Proposition de stage

Stratégie bayésienne et échantillonnage stochastique pour la comparaison de modèles

Application en astronomie

Domaines d'application: astronomie, physique (mécanique, thermique,...), sciences naturelles,...

Contexte scientifique: comparaison / sélection de modèles.

Démarche méthodologique: stratégie optimale, approche bayésienne, probabilité de modèle.

Outil et approche: évidence, espérance harmonique, approche de Chib, approximation de Laplace,...

Aspects numériques: échantillonnage stochastique, Metropolis-Hastings (Langevin / Hamilton,...)

Environnement numérique: Matlab sur PC, Automatic Differentiation and Deep Learning Toolbox.

Lieu: Groupe Signal – Image, IMS (Université de Bordeaux – CNRS – BINP), Talence, France.

Durée: cinq à six mois, à partir de janvier ou février 2024.

Contact: Jean-François Giovannelli, IMS (`Giova@IMS-Bordeaux.fr`) et Pascal Bordé, LAB (`pascal.borde@u-bordeaux.fr`)

Contexte — Le travail proposé concerne la question générique de la comparaison de modèles à partir de données expérimentales. Il s'agit de confronter des modèles mathématiques pilotés par des paramètres, à la réalité physique au travers de mesures [1–4]. Cette question constitue un sujet de recherche actuel à part entière dans la communauté « *Data Science* » et les retombées potentielles sont déterminantes. En effet, les domaines concernés sont particulièrement nombreux: on pense par exemple à la caractérisation de phénomènes physiques en mécanique, thermodynamique, électricité et électronique, en astronomie et au delà dans les autres sciences naturelles au sens large. Ce travail contribuera à la construction d'une méthodologie d'inférence et ainsi à l'objectif du projet « *Origins* » de répondre à des interrogations fondamentales concernant l'origine de la matière et de sa complexité, la formation des planètes et des étoiles, la climatologie, l'apparition de l'homme moderne,... au travers de la confrontation de modèles à des données expérimentales. Un exemple générique pourrait être celui d'un phénomène oscillatoire amorti comme dans [5] ou d'autres plus complexes comme en physique des particules [6, 7] ou encore la vélocimétrie radiale ou spectro-photométrie des transits [8, 9] pour la détection d'exoplanètes.

Méthodologie — La question qui se pose est celle de la comparaison de structures de modèles concurrents dans une liste, la sélection d'une structure dans une famille ou encore la validation ou l'invalidation d'une hypothèse. Ces modèles font intervenir deux composantes: (a) des relations souvent déterministes pour les phénomènes physiques en jeu et (b) une dimension stochastique pour inclure des incertitudes sur les inconnues et les mesures. Il s'agit le plus souvent de modèles paramétriques et la valeur des paramètres est en général inconnue, même si on peut en avoir une idée ou un ordre de grandeur. Bien sûr, ces deux aspects (structure de modèle et valeurs des paramètres)

sont intimement liés, au moins pour des modèles concurrents possédant un nombre de degrés de liberté différents et a fortiori des modèles emboîtés (rasoir d'Ockham).

Nous nous focaliserons sur les méthodes de sélection fondées sur une stratégie bayésienne [1,2,4]. Elle permet la construction de fonctions de sélection optimales en un sens clairement explicité (voir annexe). Elles reposent *in fine* sur une distribution dite a posteriori pour les modèles, c'est-à-dire conditionnelle aux données [2, 11, 12]. La construction de chaque probabilité se fonde alors sur une *évidence*, qui elle-même nécessite la marginalisation des paramètres inconnus. C'est cet aspect qui présente une difficulté majeure en général.

Aspects numériques et algorithmiques — Les évidences seront calculées par échantillonnage (production de tirages ou de réalisations) de la distribution a posteriori pour les paramètres des modèles [13–16]. Cet échantillonnage sera réalisé par un algorithme de la famille de Monte-Carlo par chaîne de Markov mixte, dit de *Metropolis-Hastings within Gibbs*, comme suit.

1. L'échantillonnage de la variance-inverse des erreurs est effectué sous une loi Gamma.
2. L'échantillonnage des autres paramètres est en général plus complexe et il est réalisé par une technique de Metropolis-Hastings, en deux étapes.
    a. *Proposition* : simuler une valeur sous une loi dite de proposition ou instrumentale.
    b. *Acceptation / Duplication* : décider, au hasard avec une probabilité bien définie, d'accepter la valeur proposée ou de dupliquer la valeur courante.

