Détermination de lois a priori

Anna Simoni[2]

[2]CREST - Ensae and CNRS

Outline

Introduction

Deux modes de pensée Bayésiens :

- **subjectiviste :** la distribution *a priori* traduit les connaissances avant l'observation des données (exemple : opinion des experts). Importance des distributions naturelles conjuguées à un modèle d'échantillonage.

- **objectif :** l'*a priori* n'est pas derivée des connaissances ex-ante de l'utilisateur. Il s'agit de rester bayésien en l'absence d'information a priori : (i) a priori non informatives ou (ii) bayésien empirique.

- Il est rare que l'information a priori soit suffisamment précise pour conduire à une détermination exacte de la loi a priori (plusieurs lois de probabilité peuvent être compatibles avec cette information) : choix souvent partiellement arbitraire.

- Il n'y a pas une façon unique de choisir une loi a priori, et le choix de cette loi a un impact sur l'inférence

- Remarque : (1) les lois a priori non fondées fournissent des inférences a posteriori non justifiées ; (2) le concept d'une loi a priori unique n'a pas de sens, sauf dans des cas très particuliers.

Introduction

Deux modes de pensée Bayésiens :

- **subjectiviste :** la distribution *a priori* traduit les connaissances avant l'observation des données (exemple : opinion des experts). Importance des distributions naturelles conjuguées à un modèle d'échantillonage.

- **objectif :** l'*a priori* n'est pas derivée des connaissances ex-ante de l'utilisateur. Il s'agit de rester bayésien en l'absence d'information a priori : (i) a priori non informatives ou (ii) bayésien empirique.

- Il est rare que l'information a priori soit suffisamment précise pour conduire à une détermination exacte de la loi a priori (plusieurs lois de probabilité peuvent être compatibles avec cette information) : choix souvent partiellement arbitraire.

- Il n'y a pas une façon unique de choisir une loi a priori, et le choix de cette loi a un impact sur l'inférence

- Remarque : (1) les lois a priori non fondées fournissent des inférences a posteriori non justifiées ; (2) le concept d'une loi a priori unique n'a pas de sens, sauf dans des cas très particuliers.

Introduction

Deux modes de pensée Bayésiens :

- **subjectiviste :** la distribution *a priori* traduit les connaissances avant l'observation des données (exemple : opinion des experts). Importance des distributions naturelles conjuguées à un modèle d'échantillonage.

- **objectif :** l'*a priori* n'est pas derivée des connaissances ex-ante de l'utilisateur. Il s'agit de rester bayésien en l'absence d'information a priori : (i) a priori non informatives ou (ii) bayésien empirique.

- Il est rare que l'information a priori soit suffisamment précise pour conduire à une détermination exacte de la loi a priori (plusieurs lois de probabilité peuvent être compatibles avec cette information) : choix souvent partiellement arbitraire.

- Il n'y a pas une façon unique de choisir une loi a priori, et le choix de cette loi a un impact sur l'inférence

- Remarque : (1) les lois a priori non fondées fournissent des inférences a posteriori non justifiées ; (2) le concept d'une loi a priori unique n'a pas de sens, sauf dans des cas très particuliers.

Introduction

Deux modes de pensée Bayésiens :

- **subjectiviste :** la distribution *a priori* traduit les connaissances avant l'observation des données (exemple : opinion des experts). Importance des distributions naturelles conjuguées à un modèle d'échantillonage.

- **objectif :** l'*a priori* n'est pas derivée des connaissances ex-ante de l'utilisateur. Il s'agit de rester bayésien en l'absence d'information a priori : (i) a priori non informatives ou (ii) bayésien empirique.

- Il est rare que l'information a priori soit suffisamment précise pour conduire à une détermination exacte de la loi a priori (plusieurs lois de probabilité peuvent être compatibles avec cette information) : choix souvent partiellement arbitraire.

- Il n'y a pas une façon unique de choisir une loi a priori, et le choix de cette loi a un impact sur l'inférence

- Remarque : (1) les lois a priori non fondées fournissent des inférences a posteriori non justifiées ; (2) le concept d'une loi a priori unique n'a pas de sens, sauf dans des cas très particuliers.

Outline

Distribution *a priori* non informatives

- Si aucune information a priori n'est disponible, il est impossible de justifier le choix d'une loi a priori sur des bases subjectives.

- Le choix d'une distribution a priori non informative conduit souvent à la spécification d'une mesure et non d'une probabilité.

- La procédure de spécification d'une mesure a priori non informative revient à definir une mesure sur Θ, à partir d'un mécanisme d'échantillonage décrit par l'échantillon $x \in \mathcal{X}$ et la probabilité d'échantillonage (qui est conditionnelle en général à la taille de l'échantillon et à des variables explicatives).

Mesures a priori. I

- Si le modèle d'échantillonage est défini par une densité (par rapport à Lebesgue) $f(x|\theta)$, θ fini-dimensionnel, une mesure à priori sera souvent caracterisée par sa densité $\pi(\theta)$ par rapport à la mesure de Lebesgue.

- $\pi(\cdot) : \Theta \to \mathbb{R}_+$ mais d'intégrale pas forcement finie.

  - Si $\int_\Theta \pi(\theta)d\theta < \infty$ on peut se ramener au cas d'une probabilité car le calcul de l'a posteriori n'en sera pas affecté.

  - Si $\int_\Theta \pi(\theta)d\theta = \infty$ alors, pour utiliser le Théorème de Bayes, il faut vérifier que $m(x) = \int_\Theta f(x|\theta)\pi(\theta)d\theta < \infty$. Si $m(x) = \infty$ alors la formule de Bayes ne peut plus se jusitifer comme le calcul d'une loi conditionnelle.

La loi a priori de Jeffreys. I

- Les lois a priori non informatives de *Jeffreys* sont fondées sur l'information de Fisher, donnée par :

$$I(\theta) = \mathbf{E}_\theta \left[\left(\frac{\partial \log \ell(\theta|X)}{\partial \theta} \right)^2 \right]$$

où $\ell(\theta|x) = f(x|\theta)$ est la vraisemblance qui caractérise le modèle d'échantillonage.

- Sous certaines conditions de régularité, cette information est aussi égale à

$$I(\theta) = -\mathbf{E}_\theta \left[\frac{\partial^2 \log f(X|\theta)}{\partial \theta \partial \theta'} \right].$$

- La loi a priori de Jeffreys est

$$\pi_J(\theta) \propto [detI(\theta)]^{1/2}$$

définie à un coefficient de normalisation près quand π est propre.

La loi a priori de Jeffreys. I

- Les lois a priori non informatives de *Jeffreys* sont fondées sur l'information de Fisher, donnée par :

$$I(\theta) = \mathbf{E}_\theta \left[\left(\frac{\partial \log \ell(\theta|X)}{\partial \theta} \right)^2 \right]$$

où $\ell(\theta|x) = f(x|\theta)$ est la vraisemblance qui caractérise le modèle d'échantillonage.

- Sous certaines conditions de régularité, cette information est aussi égale à

$$I(\theta) = -\mathbf{E}_\theta \left[\frac{\partial^2 \log f(X|\theta)}{\partial \theta \partial \theta'} \right].$$

- La loi a priori de Jeffreys est

$$\pi_J(\theta) \propto [detI(\theta)]^{1/2}$$

définie à un coefficient de normalisation près quand π est propre.

