Projet de recherche : calibration bayésienne par divergence KL discrète et distance de Wasserstein

M2 2023

1 Introduction

Dans un cadre bayésien paramétrique, la calibration de mesures *a priori* suit le choix de la forme de telles mesures (par exemple le fait que le prior soit gaussien ; la calibration étant alors l'estimation de la moyenne et la variance de cette loi). Il s'agit de l'un des aspects les plus critiqués de l'inférence bayésienne, car il laisse une large place à l'arbitraire. Il faut donc essayer de se doter de règles formelles en partant d'hypothèses sur l'interprétation du sens de grandeurs "expertes", connues *a priori*.

Il est classique de considérer que ces grandeurs ont le sens de quantiles d'une loi prédictive *a priori* sur une grandeur X dont on cherche à inférer, en présence de données $x_1, \ldots, x_n$, la loi *a posteriori*. En introduisant un vecteur de paramètre θ, de loi *a priori* $\pi(\theta)$, la loi *a priori* prédictive est

$$f(x) = \int f(x|\theta)\pi(\theta)d\theta$$

et la loi *a posteriori* prédictive est

$$f(x|x_1, \ldots, x_n) = \int f(x|\theta)\pi(\theta|x_1, \ldots, x_n)d\theta.$$

2 Première formalisation

On considère premièrement la situation où l'on cherche à représenter le comportement d'une variable aléatoire X définie sur $\{\mathbb{R}, \mathcal{B}(\mathbb{R}), \mathcal{P}\}$, de densité supposée

$$f(x) \quad = \quad \int_\Theta f(x|\theta)\pi(\theta)\, d\theta \tag{1}$$

où $x \to f(x|\theta)$ est une mesure de probabilité paramétrique connue, Θ est un espace probabilisé en dimension d, de mesure dominante $\mu(\theta)$, et $\pi(\theta)$ est une mesure intégrable sur Θ, dite *a priori*. On dispose de contraintes quantiles sur X :

$$\mathcal{P}\left(X \leq x_i\right) \quad = \quad \int_\Theta \left\{ \int_{-\infty}^{x_i} f(u|\theta)\, du \right\} \pi(\theta)\, d\theta \;=\; \alpha_i, \quad i = 1, \ldots, p \tag{2}$$

On suppose que $\pi(\theta) = \pi(\theta|\xi)$ où ξ est un ensemble d'hyperparamètres inconnus, défini dans un espace Ξ de dimension finie. On note alors $f(x) = f(x|\xi)$ la loi *a priori* prédictive (ou *marginale*) et $\mathcal{P}(X \leq x_i) = \mathcal{P}(X \leq x_i|\xi)$ sa fonction de répartition.

Connaissant la forme de $\pi(\theta|\xi)$ (ex : gaussienne, etc.), on souhaite définir une règle de calibration de ξ en fonction des contraintes (2). Une première idée, naturelle, est de proposer l'usage d'une perte quadratique éventuellement pondérée, du type

$$\xi^* \quad = \quad \arg\min_{\xi\in\Xi} \sum_{i=1}^{p} \omega_i \left(\alpha_i(\xi) - \alpha_i\right)^2 \tag{3}$$

où $\alpha_i(\xi) = \mathcal{P}(X \leq x_i|\xi)$ et ω_i est un poids indiquant la préférence "souhaitée" donnée au respect de la contrainte i. D'autres critères ont été suggérés dans la littérature, comme celui de Cooke, issu d'une discrétisation de la divergence de Kullback-Leibler entre la loi inconnue $\mathcal{F}$ de X, respectant toutes les contraintes (2), et la loi *atteignable* (ou *calibrable*) $\mathcal{P}(.|\xi)$:

$$\xi^* \quad = \quad \arg\min_{\xi\in\Xi} \sum_{i=1}^{p} (\alpha_{i+1} - \alpha_i) \log \frac{\alpha_{i+1} - \alpha_i}{\alpha_{i=1}(\xi) - \alpha_i(\xi)}$$

où $\alpha_0(\xi) = \alpha_0 = 0$ et $\alpha_{p+1}(\xi) = \alpha_{p+1} = 1$. Cependant, la littérature consacrée aux choix / métriques de calibration *a priori* en statistique bayésienne paramétrique est relativement éparse, et à ce jour aucune règle formelle claire ne se dégage.

