Plans d'expérience bayésiens: Que nous proposent Chaloner et Verdinelli?

Bayesian Experimental Design: A Review
Statistical Science, Vol. 10, No. 3

Sophie Ancelet[1,2]

[1]UMR 518 AgroParisTech/INRA, Département MIA, Equipe MORSE
[2]EDF R&D, Département MRI, Groupe T56

AppliBUGS, 17 Juin 2011

Outline

1. Introduction
2. Cadre décisionnel bayésien
3. Cas linéaire
4. Cas non linéaire : Que faire ?
5. Applications : des embûches...

1 Introduction

2 Cadre décisionnel bayésien

3 Cas linéaire

4 Cas non linéaire : Que faire ?

5 Applications : des embûches...

Qui est Kathryn Chaloner ?

- 2002- : Professor and Head, Department of Biostatistics, University of Iowa (US)
- 1978-1982 : Ph.D in Statistics : "Optimal Bayesian design for linear models", Carnegie-Mellon University (Pittsburgh, US)

Ses thèmes de recherche

biostatistique, essais cliniques, HIV, maladies infectieuses, **méthodes bayésiennes**, **plans d'expérience**, statistique

Qui est Isabella Verdinelli ?

- 2000- :
  - Professor, Carnegie-Mellon University (Pittsburgh, US), Department of Statistics
  - Professor, Sapiena University of Rome, Department of Statistical Sciences
- 1996 : Ph.D in Statistics : "Bayesian Procedures for Breast Cancer Clinical Trials", Carnegie-Mellon University

Ses thèmes de recherche

biostatistique, essais cliniques, méthodes non-paramétrique, **plans d'expérience**, **statistique bayésienne**

Contexte (1)

Soit Y **une** grandeur aléatoire d'intérêt (ou réponse) mesurée par l'expérimentateur sur n essais indépendants (n supposé fixé) :

$$Y_1, Y_2, \ldots, Y_n \sim^{i.i.d} f_\theta(x_k)$$

avec

- x_k : variables de contrôle (ou **facteurs**) (k=1,2,...,K)
- f_θ : un modèle statistique paramétré par θ

$\longrightarrow X$: Matrice de design ($Dim = n \times K$)

$\longrightarrow x_i^T \in \chi$ (compact) : Points du design ($i = 1, 2, \ldots, n$)

Contexte (2)

Hypothèse : Un choix approprié des variables de contrôle peut améliorer l'inférence statistique de certains paramètres (ou fonctions de paramètres).

Problème décisionnel

Comment choisir les valeurs des variables de contrôle i.e., x_i^T pour **optimiser** l'estimation de grandeurs inconnues d'intérêt (sous de possibles contraintes de coût) ?

$\longrightarrow$ Plans d'expérience physiques

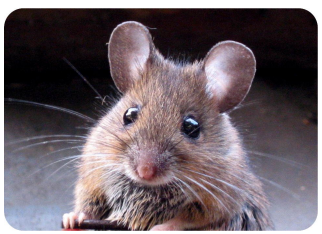
Exemple 1 : Essais cliniques (1)

Objectif

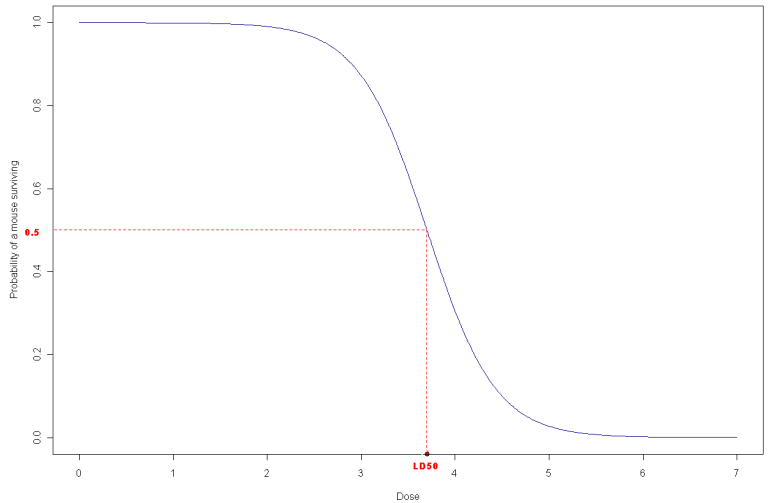
Définir le plan d'expérience qui permette d'estimer "au mieux" l'efficacité d'un médicament, mesurée par le LD50 [a]

a. LD50 : dose injectée pour laquelle un individu donné à une probabilité de mourir de 0.5.