La loi a priori de Jeffreys. I

- Les lois a priori non informatives de *Jeffreys* sont fondées sur l'information de Fisher, donnée par :

$$I(\theta) = \mathbf{E}_\theta \left[\left(\frac{\partial \log \ell(\theta|X)}{\partial \theta} \right)^2 \right]$$

où $\ell(\theta|x) = f(x|\theta)$ est la vraisemblance qui caractérise le modèle d'échantillonage.

- Sous certaines conditions de régularité, cette information est aussi égale à

$$I(\theta) = -\mathbf{E}_\theta \left[\frac{\partial^2 \log f(X|\theta)}{\partial \theta \partial \theta'} \right].$$

- La loi a priori de Jeffreys est

$$\pi_J(\theta) \propto [detI(\theta)]^{1/2}$$

définie à un coefficient de normalisation près quand π est propre.

La loi a priori de Jeffreys. II

- Elle vérifie la propriété d'invariance par reparamétrisation : pour une transformation bijective h donnée, nous avons la transformation

$$I(\theta) = I(h(\theta))(h'(\theta))^2.$$

- Le choix d'une loi a priori dépendant de l'information de Fisher se justifie par le fait que $I(\theta)$ est accepté comme un indicateur de la quantité d'information apportée par le modèle (ou l'observation) sur θ.
- Favoriser les valeurs de θ pour lesquelles $I(\theta)$ est plus grande équivaut à minimiser l'influence de la loi a priori et est donc aussi non informatif que possible.
- Si $f(x|\theta)$ appartient à une famille exponentielle, $f(x|\theta) = h(x)\exp(\theta x - \psi(\theta))$, la matrice d'information de Fisher est donnée par $I(\theta) = \partial^2\psi(\theta)/(\partial\theta\partial\theta')$ et, pour $\Theta \subset \mathbb{R}^k$,

$$\pi_J(\theta) \propto \left[\prod_{i=1}^{k} \psi''_{ii}(\theta)\right]^{1/2}$$

où $\psi''_{ii}(\theta) = \partial^2\psi(\theta)/(\partial^2\theta_i)$.

La loi a priori de Jeffreys. II

- Elle vérifie la propriété d'invariance par reparamétrisation : pour une transformation bijective h donnée, nous avons la transformation

$$I(\theta) = I(h(\theta))(h'(\theta))^2.$$

- Le choix d'une loi a priori dépendant de l'information de Fisher se justifie par le fait que $I(\theta)$ est accepté comme un indicateur de la quantité d'information apportée par le modèle (ou l'observation) sur θ.
- Favoriser les valeurs de θ pour lesquelles $I(\theta)$ est plus grande équivaut à minimiser l'influence de la loi a priori et est donc aussi non informatif que possible.
- Si $f(x|\theta)$ appartient à une famille exponentielle, $f(x|\theta) = h(x)\exp(\theta x - \psi(\theta))$, la matrice d'information de Fisher est donnée par $I(\theta) = \partial^2\psi(\theta)/(\partial\theta\partial\theta')$ et, pour $\Theta \subset \mathbb{R}^k$,

$$\pi_J(\theta) \propto \left[\prod_{i=1}^{k} \psi''_{ii}(\theta)\right]^{1/2}$$

où $\psi''_{ii}(\theta) = \partial^2\psi(\theta)/(\partial^2\theta_i)$.

La loi a priori de Jeffreys. II

- Elle vérifie la propriété d'invariance par reparamétrisation : pour une transformation bijective h donnée, nous avons la transformation

$$I(\theta) = I(h(\theta))(h'(\theta))^2.$$

- Le choix d'une loi a priori dépendant de l'information de Fisher se justifie par le fait que $I(\theta)$ est accepté comme un indicateur de la quantité d'information apportée par le modèle (ou l'observation) sur θ.

- Favoriser les valeurs de θ pour lesquelles $I(\theta)$ est plus grande équivaut à minimiser l'influence de la loi a priori et est donc aussi non informatif que possible.

- Si $f(x|\theta)$ appartient à une famille exponentielle, $f(x|\theta) = h(x)\exp(\theta x - \psi(\theta))$, la matrice d'information de Fisher est donnée par $I(\theta) = \partial^2\psi(\theta)/(\partial\theta\partial\theta')$ et, pour $\Theta \subset \mathbb{R}^k$,

$$\pi_J(\theta) \propto \left[\prod_{i=1}^{k} \psi_{ii}''(\theta)\right]^{1/2}$$

où $\psi_{ii}''(\theta) = \partial^2\psi(\theta)/(\partial^2\theta_i)$.

La loi a priori de Jeffreys. II

- Dans une échantillonage i.i.d. la mesure de Jeffrey ne dépend de la taille de l'échantillon n que par un facteur multiplicatif que l'on peut donc négliger.
- La mesure de Jeffrey n'est pas affectée par la substitution d'une statistique exhaustive à l'échantillon initial (car $I(\theta)$ n'est pas modifiée).
- Une critique de la méthode de Jeffreys est que elle ne satisfait pas au principe de vraisemblance : l'information de Fisher peut différer pour deux expériences fournissant des vraisemblances proportionnelles.

La loi a priori de Jeffreys. II

- Dans une échantillonage i.i.d. la mesure de Jeffrey ne dépend de la taille de l'échantillon n que par un facteur multiplicatif que l'on peut donc négliger.
- La mesure de Jeffrey n'est pas affectée par la substitution d'une statistique exhaustive à l'échantillon initial (car $I(\theta)$ n'est pas modifiée).
- Une critique de la méthode de Jeffreys est que elle ne satisfait pas au principe de vraisemblance : l'information de Fisher peut différer pour deux expériences fournissant des vraisemblances proportionnelles.

La loi a priori de Jeffreys. II

- Dans une échantillonage i.i.d. la mesure de Jeffrey ne dépend de la taille de l'échantillon n que par un facteur multiplicatif que l'on peut donc négliger.
- La mesure de Jeffrey n'est pas affectée par la substitution d'une statistique exhaustive à l'échantillon initial (car $I(\theta)$ n'est pas modifiée).
- Une critique de la méthode de Jeffreys est que elle ne satisfait pas au principe de vraisemblance : l'information de Fisher peut différer pour deux expériences fournissant des vraisemblances proportionnelles.

Mesures a priori de référence. I

- Proposées par Bernardo (1979). L'analyse de référence est un mode général de spécification d'une loi a priori contenant aussi peu d'information que possible.
- Modification de l'approche de Jeffreys. Une différence majeure est que cette méthode fait la distinction entre paramètres d'intérêt et paramètres de nuisance.
- Idée : soit $x \sim f(x|\theta)$ et $\theta = (\theta_1, \theta_2)$, où θ_1 est le paramètre d'intérêt. La loi de référence est obtenue en définissant d'abord $\pi(\theta_2|\theta_1)$ comme la loi de Jeffreys associée à $f(x|\theta)$ pour θ_1 fixé, puis en calculant la loi marginale

$$\widetilde{f}(x|\theta_1) = \int f(x|\theta_1, \theta_2)\pi(\theta_2|\theta_1)d\theta_2$$

et la loi de Jeffreys $\pi(\theta_1)$ associée à $\widetilde{f}(x|\theta_1)$.