L'efficacité est directement fonction de (i) l'importance du taux d'acceptation et (ii) la taille moyenne des excursions dans l'espace des paramètres, ces deux caractères étant eux-mêmes intimement liés à la pertinence de la loi de proposition relativement à la loi a posteriori. Une première option considère la distribution *a priori*: il s'agit d'un choix naturel et dont l'utilisation est aisée. Une autre option classique considère une marche aléatoire et / ou une densité gaussienne dont les paramètres sont adaptés automatiquement. Pour certains cas ces options sont tout à fait adaptées mais pour d'autres, en particulier mettant en jeu des distributions très complexes, elles demeurent valides mais deviennent inefficace à cause de la charge calculatoire (temps, convergence, encombrement mémoire, quantité de calcul,...). Pour dépasser cet limitations, il est crucial d'investir sur cette question en tant que tel et deux voies prometteuses tirent avantage d'un couplage à d'autres outils d'analyse numérique.

- On construit des propositions dirigées, fondées sur le gradient de la distribution (*e.g.,* Langevin) et sur son Hessien ou l'information de Fisher (*e.g.,* Fisher-Langevin), comme en *optimisation.*
- On construit des propositions comme solutions d'*équations différentielles* (*e.g.,* Hamilton).

On pourra consulter [17, 18], ainsi que [16, Ch. 5]. Cette combinaison, méthode d'échantillonnage fondée sur des outils d'analyse numériques (optimisation, résolution d'équations et de systèmes, ou d'équations aux dérivées partielles,...) constitue un sujet de recherche à part entière à la fois prometteur sur un plan fondamental et en phase avec les problématiques appliquées.

Un dernier élément, d'importance. En plus des probabilités pour chaque modèle, on aura aussi accès à la distribution a posteriori des paramètres et donc à la moyenne fournissant une valeur estimée, également optimale (voir annexe), des paramètres pour chaque modèle. On aura également des informations concernant des marges d'erreurs au travers des écarts-types a posteriori, et des couplages entre paramètres au travers des corrélations a posteriori.

Structuration des travaux — Pour ce qui est du déroulement chronologique, le travail pourra se structurer en trois étapes: (1) questions liées à la construction des probabilités de modèles, puis (2) celles liées au calculs numériques effectifs de ces probabilités, enfin (3) les aspects concernant l'efficacité algorithmique. Sur le plan du contenu, le travail pourra se décomposer en trois volets.

a. Étude bibliographique en particulier concernant la sélection, le calcul des probabilités et des évidences ainsi que les algorithmes de calcul numérique et les conditions de leur utilisation.

b. Mise en œuvre et codage en *Matlab*.

c. Évaluations théorique et / ou numérique des performances. Comparaison de diverses méthodes.

Ceci ne constitue qu'une possibilité qui pourra être ajustée.

References

[1] A. Gelman, J. C. Carlin, H. S. Stern et D. B. Rubin, *Bayesian Data Analysis*, Chapman & Hall/CRC, Boca Raton, USA, 2nd edition, 2004.

[2] C. P. Robert, *The Bayesian Choice. From decision-theoretic foundations to computational implementation*, Springer Texts in Statistics. Springer Verlag, New York, USA, 2007.

[3] J.-J. Droesbeke, J. Fine et G. Saporta, *Méthode bayésiennes en statistique*, Ouvrage collectif ASU-SFdS. Édition Technip, Paris, France, 2002.

[4] H. L. Harney, *Bayesian Inference: Parameter Estimation and Decision*, Physics and Astronomy. Springer, Berlin, Allemagne, 2003.

[5] A. Barbos, A. Giremus et J.-F. Giovannelli, « Bayesian noise model selection and system identification using Chib approximation based on the Metropolis-Hastings sampler », in *Actes 25 e coll. GRETSI*, Lyon, France, septembre 2015.

[6] K. E. Duffy, *First Measurement of Neutrino and Antineutrino Oscillation at T2K*, Springer, Oxford, Angleterre, theses book series edition, 2017.

[7] L. Lista, *Statistical Methods for Data Analysis in Particle Physics*, Springer, Heidelberg, Allemagne, book series: lecture notes in Physics, edition, 2017.

[8] M. Tuomi et S. Kotiranta, « Bayesian analysis of the radial velocities of HD 11506 reveals another planetary companion », *Letters, Astron. Astrophys.*, vol. 496, N° 2, pp. L13–L16, mars 2009.