- 1 médicament d'intérêt
- n=60 **souris**
- **Grandeur observée** : nombre Y_j de souris qui ont survécu entre 7 et 10 jours après avoir recu la dose d_j (j=1,2,...,J)
- **Variable de contrôle** : dose injectée (K=1)



Exemple 1 : Essais cliniques (2)



Problème décisionnel

Comment choisir les valeurs des doses d_j pour **optimiser** l'estimation du LD50 ?

- Combien de doses d'exposition testées, noté J ? Quelles valeurs d_j ? Quelle proportion de souris η_j est exposée à chaque dose d_j avec la contrainte $\sum_{j=1}^{J} \eta_j = 1$?

Exemple 1 : Essais cliniques (3)

Deux jeux de données issus d'essais cliniques pour estimer l'efficacité de différentes concentrations d'Albumine

TABLE 1

Batch 1			Batch 2		
Dose	Number exposed	Number dead	Dose	Number exposed	Number dead
2.5	10	0	2.5	10	0
3.0	10	1	3.0	10	2
3.5	10	1	3.5	10	1
4.0	10	3	4.0	10	4
4.5	10	5	4.5	10	7
5.0	10	6	5.0	10	8

Plan d'expérience pratiqué : 6 doses équiréparties avec 10 souris exposées par dose

Exemple 2 : Fiabilité industrielle (1)

Objectif

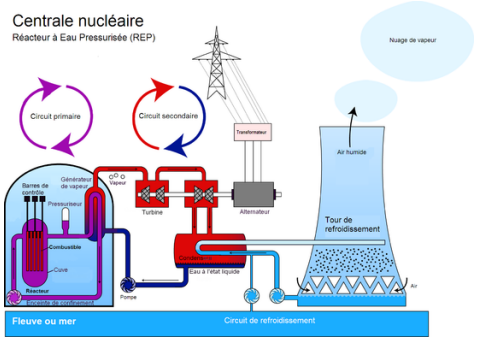
Définir le plan d'expérience qui permette d'estimer "au mieux" la ténacité minimale[a] de l'acier de cuve REP en fonction de la température et sous une contrainte de coût.

---

a. ténacité d'un acier de cuve : résistance à la rupture brutale s'initiant sur un défaut existant

Pourquoi la ténacité minimale ?

- Paramètre de position difficile à estimer
- Calculs de tenue de la cuve nucléaire
- Guide les courbes de valeurs minimales de ténacité obtenues empiriquement à partir de tests destructifs sur des éprouvettes



Exemple 2 : Fiabilité industrielle (2)

Une fonction croissante de la température

- Basse température $\longrightarrow$ ténacité faible (acier cassant) : zone de rupture "fragile"
- Haute température $\longrightarrow$ ténacité augmente (acier souple) : zone de rupture "ductile"
- Entre les deux, "zone de transition" avec variabilité de la ténacité à température fixée

- n=30 mesures sur éprouvettes-tests
- **Grandeur observée** : Y_j^i la i-ème mesure de ténacité (en $MPa/m^{1/2}$) parmi n_j mesures réalisées à température T_j $(j = 1, 2, \ldots, J)$
- **Variable de contrôle** : Température (K=1)

1 Introduction

2 Cadre décisionnel bayésien

3 Cas linéaire

4 Cas non linéaire : Que faire ?

5 Applications : des embûches...

Objectif

Rappel

Fournir une *décision* i.e., un **plan d'expérience optimal** ϵ^* permettant d'estimer "au mieux" un (des) paramètres inconnus (tout en intégrant des préoccupations d'ordre économique)

$\longrightarrow$ Besoin de spécifier un critère d'évaluation des différents plans d'expérience possibles

- prise en compte des conséquences de chaque décision
- fonction des paramètres inconnus θ

Pourquoi le choix bayésien ?

1. Permet de construire une règle de décision, i.e., un ordre total sur l'ensemble des plans d'expérience possibles, tenant compte de critères économiques potentiellement dépendant du paramètre inconnu θ.