Mesures a priori de référence. I

- Proposées par Bernardo (1979). L'analyse de référence est un mode général de spécification d'une loi a priori contenant aussi peu d'information que possible.
- Modification de l'approche de Jeffreys. Une différence majeure est que cette méthode fait la distinction entre paramètres d'intérêt et paramètres de nuisance.
- Idée : soit $x \sim f(x|\theta)$ et $\theta = (\theta_1, \theta_2)$, où θ_1 est le paramètre d'intérêt. La loi de référence est obtenue en définissant d'abord $\pi(\theta_2|\theta_1)$ comme la loi de Jeffreys associée à $f(x|\theta)$ pour θ_1 fixé, puis en calculant la loi marginale

$$\widetilde{f}(x|\theta_1) = \int f(x|\theta_1, \theta_2)\pi(\theta_2|\theta_1)d\theta_2$$

et la loi de Jeffreys $\pi(\theta_1)$ associée à $\widetilde{f}(x|\theta_1)$.

Mesures a priori de référence. I

- Proposées par Bernardo (1979). L'analyse de référence est un mode général de spécification d'une loi a priori contenant aussi peu d'information que possible.
- Modification de l'approche de Jeffreys. Une différence majeure est que cette méthode fait la distinction entre paramètres d'intérêt et paramètres de nuisance.
- Idée : soit $x \sim f(x|\theta)$ et $\theta = (\theta_1, \theta_2)$, où θ_1 est le paramètre d'intérêt. La loi de référence est obtenue en définissant d'abord $\pi(\theta_2|\theta_1)$ comme la loi de Jeffreys associée à $f(x|\theta)$ pour θ_1 fixé, puis en calculant la loi marginale

$$\widetilde{f}(x|\theta_1) = \int f(x|\theta_1, \theta_2)\pi(\theta_2|\theta_1)d\theta_2$$

et la loi de Jeffreys $\pi(\theta_1)$ associée à $\widetilde{f}(x|\theta_1)$.

Lois de référence. II

- Soit $f(x|\theta)$ un modèle d'échantillonage et $\pi(\theta)$ une loi a priori. Soit $m(x) := \int_\Theta f(x|\theta)\pi(\theta)d\theta$.
- On a deux distributions de probabilité sur $\Theta \times \mathcal{X}$: la loi jointe $\pi(\theta)f(x|\theta)$ et le produit de deux marginales $\pi(\theta)m(x)$.
- On mesure l'information apportée par un modèle statistique en utilisant la divergence de Kullback : (on note $x^{(n)} = (x_1, \ldots, x_n)$)

$$\begin{aligned} K_n(\pi) &= \int_{\mathcal{X}^n} \int_\Theta \log\left(\frac{\pi(\theta)f(x^{(n)}|\theta)}{\pi(\theta)m(x^{(n)})}\right) \pi(\theta)f(x^{(n)}|\theta)d\theta dx^{(n)} \\ &= \int_{\mathcal{X}^n} \int_\Theta \log\left(\frac{\pi(\theta|x^{(n)})}{\pi(\theta)}\right) \pi(\theta|x^{(n)})m(x^{(n)})d\theta dx^{(n)}. \end{aligned}$$

Lois de référence. II

- Soit $f(x|\theta)$ un modèle d'échantillonage et $\pi(\theta)$ une loi a priori. Soit $m(x) := \int_\Theta f(x|\theta)\pi(\theta)d\theta$.
- On a deux distributions de probabilité sur $\Theta \times \mathcal{X}$: la loi jointe $\pi(\theta)f(x|\theta)$ et le produit de deux marginales $\pi(\theta)m(x)$.
- On mesure l'information apportée par un modèle statistique en utilisant la divergence de Kullback : (on note $x^{(n)} = (x_1, \ldots, x_n)$)

$$K_n(\pi) = \int_{\mathcal{X}^n} \int_\Theta \log\left(\frac{\pi(\theta)f(x^{(n)}|\theta)}{\pi(\theta)m(x^{(n)})}\right) \pi(\theta)f(x^{(n)}|\theta)d\theta dx^{(n)}$$

$$= \int_{\mathcal{X}^n} \int_\Theta \log\left(\frac{\pi(\theta|x^{(n)})}{\pi(\theta)}\right) \pi(\theta|x^{(n)})m(x^{(n)})d\theta dx^{(n)}.$$

Lois de référence. II

- Soit $f(x|\theta)$ un modèle d'échantillonage et $\pi(\theta)$ une loi a priori. Soit $m(x) := \int_\Theta f(x|\theta)\pi(\theta)d\theta$.
- On a deux distributions de probabilité sur $\Theta \times \mathcal{X}$: la loi jointe $\pi(\theta)f(x|\theta)$ et le produit de deux marginales $\pi(\theta)m(x)$.
- On mesure l'information apportée par un modèle statistique en utilisant la divergence de Kullback : (on note $x^{(n)} = (x_1, \ldots, x_n)$)

$$\begin{aligned} K_n(\pi) &= \int_{\mathcal{X}^n} \int_\Theta \log\left(\frac{\pi(\theta)f(x^{(n)}|\theta)}{\pi(\theta)m(x^{(n)})}\right) \pi(\theta)f(x^{(n)}|\theta)d\theta dx^{(n)} \\ &= \int_{\mathcal{X}^n} \int_\Theta \log\left(\frac{\pi(\theta|x^{(n)})}{\pi(\theta)}\right) \pi(\theta|x^{(n)})m(x^{(n)})d\theta dx^{(n)}. \end{aligned}$$

Lois de référence. III

- On appellera a priori de référence la probabilité a priori π_r qui maximise $K_n(\pi)$ (pour $n < \infty$, on a en général $0 \leq K_n(\pi) < \infty$)

- Ce problème de minimisation n'a pas de solution générale. On peut vérifier que, si π_r est l'a priori de référence et si $\pi_r(\theta|x^{(n)}) \propto \pi_r(\theta) f(x^{(n)}|\theta)$, ces deux densités doivent vérifier :

$$\pi_r(\theta) \propto \exp\left\{ \int \ln \pi_r(\theta|x^{(n)}) f(x^{(n)}|\theta) dx \right\}.$$

Lois de référence. III

- On appellera a priori de référence la probabilité a priori π_r qui maximise $K_n(\pi)$ (pour $n < \infty$, on a en général $0 \leq K_n(\pi) < \infty$)

- Ce problème de minimisation n'a pas de solution générale. On peut vérifier que, si π_r est l'a priori de référence et si $\pi_r(\theta|x^{(n)}) \propto \pi_r(\theta) f(x^{(n)}|\theta)$, ces deux densités doivent vérifier :

$$\pi_r(\theta) \propto \exp\left\{ \int \ln \pi_r(\theta|x^{(n)}) f(x^{(n)}|\theta) dx \right\}.$$

Outline

Spécification subjective de l'a priori à partire de la marginale. I

- Quand l'espace des paramètres Θ est fini, il est souvent possible d'obtenir une évaluation subjective des probabilités des différentes valeurs de θ (e.g. en utilisant des expériences précédentes du même type si possible).

- Quand l'espace des paramètres Θ n'est pas dénombrable la détermination subjective de la loi a priori π est beaucoup plus compliquée.

- Une difficulté majeure se présente lorsque Θ n'est pas borné.

- Parfois, le practicien est capable de fournir la distribution de probabilité d'une des caracteristiques d'un evenement, par exemple le coût x d'un equipement industriel. Ceci se résume en la spécification d'un modèle statistique $f(x|\theta) = \prod_{i=1}^{n} f(x_i|\theta)$ accompagné de la connaissance de la distribution marginale :

$$m(x) = \int_{\Theta} f(x|\theta)\pi(\theta)d\theta$$

fournie par l'évaluation technique du practicien.

