ECN 6338 Cours 3

Quelques sujets préalables

William McCausland

2022-01-27

Survol du Cours 2

Maximisation sous contraintes

- une contrainte d'égalité
- plusieures contraintes d'égalité
- plusieures contraintes d'égalité plus non-négativité
- plusieures contraintes d'égalité et d'inégalité
- exemple

Maximisation de la vraisemblance

- la vraisemblance
- l'estimateur maximum de vraisemblance et ses propriétés
- les problèmes d'optimisation à effectuer

Inférence bayésienne

- les lois *a priori* et *a posteriori*
- les problèmes d'intégration à effectuer

Problème de maximisation avec une contrainte d'égalité

Problème :

$$\max_{x \in \mathbb{R}^n} f(x) \quad \text{t.q.} \quad g(x) = c,$$

où

- $f : \mathbb{R}^n \to \mathbb{R}$, $g : \mathbb{R}^n \to \mathbb{R}$,
- $c \in \mathbb{R}$,
- $f, g \in C^2$, l'espace de fonctions avec deux dérivées continues.

Fonction de Lagrange, en $x \in \mathbb{R}^n$, $\lambda \in \mathbb{R}$:

$$L(x, \lambda) = f(x) + \lambda[c - g(x)].$$

Théorème : Si x^* est une solution et que $g_j(x^*) \neq 0$ pour au moins un j, il existe $\lambda^* \in \mathbb{R}$ tel que

- $L_j(x^*, \lambda^*) = 0$, $j = 1, \ldots, n$, et $L_\lambda(x^*, \lambda^*) = 0$.

Problème avec plusieurs contraintes d'égalité

Problème :

$$\max_{x \in \mathbb{R}^n} f(x) \quad \text{t.q.} \quad g(x) = c,$$

où

- $f : \mathbb{R}^n \to \mathbb{R}$, $g : \mathbb{R}^n \to \mathbb{R}^m$; $m < n$; $f, g \in C^2$,
- $c \in \mathbb{R}^m$.

Fonction de Lagrange, en $x \in \mathbb{R}^n$, $\lambda \in \mathbb{R}^m$ est :

$$L(x, \lambda) = f(x) + \lambda^\top [c - g(x)] = f(x) + \sum_{i=1}^{m} \lambda_i [c_i - g^i(x)].$$

Théorème : Si x^* est une solution et le rang du jacobien $g_x(x^*)$ est m, il existe $\lambda^* \in \mathbb{R}^m$ tel que

- $L_x(x^*, \lambda^*) = 0_n$, $L_\lambda(x^*, \lambda^*) = 0_m$.

Plusieurs contraintes d'égalité, non-négativité

Problème :

$$\max_{x \in \mathbb{R}^n} f(x) \quad \text{t.q.} \quad g(x) = c,\ x \geq 0,$$

où f, g et c sont comme dans le problème précédent.

Fonction de Lagrange, comme dans le dernier problème :

$$L(x, \lambda) = f(x) + \lambda[c - g(x)].$$

Théorème : Si x^* est une solution et le rang du jacobien $g_x(x^*)$ est m, il existe $\lambda^* \in \mathbb{R}^m$ tel que

- $L_x(x^*, \lambda^*) \leq 0$, $x^* \geq 0$ avec écarts complémentaires,
- $L_\lambda(x^*, \lambda^*) = 0$.

Exemple, utilité quasi-linéaire I (exemple 3.1 de Dixit)

Le problème : pour prix $p > 0$ et $q > 0$, revenu $I > 0$ et $a > 0$,

$$\max_{x,y \in \mathbb{R}} y + a \ln x \quad \text{t.q.} \quad px + qy = I.$$

La fonction de Lagrange :

$$L(x, y, \lambda) = y + a \ln x + \lambda(I - px - qy).$$

Les conditions nécessaires pour un maximum :

$$L_x = \frac{a}{x} - \lambda p \leq 0, \; x \geq 0;$$

$$L_y = 1 - \lambda q \leq 0, \; y \geq 0;$$

$$I - px - qy = 0.$$

- Cas $x = 0$, $y = 0$: $I - px - qy = 0$ est impossible.
- Cas $x = 0$, $y > 0$: $L_x \leq 0$ est impossible.