2. Permet d'intégrer **avant l'acquisition des données**, des connaissances externes préalables (études antérieures, littérature, avis d'experts...) à travers la spécification de lois *a priori*
   $\longrightarrow$ Plans d'expérience possiblement moins coûteux (par rapport au cadre classique)

Les ingrédients de base à la recherche d'un plan d'expérience optimal

1. L'ensemble des plans d'expérience possibles, $\mathcal{E}$, associés à une procédure de collecte d'informations en univers uncertain
2. Un modèle probabiliste $[y|\theta, \epsilon]$ associé à des observations y issues du plan d'expérience $\epsilon \in \mathcal{E}$ et dont les réalisations vont renseigner sur θ
3. La loi *a priori* sur $\theta \in \Theta$
4. Une fonction d'utilité $u(\epsilon, \theta, y)$ qui, à chaque réalisation d'observations y issues du plan d'expérience ϵ, associe un niveau d'utilité, encore appelé gain, quand les paramètres prennent la valeur θ.

Du risque classique au risque bayésien...

Le risque classique

$$U_C(\epsilon, \theta) = \int_y u(\epsilon, \theta, y)[y|\theta, \epsilon] \qquad \longrightarrow \text{Optimalité locale}$$

Le bayésien intègre l'incertitude sur θ !

Le risque bayésien (ou utilité espérée)

$$U(\epsilon) = \int_\theta \int_{y|\theta,\epsilon} u(\epsilon, \theta, y)[y|\theta, \epsilon][\theta] dy d\theta$$

$$U(\epsilon) = \int_{y|\epsilon} \int_{\theta|y,\epsilon} u(\epsilon, \theta, y)[\theta|y, \epsilon][y|\epsilon] d\theta dy \qquad \longrightarrow \text{Ordre total}$$

Un problème d'optimisation

Que cherche-t-on ?

$$\epsilon^* = argmax_{\epsilon \in \mathcal{E}} U(\epsilon)$$

Deux utilités espérées classiques pour l'estimation

"Selecting a utility function that appropriately describes the goals of a given experiment is very important" (Chaloner & Verdinelli)

2- Information de Shannon espérée

$$U_1(\epsilon) = \int log[\theta|y,\epsilon][y,\theta|\epsilon]d\theta dy$$

2- Opposée de la variance a posteriori espérée

$$U_2(\epsilon) = -\int (\theta - \widehat{\theta})^T A(\theta - \widehat{\theta})[y,\theta|\epsilon]d\theta dy$$

avec A une matrice symmétrique définie positive

1 Introduction

2 Cadre décisionnel bayésien

3 Cas linéaire

4 Cas non linéaire : Que faire ?

5 Applications : des embûches...

Contexte

Modèle de regression linéaire normal à σ^2 fixé

$$y|\theta, \sigma^2, \epsilon \sim \mathcal{N}_n(X\theta, \sigma^2 I)$$

avec θ de dimension l, I la matrice identité $n \times n$

Loi a priori sur θ

$$\theta \sim \mathcal{N}_l(\theta_0, \sigma^2 R^{-1})$$

avec $\sigma^2 R^{-1}$ la matrice $l \times l$ de variance-covariance *a priori* est **connue**

$\longrightarrow \theta|y, \epsilon \sim \mathcal{N}_l(\theta^*, \Sigma(\theta, \epsilon))$ (par conjugaison)

Les critères d'optimalité alphabétiques bayésiens (1)

1- Information de Shannon espérée

$$U_1(\epsilon) = -\frac{I}{2}log(2\pi) - \frac{I}{2} + \frac{1}{2}logdet(\Sigma^{-1}(\theta, \epsilon))$$

avec $\Sigma(\theta, \epsilon) = \sigma^2(X^TX + R)^{-1}$.

$\Rightarrow$ Critère bayésien de D-optimalité

$$\phi_1(\epsilon) = \det\{X^TX + R\}$$

Remarques :

- $+$ petite variance possible sur l'ensemble des $\theta=$ termes diagonaux de $X^TX + R$ les plus petits possibles
- Analogue bayésien du critère de D-optimalité classique

Les critères d'optimalité alphabétiques bayésiens (2)

2- Opposée de la variance a posteriori espérée

$$U_2(\epsilon) = -trA\Sigma(\theta, \epsilon)$$

avec A une matrice symmétrique définie positive et $\Sigma(\theta, \epsilon) = \sigma^2(X^TX + R)^{-1}$.