Spécification subjective de l'a priori à partire de la marginale. I

- Quand l'espace des paramètres Θ est fini, il est souvent possible d'obtenir une évaluation subjective des probabilités des différentes valeurs de θ (e.g. en utilisant des expériences précédentes du même type si possible).

- Quand l'espace des paramètres Θ n'est pas dénombrable la détermination subjective de la loi a priori π est beaucoup plus compliquée.

- Une difficulté majeure se présente lorsque Θ n'est pas borné.

- Parfois, le practicien est capable de fournir la distribution de probabilité d'une des caracteristiques d'un evenement, par exemple le coût x d'un equipement industriel. Ceci se résume en la spécification d'un modèle statistique $f(x|\theta) = \prod_{i=1}^{n} f(x_i|\theta)$ accompagné de la connaissance de la distribution marginale :

$$m(x) = \int_{\Theta} f(x|\theta)\pi(\theta)d\theta$$

fournie par l'évaluation technique du practicien.

Spécification subjective de l'a priori à partire de la marginale. I

- Quand l'espace des paramètres Θ est fini, il est souvent possible d'obtenir une évaluation subjective des probabilités des différentes valeurs de θ (e.g. en utilisant des expériences précédentes du même type si possible).
- Quand l'espace des paramètres Θ n'est pas dénombrable la détermination subjective de la loi a priori π est beaucoup plus compliquée.
- Une difficulté majeure se présente lorsque Θ n'est pas borné.
- Parfois, le practicien est capable de fournir la distribution de probabilité d'une des caracteristiques d'un evenement, par exemple le coût x d'un equipement industriel. Ceci se résume en la spécification d'un modèle statistique $f(x|\theta) = \prod_{i=1}^{n} f(x_i|\theta)$ accompagné de la connaissance de la distribution marginale :

$$m(x) = \int_{\Theta} f(x|\theta)\pi(\theta)d\theta$$

fournie par l'évaluation technique du practicien.

Spécification subjective de l'a priori à partire de la marginale. I

- Quand l'espace des paramètres Θ est fini, il est souvent possible d'obtenir une évaluation subjective des probabilités des différentes valeurs de θ (e.g. en utilisant des expériences précédentes du même type si possible).

- Quand l'espace des paramètres Θ n'est pas dénombrable la détermination subjective de la loi a priori π est beaucoup plus compliquée.

- Une difficulté majeure se présente lorsque Θ n'est pas borné.

- Parfois, le practicien est capable de fournir la distribution de probabilité d'une des caracteristiques d'un evenement, par exemple le coût x d'un equipement industriel. Ceci se résume en la spécification d'un modèle statistique $f(x|\theta) = \prod_{i=1}^{n} f(x_i|\theta)$ accompagné de la connaissance de la distribution marginale :

$$m(x) = \int_{\Theta} f(x|\theta)\pi(\theta)d\theta$$

fournie par l'évaluation technique du practicien.

Spécification subjective de l'a priori à partire de la marginale. II

- Ce dernier point montre que il se peut que des informations sujectives sur θ ne s'expriment pas naturellement en terme d'a priori mais en terme de distribution marginale de l'échantillon.

- Alors, on determine l'a priori à partir de la loi margniale.

- Une solution consiste à chercher l'a priori au sein d'une famille paramétrique $\pi(\theta|\gamma)$, $\gamma \in \mathbb{R}^k$, où γ est tel que $\int_\Theta f(x|\theta)\pi(\theta|\gamma)d\theta$ est proche de $m(x)$ au sens, par exemple, de la divergence de Kullback. On choisira alors

$$\gamma_0 = \arg\min \int_{\mathcal{X}^n} \ln \frac{m(x)}{\int_\Theta f(x|\theta)\pi(\theta|\gamma)} m(x)dx.$$

  Ce problème n'a en général pas de solution analytique et doit etre résolu numériquement.

- La résolution de l'équation fonctionnelle en $\pi(\theta)$ est beaucoup plus complexe (équation de Fredholm de type I).

Spécification subjective de l'a priori à partire de la marginale. II

- Ce dernier point montre que il se peut que des informations sujectives sur θ ne s'expriment pas naturellement en terme d'a priori mais en terme de distribution marginale de l'échantillon.

- Alors, on determine l'a priori à partir de la loi margniale.

- Une solution consiste à chercher l'a priori au sein d'une famille paramétrique $\pi(\theta|\gamma)$, $\gamma \in \mathbb{R}^k$, où γ est tel que $\int_\Theta f(x|\theta)\pi(\theta|\gamma)d\theta$ est proche de $m(x)$ au sens, par exemple, de la divergence de Kullback. On choisira alors

$$\gamma_0 = \arg\min \int_{\mathcal{X}^n} \ln \frac{m(x)}{\int_\Theta f(x|\theta)\pi(\theta|\gamma)} m(x)dx.$$

  Ce problème n'a en général pas de solution analytique et doit etre résolu numériquement.

- La résolution de l'équation fonctionnelle en $\pi(\theta)$ est beaucoup plus complexe (équation de Fredholm de type I).

Spécification subjective de l'a priori à partire de la marginale. II

- Ce dernier point montre que il se peut que des informations sujectives sur θ ne s'expriment pas naturellement en terme d'a priori mais en terme de distribution marginale de l'échantillon.

- Alors, on determine l'a priori à partir de la loi margniale.

- Une solution consiste à chercher l'a priori au sein d'une famille paramétrique $\pi(\theta|\gamma)$, $\gamma \in \mathbb{R}^k$, où γ est tel que $\int_\Theta f(x|\theta)\pi(\theta|\gamma)d\theta$ est proche de $m(x)$ au sens, par exemple, de la divergence de Kullback. On choisira alors

$$\gamma_0 = \arg\min \int_{\mathcal{X}^n} \ln \frac{m(x)}{\int_\Theta f(x|\theta)\pi(\theta|\gamma)} m(x)dx.$$

  Ce problème n'a en général pas de solution analytique et doit etre résolu numériquement.

- La résolution de l'équation fonctionnelle en $\pi(\theta)$ est beaucoup plus complexe (équation de Fredholm de type I).

Spécification subjective de l'a priori à partire de la marginale. II

- Ce dernier point montre que il se peut que des informations sujectives sur θ ne s'expriment pas naturellement en terme d'a priori mais en terme de distribution marginale de l'échantillon.

- Alors, on determine l'a priori à partir de la loi margniale.

- Une solution consiste à chercher l'a priori au sein d'une famille paramétrique $\pi(\theta|\gamma)$, $\gamma \in \mathbb{R}^k$, où γ est tel que $\int_\Theta f(x|\theta)\pi(\theta|\gamma)d\theta$ est proche de $m(x)$ au sens, par exemple, de la divergence de Kullback. On choisira alors

$$\gamma_0 = \arg\min \int_{\mathcal{X}^n} \ln \frac{m(x)}{\int_\Theta f(x|\theta)\pi(\theta|\gamma)} m(x)dx.$$

  Ce problème n'a en général pas de solution analytique et doit etre résolu numériquement.

- La résolution de l'équation fonctionnelle en $\pi(\theta)$ est beaucoup plus complexe (équation de Fredholm de type I).