Exemple, utilité quasi-linéaire II

Les conditions nécessaires pour un maximum, encore :

$$L_x = \frac{a}{x} - \lambda p \leq 0,\ x \geq 0;$$

$$L_y = 1 - \lambda q \leq 0,\ y \geq 0;$$

$$I - px - qy = 0.$$

- Cas $x > 0$, $y = 0$:
  - dans ce cas $x = I/p$ et $\lambda = a/I$,
  - il faut que $1 - aq/I \leq 0$, $I \leq aq$.
- Cas $x > 0$, $y > 0$:
  - dans ce cas $\lambda = a/(px) = 1/q$,
  - $x = aq/p$,
  - Il faut que $I - px = I - aq > 0$, auquel cas $y = I/q - a$.

Exemple, utilité quasi-linéaire III

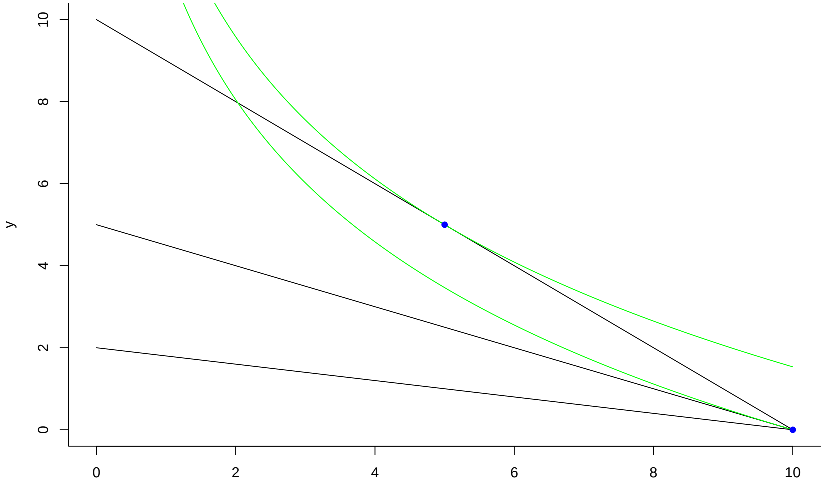
```r
p <- 1; I <- 10  # Prix de x et revenu
a <- 5;          # Paramètre d'utilité
q1 <- 1; q2 <- 2; q3 <- 5   # Trois valeurs du prix de y
x = seq(0, 10, length.out = 1000)  # Grille, valeurs de x

# Trois budgets
b1 = (I - p*x)/q1;
b2 = (I - p*x)/q2;
b3 = (I - p*x)/q3;

# Deux courbes d'indifférence
y1 = 5*log(5)+5 - 5*log(x)  # Qui passe par (5, 5)
y2 = 5*log(10) - 5*log(x)   # Qui passe par (10, 0)
```

Exemple, utilité quasi-linéaire IV

```
plot(x, b1, type='l', xlab='x', ylab='y', bty='l');
lines(x, b2); lines(x, b3)
lines(x, y1, col='green'); lines(x, y2, col='green')
points(c(5, 10), c(5, 0), col='blue', pch=16)
```



Plusieurs contraintes d'égalité et d'inégalité

Problème :

$$\max_{x \in \mathbb{R}^n} f(x) \quad \text{t.q.} \quad g(x) = c,\ h(x) \leq d,$$

où f, g et c sont comme dans le problème précédent,

- $h \colon \mathbb{R}^n \to \mathbb{R}^l$, $h \in C^2$, $d \in \mathbb{R}^l$.