$\Rightarrow$ Critère bayésien de A-optimalité

$$\phi_2(\epsilon) = -trA(X^TX + R)^{-1}$$

Remarques :

- $+$ petite somme des variances des θ
- Analogue bayésien du critère de A-optimalité classique

Oui mais en pratique...

- $[y|\theta, \epsilon]$ n'est pas un modèle de regression linéaire normal
- $u(\theta, \epsilon, y)$ a une forme quelconque
- $\mathcal{E}$ et Θ sont des espaces de grande dimension

$\longrightarrow$ Calcul exact de $U(\epsilon)$ et recherche de ϵ^*
$=$ **Challenge numérique** très difficile !

1 Introduction

2 Cadre décisionnel bayésien

3 Cas linéaire

4 **Cas non linéaire : Que faire ?**

5 Applications : des embûches...

Que nous proposent Chaloner et Verdinelli (1) ?

Principe : Se rapprocher du cadre linéaire via deux approximations asymptotiques successives

Approximation asymptotique 1

Hypothèse : la matrice d'information de Fisher n'est pas singulière

$$\theta|y,\epsilon \sim N_l\left(\widehat{\theta},[nI(\widehat{\theta},\epsilon)+R]^{-1}\right)$$

où

- $\widehat{\theta}$: Mode de la loi de probabilité jointe *a posteriori* de θ (EMV généralisé)
- $I(\widehat{\theta},\epsilon)$: Matrice d'information de Fisher observée normalisée
- R : Matrice de précision *a priori*

Que nous proposent Chaloner et Verdinelli (2)

Approximation asymptotique 2

Contexte : L'utilité bayésienne prédictive ne dépend des données y qu'à travers $\widehat{\theta}$
Comme $\widehat{\theta}$ est un estimateur consistent de θ :

$$\widehat{\theta} = T(y) \longrightarrow^{p.s.} \theta$$
$$\Longrightarrow \widehat{\theta} = T(y) \longrightarrow^{Loi} \theta$$

Dans le contexte prédictif, $[\theta]$ porte toute l'information nécessaire pour générer des observations y !

Quels critères peut-on optimiser (1) ?

1- Approximation de l'information de Shannon espérée

$$\widetilde{U}_1(\epsilon) = -\frac{I}{2} log(2\pi) - \frac{I}{2} + \frac{1}{2} \int logdet\{nI(\theta, \epsilon) + R\}[\theta] d\theta$$

avec $I(\theta, \epsilon)$ la matrice d'information de Fisher attendue normalisée

$\Rightarrow$ Critère bayésien de D-optimalité approché

$$\widetilde{\phi}_1(\epsilon) = \mathbb{E}_\theta \left(logdet\{nI(\theta, \epsilon) + R\} \right)$$

Quels critères peut-on optimiser (2) ?

2- Approximation de l'opposée de la variance a posteriori espérée

$$\widetilde{U}_2(\epsilon) = -\int tr\left(A\{nI(\theta,\epsilon)+R\}^{-1}\right)[\theta]d\theta$$

avec A une matrice symmétrique définie positive et $I(\theta,\epsilon)$ la matrice d'information de Fisher attendue normalisée

$\Rightarrow$ Critère bayésien de A-optimalité approché

$$\widetilde{\phi}_2(\epsilon) = -\mathbb{E}_\theta\left(tr\left(A\{nI(\theta,\epsilon)+R\}^{-1}\right)\right)$$

1 Introduction

2 Cadre décisionnel bayésien

3 Cas linéaire

4 Cas non linéaire : Que faire ?

5 **Applications : des embûches...**

Exemple 2 : Essais cliniques souris (1)

Plan d'expérience ϵ

$\epsilon = (J, d_1, d_2, \ldots, d_J, \eta_1, \eta_2, \ldots, \eta_J)$

Modèle d'observations

$$Y_j \sim Binomial(n_j, p_j) \qquad \forall j = 1, 2, ..., J$$
$$logit(p_j) = \beta(d_j - \gamma)$$

avec p_j la probabilité d'avoir survécu entre 7 et 10 jours après avoir eu la dose d'exposition d_j.
Remarque : $\gamma = LD50$

Exemple 2 : Essais cliniques souris (2)

Vraisemblance

$$[Y_1, \ldots, Y_J | \beta, \gamma, \epsilon] = \prod_{j=1}^{J} [Y_j | \beta, \gamma, d_j, n_j] = \prod_{j=1}^{J} C_{n_j}^{y_j} p_j^{y_j} (1 - p_j)^{n_j - y_j}$$

avec $n_j = \eta_j n$ et $p_j = logit^{-1}(\beta(d_j - \gamma))$.