Spécification subjective de l'a priori à partire de la marginale. III

Sources possibles d'information subjective sur $m(x)$:

- Parfois, le paramètre θ n'a pas une interpretation au sens d'une quantité physique. Néanmoins, le practicien peut predire partiellement le résultat de l'experience physique.

- On peut utiliser les données pour calculer un estimateur de $m(x)$ (Bayesien empirique). Exemple : $x =$ résultat d'un test, $\theta =$ aptitude (non observée). Alors, $x|\theta \sim f(x|\theta)$. $m(x) =$ distribution observée du résultat d'un test.

L'approche ML-II

La marginale (ou vraisemblance marginale) $m(x)$ incorpore la plausibilité de $f(\cdot|\theta)$ et π en terme de données. Si on traite $f(\cdot|\theta)$ comme connu (sauf θ), alors $m(x)$ reflète la plausibilité de π.

Quand, pour les données observées, $m(x|\pi_1) > m(x|\pi_2)$, alors les données fournissent plus support en faveur de π_1 que de π_2.

Alors, $m(x|\pi)$ peut être vue comme une fonction de vraisemblance pour π.

Définition

Soit Γ une classe de distributions a priori, et soit $\hat{\pi} \in \Gamma$ telle que

$$m(x|\hat{\pi}) = \sup_{\pi \in \Gamma} m(x|\pi).$$

Alors, $\hat{\pi}$ est appellée a priori de Maximum de Vraisemblance de type II (ou a priori ML-II).

L'approche ML-II : Exemple Random Effects

Berger, Liseo nad Wolpert (1999, Statistical Science)

- $X_i \sim \mathcal{N}(\mu_i, 1)$, $\mu_i \sim \mathcal{N}(\xi, \tau^2)$. On veut faire inference sur $\theta = (\xi, \tau^2)$ en ignorant $\mu = (\mu_1, \ldots, \mu_n)$.
- Vraisemblance marginale :

$$\begin{aligned} m(x|\xi,\tau) &= \int_{\mathbb{R}^n} (2\pi)^{-n/2} \exp\left(-\sum_{i=1}^n \frac{(x_i-\mu)^2}{2}\right) (2\pi\tau^2)^{-n/2} \times \\ &\qquad \exp\left(-\sum_{i=1}^n \frac{(\mu_i-\xi)^2}{2\tau^2}\right) d\mu_1 \ldots d\mu_n \\ &= (1+\tau^2)^{-n/2} \exp\left(-n[s^2+(\bar{x}-\xi)^2]/(2(1+\tau^2))\right), \end{aligned}$$

  where $\bar{x}$ is the sample mean and $s^2 = \sum_i (x_i - \bar{x})/n$.

- Vraisemblance profilée :

$$\begin{aligned} \hat{m}(x|\xi,\tau) &= \sup_{\mu\in\mathbb{R}^p} (2\pi)^{-n/2} \exp\left(-\sum_{i=1}^n \frac{(x_i-\mu)^2}{2}\right) (2\pi\tau^2)^{-n/2} \exp\left(-\sum_{i=1}^n \frac{(\mu_i-\xi)^2}{2\tau^2}\right) \\ &= (\tau)^{-n} \exp\left(-n[s^2+(\bar{x}-\xi)^2]/(2(1+\tau^2))\right). \end{aligned}$$

On compare $\hat{L}(\tau^2) = \hat{m}(x|\xi = \bar{x}, \tau)$ et $L(\tau^2) = m(x|\xi = \bar{x}, \tau)$.

L'approche ML-II : Exemple Random Effects

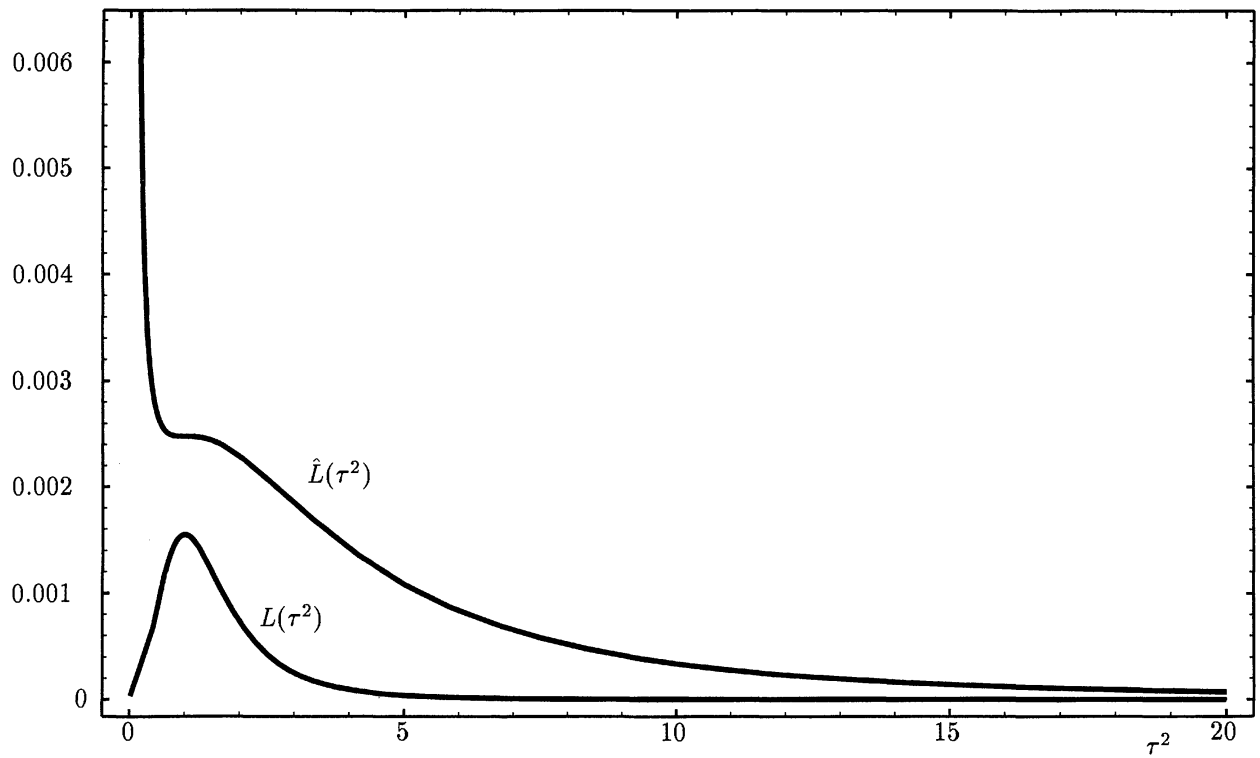

FIG. 1. *Integrated and profile likelihoods for the random effects model.*

Loi a priori d'entropie maximale. I

- Si certaines caractéristiques de la loi a priori sont connues (moments, quantiles, etc.), en supposant qu'elles peuvent s'écrire comme des espérances a priori $(k = 1, \ldots, K)$,

$$\mathbf{E}^{\pi}[g_k(\theta)] = \omega_k, \tag{1}$$

une façon de choisir un a priori qui satisfait ces contraintes est la méthode de l'entropie maximale (Jaynes 1980, 1983).