Fonction de Lagrange :

$$L(x, \lambda) = f(x) + \lambda^\top [c - g(x)] + \mu^\top [d - h(x)].$$

Théorème (Karush-Kuhn-Tucker) : Si x^* est une solution, le rang des jacobiens $g_x(x^*)$ et $h_x(x^*)$ sont m et l, il existe $\lambda^* \in \mathbb{R}^m$ et $\mu^* \in \mathbb{R}^l$ tels que

- $L_x(x^*, \lambda^*, \mu^*) = 0_n$,
- $L_\lambda(x^*, \lambda^*, \mu^*) = 0_m$,
- $L_\mu(x^*, \lambda^*, \mu^*) \geq 0_l$, $\mu \geq 0_l$ avec écarts complémentaires.

Comparaison avec la page 122 dans Judd

Dans Judd :

- $g(x) = 0$ et $h(x) \leq 0$, pas $g(x) = c$ et $h(x) \leq d$.
  - aucune perte de généralité : défine $\tilde{g}(x) = g(x) - c$, $\tilde{h}(x) = h(x) - d$
- moins de détail sur les conditions de rang ("constraint qualification")
- problème de minimisation, pas de maximisation
  - le relâchement d'une contrainte *réduit* la valeur optimale $f(x^*)$
- $L(x, \lambda, \mu) = f(x) + \lambda^\top g(x) + \mu^\top h(x)$ (signe opposé des deux derniers termes)
- Les conditions pour μ et L_μ sont $\mu \leq 0$, $L_\mu \leq 0$.

Le problème de maximisation est plus naturel pour les économistes mais les logiciels exigent souvent des fonctions à minimiser. Il faut "traduire" la spécification du problème et bien intérpreter la signe des prix d'ombre et autres résultats.

Exemple, chomage technique (exemple 3.2 de Dixit)

Le problème :

- Une économie a 300 unités de L (main d'oeuvre) et 450 unités de T, pour la production de blé et de boeuf.
- Produire une unité de blé prend 2 unités de L, 1 unité de T.
- Produire une unité de boeuf prend 1 unité de L, 2 unités de T.
- On veut maximiser $W(x, y) = (1 - \beta) \ln x + \beta \ln y$, où x et y sont les quantités de blé et de boeuf.
- On écarte d'emblée la possibilité des valeurs $x < 0$, $y < 0$.

La fonction de Lagrange :

$$L(x, y, \mu_L, \mu_T) = (1-\beta) \ln x + \beta \ln y + \mu_L[300-2x-y] + \mu_T[450-x-2y]$$

Chomage technique, conditions de première ordre

$$L(x, y, \mu_L, \mu_T) = (1-\beta)\ln x + \beta \ln y + \mu_L[300-2x-y] + \mu_T[450-x-2y]$$

Les conditions de première ordre sont

$$\frac{1-\beta}{x} - 2\mu_L - \mu_T = 0, \quad \frac{\beta}{y} - \mu_L - 2\mu_T = 0,$$

