Partiel d'Économétrie Bayésienne

Stéphane Adjemian

Avril 2007

QUESTION. Nous avons vu en cours plusieurs approches pour formaliser l'ignorance sur un paramètre. On peut par exemple utiliser le prior « plat » qui stipule que :

1. si le paramètre d'intérêt peut prendre des valeurs entre $-\infty$ et ∞, alors la distribution *a priori* non-informative est uniforme entre $-\infty$ et ∞.
2. si le paramètre d'intérêt peut prendre des valeurs entre 0 et ∞, alors la distribution *a priori* non-informative du logarithme de ce paramètre est uniforme entre $-\infty$ et ∞.

Expliquez en quoi un prior « plat » est non informatif (dans les cas 1 et 2).

Deux propriétés permettent d'interpréter le prior « plat » comme un prior non informatif. La plus évidente est la propriété d'uniformité. Les probabilités que μ soit dans $[a, a+h]$ ou dans $[b, b+h]$ sont identiques. Mais l'uniformité ne suffit pas à caractériser l'absence d'information. En effet, si $h > h'$ alors un prior uniforme (c'est à dire l'utilisation d'une densitée de probabilité uniforme sur un support borné) dit que la probabilité de l'événement $\mu \in [a, a+h]$ est supérieure à la probabilité de l'événement $\mu \in [b, b+h']$. Il y a donc de l'information dans un prior uniforme. Le prior « plat » évoqué plus haut suppose une distribution uniforme entre $-\infty$ et ∞. Il s'agit donc d'une densité qui ne somme pas à un : l'intégrale de la densité de probabilité par rapport à μ entre $-\infty$ et ∞ n'est pas définie. On parle alors de prior impropre. Cette seconde propriété est intéréssante, car on ne peut plus alors comparer les évènements $\mu \in [a, a+h]$ et $\mu \in [b, b+h']$ puisque tout deux sont de mesures nulles. Ainsi on ne peut dire si un évènement est plus probable qu'un autre. En ce sens le prior est non informatif.

Pour un paramètre, disons σ, ne pouvant prendre que des valeurs positives, par exemple un écart-type, dire que le $\log$ du paramètre est uniforme entre $-\infty$ et ∞ revient à dire que la densité de ce paramètre est proportionnelle à $1/\sigma$.

$$p(\log\sigma) \propto 1$$

$$\Leftrightarrow p(\sigma) \propto \frac{1}{\sigma}$$

On montre facilement que :

1. $\int_0^\infty p(\sigma)\mathrm{d}\sigma = \infty$
2. $\int_0^a p(\sigma)\mathrm{d}\sigma = \infty$
3. $\int_a^\infty p(\sigma)\mathrm{d}\sigma = \infty$

Le premier point nous dit que le prior est impropre. Nous avons vu que cette propriété est intéressante si nous cherchons à caractériser le non savoir. les deux point suivants sont tout aussi intéressants. Le fait que ces deux intégrales soient non définies nous dit que l'évènement « σ petit » n'est pas plus (ou moins) probable que l'évènement « σ grand ». La comparaison des deux évènement est indéterminée. En ce sens, le prior est non informatif.

EXERCICE. On considère le processus générateur des données (DGP) suivant :

$$y_t = \rho y_{t-1} + \varepsilon_t$$

avec $\rho \in \mathbb{R}$, $\varepsilon_t \underset{iid}{\sim} \mathcal{N}\left(0, \sigma^2\right)$ et $y_0 = 0$, pour $t = 1, \ldots, T$. On notera $\mathcal{Y}_T \equiv (y_1, y_2, \ldots, y_T)$ l'échantillon.

(1.1) Écrivez la densité de y_1 conditionnellement à $y_0 = 0$, ϕ et σ^2.

Nous savons que $y_1 = \rho y_0 + \varepsilon_1 = \varepsilon_1$. Ainsi la distribution de y_1 conditionnelle à ρ, σ^2 et y_0 est gaussienne d'espérance nulle et de variance σ^2. Nous avons donc :

$$p(y_1|y_0, \rho, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2} y_1^2}$$

(1.2) Écrivez la densité de y_t conditionnellement à y_{t-1}, ϕ et σ^2.

