rejet
  - 2003 - Beaumont el al. - Introduit le terme "Approximate Bayesian Computation" ; ABC avec régression linéaire locale
  - 2003 - Marjoram et al. - Oriente l'exploration de l'espace des paramètres dans l'ABC ; MCMC sans vraisemblance
  - Depuis : foisonement des améliorations et des nouvelles procédures

La voie statisticienne

Pourtant, tout était là...

- Méthodes des contrastes ou équations estimantes : contourner l'utilisation de la vraisemblance pour faire de l'inférence (cf. en particulier la méthode des moments)
- Statistiques exhaustives
- 1962 - Parzen - Estimation de la densité par noyau
- 1964 - Nadaraya ; Watson - Régression par noyau
- 1984 - Diggle and Gratton - Estimation (simulation + lissage) de la vraisemblance pour les modèles stochastiques implcites
- 1984 - Durbin - ABC acceptation-rejet avec seuil de tolérence mais directement sur les jeux de données (n'utilise pas de statistiques résumées)

La jonction des deux voies

Pour justifier l'ABC et proposer de nouvelles procédures

- Comportement asymptotique des lois a posteriori issues de l'ABC
- Sélection automatique du seuil d'acceptation
- Lien entre ABC et régression par noyau (cf. Blum, 2009)
- ABC et solutions contre le fléau de la dimension
- ABC et algorithmes d'exploration efficace de l'espace des paramètres (dont les algorithmes adaptatifs)

Plan

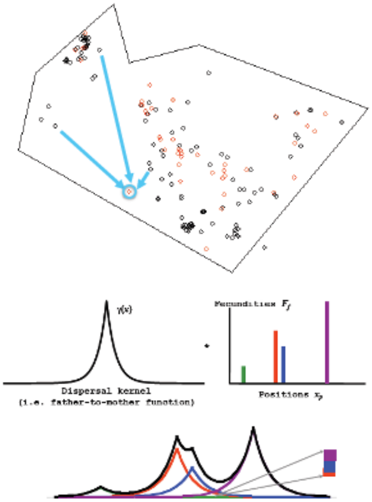
1. Introduction
2. Quelques algorithmes de calcul bayésien approché
3. L'émergence de l'ABC
4. **Travaux menés par BioSP sur l'ABC**

Quelques travaux sur l'ABC menés au sein de BioSP

- Travaux s'inscrivant dans le projet ANR EMILE (2009–2013)
  - Etudes de méthodes inférentielles et logiciels pour l'évolution
  - Etude des aspects théoriques de l'ABC
  - Extension de l'ABC à des applications autres que l'évolution
- Recherche des statistiques résumées informatives pour des dynamiques démo-génétiques (applis en écologie et épidémiologie spatiale)
- ABC conditionnel aux estimations ponctuelles des paramètres
- Etude du comportement asymptotique des lois a posteriori issues de l'ABC
- Loi a posteriori basée sur un contraste
- Contrast-based ABC

Statistiques résumées informatives pour des dynamiques démo-génétiques (Carpentier, Guiton & Klein)

- Modèle dynamique (stochastique implicite) : convolution entre un processus de sources (les pères) marquées (génétique, fertilité) et une fonction de dispersion
- Données : génotypes de graines échantillonnées sur les mères (récepteurs de pollen)
- Statistiques résumées pour l'ABC : indices de différentiation génétique (2 à 2) à différentes distances
- Gérer le très grand nombre de statistiques



ABC conditionnel aux estimations ponctuelles des paramètres (Soubeyrand, Martin & Carpentier)

ABC basique (acceptation/rejet) :

- $\theta_i \sim \pi(\cdot)$, $i = 1, \ldots, I$, I très grand
- $S_i = s(\mathcal{M}_{\theta_i}(\omega_i))$ où $\mathcal{M}_{\theta_i}$ est un modèle implicite
- Classiquement, on estime la loi a posteriori $p(\theta \mid d(S_i, S_{obs}) \leq \delta)$
- Généralement, d standardisation + distance euclidienne
- Problèmes : stats redondantes, stats inutiles...

Solution proposée :

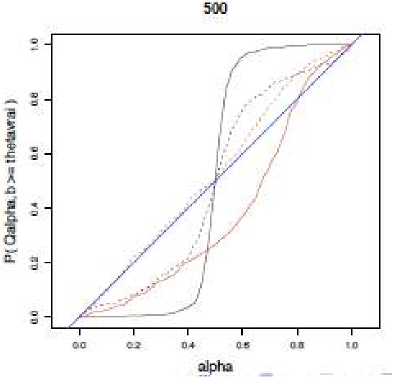
- On introduit un modèle empirique (e.g. ppr ; un métamodèle) qui permet d'obtenir un estimateur $\hat{\theta}_i$ de θ_i sachant S_i et on estime la loi a posteriori suivante : $p(\theta \mid d_\Theta(\hat{\theta}_i, \hat{\theta}_{obs}) \leq \delta_\Theta)$

ABC conditionnel aux estimations ponctuelles des paramètres (Soubeyrand, Martin & Carpentier)

Exemple :

- Données : $X_1, \ldots, X_{50} \sim N(0,1)$
- Stats : $S = (\bar{X}^3_{1:40}, SD^3_{1:40}, \bar{X}^3_{11:20}, SD^3_{11:20}, \bar{X}^3_{21:30}, SD^3_{21:30}, \beta(0.1, 0.1))$
- $B = 500$ répétitions, $I = 50\ 000$ simulations pour chaque répét., taux d'acceptation $\tau = 0.005$
- QQ-plot : $P(\theta_{vrai} \leq Q^{(b)}_{\alpha}) = \alpha$

noir : μ
rouge : σ
continu : $\cdot \mid S$
tirets : $\cdot \mid \hat{\theta}$



Loi a posteriori basée sur un contraste (Soubeyrand, Carpentier, Desassis & Chadœuf, 2009)