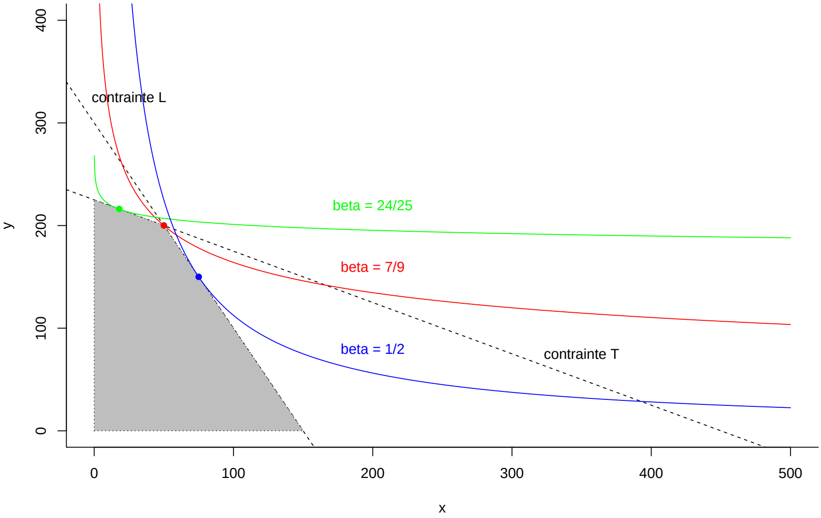
et avec écarts complémentaires,

$$300 - 2x - y \geq 0, \quad \mu_L \geq 0;$$

$$450 - x - 2y \geq 0, \quad \mu_T \geq 0.$$

- $\mu_L = 0$, $\mu_T = 0$ ne vérifie pas les deux premières équations.
- $\mu_L > 0$, $\mu_T > 0$ donne le plan unique sans chomage, mais il faut vérifier $\mu_L > 0$ et $\mu_T > 0$: il faut que $2/3 < \beta < 8/9$.
- $\mu_L = 0$ et $\mu_T > 0$ (chomage de L) requiert $\beta \geq 8/9$.
- $\mu_L > 0$ et $\mu_T = 0$ (chomage de T) requiert $\beta \leq 2/3$.

Trois solutions, selon la valeur de β (‘chomage.R’)



Éléments de l'analyse maximum de vraisemblance

- Quantités pertinentes :
  - θ, un vecteur de paramètres inconnus,
  - $y = (y_1, \dots, y_T)$, un vecteur aléatoire des variables observables,
  - y°, le vecteur observé.
- Fonctions pertinentes :
  - $f(y|\theta)$, la densité conditionnelle des données (modèle),
  - $\mathcal{L}(\theta; y) = f(y|\theta)$, la vraisemblance,
  - $\mathcal{L}(\theta; y^\circ) = f(y^\circ|\theta)$, la vraisemblance réalisée.

Le modèle Bernoulli

- Supposez que les y_i sont iid Bernoulli avec probabilité $\theta \in [0,1]$: $y_i = 1$ avec probabilité θ, $y_i = 0$ avec probabilité $(1-\theta)$.
- Alors la fonction de masse de probabilité de y_i est

$$\begin{aligned} f(y_i|\theta) &= \begin{cases} \theta & y_i = 1 \\ (1-\theta) & y_i = 0 \end{cases} \\ &= \theta^{y_i}(1-\theta)^{1-y_i}. \end{aligned}$$

- On observe le vecteur aléatoire $y = (y_1, \ldots, y_n)$; la fonction de masse de probabilité de y est

$$f(y|\theta) = \prod_{i=1}^{n} f(y_i|\theta) = \prod_{i=1}^{n} \theta^{y_i}(1-\theta)^{1-y_i} = \theta^{n_1}(1-\theta)^{n_0},$$

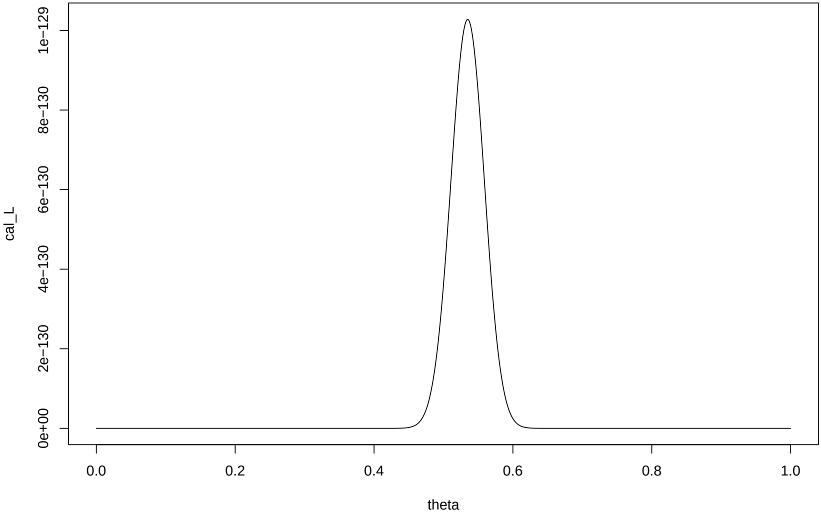
  où
  - $n_1 = \sum_{i=1}^{n} y_i$ est le nombre de fois qu'on observe 1, et
  - $n_0 = n - \sum_{i=1}^{n} y_i$ est le nombre de fois qu'on observe 0.

