Prior de référence sous contraintes selon différentes mesures de dissimilarité

Antoine Van Biesbroeck [1,2], Clément Gauchy [3], Cyril Feau [2], Josselin Garnier [1]

[1] *CMAP, CNRS, École polytechnique, Institut polytechnique de Paris, 91120 Palaiseau, France ; email :* antoine.van-biesbroeck@polytechnique.edu.

[2] *Université Paris-Saclay, CEA, Service d'Études Mécaniques et Thermiques, 91191 Gif-sur-Yvette, France.*

[3] *Université Paris-Saclay, CEA, Service de Génie Logiciel pour la Simulation, 91191 Gif-sur-Yvette, France.*

Résumé. La théorie des priors de référence propose une solution à la question du choix du prior en analyse bayésienne, en construisant ce dernier comme celui qui minimise son influence a posteriori. Bien que permettant la construction d'un cadre alors qualifié d'objectif, l'expression du prior de référence prend souvent une forme proche d'un prior non informatif de Jeffreys. Cette dernière est parfois perçue comme encombrante à l'implémentation et rigide quant à l'introduction d'un jugement a priori. Dans cette communication, nous nous appuyons sur des travaux récents de la littérature, qui étendent la théorie des priors de référence. Leur formalisme et leurs résultats nous permettent en effet de définir un prior à la fois dit de référence et qui satisfait un certains nombres de contraintes pour lesquels nous définissont un cadre adéquat. Le résultat théorique que nous démontrons propose l'expression de ce prior de référence sous des contraintes linéaires, dont le choix pratique reste large et inclusif.

Mots-clés. Prior de référence, information mutuelle, f-divergence, contraintes.

Abstract. The reference prior theory porposes a solution to the prior choice issue in Bayesian analysis. It suggests the construction of that latter as the one that minimizes its *a posteriori* influence. While it allows the construction of a framework that one could call objective, the reference prior expression often takes a form close to a non-informative Jeffreys prior. That latter is sometimes perceived as a cumberstone to implement and to be rigid with any possible introduction of prior judgements. In this paper, we take as support an extention of the reference prior theory unveiled in recent works from the litterature. Their formalism and their results permit us indeed to define a prior which both (i) is called a reference prior and (ii) satisfies certain constraints for which we difine an appropriate framework. The theoretical result that we prove expresses that reference prior under linear constraints, whose choice stays wide and inclusive.

Keywords. Reference prior, mutual information, f-divergence, constraints.

1 Introduction

La théorie des priors de référence s'inscrit dans le domaine de l'analyse bayésienne par la définition et l'usage d'outils de la théorie de l'information pour répondre à la question du choix de la distribution a priori.

La définition originelle des priors de référence est due à Bernardo (1979), qui, en proposant la maximisation de l'information mutuelle comme critère de choix, a orienté la question vers la recherche du prior qui minimise son influence sur la distribution a posteriori qu'il induit. Le travail de Clarke & Barron (1994) démontre rigoureusement que sous cette définition classique, le prior de référence est celui de Jeffreys, déjà plébiscité pour sa propriété d'invariance par re-paramétrisation du modèle étudié. L'emploi et la considération de cette théorie reste au cœur de la littérature. Son étude constitue une problématique récente, que ce soit son expression dans différents contextes (Muré, 2021; Keefe et al., 2019), son approximation (Gu et al., 2018) ou son implémentation (Gauchy et al., 2023; Nalisnick & Smyth, 2017).

De récents travaux ont exploré de possibles extensions de la définition des priors de référence (Van Biesbroeck, 2023). Ceux-ci, qui s'appuient sur le point de vue de l'analyse de sensibilité globale, proposent une définition plus générale du prior de référence, qui est celui qui, en un sens, maximise une mesure de l'impact qu'induit la connaissance du paramètre d'intérêt sur la distribution des observations du système étudié. Leurs conclusions appuient la robustesse de la prise en compte du prior de Jeffreys en inférence bayésienne.