[9] M. Tuomi, « Bayesian re-analysis of the radial velocities of Gliese 581 », *Letters, Astron. Astrophys.*, vol. 528, 2011.

[10] Y. Yang, N. Zou, E. Lin, F. Suo et Z. Chen, « A neural network method for nonconvex optimization and its application on parameter retrieval », *IEEE Trans. Signal Processing*, vol. 69, pp. 3383–3398, 2021.

[11] T. Ando, *Bayesian model selection and statistical modeling*, Chapman & Hall/CRC, Boca Raton, USA, 2010.

[12] J. Ding, V. Tarokh et Y. Yang, « Model selection techniques: An overview », *IEEE Signal Proc. Mag.*, vol. 35, N° 6, pp. 16–34, novembre 2018.

[13] J.-M. Marin et C. P. Robert, *Bayesian Core. A Practical Approach to Computational Bayesian Statistics*, Texts in statistics. Springer, Paris, France, 2007.

[14] J. Albert, *Bayesian Computation With R*, Springer-Verlag New York Inc., New York, NY, USA, 2009.

[15] D. Gamerman et H. F. Lopes, *Markov Chain Monte Carlo: stochastic simulation for Bayesian inference*, Chapman & Hall/CRC, Boca Raton, USA, 2nd edition, 2006.

[16] S. Brooks, A. Gelman, G. L. Jones et X.-L. Meng, *Handbook of Markov Chain Monte Carlo*, Chapman & Hall / CRC, Boca Raton, USA, 2011.

[17] M. Girolami et B. Calderhead, « Riemannian manifold Hamiltonian Monte Carlo (with discussion) », *J. R. Statist. Soc. B*, vol. 73, pp. 123–214, 2011.

[18] C. Vacar, J.-F. Giovannelli et Y. Berthoumieu, « Langevin and Hessian with Fisher approximation stochastic sampling for parameter estimation of structured covariance », in *Proc. IEEE ICASSP*, Prague, Czech Republic, mai 2011, pp. 3964–3967.

[19] G. Roberts et O. Stramer, « Langevin Diffusions and Metropolis-Hastings Algorithms », *Methodology and Computing in Applied Probability*, vol. 4, pp. 337–358, 2003.

Annexe sélection optimale — Afin de formaliser les choses, on note $\boldsymbol{y}$ un vecteur de données dans un espace $\mathcal{Y}$ et on note m un modèle dans une liste de modèles candidats $\mathcal{M}$. La question générique de la sélection de modèle est celle de la construction d'une fonction de décision Δ, de l'espace des données $\mathcal{Y}$ vers la liste des modèles $\mathcal{M}$ (l'espace des décisions), qui associe un modèle sélectionné à un jeu de données et on écrit $\widehat{m} = \varphi(\boldsymbol{y})$.

$$\begin{array}{rccc} \Delta : & \mathcal{Y} & \longrightarrow & \mathcal{M} \\ & \boldsymbol{y} & \longmapsto & \widehat{m} = \Delta\left(\boldsymbol{y}\right) \end{array}$$

Il existe de plusieurs manières d'aborder les choses et on se focalise sur une approche reposant sur l'idée d'optimalité. Pour cela on définit un coût (*cost function*), noté $\mathcal{C}$

$$\begin{array}{rccc} \mathcal{C} : & \mathcal{M} \times \mathcal{M} & \longrightarrow & \mathbb{R} \\ & (m\,,\,m') & \longmapsto & \mathcal{C}\left[\,m\,,\,m'\,\right] \end{array}$$

qui quantifie le coût lié à une erreur de décision sur le modèle, typiquement un coût binaire (coût nul si $m = m'$ et coût unitaire si $m \neq m'$). Dans un second temps, on définit un risque (*risk function*), notée $\mathcal{R}$, comme un coût moyen (sur les modèles et les données):

$$\mathcal{R}(\Delta) = \mathbb{E}_{M,\boldsymbol{Y}}\left\{\,\mathcal{C}\left[\,M\,,\,\Delta(\boldsymbol{Y})\,\right]\right\}$$

associé à la fonction de décision Δ. Naturellement, la fonction de sélection choisie Δ_{opt} est celle qui minimise le risque:

$$\Delta_{\text{opt}} = \arg\min_{\Delta} \mathcal{R}(\Delta)$$

et qui est optimale en ce sens. On est alors en capacité d'extraire des données, de manière optimale, des informations à propos des modèles (puis de leurs paramètres par une stratégie similaire), c'est-à-dire en un sens d'interpréter les données.