Deux intérpretations de la même expression

- Deux intérpretations de l'expression $\theta^{n_1}(1-\theta)^{n_0}$:
  - Fonction de masse de probabilité $f(y|\theta) = \theta^{n_1}(1-\theta)^{n_0}$.
  - Fonction de vraisemblance $\mathcal{L}(\theta; y) = \theta^{n_1}(1-\theta)^{n_0}$.
- $f(y|\theta)$ donne, pour θ fixe, les probabilités relatives des séquences possibles $(y_1, \ldots, y_n)$.
- $\mathcal{L}(\theta; y)$ donne, pour y fixe (notamment $y = y^\circ$) une note (ou évaluation) à chaque valeur θ pour la qualité de sa prévision des données observées.
- Soit $L(\theta; y) = \log \mathcal{L}(\theta; y)$, la log-vraisemblance.

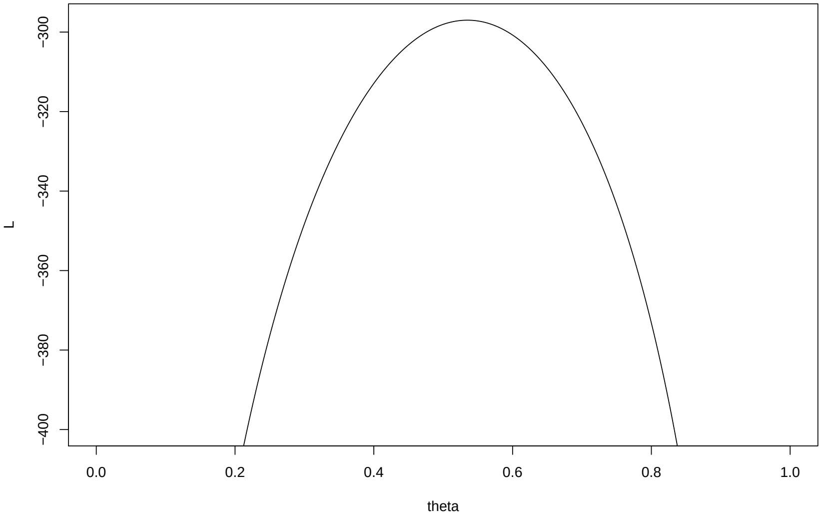
La vraisemblance Bernoulli pour $n_0 = 200$, $n_1 = 230$

```
n_0 = 200; n_1 = 230; theta = seq(0, 1, by=0.001)
cal_L = theta^n_1 * (1-theta)^n_0
plot(theta, cal_L, type='l')
```



La log vraisemblance Bernoulli pour $n_0 = 200$, $n_1 = 230$

```
L = n_1 * log(theta) + n_0 * log(1-theta)
plot(theta, L, type='l', ylim=c(-400, max(L)))
```



Maximum de la vraisemblance Bernoulli

- Vraisemblance : $\mathcal{L}(\theta; y) = \theta^{n_1}(1-\theta)^{n_0}$.
- Log vraisemblance : $L(\theta; y) = n_1 \log(\theta) + n_0 \log(1-\theta)$
- Deux dérivées de la log vraisemblance :

$$\frac{\partial L(\theta; y)}{\partial \theta} = \frac{n_1}{\theta} - \frac{n_0}{1-\theta}$$

$$\frac{\partial^2 L(\theta; y)}{\partial \theta^2} = -\frac{n_1}{\theta^2} - \frac{n_0}{(1-\theta)^2} < 0.$$

- La valeur $\hat{\theta}$ (souvent vue comme une variable aléatoire) qui maximise la vraisemblance et la log-vraisemblance est

$$\hat{\theta} = \frac{n_1}{n_0 + n_1} = \frac{n_1}{n}.$$

- Pour $n_0 = 200$ et $n_1 = 230$, $\hat{\theta} = \frac{23}{43} \approx 0.5349$.