- Let $(X_i)_{i\leq t}$ be a sample of size t with distribution P_α. Then, the posterior distribution of α is

$$p(\alpha \mid X_i, i \leq t) = \frac{P_\alpha(X_i, i \leq t)\pi(\alpha)}{\int_\Theta P_\beta(X_i, i \leq t)\pi(\beta)d\beta} = \frac{\exp(-tU_t^{lik}(\alpha))\pi(\alpha)}{\int_\Theta \exp(-tU_t^{lik}(\beta))\pi(\beta)d\beta}$$

- Substituting $U_t^{lik}(\alpha)$ with any contrast $U_t(\alpha)$ leads to a contrast-based (CB) posterior distribution :

$$p_t(\alpha) = \frac{\exp(-tU_t(\alpha))\pi(\alpha)}{\int_\Theta \exp(-tU_t(\beta))\pi(\beta)d\beta}.$$

- Utilité :
  - Dans une perspective fréquentiste : on obtient la loi du MAP en s'affranchissant de certains calculs
  - Dans une perspective bayésienne : si $I_\theta = \Gamma_\theta$ la CB-posterior peut être utiliser à la mode bayésienne
  - Mais comme en ABC, c'est une posterior conditionnelle à l'information contenue dans le contraste utilisé

Exemple : Modèle de rugosité du sol

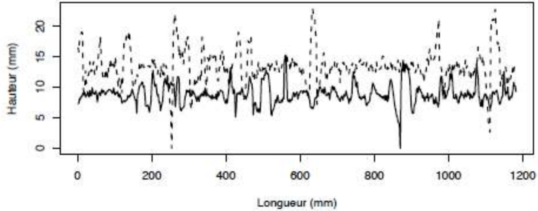
- Modèle : convolution entre un processus ponctuel poissonien d'intensité α/β et des cylindres aléatoires ($rayon = hauteur \sim Exp(\beta)$)

$$Y_M = \sum_{(x,r)\in(X,R)} f(x-M,r) \quad \text{with} \quad f(x,r) = r1_{\{||x||<r\}}$$

- Contraste, basé sur les deux premiers moments $\hat{\mu}_A = (\frac{1}{\nu(A)}\int_A Y_M dM, \frac{1}{\nu(A)}\int_A Y_M^2 dM)$, pour lequel $I_\theta = \Gamma_\theta$:

$$U_A(\theta) = (\hat{\mu}_A - E(\hat{\mu}_A))' V(\alpha,\beta))^{-1}(\hat{\mu}_A - E(\hat{\mu}_A))/2$$

- Données :



Exemple : Modèle de rugosité du sol

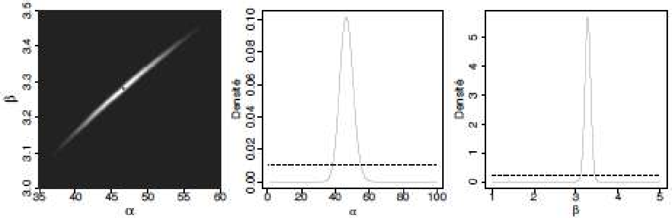
- Modèle : convolution entre un processus ponctuel poissonien d'intensité α/β et des cylindres aléatoires $(\textit{rayon} = \textit{hauteur} \sim \textit{Exp}(\beta))$

$$Y_M = \sum_{(x,r)\in(X,R)} f(x - M, r) \qquad \textit{with} \qquad f(x,r) \ = r 1_{\{||x||<r\}}$$

- Contraste, basé sur les deux premiers moments $\hat{\mu}_A = (\frac{1}{\nu(A)} \int_A Y_M dM, \frac{1}{\nu(A)} \int_A Y_M^2 dM)$, pour lequel $I_\theta = \Gamma_\theta$:

$$U_A(\theta) = (\hat{\mu}_A - E(\hat{\mu}_A))' V(\alpha,\beta))^{-1} (\hat{\mu}_A - E(\hat{\mu}_A))/2$$

- CB-posterior utilisable à la mode bayésienne (car $I_\theta = \Gamma_\theta$) :



Contrast-based ABC ou le choix raisonné d'une distance

(Carpentier, Chadœuf & Klein, 2010)

- Idée : Faire de l'ABC en utilisant les résultats obtenus pour la CB-posterior
- Grandes lignes de l'algorithme :
  - Tirer $\theta_i \sim_{\text{iid}} \pi$, $\mathcal{D}_i \sim_{\text{indép.}} \mathcal{M}_{\theta_i}$, calculer $S_i = s(\mathcal{D}_i)$
  - Retenir le mode θ^* de la CB-posterior

$$\hat{p}^*(\theta \mid S) = \frac{\sum_{i=1}^{I} \exp\{-n(S_i - S)'(S_i - S)\} K_h(\theta_i - \theta)}{\sum_{i=1}^{I} \exp\{-n(S_i - S)'(S_i - S)\}}$$

  - Estimer $V = V_{\theta^*}(S_i)$ en simulant sous θ^*
  - Calculer la posterior

$$\hat{p}(\theta \mid S) = \frac{\sum_{i=1}^{I} \exp\{-(n/2)(S_i - S)' \hat{V}^{-1}(S_i - S)\} K_h(\theta_i - \theta)}{\sum_{i=1}^{I} \exp\{-(n/2)(S_i - S)' \hat{V}^{-1}(S_i - S)\}}$$

Conclusion

- Utiliser des statistiques résumées contenant de l'information sur les paramètres d'intérêt, extraire au mieux cette information (c'est de la stat !) et le faire vite

Merci.