Stage/These 2022

Title **Likelihood-free cosmological parameter inference using theoretical high-order statistics predictions**

Laboratory: IRFU/DAP/CosmoStat**,** CEA Saclay
Supervisors: Jean-Luc Starck & Sandrine Codis
Contact: jstarck@cea.fr ☏ : 01 69 08 57 64 http://jstarck.cosmostat.org
sandrine.codis@cea.fr ☏ : 01 69 08 35 87 https://sandrinecodis.wixsite.com/sandrinewebsite

Keywords: weak lensing, Euclid, cosmological parameter likelihood free inference.

Subject:

The Euclid satellite, to be launched in 2023, will observe the sky in the optical and infrared, and will be able to map large scale structures and weak lensing distortions out to high redshifts. Weak gravitational lensing is thought to be one of the most promising tools of cosmology to constrain models. Weak lensing probes the evolution of dark-matter structures and can help distinguish between dark energy and models of modified gravity. Cosmological parameters are traditionally estimated using a Gaussian likelihood based on theoretical predictions of second order statistic such as the power spectrum or the two point correlation functions. This requires to build a covariance matrices, and therefore need a lot of very heavy n-body simulations. This approach presents also several additional drawbacks: First, second order statistics captures all available information in the data only in the case of Gaussian Random Fields, while matter distribution is highly non-Gaussian showing many features such filaments, walls or clusters. Second, the covariance matrix is cosmology dependent and the noise it generally not Gaussian, both aspects being generally poorly taken into account. Finally, all systematic effects such as masks, intrinsic alignement, baryonic feedback are very difficult to take into account. For all these reasons, a new approach has recently emerged, called likelihood-free cosmological parameter inference which are based on a forward modelling. It has the great advantage to not need covariance matrices anymore, avoiding the storage of huge simulated data set (we typically need 10000 n-body realisations for each set of cosmological parameters). Furthermore, it opens us the door to use high order statistical information and it is relatively straightforward to include all systematics effect. It has however two serious drawbacks, the firsts one is the need of huge GPU ressources to process surveys such as Euclid and the second is that the solution relies on the accuracy of simulations, which could lead to infinite discussion in case the results are different from what is expected. Thanks to a recent breakthrough (Codis, 2021), we have now theoretical tools to predict, for a given set of cosmological parameters, the multi-scale density probability function (pdf) of convergence maps such as the one that will be observed with Euclid.

The goal of this PhD work is to develop an hybrid approach, consisting in a likelihood-free cosmological parameter inference which would be based on the high order statics theoretical prediction rather than n-body simulations. It would therefore have the advantage of both previously described approaches, as it **will not need to store huge data** set to compute a covariance matrix and **it will not require huge CPU/ GPU** ressources as the forward modelling method. This intense frugality will make this approach highly competitive to constraint the cosmological model using high order statistics in future surveys.

To achieve this goal, the first step will be to build a map emulator, similar to what has been done for 2 point statistics (i.e. the flask method), but which respects accurately the high order predictions. Using this emulator, it will then be possible to use it as a bypass in a recently developed Likelihood Free Inference code. This will allow the use of high order statics such as the l1-norm of the wavelet transform of the convergence to constrain the cosmological parameters, which is an extremely powerful summary statistic (Ajani et al, 2021). The method will be used first on the CFIS survey, and then on Euclid.

References

1. Barthelemy A., **Codis S**. and Bernardeau F., "Probability distribution function of the aperture mass field with large deviation theory", 2021, MNRAS, 503, 5204;
2. V. Ajani, J.-L. Starck and V. Pettorino, "Starlet l1-norm for weak lensing cosmology", **Astronomy and Astrophysics**, 645, L11, 2021.

Résumé:

Le satellite Euclid, qui sera lancé en 2023, observera le ciel dans les domaines optique et infrarouge, et mesurera les distorsions gravitationnelles jusqu'à des redshifts très élevés. L'effet de lentille gravitationnelle faible est considérée comme l'un des outils les plus prometteurs de la cosmologie pour contraindre les modèles. Les lentilles faibles sondent l'évolution des structures de la matière noire et peuvent aider à distinguer l'énergie noire des modèles de gravité modifiée. Grâce aux mesures de cisaillement, nous pourrons reconstruire une carte de masse de matière noire de 15 000 degrés carrés. La cartographie de masse implique la construction de cartes bidimensionnelles utilisant des mesures de forme de galaxie, représentant la densité de matière totale intégrée le long de la ligne de visée. Les cartes de masse sur des petits champs ont souvent été utilisées pour étudier la structure et la distribution en masse des amas de galaxies, alors que les cartes à grand champ ne sont possibles que depuis peu, en raison des stratégies d'observation de relevés de galaxies tels que CFHTLenS, HSC, DES et KiDS. Les cartes de masse contiennent des informations cosmologiques non gaussiennes significatives et peuvent être utilisées pour identifier des amas massifs ainsi que pour effectuer une corrélation croisée entre le signal de lentille et les structures d'avant plan.
Les paramètres cosmologiques sont traditionnellement estimés à l'aide d'une vraisemblance gaussienne basée sur des prédictions théoriques de statistiques de second ordre telles que le spectre de puissance ou les fonctions de corrélation à deux points. Cela nécessite de construire des matrices de covariance, et donc de nombreuses simulations à n corps très lourdes. Cette approche présente également plusieurs inconvénients supplémentaires : premièrement, les statistiques de second ordre capturent toutes les informations disponibles dans les données uniquement dans le cas des champs aléatoires gaussiens, tandis que la distribution de la matière est hautement non gaussienne et présente de nombreuses caractéristiques telles que des filaments, des feuillets ou des amas. Deuxièmement, la matrice de covariance est dépendante de la cosmologie et le bruit n'est généralement pas gaussien, ces deux aspects étant généralement mal pris en compte. Enfin, tous les effets systématiques tels que les masques, l'alignement intrinsèque, les effets baryoniques sont très difficiles à prendre en compte. Pour toutes ces raisons, une nouvelle approche a récemment émergé, appelée inférence de paramètres cosmologiques sans vraisemblance, basée sur une modélisation "forward". Il a le grand avantage de ne plus avoir besoin de matrices de covariance, évitant le stockage d'un énorme ensemble de données simulées (nous avons généralement besoin de 10 000 réalisations à n corps pour chaque ensemble de paramètres cosmologiques). De plus, cela nous ouvre la porte à l'utilisation d'informations statistiques d'ordre élevé et il est relativement simple d'inclure tous les effets systématiques. Il présente cependant deux inconvénients sérieux, le premier est le besoin d'énormes ressources GPU pour traiter des relevés tels qu'Euclid et le second est que la solution repose sur la précision des simulations, ce qui pourrait conduire à des discussions infinies au cas où les résultats seraient différents de ce qui est attendu. Grâce à une percée récente (Codis, 2021), nous disposons désormais d'outils théoriques pour prédire, pour un ensemble donné de paramètres cosmologiques, la fonction de probabilité de densité multi-échelle (pdf) de cartes de convergence comme celle qui sera observée avec Euclid .