Maximum de vraisemblance : conditions de régularité

- Définitions :
  - θ est le vecteur des paramètres ; Θ, l'ensemble de toutes les valeurs possibles de θ.
  - y est le vecteur (aléatoire) des données.
- Conditions informelles de regularité :
  1. Le modèle est correct pour une valeur $\theta = \theta_0 \in \Theta$.
  2. La vraie valeur θ_0 est dans l'intérieur de Θ.
  3. Identification :

  $$\theta \neq \theta_0 \Rightarrow f(\cdot|\theta) \neq f(\cdot|\theta_0).$$

  4. $L(\theta; y) \equiv \log f(y|\theta)$ a toujours un maximum global unique.
  5. Le gradient de $L(\theta; y)$ (par rapport à θ) est toujours borné.
  6. La matrice $\mathcal{I}(\theta)$ suivante (matrice d'information de Fisher) est définie positive:

  $$\mathcal{I}(\theta) = E_{y|\theta}\left[\frac{\partial L(\theta; y)}{\partial \theta^\top}\frac{\partial L(\theta; y)}{\partial \theta}\right].$$

Maximum de vraisemblance : résultats

Résultats : (Soit $\hat{\theta} \equiv \arg\max_\theta L(\theta; y)$, qui existe et est unique.)

1. $\hat{\theta} \to_p \theta_0$ (loi de grands nombres)
2. $\sqrt{n}(\hat{\theta} - \theta_0) \to_d N(0, n\mathcal{I}(\theta_0)^{-1})$ (théorème central limite)
3. $\mathcal{I}(\theta) = E_{y|\theta}\left[-\frac{\partial^2 L(\theta;y)}{\partial\theta\partial\theta^\top}\right]$.

Problèmes restants :

1. Il faut trouver $\hat{\theta}$.
2. La variance asymptotique $\mathcal{I}(\theta_0)^{-1}$ de $\hat{\theta}$ dépend de θ_0, qui est inconnu.
3. L'espérance dans les deux expressions pour $\mathcal{I}(\theta)$ sont difficiles à évaluer analytiquement.

Exemple Bernoulli

- Un cas rare où les calculs analytiques sont faisables.
- La matrice d'information de Fisher :

$$\mathcal{I}(\theta) = E_{y|\theta}\left[-\frac{\partial^2 L}{\partial\theta^2}\right] = E_{y|\theta}\left[\frac{n_1}{\theta^2} + \frac{n_0}{(1-\theta)^2}\right]$$
$$= \frac{n\theta}{\theta^2} + \frac{n(1-\theta)}{(1-\theta)^2} = \frac{n}{\theta(1-\theta)}.$$

- La variance de $\hat{\theta}$ (exacte, pas asymptotique) :

$$\mathrm{Var}[\hat{\theta}] = \mathrm{Var}\left[\frac{n_1}{n}\right] = \frac{1}{n^2} n \mathrm{Var}[y_i] = \frac{1}{n}(\theta - \theta^2) = \frac{\theta(1-\theta)}{n}.$$

- Pour $n_0 = 200$ et $n_1 = 230$, $\mathrm{Var}[\hat{\theta}]$ est de $(0.02411)^2$ pour $\theta = 1/2$ et $(0.02405)^2$ pour $\theta = \hat{\theta} \approx 0.5349$

Éléments de l'analyse bayésienne

- Quantités pertinentes :
  - θ, un vecteur de paramètres inconnus *aléatoire*
  - $y = (y_1, \ldots, y_T)$, un vecteur aléatoire des variables observables,
  - y°, le vecteur observé.
- Fonctions pertinentes :
  - $f(y|\theta)$, la densité conditionnelle des données (modèle),
  - $\mathcal{L}(\theta; y^\circ) = f(y^\circ|\theta)$, la vraisemblance réalisé,
  - $f(\theta)$, la densité *a priori*,
  - $f(\theta, y)$, la densité conjointe,
  - $f(\theta|y)$, la densité *a posteriori*,
  - $f(y)$, la densité marginale des données,
  - $f(y^\circ)$, la vraisemblance marginale (un nombre).