Néanmoins, dans des usages pratiques, le canevas bayésien et la question ouverte du choix de la distribution a priori est parfois plébiscitée pour sa permission d'introduction d'information, que ce soit à des fins d'optimisation de ses capacités d'estimation, ou de prise en compte de jugements d'experts.

Cette communication répond à cette problématique en proposant un résultat permissif à l'introduction de contraintes variées sur une loi a priori, tout en conservant son caractère de prior de référence. En prenant pour appui le cadre étendu des priors de référence de (Van Biesbroeck, 2023) et ses résultats, nous démontrons un théorème qui donne l'expression que doit avoir, au sens de la théorie des priors de référence, un prior qui satisfait les contraintes désirées. Notre résultat, qui se limite à des contraintes prenant la forme de projections linéaires saurait s'appliquer sur l'implémentation de contraintes sur les moments a priori ou bien sur la distribution marginale, suivant des idées proposées dans (Bousquet, 2023) relatives à l'ellicitation prédictive par exemple.

Le prochaine section propose une explicitation du canevas bayésien classique que nous considérons dans ce travail. Ensuite, après un rappel du contexte de la théorie des priors de référence généralisée telle que nous la considérons, nous dévouons la section 3 à la présentation de notre résultat principal et de sa démonstration. Finalement, nous concluons notre travail en section 4.

2 Canevas bayésien

Le cadre bayésien considéré dans ce travail est classiquement construit. Nous en proposons une rapide revue dans cette section.

Soit $k > 1$, considérons un espace probabilisé $(\Omega, \mathscr{P}, \mathbb{P})$ sur lequel sont définis la variable aléatoire T, à valeur dans l'espace mesurable $(\Theta, \mathscr{T})$, et le vecteur aléatoire $\mathbf{Y} = (Y_i)_{i=1}^k$, à valeur dans l'espace mesurable $(\mathcal{Y}^k, \mathscr{Y}^{\otimes k})$.

La variable aléatoire $\mathbf{Y}$ représente les observations itératives d'un système étudié, elles sont considérées comme indépendantes et identiquement distribuées conditionnellement à T :

$$\text{pour tout } B_1, \dots, B_k \in \mathscr{Y},\ \mathbb{E}[\prod_{i=1}^k \mathbb{1}_{Y_i \in B_i} | T] = \prod_{i=1}^k \mathbb{E}[\mathbb{1}_{Y_1 \in B_i} | T] \text{ p.s.} \tag{1}$$

Généralement et dans ce travail, Θ est un sous-ensemble de $\mathbb{R}^d$, $d \geq 1$, et la distribution π de T est appelée le prior. Les distributions conditionnelles $\mathbb{P}_{Y_i|T}$ pour tout Y_i existent comme suit :

$$\forall \theta \in \Theta,\ \mathbb{P}_{Y_i|T=\theta} \text{ est une probabilité sur } (\mathcal{Y}, \mathscr{Y}), \tag{2}$$

$$\forall B \in \mathscr{Y},\ \mathbb{P}_{Y_i|T=\cdot}(B) \text{ est mesurable sur } \Theta, \tag{3}$$

$$\forall A \in \mathscr{T},\ B \in \mathscr{Y},\ \mathbb{P}(T \in A, Y_i \in B) = \int_A \mathbb{P}_{Y_i|T=\theta}(B) d\pi(\theta), \tag{4}$$

avec $\mathbb{P}_{\mathbf{Y}|T} = \mathbb{P}_{Y_1|T}^{\otimes k} = \mathbb{P}_{Y|T}^{\otimes k}$ résultant de (1).