Le problème à résoudre reste donc le suivant : **comment bien poser le problème de la calibration (vue comme une inversion stochastique)** ? Quels choix / quelles conditions peut-on mettre en lumière sur $f(x|\theta)$, $\pi(\theta|\xi)$ et la fonction de coût de façon à obtenir la convexité ou la quasi-convexité de la fonction à optimiser (minimiser) en ξ ? Il semble aisé de montrer cette convexité dans des cas où $p = 2$, pour un choix de fonction de coût L2 ou KL, mais peut-on faire mieux ? En particulier, peut-on utiliser la distance de Wasserstein $W2$?

3 Programme de travail

Dans ce travail, on s'intéressera donc à (dans l'ordre) :

- Formaliser le problème de la calibration d'un point de vue général, en considérant que X est unidimensionel mais pas forcément θ ;
- Proposer un ou des algorithmes de calibration en fonction de la discrépance utilisée (KL ou W2). La mise en oeuvre pratique du calcul nécessite certainement des calculs d'optimisation stochastique.
- Tester l'application sur des familles de prior :
  - Les modèles conjuguées (famille exponentielle naturelle) dont le JCP décrit au § 4 ;
  - Le prior non conjugué proposé pour la loi de Weibull dans [2].
  - Si le temps le permet, le prior pour les modèles de Fréchet et de Gumbel proposé au § 5.

Pour tous ces priors, la calibration doit être effectuée conditionnellement à la valeur d'une taille virtuelle m correspondant à la "force" de l'information *a priori* (voir le cours sur les priors conjugués).

On aura à coeur de rédiger un document sur LateX retraçant cette recherche, accompagné de codes R ou Python documenté. Il est envisageable de placer du code LateX sur un notebook R ou Python dont le code illustre les aspects théoriques.

4 Calibration conditionnelle

Une situation importante est celle où la forme de la mesure *a priori* $\pi(\theta)$ est définie par

$$\pi(\theta) \quad \simeq \quad \pi^J(\theta|\tilde{x}_1, \ldots, \tilde{x}_m)$$

où $\tilde{x}_1, \ldots, \tilde{x}_m$ est un échantillon *virtuel* censé suivre $f(x|\theta)$ de façon iid, et $\pi^J(\theta)$ est une mesure *a priori* de référence. Dans ce cas, il est parfois possible de séparer les hyperparamètres ξ en deux groupes :

— des hyperparamètres ξ_s reliés à des statistiques inconnues de $\tilde{x}_1, \ldots, \tilde{x}_m$;
— la taille m de cet échantillon virtuel.

Un exemple est issu de la théorie des *Jeffreys Conjugate Priors* (JCP) (voir § 3.1.4.2 dans [1] : $X|\theta$ appartient à la famille exponentielle

$$f(x|\theta) = \exp(\theta.t(x) - \phi(\theta))h(x)$$

où Θ est un domaine ouvert de $\mathbb{R}^d$. On suppose que $\phi(\theta)$ et que l'information de Fisher $I(\theta)$ sont continues. Soit $\pi^J(\theta)$ le prior de Jeffreys par rapport à la mesure de Lebesgue

$$\pi^J(\theta) \quad \propto \quad |I(\theta)|^{1/2} \tag{4}$$

et soit

$$\pi(\theta|\alpha, \beta) \quad \propto \quad \exp(\alpha.\theta - \beta\phi(\theta))\, |I(\theta)|^{1/2} \tag{5}$$

un représentant de la classe des JCP. Alors celui-ci, outre converger vers $\pi^J(\theta)$ au sens de la convergence $q-$vague (voir Prop. 3.1.28 dans [1]), peut également être interprété comme la loi a posteriori de Jeffreys pour un échantillon virtuel $\tilde{x}_1, \ldots, \tilde{x}_m$ tel que

$$\alpha \quad = \quad \sum_{i=1}^{m} t(\tilde{x}_i) \tag{6}$$

et de taille $m = \beta$.