Inférence bayésienne

- Par la règle de Bayes,

$$f(\theta|y^\circ) = \frac{f(\theta, y^\circ)}{f(y^\circ)} = \frac{f(\theta)f(y^\circ|\theta)}{f(y^\circ)} \propto f(\theta)f(y^\circ|\theta).$$

- $f(\theta)$ représente notre incertitude sur θ avant l'observation de y.
- $f(\theta|y^\circ)$ resprésente notre incertitude sur θ après qu'observe $y = y^\circ$.
- Un point important à retenir : $f(\theta|y^\circ) \propto f(\theta, y^\circ)$.

Reprise et extension de l'exemple Bernoulli

- Si y_i est Bernoulli avec probabilité θ, $f(y|\theta) = \theta^{n_1}(1-\theta)^{n_0}$.
- Mettons qu'on choisit la loi *a priori* $\theta \sim \mathrm{Beta}(\alpha, \beta)$ sur $[0, 1]$:

$$f(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}.$$

- La densité conjointe est

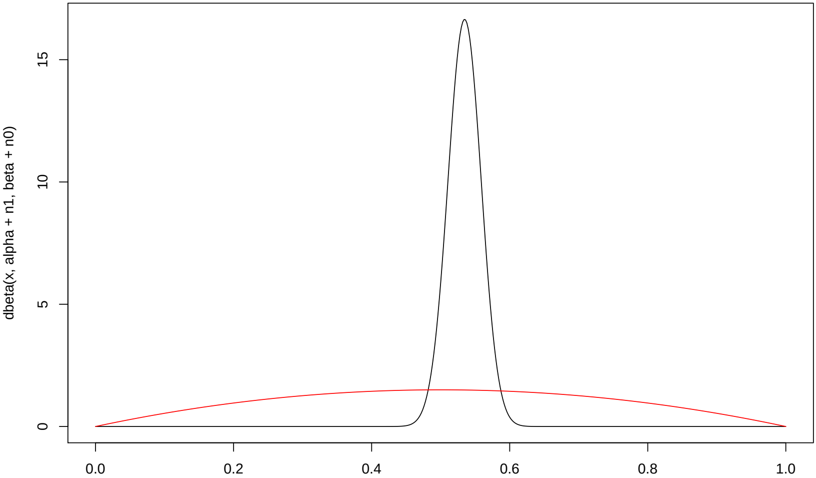
$$f(\theta, y) = f(\theta)f(y|\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha+n_1-1}(1-\theta)^{\beta+n_0-1}.$$

- La loi *a posteriori* doit être $\theta \sim \mathrm{Beta}(\alpha + n_1, \beta + n_0)$.
- La vraisemblance marginale est $f(\theta, y)/f(\theta|y)$:

$$f(y) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\frac{\Gamma(\alpha+n_1)\Gamma(\beta+n_0)}{\Gamma(\alpha+\beta+n)}.$$

Graphique pour l'exemple Bernoulli

```
n0 = 200; n1 = 230; alpha=2; beta=2
x = seq(0, 1, by=0.002)
plot(x, dbeta(x, alpha+n1, beta+n0), type='l')
lines(x, dbeta(x, alpha, beta), col='red')
```