L'objectif de ce travail de thèse est de développer une approche hybride, consistant en une inférence de paramètres cosmologiques sans vraisemblance qui serait basée sur la prédiction théorique statique d'ordre élevé plutôt que sur des simulations à n corps. Il aurait donc l'avantage des deux approches décrites précédemment, car il n'aura pas besoin de stocker un énorme ensemble de données pour calculer une matrice de covariance et il ne nécessitera pas d'énormes ressources CPU/GPU comme méthode de modélisation avancée. Cette frugalité intense rendra cette approche hautement compétitive pour contraindre le modèle cosmologique en utilisant des statistiques d'ordre élevé dans les futurs relevés.
Pour atteindre cet objectif, la première étape sera de construire un émulateur de carte, similaire à ce qui a été fait pour les statistiques à 2 points (c'est-à-dire la méthode flask), mais qui respecte avec précision les prédictions d'ordre élevé. En utilisant cet émulateur, il sera alors possible de l'utiliser comme contournement dans un code d'inférence récemment développé. Cela permettra l'utilisation de statiques d'ordre élevé telles que la norme l1 de la transformée en ondelettes de la carte convergence pour contraindre les paramètres cosmologiques, la norme l1 étant qui une statistique extrêmement puissante (Ajani et al, 2021). La méthode sera d'appliqué sur le relevé CFIS, puis sur Euclid.

1. Barthelemy A., **Codis S**. and Bernardeau F., "Probability distribution function of the aperture mass field with large deviation theory", 2021, MNRAS, 503, 5204;
2. V. Ajani, J.-L. Starck and V. Pettorino, "Starlet l1-norm for weak lensing cosmology", **Astronomy and Astrophysics**, 645, L11, 2021.

Méthodologie:

L'objectif de la these est de disposer d'une nouvelle approche permettant de contraindre les paramètres cosmologiques avec les statistiques supérieurs, en utilisant les derniers résultats théoriques de prediction cosmologique, et le moteur d'inférence développé à CosmoStat. Celui ci est implémenté en TensorFlow, permettant d'avoir un modèle physique "forward" entièrement différentiable. Pour intégrer les predictions théoriques dans notre code, il faut disposer en TensorFlow de deux outils:

- La prédiction théorique, qui est actuellement codée en mathematica.
- Un émulateur de carte.

La première étape consistera alors re-coder en TensorFlow les codes de prédiction. Pour optimiser l'extraction d'information, il faudra adapter le code pour qu'il fasse une prediction sur la distribution des coefficients d'ondelettes, car la prédiction actuelle concernent des coefficients lissés à une échelle donnée avec une fonction "box". Il faudra aussi dériver la norme l1 des coefficients de la densité de probabilité. On pourra valider la qualité des prédictions avec des simulations hautes résolutions existantes.

Une fois que cet outil sera opérationnel, on pourra alors utiliser la puissance de TensorFlow et en particulier de la differentiation automatique pour bâtir un émulateur de carte. L'idée est d'utiliser un émulateur comme la méthode *flask* qui permet de simuler des cartes sphériques avec une distribution loi-normale, et de poser un problème inverse pour modifier la simulation afin qu'elle vérifie toute les propriétés statistiques requises. Grace à la differentiation automatique, on sait maintenant résoudre aisément ce type de problèmes. Si la generation des cartes reste gourmande en CPU/GPU, on aura la possibilité d'intégrer un U-net qui sera entrainé pour reproduire les résultats de la méthode.

Une fois l'émulateur développé, on pourra alors l'intégrer dans le code TensorFlow de "forward modelling", qui gère déjà tous les systématiques majeures des cartes de masse de weak lensing (alignement intrinsèque, erreur sur les mesures de redshifts, bruit instrumental, masques, etc).

La dernière étape sera d'appliquer la nouvelle méthode aux données des relevés CFIS et sur du telescope spatial Euclid qui sera lancé debut 2023.

Plusieurs publications sont envisagées:

- Prediction théorique de la norme l1 des coefficients rondelettes pour une cosmologie donnée.
- Emulateur de carte de masse respectant les predictions théoriques d'ordre supérieur.
- Likelihood free inference code en utilisant l'émulateur développé, et validation de la méthode.
- Application sur les données du relevé CFIS, le catalogue sera disponible dès la fin 2022.