Nous savons que $y_t = \rho y_{t-1} + \varepsilon_t$. Ainsi la distribution de y_t conditionnelle à ρ, σ^2 et y_{t-1} est gaussienne d'espérance ρy_{t-1} et de variance σ^2. Nous avons donc :

$$p(y_t|y_{t-1}, \rho, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y_t - \rho y_{t-1})^2}$$

(1.3) Montrez que la vraisemblance associée à ce modèle est :

$$\mathcal{L}(\rho, \sigma^2; \mathcal{Y}_T) = \left(\sigma^2\right)^{-\frac{T}{2}} (2\pi)^{-\frac{T}{2}} e^{-\frac{1}{2\sigma^2}\sum_{t=1}^{T}(y_t - \rho y_{t-1})^2}$$

La vraisemblance est la densité jointe de l'échantillon :

$$\mathcal{L}(\rho, \sigma^2; \mathcal{Y}_T) = p(y_T, y_{T-1}, \ldots, y_2, y_1|\rho, \sigma^2)$$

En utilisant T fois le théorème de Bayes, et en notant que la condition initiale est déterministe, on peut réécrire cette densité jointe sous la forme :

$$\mathcal{L}(\rho, \sigma^2; \mathcal{Y}_T) = \prod_{t=1}^{T} p(y_t|y_{t-1}|\rho, \sigma^2)$$

En substituant les précédantes réponses on obtient directement ce qu'il fallait montrer.

(2) Supposons que le paramètre σ^2 soit connu. Donnez l'estimateur du maximum de vraisemblance de ρ, que nous noterons $\hat{\rho}_T$. Calculez la variance, $\mathbb{V}[\hat{\rho}_T]$, de cet estimateur. Comment la variance de cet estimateur est-elle affectée par une augmentation de la taille de l'échantillon ? Que pouvez-vous dire de la distribution de $\hat{\rho}_T$ conditionnellement à ρ ?

Dans ce modèle, l'estimateur du MV est identique à l'estimateur des MCO. Celui-ci est donné par :

$$\widehat{\rho}_T = \frac{\sum_{t=1}^{T} y_t y_{t-1}}{\sum_{t=1}^{T} y_{t-1}^2}$$

La variance de l'estimateur du maximum de vraisemblance est donnée par :

$$\mathbb{V}[\widehat{\rho}_T] = \sigma^2 \left(\sum_{t=1}^{T} y_{t-1}^2 \right)^{-1}$$

En notant que pour tout t y_{t-1}^2 est positif, on voit que lorsque la taille de l'échantillon augmente la somme des y_{t-1}^2 augmente. Ainsi, lorsque la taille de l'échantillon augmente, la variance de l'estimateur du MV diminue. Plus l'échantillon est informatif plus la précision de l'estimateur est importante. La distribution de $\widehat{\rho}_T$ sachant ρ dépend de la valeur de ρ. Si le DGP est stationnaire ($|\rho| < 1$) la distribution de l'estimateur est gaussienne (centrée sur la vraie valeur). En présence d'une racine unitaire $\rho = 1$ la distribution de l'estimateur est non standard, il s'agit d'une fonctionnelle de processus de Wiener (voir un cours de Séries Temporelles et la distribution asymptotique du test de Dickey-Fuller).

(3) Nous n'avons pas d'*a priori* sur la vraie valeur de ρ qui peut prendre des valeurs entre $-\infty$ et ∞. Nous utilisons donc un prior « plat ». Montrez que la distribution *a posteriori* est gaussienne. Donnez l'espérance et la variance *a posteriori*. Commentez.

Puisque nous considérons un prior «plat» pour ρ et que nous supposons σ^2 connu, la densité *a priori* pour ρ est proportionnelle à une constante entre $-\infty$ et ∞. La densité postérieure est donc proportionnelle à la fonction de vraisemblance :

$$p(\rho|\mathcal{Y}_T) = \mathcal{L}(\rho; \sigma^2, \mathcal{Y}_T)$$

Réécrivons la vraisemblance en faisant apparaître l'estimateur du MV (c'est-à-dire une statistique résumant l'information amenée par l'échantillon)

obtenu plus haut. Nous avons :

$$\begin{aligned}\sum_{t=1}^{T}(y_t-\rho y_{t-1})^2 &= \sum_{t=1}^{T}(y_t-\widehat{\rho}_T y_{t-1}+\widehat{\rho}_T y_{t-1}-\rho y_{t-1})^2\\ &= \sum_{t=1}^{T}(y_t-\widehat{\rho}_T y_{t-1})^2+(\rho-\widehat{\rho}_T)^2\sum_{t=1}^{T}y_{t-1}^2\\ &\quad +2(\rho-\widehat{\rho}_T)\sum_{t=1}^{T}(y_t-\widehat{\rho}_T y_{t-1})\,y_{t-1}\end{aligned}$$

où le dernier terme est nul par construction de l'estimateur du MV. Ainsi la densité postérieure est caractérisée par :

$$\begin{aligned}p(\rho|\mathcal{Y}_T) &\propto \left(\sigma^2 2\pi\right)^{-\frac{T}{2}}\exp\left\{-\frac{1}{2\sigma^2 s^2}\left(\rho-\widehat{\rho}_T\right)^2\right\}\\ &\quad\times\exp\left\{-\frac{1}{2\sigma^2}\sum_{t=1}^{T}\left(y_t-\widehat{\rho}_T y_{t-1}\right)^2\right\}\end{aligned}$$