On suppose également que le problème admette une vraisemblance : il existe des densités $(\ell(\cdot|\theta))_{\theta \in \Theta}$ par rapport à une mesure commune μ sur $\mathcal{Y}$ telles que

$$\forall B \in \mathscr{Y},\ \mathbb{P}_{Y|T=\theta}(A) = \int_B \ell(y|\theta) d\mu(y) \text{ pour } \pi\text{-presque tout } \theta. \tag{5}$$

Ceci permet la définition des densités marginales et a posteriori, respectivement définies par rapport à $\mu^{\otimes k}$ et à π :

$$\forall \mathbf{y} \in \mathcal{Y}^k,\ p_{\mathbf{Y}}(\mathbf{y}) = \int_\Theta \prod_{i=1}^k \ell(y_i|\theta) d\pi(\theta) = \int_\Theta \ell_k(\mathbf{y}|\theta) d\pi(\theta), \tag{6}$$

$$\forall \theta \in \Theta,\ \mathbf{y} \in \mathcal{Y}^k,\ p(\theta|\mathbf{y}) = \frac{\ell_k(\mathbf{y}|\theta)}{p_{\mathbf{Y}}(\mathbf{y})} \text{ si } p_{\mathbf{Y}}(\mathbf{y}) \neq 0,\ p(\theta|\mathbf{y}) = 1 \text{ sinon.} \tag{7}$$

Dans un tel cadre, des hypothèses classiques de régularité de la vraisemblance par rapport à θ (voir par ex. Lehmann (1999)) permettent à la matrice d'information de Fisher $\mathcal{I}(\theta) = (\mathcal{I}(\theta)_{i,j})_{i,j=1}^d$ d'être bien définie selon :

$$\mathcal{I}(\theta)_{i,j} = -\int_{\mathcal{Y}} [\partial^2_{\theta_i \theta_j} \log \ell(y|\theta)]\, \ell(y|\theta) d\mu(y). \tag{8}$$

Notons enfin que le théorème d'extension de Kolmogorov propose un cadre qui inclut le travail qui précède et qui l'étend à toute valeur de k : il existe un espace probabilisé $(\overline{\Omega}, \overline{\mathscr{P}}, \overline{\mathbb{P}})$ tel que pour tout k, et tout $A \in \sigma_T$, $B_1 \in \sigma_{Y_1}, \ldots, B_k \in \sigma_{Y_k}$,

$$\overline{\mathbb{P}}(A, B_1, \ldots, B_k) = \mathbb{P}(A, B_1, \ldots, B_k) = \int_{T(A)} \mathbb{P}^{\otimes k}_{\mathbf{Y}|T=\theta}(\mathbf{Y}(B_1 \times \cdots \times B_k)) d\pi(\theta). \tag{9}$$

3 Prior de référence sous contraintes

3.1 Information mutuelle et prior de référence généralisé

Cette communication prend pour support la théorie étendue des priors de références telle que proposée et décrite dans (Van Biesbroeck, 2023). Dans notre travail, nous considérons l'information mutuelle I_{D_f}, définie à partir d'une f-divergence D_f comme mesure de dissimilarité :

$$I_{D_f}(\pi|k) = \int_\Theta D_f(\mathbb{P}_{\mathbf{Y}|T=\theta}||\mathbb{P}_{\mathbf{Y}}) d\pi(\theta), \tag{10}$$

$$\text{avec} \quad D_f(\mathbb{P}_{\mathbf{Y}|T=\theta}||\mathbb{P}_{\mathbf{Y}}) = \int_{\mathcal{Y}^k} f\left(\frac{\ell_k(\mathbf{y}|\theta)}{p_{\mathbf{Y}}(\mathbf{y})}\right) d\mu^{\otimes k}(\mathbf{y}). \tag{11}$$

Le cadre des f-divergences laisse plutôt ouvert le choix de fonction f. A titre d'exemple, si $f = -\log$, alors D_f n'est autre que la célèbre divergence de Kullback-Leibler. Dans notre travail, nous nous limitons à des fonctions f mesurables, localement bornées et dont les comportements asymptotiques en 0 et en $+\infty$ sont contrôlés selon :

$$f(x) \underset{x\to 0^+}{=} \alpha x^\beta + o(x^\beta), \quad f(x) \underset{x\to\infty}{=} O(x), \tag{12}$$

avec $\alpha < 0$, $\beta \in]0, 1[$.

Nous supposons également Θ compact. Dans ce cadre, nous adaptons et ré-exprimons ci-dessous la définition du prior de référence généralisé sous la considération de la mesure de dissimilarité D_f.