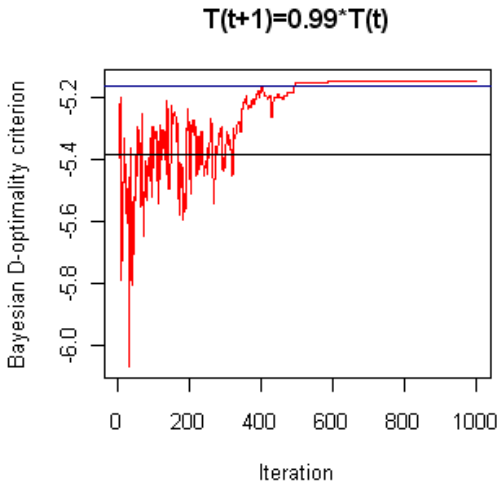
Lois a priori Beta Pert

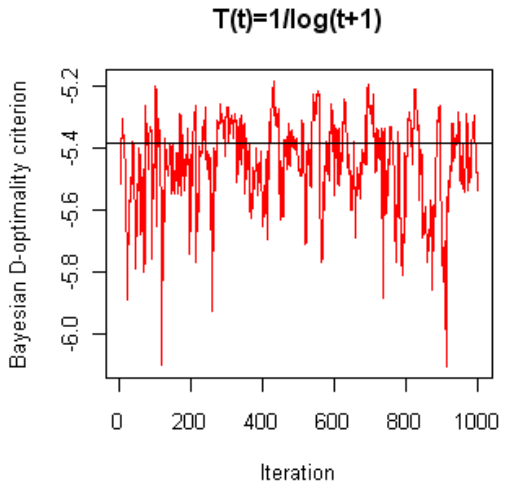
$$\gamma = 3.2 + Z_1$$
$$\beta = -4 + 2.5 Z_2$$

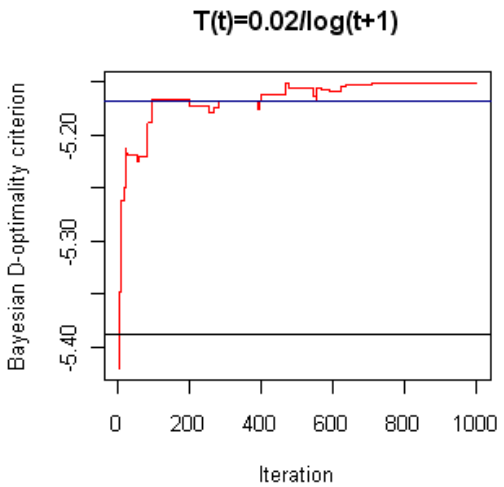
avec Z_1 et Z_2 deux lois beta(4,4). $\gamma \in [3.2, 4.2]$ et $\beta \in [-4, -1.5]$.

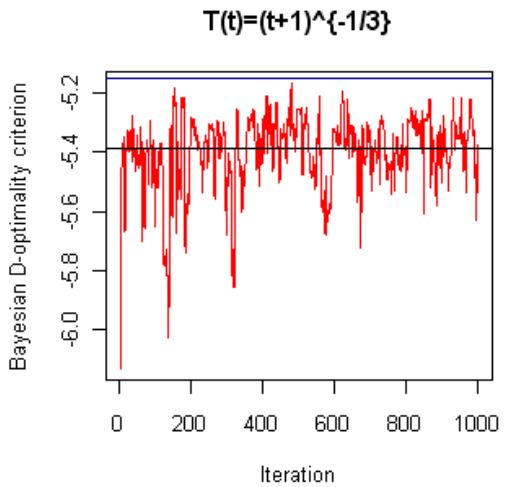
Exemple 2 : Essais cliniques souris (3)

Implémentation d'un algorithme de recuit simulé avec espérance a priori approchée par simulations Monte-Carlo (10000 valeurs)









Exemple 2 : Essais cliniques souris (4)

Plan d'expérience global optimal, valeur du critère bayésien de A-optimalité, obtenus pour les 2 "meilleurs" runs. Lois a priori **beta Pert**, et de support [3.2,4.2] pour le LD50.