Si on concentre sur la première exponentielle (le seul terme où apparaît le paramètre ρ) on voit que la distribution postérieure est gaussienne d'espérance l'estimateur du MV et de variance la variance de l'estimateur du MV ($\sigma^2 s^2$). La distribution postérieure est d'autant plus concentrée que la taille de l'échantillon est grande ou que la taille de l'innovation est faible.

(4) En travaillant avec une fonction de perte quadratique, donnez une estimation ponctuelle de ρ.

Si la fonction de perte est quadratique l'estimation ponctuelle est l'espérance postérieure, c'est-à-dire dans notre cas l'estimateur du maximum de vraisemblance. En effet, l'estimation ponctuelle est obtenue en minimisant la perte espérée :

$$\rho^\star = \arg\min_a \int (a-\rho)^2 p(\rho|\mathcal{Y}_T)\mathrm{d}\rho$$

La condition nécessaire (et suffisante puisque l'objectif est quadratique) d'optimalité est :

$$2\int(\rho^\star-\rho)p(\rho|\mathcal{Y}_T)\mathrm{d}\rho = 0$$

soit de façon équivalente :

$$\rho^\star = \int \rho p(\rho|\mathcal{Y}_T)\mathrm{d}\rho$$

l'espérance postérieure de ρ. L'estimation ponctuelle bayésienne n'est pas différente de l'estimateur du maximum de vraisemblance car nous avons considéré un *a priori* non informatif.

(5) Calculez la densité marginale de l'échantillon. À quoi peut servir cette densité ?

La densité marginale de l'échantillon est l'intégrale du noyau postérieur par rapport au paramètre ρ. Cette quantité mesure la qualité d'ajustement du modèle, et permet la comparaison de modèles éventuellement non imbriqués. Le noyau postérieur est le membre de droite de la dernière équation dans la réponse à la question (3). L'intégration par rapport à ρ, en utilisant la définition de la densité gaussienne, donne directement :

$$\begin{aligned}p\left(\mathcal{Y}_T\right) &= \left(\sigma^2 2\pi\right)^{-\frac{T}{2}}\left(\sigma^2 s^2 2\pi\right)^{-\frac{1}{2}}\\ &\quad\times\exp\left\{-\frac{1}{2\sigma^2}\sum_{t=1}^{T}\left(y_t-\widehat{\rho}_T y_{t-1}\right)^2\right\}\end{aligned}$$

(6) On se donne maintenant un prior plus informatif, en supposant que $\rho\sim\mathcal{N}\left(\rho_0,\sigma_0^2\right)$. Caractérisez la distribution postérieure. Interprétez les résultats.

La densité postérieure est maintenant caractérisée par :

$$\begin{aligned}p(\rho|\mathcal{Y}_T) &\propto \left(\sigma^2 2\pi\right)^{-\frac{T}{2}}\left(\sigma_0^2 2\pi\right)^{-\frac{1}{2}}\\ &\quad\times\exp\{-\frac{1}{2}\frac{\nu\widehat{\sigma}^2}{\sigma^2}\}\\ &\quad\times\exp\left\{-\frac{1}{2\sigma^2 s^2}\left(\rho-\widehat{\rho}_T\right)^2\right\}\\ &\quad\times\exp\left\{-\frac{1}{2\sigma_0^2}\left(\rho-\rho_0\right)^2\right\}\end{aligned}$$

où $\nu = T-1$ et $\widehat{\sigma}^2 = (T-1)^{-1}\sum_{t=1}^{T}\left(y_t-\widehat{\rho}_T y_{t-1}\right)^2$. Réécrivons les deux derniers termes de façon à faire apparaître une unique forme quadratique en ρ. Posons :

$$A(\rho) = \frac{\left(\rho-\widehat{\rho}_T\right)^2}{\sigma^2 s^2}+\frac{\left(\rho-\rho_0\right)^2}{\sigma_0^2}$$