Définition 1 (D_f-prior de référence (Van Biesbroeck, 2023)). Soit $\mathcal{P}$ une classe de priors sur Θ. Un prior $\pi^* \in \mathcal{P}$ est appelé D_f-prior de référence sur la classe $\mathcal{P}$ au taux $\varphi(k)$ si

$$\forall \pi \in \mathcal{P}, \lim_{k\to\infty} \varphi(k)(I_{D_f}(\pi^*|k) - I_{D_f}(\pi|k)) \geq 0. \tag{13}$$

Le théorème ci-après est extrait de (Van Biesbroeck, 2023, Theorem 1). Sous des hypothèses formelle que l'auteur décrit, il donne une expression de la limite impliquée dans l'équation (13).

Théorème 1 (Van Biesbroeck (2023)). *Soit π un prior qui admet une densité continue par rapport à la mesure de Lebesgue sur Θ, que l'on note également π ici sans ambiguïté. Alors,*

$$\lim_{k\to\infty} k^{d\beta/2} I_{D_f}(\pi|k) = l(\pi) = \alpha C_\beta \int_\Theta \pi(\theta)^{1+\beta} |\mathcal{I}(\theta)|^{-\beta/2} d\theta \tag{14}$$

avec $C_\beta = (2\pi)^{d\beta/2}(1-\beta)^{-d/2}$.

Le choix de la classe $\mathcal{P}$ dans la définition 1 reste ouvert, et bien que le théorème ci-dessus amène à se restreindre à la classe —encore large— des priors à densités continues, des restrictions additionnelles pourraient très bien y être ajoutées. C'est cette idée que nous explorons dans la section qui suit, avec l'introduction de contraintes linéaires sur le prior.

3.2 Résultat principal

La principale contribution de ce travail constitue le théorème qui suit, en proposant une expression d'un prior de référence sous un certain nombre de contraintes linéaires.

Le cadre de travail général se limite ici à celui des priors qui admettent une densité continue par rapport à la mesure de Lebesgue. A ce titre et à dessein de simplification, la lettre π servira dorénavant plutôt la désignation générique d'une densité plutôt que d'un prior. E désigne alors l'espace des fonctions continues de Θ dans $\mathbb{R}$, il est muni de la norme infinie : $\|p\| = \sup_\Theta |p|$ pour tout $p \in E$.

Hypothèse 1. Une famille de fonctions mesurables $g_1, \dots, g_p : \Theta \longrightarrow \mathbb{R}$ est dite satisfaisant l'hypothèse 1 si elles sont intégrables sur Θ et que les fonctions $g_0, \dots, g_p$ sont linéairement indépendantes, en définissant $g_0 : \theta \mapsto 1$.

Théorème 2. *Soient $g_1, \dots, g_p$ des applications mesurables de Θ vers $\mathbb{R}$ qui satisfont l'hypothèse 1. On définit $\mathcal{P}$ comme la classe des priors admettant une densité $\pi \in E$ positive et telle que $\forall j = 1, \dots, p$, $\int_\Theta g_i(\theta)\pi(\theta)d\theta = c_j$ pour des certains $c_j \in \mathbb{R}$. Si $\mathcal{P}$ est non vide, alors il existe un unique D_f-prior de référence sur $\mathcal{P}$. S'il est strictement positif sa densité π^* s'écrit*

$$\pi^* = J \cdot (\lambda_0 + \sum_{i=1}^{p} \lambda_i g_i)^{1/\beta}, \tag{15}$$

pour certains scalaires $\lambda_0, \dots, \lambda_p \in \mathbb{R}$. Réciproquement, un prior de $\mathcal{P}$ dont la densité s'exprime sous la forme ci-dessus est l'unique D_f-prior de référence. J désigne la densité du prior de Jeffreys : $J(\theta) = |\mathcal{I}(\theta)|^{1/2} / \int_\Theta |\mathcal{I}(\tilde{\theta})|^{1/2} d\tilde{\theta}$.