- Pour une densité π, l'entropie $H(\pi)$ est :

$$H(\pi) = -\int \pi(\theta) \ln \pi(\theta) d\theta. \tag{2}$$

- L'a priori π qui maximise l'entropie minimise l'information a priori apportée par π sur θ.
- Si Θ est discret, la distribution d'entropie maximale, sous les contraintes de moments (1), est la distribution associée à la densité

$$\pi^*(\theta_i) = \frac{\exp\left(\sum_{k=1}^{K} \lambda_k g_k(\theta_i)\right)}{\sum_j \exp\left(\sum_{k=1}^{K} \lambda_k g_k(\theta_j)\right)}$$

où les λ_k sont les multiplicateurs de Lagrange.

Loi a priori d'entropie maximale. I

- Si certaines caractéristiques de la loi a priori sont connues (moments, quantiles, etc.), en supposant qu'elles peuvent s'écrire comme des espérances a priori $(k = 1, \ldots, K)$,

$$\mathbf{E}^{\pi}[g_k(\theta)] = \omega_k, \tag{1}$$

une façon de choisir un a priori qui satisfait ces contraintes est la méthode de l'entropie maximale (Jaynes 1980, 1983).

- Pour une densité π, l'entropie $H(\pi)$ est :

$$H(\pi) = -\int \pi(\theta) \ln \pi(\theta) d\theta. \tag{2}$$

- L'a priori π qui maximise l'entropie minimise l'information a priori apportée par π sur θ.
- Si Θ est discret, la distribution d'entropie maximale, sous les contraintes de moments (1), est la distribution associée à la densité

$$\pi^*(\theta_i) = \frac{\exp\left(\sum_{k=1}^{K} \lambda_k g_k(\theta_i)\right)}{\sum_j \exp\left(\sum_{k=1}^{K} \lambda_k g_k(\theta_j)\right)}$$

où les λ_k sont les multiplicateurs de Lagrange.

Loi a priori d'entropie maximale. I

- Si certaines caractéristiques de la loi a priori sont connues (moments, quantiles, etc.), en supposant qu'elles peuvent s'écrire comme des espérances a priori $(k = 1, \ldots, K)$,

$$\mathbf{E}^{\pi}[g_k(\theta)] = \omega_k, \tag{1}$$

une façon de choisir un a priori qui satisfait ces contraintes est la méthode de l'entropie maximale (Jaynes 1980, 1983).

- Pour une densité π, l'entropie $H(\pi)$ est :

$$H(\pi) = -\int \pi(\theta) \ln \pi(\theta) d\theta. \tag{2}$$

- L'a priori π qui maximise l'entropie minimise l'information a priori apportée par π sur θ.
- Si Θ est discret, la distribution d'entropie maximale, sous les contraintes de moments (1), est la distribution associée à la densité

$$\pi^*(\theta_i) = \frac{\exp\left(\sum_{k=1}^{K} \lambda_k g_k(\theta_i)\right)}{\sum_j \exp\left(\sum_{k=1}^{K} \lambda_k g_k(\theta_j)\right)}$$

où les λ_k sont les multiplicateurs de Lagrange.

Loi a priori d'entropie maximale. I

- Si certaines caractéristiques de la loi a priori sont connues (moments, quantiles, etc.), en supposant qu'elles peuvent s'écrire comme des espérances a priori $(k = 1, \ldots, K)$,

$$\mathbf{E}^{\pi}[g_k(\theta)] = \omega_k, \tag{1}$$

une façon de choisir un a priori qui satisfait ces contraintes est la méthode de l'entropie maximale (Jaynes 1980, 1983).

- Pour une densité π, l'entropie $H(\pi)$ est :

$$H(\pi) = -\int \pi(\theta) \ln \pi(\theta) d\theta. \tag{2}$$

- L'a priori π qui maximise l'entropie minimise l'information a priori apportée par π sur θ.
- Si Θ est discret, la distribution d'entropie maximale, sous les contraintes de moments (1), est la distribution associée à la densité

$$\pi^*(\theta_i) = \frac{\exp\left(\sum_{k=1}^{K} \lambda_k g_k(\theta_i)\right)}{\sum_j \exp\left(\sum_{k=1}^{K} \lambda_k g_k(\theta_j)\right)}$$

où les λ_k sont les multiplicateurs de Lagrange.

Loi a priori d'entropie maximale. II

- Sans contrainte sur π, la distribution d'entropie maximale est la distribution uniforme sur Θ.
- L'extension au cas continu est plus délicate, car elle implique le choix d'une mesure de référence π_0 (π_0 peut être vue comme la distribution complétement non informative)
- Une fois la mesure de référence π_0 choisie, l'entropie de π est définie par

$$H(\pi) = -\int \pi(\theta) \ln\left(\frac{\pi(\theta)}{\pi_0(\theta)}\right) d\theta \tag{3}$$

qui est aussi la distance de Kullback-Leibler entre π et π_0. Dans ce cas, la distribution d'entropie maximale sous (1) est donnée par la densité

$$\pi^*(\theta) = \frac{\pi_0(\theta) \exp\left(\sum_{k=1}^{K} \lambda_k g_k(\theta_i)\right)}{\int_\Theta \exp\left(\sum_{k=1}^{K} \lambda_k g_k(\eta)\right) \pi_0(d\eta)}.$$

Loi a priori d'entropie maximale. II

- Sans contrainte sur π, la distribution d'entropie maximale est la distribution uniforme sur Θ.

- L'extension au cas continu est plus délicate, car elle implique le choix d'une mesure de référence π_0 (π_0 peut être vue comme la distribution complétement non informative)

- Une fois la mesure de référence π_0 choisie, l'entropie de π est définie par

$$H(\pi) = -\int \pi(\theta) \ln\left(\frac{\pi(\theta)}{\pi_0(\theta)}\right) d\theta \tag{3}$$

qui est aussi la distance de Kullback-Leibler entre π et π_0. Dans ce cas, la distribution d'entropie maximale sous (1) est donnée par la densité

$$\pi^*(\theta) = \frac{\pi_0(\theta) \exp\left(\sum_{k=1}^K \lambda_k g_k(\theta_i)\right)}{\int_\Theta \exp\left(\sum_{k=1}^K \lambda_k g_k(\eta)\right) \pi_0(d\eta)}.$$

Loi a priori d'entropie maximale. II

- Sans contrainte sur π, la distribution d'entropie maximale est la distribution uniforme sur Θ.

- L'extension au cas continu est plus délicate, car elle implique le choix d'une mesure de référence π_0 (π_0 peut être vue comme la distribution complétement non informative)

- Une fois la mesure de référence π_0 choisie, l'entropie de π est définie par

$$H(\pi) = -\int \pi(\theta) \ln \left(\frac{\pi(\theta)}{\pi_0(\theta)} \right) d\theta \tag{3}$$

qui est aussi la distance de Kullback-Leibler entre π et π_0. Dans ce cas, la distribution d'entropie maximale sous (1) est donnée par la densité

$$\pi^*(\theta) = \frac{\pi_0(\theta) \exp\left(\sum_{k=1}^K \lambda_k g_k(\theta_i)\right)}{\int_\Theta \exp\left(\sum_{k=1}^K \lambda_k g_k(\eta)\right) \pi_0(d\eta)}.$$

Loi a priori conjuguées. I

Soit P_θ, $\theta \in \Theta$ une famille de probabilités d'échantillonage sur $\mathcal{X}$. (θ non nécessairement de dimension finie). Soit $\mathscr{M}$ une famille de lois de probabilité sur Θ.
On s'intéresse aux deux propriétés suivantes de $\mathscr{M}$:

- On dira que $\mathscr{M}$ est fermée si, pour toute probabilité de $\mathscr{M}$ choisi comme loi a priori et pour tout échantillon observé, la loi a posteriori déduite est encore un élément de $\mathscr{M}$.