En développant les numérateurs et en regroupant les termes « qui vont bien ensemble » nous obtenons :

$$A(\rho) = \left(\frac{1}{s^2\sigma^2} + \frac{1}{\sigma_0^2}\right)\rho^2 - 2\rho\left(\frac{\widehat{\rho}_T}{s^2\sigma^2} + \frac{\rho_0}{\sigma_0^2}\right) + \left(\frac{\widehat{\rho}_T^2}{s^2\sigma^2} + \frac{\rho_0^2}{\sigma_0^2}\right)$$

soit :

$$A(\rho) = \left(\frac{1}{s^2\sigma^2} + \frac{1}{\sigma_0^2}\right)\left[\rho - \frac{\frac{\widehat{\rho}_T}{s^2\sigma^2} + \frac{\rho_0}{\sigma_0^2}}{\frac{1}{s^2\sigma^2} + \frac{1}{\sigma_0^2}}\right] + \widetilde{A}$$

où $\widetilde{A}$ contient des termes qui ne dépendent pas de ρ. En notant $B(\rho) = A(\rho) - \widetilde{A}$, la densité postérieure de ρ est caractérisée par :

$$p(\rho|\mathcal{Y}_T) \propto \exp\left\{-\frac{1}{2}B(\rho)\right\}$$

la densité postérieure est donc gaussienne les moments d'ordre un et deux sont :

$$\mathbb{E}[\rho] = \frac{\frac{\widehat{\rho}_T}{s^2\sigma^2} + \frac{\rho_0}{\sigma_0^2}}{\frac{1}{s^2\sigma^2} + \frac{1}{\sigma_0^2}}$$

$$\mathbb{V}[\rho] = \frac{1}{\frac{1}{s^2\sigma^2} + \frac{1}{\sigma_0^2}}$$

La variance postérieure tend vers la variance de l'estimateur du MV lorsque la variance du prior tend vers l'infini, c'est-à-dire lorsque le prior devient de moins en moins informatif. La variance postérieure tend vers 0 lorsque la variance *a priori* tend vers 0, c'est-à-dire lorsque l'on se rapproche d'un étalonnage du modèle. L'espérance *a posteriori* est une combinaison linéaire convexe de l'estimateur du maximum de vraisemblance et de l'espérance *a priori*. L'espérance postérieure est d'autant plus proche de l'estimateur du MV que l'échantillon est informatif (*ie* $s^2\sigma^2$ petit relativement à σ_0^2), lorsque la taille de l'échantillon tend vers l'infini l'espérance postérieure tend donc vers l'estimateur du MV.

(7) On suppose que le prior sur ρ est spécifié par une densité de probabilité quelconque $\varphi(\rho)$ (φ est une fonction positive qui somme à un). Que pouvez vous dire de la distribution postérieure ? Proposez une approximation asymptotique de la densité marginale.

La densité postérieure est proportionnelle au produit de $\varphi(\rho)$ par la vraisemblance. Cette fois-ci on ne peut mettre aussi facilement un nom sur la forme de la distribution postérieure, il y a peu de chance qu'il s'agisse d'une distribution gaussienne... Le calcul des moments postérieurs ou de la densité marginale de l'échantillon est donc moins évident. Néanmoins, nous avons vu en cours qu'asymptotiquement la distribution postérieure tend vers une distribution gaussienne. Si l'échantillon est de taille raisonnable nous pouvons donc approximer la distribution postérieure avec une distribution normale. En approximant, à l'ordre deux, le noyau postérieur autour de son mode, il vient :

$$\mathcal{K}(\rho) \doteq \mathcal{K}(\rho^*)e^{-\frac{1}{2}[\mathcal{K}''(\rho^*)]^{-1}(\rho-\rho^*)^2} \equiv \widetilde{\mathcal{K}}(\rho)$$

où $\mathcal{K}$ est le noyau postérieur et ρ^* la valeur de ρ au mode. On reconnaît la densité d'une loi normale, à une constante d'intégration près. Notons c cette constante d'intégration, c'est-à-dire la constante telle que $\int c^{-1}\widetilde{\mathcal{K}}(\rho)\mathrm{d}\rho = 1$. Cette constante est une approximation de la densité marginale, $p(\mathcal{Y}_T)$. Par définition de la densité d'une loi normale, on a :

$$c = \mathcal{K}(\rho^*)(2\pi)^{\frac{1}{2}}\mathcal{K}''(\rho^*)^{\frac{1}{2}}$$

On dit que c est l'approximation de Laplace de la densité marginale.