L'hyperparamètre m reflète donc la taille de l'information *a priori*, qui peut être comparée à celle de données réelles. La calibration de cet hyperparamètre répond à une logique différente ; on peut ainsi envisager de minimiser ξ_s, conditionnellement à m, puis de produire des choix de m par d'autres types de règle.

5 Priors de Gumbel et Fréchet

On considère le cas des lois extrêmes de Fréchet et Gumbel, définies par leurs distributions respectives de la façon suivante : soit $\theta = (\mu, \sigma, \xi)$; pour $\xi > 0$, la fonction de répartition de Fréchet est

$$P(X < x|\theta) \quad = \quad \exp\left\{-\left(\frac{x-\mu}{\sigma}\right)^{-1/\xi}\right\}, \tag{7}$$

avec $\sigma > 0$, $\mu \in I\!R$ et $x \geq \mu$. Pour $\xi = 0$, on obtient la loi de Gumbel, telle que

$$P(X < x|\theta) \quad = \quad \exp\left\{-\exp\left(-\frac{x-\mu}{\sigma}\right)\right\} \tag{8}$$

avec $\sigma > 0$, $\mu \in I\!R$ et $x \in I\!R$. On considère alors les lois *a priori* suivantes, où m est une taille d'échantillon virtuelle.

5.1 Loi de Fréchet

On reparamétrise

$$P(X < x|\theta) \quad = \quad \exp\left\{-\nu\,(x-\mu)^{-1/\xi}\right\}$$

et on appelle à présent $\theta = (\mu, \nu, \xi)$ avec $\nu = \sigma^{1/\xi} > 0$. Le prior choisi est défini par

$$\begin{aligned} \nu|\mu,\xi &\sim \mathcal{G}\left(m, s_1(\mu,\xi)\right), \\ \xi|\mu &\sim \mathcal{IG}\left(m, s_2(\mu)\right), \\ \pi(\mu) &\propto \frac{\mathbb{1}_{\{\mu \leq x_{e_1}\}}}{(x_{e_2}-\mu)^m s_2^m(\mu)} \end{aligned} \tag{9}$$

où $\mu < x_{e_1} < x_{e_2}$ et

$$\begin{aligned} s_1(\mu,\xi) &= m(x_{e_1}-\mu)^{-1/\xi}, \\ s_2(\mu) &= m\log\left(\frac{x_{e_2}-\mu}{x_{e_1}-\mu}\right). \end{aligned}$$

5.2 Prior de Gumbel (issu de [3])

On considère le prior

$$\pi(\mu,\sigma) \quad \propto \quad \sigma^{-m}\exp\left(m\frac{(\mu-\bar{\tilde{x}}_m)}{\sigma} - \sum_{i=1}^{m}\exp\left\{-\frac{\tilde{x}_i-\mu}{\sigma}\right\}\right) \tag{10}$$

qui est propre (intégrable) pour $m \geq 3$

Les hyperparamètres $(m, \bar{\tilde{x}}_m, \tilde{x}_1, \ldots, \tilde{x}_m)$ correspondent respectivement à la taille d'un échantillon virtuel, sa moyenne et les données virtuelles elles-mêmes.

Références

[1] C. Bioche. *Approximation de lois impropres et applications*. PhD thesis, 2015.

[2] N. Bousquet. Elicitation of Weibull priors. *https://arxiv.org/abs/1007.4740*, 2010.

[3] R.A. Chechile. Bayesian analysis of Gumbel distributed data. *Communications in Statistics - Theory and Methods*, 30(3) :485–496, 2001.