- On va supposer que les éléments de $\mathscr{M}$ sont paramétrés par un hyperparamètre γ. Le passage de distribution a priori à distribution a posteriori se réduit dans ce cas à une mise à jour des hyperparamètres γ correspondants.

Définition

Une famille $\mathscr{M}$ de distributions de probabilité sur Θ est dite conjuguée (ou fermée par échantillonnage) à une famille de probabilités d'échantillonage P_θ si, $\forall \pi \in \mathscr{M}$, la distribution a posteriori $\pi(\theta|x)$ appartient également à $\mathscr{M}$.

Loi a priori conjuguées. I

Soit P_θ, $\theta \in \Theta$ une famille de probabilités d'échantillonage sur $\mathcal{X}$. (θ non nécessairement de dimension finie). Soit $\mathscr{M}$ une famille de lois de probabilité sur Θ. On s'intéresse aux deux propriétés suivantes de $\mathscr{M}$:

- On dira que $\mathscr{M}$ est fermée si, pour toute probabilité de $\mathscr{M}$ choisi comme loi a priori et pour tout échantillon observé, la loi a posteriori déduite est encore un élément de $\mathscr{M}$.

- On va supposer que les éléments de $\mathscr{M}$ sont paramétrés par un hyperparamètre γ. Le passage de distribution a priori à distribution a posteriori se réduit dans ce cas à une mise à jour des hyperparamètres γ correspondants.

Définition

Une famille $\mathscr{M}$ de distributions de probabilité sur Θ est dite conjuguée (ou fermée par échantillonnage) à une famille de probabilités d'échantillonage P_θ si, $\forall \pi \in \mathscr{M}$, la distribution a posteriori $\pi(\theta|x)$ appartient également à $\mathscr{M}$.

Loi a priori conjuguées. I

Soit P_θ, $\theta \in \Theta$ une famille de probabilités d'échantillonage sur $\mathcal{X}$. (θ non nécessairement de dimension finie). Soit $\mathscr{M}$ une famille de lois de probabilité sur Θ. On s'intéresse aux deux propriétés suivantes de $\mathscr{M}$:

- On dira que $\mathscr{M}$ est fermée si, pour toute probabilité de $\mathscr{M}$ choisi comme loi a priori et pour tout échantillon observé, la loi a posteriori déduite est encore un élément de $\mathscr{M}$.

- On va supposer que les éléments de $\mathscr{M}$ sont paramétrés par un hyperparamètre γ. Le passage de distribution a priori à distribution a posteriori se réduit dans ce cas à une mise à jour des hyperparamètres γ correspondants.

Définition

Une famille $\mathscr{M}$ de distributions de probabilité sur Θ est dite conjuguée (ou fermée par échantillonnage) à une famille de probabilités d'échantillonage P_θ si, $\forall \pi \in \mathscr{M}$, la distribution a posteriori $\pi(\theta|x)$ appartient également à $\mathscr{M}$.

Loi a priori conjuguées. II

- L'approche a priori conjuguée peut être justifiée partiellement par un raisonnement d'invariance : quand l'observation de $x \sim f(x|\theta)$ modifie $\pi(\theta)$ en $\pi(\theta|x)$, l'information transmise par x sur θ est limitée ; par conséquent, elle ne devrait pas entraîner une modification de toute la structure de $\pi(\theta)$, mais simplement de ses paramètres.

- Les lois a priori conjuguées sont surtout utilisées dans des environnements où l'information est limitée, car elles ne nécessitent la détermination que de quelques paramètres.

- Mais, la principale motivation pour utiliser les lois a priori conjuguées reste la commodité de traitement.

- Alors, on peut aussi voir le rôle des lois a priori conjuguées comme de fournir une première approximation de la distribution a priori adéquate, qui devrait être suivie d'une analyse de robustesse.

Loi a priori conjuguées. II

- L'approche a priori conjuguée peut être justifiée partiellement par un raisonnement d'invariance : quand l'observation de $x \sim f(x|\theta)$ modifie $\pi(\theta)$ en $\pi(\theta|x)$, l'information transmise par x sur θ est limitée ; par conséquent, elle ne devrait pas entraîner une modification de toute la structure de $\pi(\theta)$, mais simplement de ses paramètres.

- Les lois a priori conjuguées sont surtout utilisées dans des environnements où l'information est limitée, car elles ne nécessitent la détermination que de quelques paramètres.

- Mais, la principale motivation pour utiliser les lois a priori conjuguées reste la commodité de traitement.

- Alors, on peut aussi voir le rôle des lois a priori conjuguées comme de fournir une première approximation de la distribution a priori adéquate, qui devrait être suivie d'une analyse de robustesse.

Loi a priori conjuguées. II

- L'approche a priori conjuguée peut être justifiée partiellement par un raisonnement d'invariance : quand l'observation de $x \sim f(x|\theta)$ modifie $\pi(\theta)$ en $\pi(\theta|x)$, l'information transmise par x sur θ est limitée ; par conséquent, elle ne devrait pas entraîner une modification de toute la structure de $\pi(\theta)$, mais simplement de ses paramètres.

- Les lois a priori conjuguées sont surtout utilisées dans des environnements où l'information est limitée, car elles ne nécessitent la détermination que de quelques paramètres.

- Mais, la principale motivation pour utiliser les lois a priori conjuguées reste la commodité de traitement.

- Alors, on peut aussi voir le rôle des lois a priori conjuguées comme de fournir une première approximation de la distribution a priori adéquate, qui devrait être suivie d'une analyse de robustesse.

Loi a priori conjuguées. II

- L'approche a priori conjuguée peut être justifiée partiellement par un raisonnement d'invariance : quand l'observation de $x \sim f(x|\theta)$ modifie $\pi(\theta)$ en $\pi(\theta|x)$, l'information transmise par x sur θ est limitée ; par conséquent, elle ne devrait pas entraîner une modification de toute la structure de $\pi(\theta)$, mais simplement de ses paramètres.

- Les lois a priori conjuguées sont surtout utilisées dans des environnements où l'information est limitée, car elles ne nécessitent la détermination que de quelques paramètres.

- Mais, la principale motivation pour utiliser les lois a priori conjuguées reste la commodité de traitement.

- Alors, on peut aussi voir le rôle des lois a priori conjuguées comme de fournir une première approximation de la distribution a priori adéquate, qui devrait être suivie d'une analyse de robustesse.

Familles naturelles conjuguées. I

Les lois a priori conjuguées sont généralement associées à un *type particulier de lois d'échantillonnage* qui permet toujours leur obtention. Ces lois constituent des *familles exponentielles* (Brown 1986).

Définition

Soient λ une mesure σ-finie sur $\mathcal{X}$, Θ l'espace des paramètres, $C(\cdot)$ et $h(\cdot)$ des fonctions respectivement de $\mathcal{X}$ et Θ dans $\mathbb{R}_+$, et $R(\cdot)$ et $T(\cdot)$ des fonctions de Θ et $\mathcal{X}$, respectivement, dans $\mathbb{R}^k$. La famille des distributions de densité (par rapport à λ)

$$f(x|\theta) = C(\theta)h(x)\exp\{R(\theta)T(x)\} \tag{4}$$

est dite famille exponentielle de dimension k. Dans le cas particulier où $\Theta \subset \mathbb{R}^k$, $\mathcal{X} \subset \mathbb{R}^k$ et

$$f(x|\theta) = C(\theta)h(x)\exp\{\theta' x\}$$

la famille est dite naturelle.