Les contraintes linéaires auxquelles se limitent ce théorème sauraient intuitivement servir de contraintes de moment. Un exemple simple pourrait être de choisir $p = d$ et $g_i = \theta_i$ pour fixer les espérances a priori. Par ailleurs, ce format de contraintes peut aussi s'appliquer aux paramètres d'ellicitation prédictive (Bousquet, 2023) : soit des percentiles prédictifs $(t_{\delta_i}, \delta_i)_i$, on contraint le prior π à satisfaire :

$$\mathbb{P}(h(\mathbf{Y}) \leq t_{\delta_i}) = \int_\Theta \int_{\mathcal{Y}^k} \mathbb{1}_{h(\mathbf{y}) \leq t_{\delta_i}} \ell_k(\mathbf{y}|\theta) d\mu^{\otimes k}(\mathbf{y}) \pi(\theta) d\theta \tag{16}$$

pour une certaine fonction h.

3.3 Démonstration

Rappelons la notation l issue de l'équation (14) :

$$l(\pi) = \alpha C_\beta \int_\Theta \pi(\theta)^{1+\beta} |\mathcal{I}(\theta)|^{-\beta/2} d\theta. \tag{17}$$

Le théorème 1 fait le lien entre l'optimisation de l et le D_f-prior de référence. Ce dernier est en effet le point en lequel l atteint son maximum.

Remarquons que Θ étant compact, la paire $(E, \|\cdot\|)$ constitue un espace de Banach dont la restriction U composée des fonctions strictement positives est un sous-ensemble ouvert et convexe, sur lequel l définit une fonction concave à valeurs dans $\mathbb{R}$.

Différentions l sur U. Pour ceci on écrit $l = \phi_1 \circ \phi_2$ avec

$$\phi_1 : \pi \in E \longmapsto \pi^{1+\beta} \in E; \qquad \phi_2 : \pi \in E \longmapsto \alpha C_\beta \int_\Theta \pi(\theta)|\mathcal{I}(\theta)|^{-\beta/2} d\theta. \tag{18}$$

Comme ϕ_2 est une application linéaire continue de E dans $\mathbb{R}$, l est différentiable tant que ϕ_1 l'est, avec :

$$dl(\pi) = d\phi_1(\phi_2(\pi)) \circ \phi_2. \tag{19}$$

Soit $\pi \in U$. Pour tout $\varepsilon > 0$ il existe $\tilde{\varepsilon} > 0$ tel que tant que $|x| < \|\pi\|$ et $|u| < \tilde{\varepsilon}$ alors $|(x+u)^{1+\beta} - x^\beta - (1+\beta)x^\beta u| < \varepsilon|u|$. Ainsi, pour tout $h \in E$ tel que $\|h\| < \tilde{\varepsilon}$, on peut écrire

$$\|\phi_1(\pi + h) - \phi_1(\pi) - (1+\beta)\pi^\beta h\| < \varepsilon\|h\|. \tag{20}$$

On conclut que ϕ_1 est différentiable sur U avec $d\phi_1(\pi)h = (1+\beta)\pi^\beta h$, pour tout $\pi \in U$, $h \in E$. Cette différentielle est de plus continue, dont on déduit que l est également continûment différentiable.

Ainsi, considérant la contrainte additionnelle $\int_\Theta \pi(\theta)d\theta = 1$, ce problème peut être traité en appliquant le théorème des extrema liés (cf. par ex. Cartan (2007)) selon lequel il existe $\lambda_0, \ldots, \lambda_p \in \mathbb{R}^{p+1}$ tels que tout extremum $\pi^* \in U \cap C$ (où C désigne les fonctions qui satisfont les contraintes) de l et satisfaisant les contraintes vérifie

$$dl(\pi^*)h - \lambda_0 \int_\Theta h(\theta)d\theta - \sum_{i=1}^{p} \lambda_i \int_\Theta h(\theta)g_i(\theta)d\theta = 0 \tag{21}$$

pour tout $h \in E$. Enfin, comme $dl(\pi)h = \alpha C_\beta(1+\beta)\int_\Theta \pi(\theta)^\beta |\mathcal{I}(\theta)|^{-\beta/2} h(\theta)d\theta$, on obtient, quitte à renommer les λ_i,

$$\pi^*(\theta) = J(\theta)\left(\lambda_0 + \sum_{i=1}^{p} \lambda_i g_i(\theta)\right)^{1/\beta}. \tag{22}$$

Aussi, la stricte concavité de l implique que, (i) s'il existe, π^* est l'unique argument maximal de l sur $U \cap C$, et que (ii) si réciproquement un prior $\pi^* \in U \cap C$ satisfait l'équation (22) alors il est l'unique l'argument maximal de l sur $U \cap C$.