Exemple gaussien I

- Considérez les modèle $y_t \sim \mathrm{iid}\, N(\mu, h^{-1})$.
- Le vecteur de paramètres est $\theta = (\mu, h)$.
- Le vecteur d'observables est $y = (y_1, \ldots, y_T)$.
- La densité des données est

$$f(y|\theta) = \prod_{t=1}^{T} \sqrt{\frac{h}{2\pi}} \exp\left[-\frac{h}{2}(y_t - \mu)^2\right]$$
$$= \left(\frac{h}{2\pi}\right)^{T/2} \exp\left[-\frac{h}{2}\sum_{t=1}^{T}(y_t - \mu)^2\right].$$

Exemple gaussien II

- Mettons qu'on choisit une loi *a priori* où h et μ sont indépendents, avec

$$\mu \sim N(\bar{\mu}, \bar{\omega}^{-1}), \quad \bar{s}^2 h \sim \chi^2(\bar{\nu}),$$

où $\bar{\mu}$, $\bar{\omega}$, $\bar{s}$ et $\bar{\nu}$ sont des hyperparamètres constants choisis par l'investigateur.
- La densité *a priori* est

$$f(\theta) \propto \exp\left[-\frac{\bar{\omega}}{2}(\mu - \bar{\mu})^2\right] \cdot h^{(\bar{\nu}-2)/2} \exp\left[-\frac{1}{2}\bar{s}^2 h\right].$$

- La densité conjointe est

$$f(\theta, y) \propto h^{(\bar{\nu}+T-2)/2} \exp\left[-\frac{\bar{\omega}}{2}(\mu - \bar{\mu})^2 - \frac{h}{2}\left(\bar{s}^2 + \sum_{t=1}^{T}(y_t - \mu)^2\right)\right].$$

L'intégration et les objectifs de l'analyse bayésienne

- Plusieurs problèmes d'inférence bayésienne ont, comme solution, une intégrale par rapport à la densité *a posteriori*.
- Exemple 1, estimation ponctuelle de θ_k sous perte quadratique:

$$\hat{\theta}_k = E[\theta_k|y^\circ] = \int \theta_k f(\theta|y^\circ)\, d\theta.$$

- Exemple 2, quantification de l'incertitude sur θ_k :

$$\mathrm{Var}[\theta|y^\circ] = E[(\theta_k - E[\theta_k|y^\circ])^2|y^\circ].$$

- Exemple 3, densité prédictive (valeurs de y_{T+1} sur une grille) :

$$f(y_{T+1}|y^\circ) = E[f(y_{T+1}|\theta, y^\circ)|y^\circ].$$

Preuve de l'exemple 3

$$
\begin{aligned}
E[f(y_{T+1}|y_1,\ldots,y_T,\theta)|y_1,\ldots,y_T] \\
&= \int f(y_{T+1}|y_1,\ldots,y_T,\theta)f(\theta|y_1,\ldots,y_T)\,d\theta \\
&= \int f(y_{T+1},\theta|y_1,\ldots,y_T)\,d\theta \\
&= f(y_{T+1}|y_1,\ldots,y_T)
\end{aligned}
$$

Méthodes pour trouver $E[g(\theta)|y^\circ]$

- Calcul analytique : élégant, exacte, presque toujours insoluble.
- Simulation Monte Carlo indépendante :
  - Si on peut simuler $\theta^m \sim \mathrm{iid}\, \theta|y^\circ$,
  $$\frac{1}{M}\sum_{m=1}^{M} g(\theta^m) \to_p E[g(\theta)|y^\circ].$$
  - Cependant, cette simulation est rarement faisable.
- Simulation Monte Carlo chaîne de markov (MCMC) :
  - On choisit un processus markovien avec densité de transition $f(\theta^m|\theta^{m-1})$ telle que la loi *a posteriori* $\theta|y^\circ$ est la loi stationnaire du processus. C'est à dire :
  $$\theta^{m-1} \sim f(\theta|y^\circ) \Rightarrow \theta^m \sim f(\theta|y^\circ).$$
  - Sous quelques conditions techniques, la loi de θ^m converge à la loi *a posteriori* et
  $$\frac{1}{M}\sum_{m=1}^{M} g(\theta^m) \to_p E[g(\theta)|y^\circ].$$