Familles naturelles conjuguées. II

Les familles exponentielles ont certaines caractéristiques intéressantes. En particulier, elles sont telles que, pour tout échantillon de (4), il existe une statistique exhaustive de dimension constante.

Théorème (Koopman (1936) et Pitman (1936))

Si une famille de lois $f(x|\theta)$ à support constant est telle que, à partir d'une taille d'échantillon suffisamment grande, il existe une statistique exhaustive de taille fixe, la famille est exponentielle.

Les familles exponentielles naturelles peuvent aussi être réécrites sous la forme

$$f(x|\theta) = h(x)e^{\theta' x - \psi(\theta)}$$

où $\psi(\theta)$ est dite fonction cumulante des moments.

Familles naturelles conjuguées. III

Tab. 3.4. Lois a priori conjuguées naturelles pour quelques familles exponentielles usuelles.

$f(x\|\theta)$	$\pi(\theta)$	$\pi(\theta\|x)$
Normale $\mathscr{N}(\theta, \sigma^2)$	Normale $\mathscr{N}(\mu, \tau^2)$	$\mathscr{N}(\varrho(\sigma^2\mu + \tau^2 x), \varrho\sigma^2\tau^2)$ $\varrho^{-1} = \sigma^2 + \tau^2$
Poisson $\mathscr{P}(\theta)$	Gamma $\mathscr{G}(\alpha, \beta)$	$\mathscr{G}(\alpha + x, \beta + 1)$
Gamma $\mathscr{G}(\nu, \theta)$	Gamma $\mathscr{G}(\alpha, \beta)$	$\mathscr{G}(\alpha + \nu, \beta + x)$
Binomiale $\mathscr{B}(n, \theta)$	Bêta $\mathscr{B}e(\alpha, \beta)$	$\mathscr{B}e(\alpha + x, \beta + n - x)$
Binomiale Négative $\mathscr{N}eg(m, \theta)$	Bêta $\mathscr{B}e(\alpha, \beta)$	$\mathscr{B}e(\alpha + m, \beta + x)$
Multinomiale $\mathscr{M}_k(\theta_1, \ldots, \theta_k)$	Dirichlet $\mathscr{D}(\alpha_1, \ldots, \alpha_k)$	$\mathscr{D}(\alpha_1 + x_1, \ldots, \alpha_k + x_k)$
Normale $\mathscr{N}(\mu, 1/\theta)$	Gamma $\mathscr{G}a(\alpha, \beta)$	$\mathscr{G}(\alpha + 0.5, \beta + (\mu - x)^2/2)$

Familles naturelles conjuguées. IV

Le lois conjuguées des familles exponentielles sont données par la proposition suivante :

Proposition

Soit $f(x|\theta) = h(x)e^{\theta' x - \psi(\theta)}$. *Une famille conjuguée pour* $f(x|\theta)$ *est donnée par*

$$\pi(\theta|\mu,\gamma) = K(\mu,\gamma)e^{\theta'\mu - \gamma\psi(\theta)}, \tag{5}$$

où $K(\mu,\gamma)$ *est la constante de normalisation de la densité. La loi a posteriori correspondante est* $\pi(\theta|\mu + x, \gamma + 1)$.

Un apport subjectif via la détermination de valeurs des hyperparamètres (μ,γ) est nécessaire pour l'inférence bayésienne.
Les lois a priori conjuguées ont un attrait supplémentaire : si $\xi(\theta)$ est l'espérance de $x \sim f(x|\theta)$, l'espérance a posteriori de $\xi(\theta)$ est linéaire en x pour une loi a priori conjuguée.

Proposition

Si Θ *est un ensemble ouvert dans* $\mathbb{R}^k$ *et* θ *a pour loi a priori*

$$\pi(\theta|\mu_0,\gamma) \propto e^{\theta'\mu_0 - \gamma\psi(\theta)},$$

avec $\mu_0 \in \mathcal{X}$, *alors* $\mathbb{E}^\pi[\xi(\theta)] = \mathbb{E}^\pi[\nabla\psi(\theta)] = \frac{\mu_0}{\gamma}$.

A priori hiérarchique et analyse bayésienne empirique. I

- A priori hiérarchique : quand l'a priori sur θ est fonction d'un hyperparamètre γ sur lequel une distribution a priori sera specifiée.
- On construit donc une probabilité jointe : $\pi(\gamma)\pi(\theta|\gamma)f(x|\theta)$.

Cette construction présente plusieurs intérêts :

1) Elle fournit un cadre théorique au traitement de familles d'a priori en interprétant γ comme l'index de cette famille $\{\pi(\theta|\gamma)\}_\gamma$.
2) L'introduction de paramètres incidents (i.e. de dimension liée à la taille de l'échantillon) $\theta = (\theta_1, \ldots, \theta_n)$ a des intérêts en statistique car les variables non observables (variables d'hétérogénéité, effets aléatoires dans les panels, ...) sont traitées comme des paramètres incidents.

A priori hiérarchique et analyse bayésienne empirique. I

- A priori hiérarchique : quand l'a priori sur θ est fonction d'un hyperparamètre γ sur lequel une distribution a priori sera specifiée.
- On construit donc une probabilité jointe : $\pi(\gamma)\pi(\theta|\gamma)f(x|\theta)$.

Cette construction présente plusieurs intérêts :

1) Elle fournit un cadre théorique au traitement de familles d'a priori en interprétant γ comme l'index de cette famille $\{\pi(\theta|\gamma)\}_\gamma$.
2) L'introduction de paramètres incidents (i.e. de dimension liée à la taille de l'échantillon) $\theta = (\theta_1, \ldots, \theta_n)$ a des intérêts en statistique car les variables non observables (variables d'hétérogénéité, effets aléatoires dans les panels, ...) sont traitées comme des paramètres incidents.

A priori hiérarchique et analyse bayésienne empirique. I

- A priori hiérarchique : quand l'a priori sur θ est fonction d'un hyperparamètre γ sur lequel une distribution a priori sera specifiée.
- On construit donc une probabilité jointe : $\pi(\gamma)\pi(\theta|\gamma)f(x|\theta)$.

Cette construction présente plusieurs intérêts :

1) Elle fournit un cadre théorique au traitement de familles d'a priori en interprétant γ comme l'index de cette famille $\{\pi(\theta|\gamma)\}_\gamma$.
2) L'introduction de paramètres incidents (i.e. de dimension liée à la taille de l'échantillon) $\theta = (\theta_1, \ldots, \theta_n)$ a des intérêts en statistique car les variables non observables (variables d'hétérogénéité, effets aléatoires dans les panels, ...) sont traitées comme des paramètres incidents.

A priori hiérarchique et analyse bayésienne empirique. II

- L'introduction d'a priori hiérarchique peut aider la specification a priori en permettant, par exemple, de melanger a priori informatives et non informatives :

$$\theta = (\theta_1, \ldots, \theta_n), \qquad \pi(\theta|\gamma) = \prod_{i=1}^{n} h(\theta_i|\gamma)$$

a priori non informative sur γ.

- La structure hiérarchique permet aussi une justification du point de vue Bayésien empirique.
- Le Bayésien empirique consiste à utiliser les données pour specifier l'a priori. Par exemple, on remplace γ par son estimateur $\hat{\gamma}$.