Pour conclure, on remarque que les densités des priors positifs satisfaisant les contraintes constituent la fermeture de l'ensemble $U \cap C$. Puisque π^* tel que défini ci-dessus maximise l sur $U \cap C$, la continuité de l sur sa fermeture induit le caractère maximal de π^* sur celle-ci également. Autrement dit, le prior auquel il est associé est le D_f-prior de référence sur $\mathcal{P}$.

4 Conclusion

Le question du choix du prior en inférence bayésienne reste ouverte et complexe. D'une part, le point de vue de la théorie des priors de référence suggère la minimisation de l'influence de toute information a priori ; d'autre part, de nombreuses applications pratiques plébiscitent l'emploi du canevas bayésien pour l'introduction d'un jugement a priori.

Notre communication propose une réconcilitation de ces deux mondes en définissant un cadre sur lequel leur intersection est possible. Notre prior de référence se construit ici sous contraintes, et explicite en notre sens le bon cadre pour l'introduction d'information a priori.

Ce dernier formalisme s'appuie sur une définition plus étendue de la théorie des priors de référence qui propose la considération de mesures de dissimilarité plus large, en complément à l'historique divergence de Kullback-Leibler. La souplesse de leur cadre et de leur résultat est appuyé par notre travail et ses ouvertures.

Références

Bernardo, J. M. (1979). Reference posterior distributions for Bayesian inference. *Journal of the Royal Statistical Society. Series B*, 41(2), 113–147. doi:10.1111/j.2517-6161.1979.tb01066.x.

Bousquet, N. (2023). Discussion of "specifying prior distributions in reliability applications" : Towards new formal rules for informative prior elicitation ? *Applied Stochastic Models in Business and Industry*. doi:10.1002/asmb.2794.

Cartan, H. (2007). *Cours de calcul différentiel*. Sciences et techniques. Hermann, 2ème édition.

Clarke, B. S. & Barron, A. R. (1994). Jeffreys' prior is asymptotically least favorable under entropy risk. *Journal of Statistical Planning and Inference*, 41(1), 37–60. doi:10.1016/0378-3758(94)90153-8.

Gauchy, C., Van Biesbroeck, A., Feau, C., & Garnier, J. (2023). Inférence variationelle de lois a priori de référence. Dans *Proceedings des 54èmes Journées de Statistiques de la SFDS (JdS)*.

Gu, M., Wang, X., & Berger, J. O. (2018). Robust Gaussian stochastic process emulation. *The Annals of Statistics*, 46(6A), 3038–3066. doi:10.1214/17-AOS1648.

Keefe, M. J., Ferreira, M. A. R., & Franck, C. T. (2019). Objective Bayesian Analysis for Gaussian Hierarchical Models with Intrinsic Conditional Autoregressive Priors. *Bayesian Analysis*, 14(1), 181–209. doi:10.1214/18-BA1107.

Lehmann, E. L. (1999). *Elements of Large-Sample Theory*. Springer Texts in Statistics. Springer New York, NY, 1 édition.

Muré, J. (2021). Propriety of the reference posterior distribution in Gaussian process modeling. *The Annals of Statistics*, 49(4), 2356–2377. doi:10.1214/20-AOS2040.

Nalisnick, E. & Smyth, P. (2017). Learning approximately objective priors. Dans *Proceedings of the 33rd Conference on Uncertainty in Artificial Intelligence (UAI)*.

Van Biesbroeck, A. (2023). Generalized mutual information and their reference prior under Csizar f-divergence. arXiv.2310.10530. doi:10.48550/arXiv.2310.10530.