J	η^*	d^*	$\widetilde{\phi}_2(\epsilon^*)$
J=2	(0.56,0.44) (0.40,0.60)	(3.39,4.12) (3.21,3.93)	-0.7787 -0.7833
J=3	(0.54,0.07,0.39) (0.47,0.35,0.18)	(3.37,3.95,4.11) (3.30,4.00,4.07)	-0.7814 -0.7849
J=4	(0.07,0.39,0.53,0.01) (0.48,0.11,0.41,0.00)	(3.13,3.36,4.03,4.78) (3.33,3.90,4.08,4.53)	-0.7752 -0.7832
J=5	(0.40,0.05,0.26,0.17,0.11) (0.40,0.29,0.04,0.24,0.02)	(3.30,3.38,3.95,3.98,4.21) (3.37,3.75,4.01,4.28,4.39)	-0.7853 -0.7879
J=6	(0.43,0.10,0.07,0.15,0.18,0.07) (0.49,0.03,0.06,0.39,0.01,0.02)	(3.38,3.54,3.62,3.99,4.29,4.50) (3.40,3.43,3.78,4.12,4.21,4.31)	-0.7857 -0.7873
J=7	(0.03,0.45,0.01,0.17,0.29,0.03,0.02) (0.09,0.34,0.04,0.03,0.22,0.27,0.01)	(3.06,3.39,3.68,3.73,4.15,4.59,4.79) (3.26,3.33,3.60,3.70,3.97,4.12,4.67)	-0.7778 -0.7783
J=8	(0.08,0.09,0.38,0.39,0.05,0.0,0.0,0.01) (0.36,0.17,0.40,0.06,0.01,0.0,0.0,0.0)	(3.13,3.41,3.42,4.05,4.31,4.68,4.91,5.19) (3.24,3.75,3.95,4.31,4.39,4.84,5.01,5.3)	-0.7905 -0.7926

Chaloner & Verdinelli : $J = 4$, $d = (3.07, 3.47, 3.73, 4.13)$, $\eta = (0.3, 0.2, 0.2, 0.3)$ et $\widetilde{\phi}_2(\epsilon) = -0.8320$

Exemple 2 : Essais cliniques souris (5)

- Résultats très instables...
- Recuit simulé inefficace ? Surface d'utilité plate au voisinage de son maximum ? Support multidimensionnel ?

$\longrightarrow$ Transformer le problème d'optimisation en un problème de simulation (Peter Muller (1999), Amzal et al. (2006)
$\longrightarrow$ Spécification de nouvelles fonctions d'utilité permettant d'utiliser les idées de Muller
$\longrightarrow$ (En cours) Implémentation d'un algorithme particulaire avec recuit simulé par apprentissage séquentiel (Amzal et al., 2006)

Exemple 3 : Fiabilité industrielle

Plan d'expérience ϵ

$$\epsilon = (J, T_1, T_2, \ldots, T_J, \eta_1, \eta_2, \ldots, \eta_J)$$

Loi de Weibull à 3 paramètres

$$Y_j^i - K_{min}(T_j) | T_j \sim^{i.i.d} \mathcal{W}(\mu(T_j), \beta)$$

$\longrightarrow$ Problème : Matrice d'information de Fisher non définie ....

$\longrightarrow$ Des pistes possibles ?

- Amzal, B. and Bois, F.Y. and Parent, E. and Robert, C.P. (2006). Bayesian-Optimal Design via Interacting Particle Systems. *Journal of the American Statistical Association*, Vol. 101, No. 474.
- Berger, J.O. (1985). *Statistical Decision Theory and Bayesian Analysis.*, Berlin : Springer-Verlag
- Chaloner,K. and Verdinelli, I. (1995). Bayesian Experimental Design : a Review. *Statistical Science*, Vol. 10, No. 3 pp 273-304.
- Muller, P. (1999). Simulation-Based Optimal Design. *Bayesian statistics 6*, pp 459-474
- Robert, C.P. (2006). *Le choix bayésien. Principes et pratique.*, Springer
- Rubin, D.B. (1988). Using the SIR algorithm to Simulate Posterior Distributions. Pages 395-402 of :*Bayesian Statistics 3*, vol. 3. Oxford University Press.