Ecole Doctorale des Sciences Fondamentales

Title of the thesis: Mathematical aspects in statistical inference of initial cosmological parameters through forward modelling

Supervisor : A. Guillin and M. Michel
Laboratory : Laboratoire de Mathématiques Blaise Pascal
University : Université Clermont-Auvergne
Email and Phone : arnaud.guillin@uca.fr (04 73 40 70 90) manon.michel@uca.fr (04 73 40 70 88)
Possible co-supervisor :
Laboratory : Laboratoire de Mathématiques Blaise Pascal
University : Université Clermont-Auvergne

Summary :
Motivation : The BORG algorithm (Bayesian Origin Reconstruction from Galaxies) [1,2] aims to improve our understanding of dark matter and dark energy by providing a full characterization of the three-dimensional cosmic large-scale structure, including the non-linear evolving ones. These nonlinear structures will indeed add up to 80% of the information provided in the next-generation cosmological surveys [3]. They are ruled by complex higher order statistics than two-points ones, which requires novel data models capable of accounting for the properties of non-linear gravitational structure formation [4,5]. The BORG algorithm develops an original solution which transforms this issue into a statistical initial condition problem. Inside the BORG framework, the primordial initial conditions are recovered from observational data through Bayesian inference, physical forward modelling and Markov chain Monte Carlo (MCMC) sampling.
Following a Bayesian analysis, the a priori information on the initial conditions are modelled through a probability distribution and gives rise to different possible realisations of the initial Gaussian density perturbations. The final density fields obtained from these evolved initial perturbations can be compared to the available observational surveys and, through Bayes formula, this comparison yields a posterior probability distribution of the initial parameters. By implementing a full Bayesian probabilistic setting, one then obtains a more global picture than a pointwise parameter estimate and in particular uncertainty estimations.
This collection of samples of the initial conditions is obtained through a MCMC scheme. Such a method allows to explore the parameter space, i.e. the complete set of values of the parameters ruling the initial conditions, through a Markov chain. The large dimensionality (more than ten millions parameters) and complexity of this inference problem require however fast sampling MCMC methods. Traditional MCMC algorithms [6] indeed rely on accept-reject mechanism for correctness, leading to the waste of CPU times into computation of rejected steps and most of the time to a strong and slow random walk behavior. The BORG algorithm currently addresses this problem by using an efficient implementation of the Hamiltonian Monte Carlo (HMC) algorithm [7,8]. After extending the state space with auxiliary variables playing the role of velocities, HMC schemes rely on a discrete integration of an artificial Newtonian evolution of the extended state.

It allows to inject some persistence in the sequential positions visited by the Markov chain but is still using an acceptreject mechanism for correctness. Still, numerical costs of this algorithm, as well as its sensitivity to unstable floating points, pose a challenge for the analysis of next-generation cosmological data as provided by the Large Synoptic Survey Telescope or the European Euclid satellite mission.

Goal : To address this issue, we propose to implement Piecewise deterministic Monte Carlo (PDMC) approaches into the BORG algorithm, as speedups of several orders of magnitude could be obtained in comparison to the state of the art [9-11], including HMC. PDMC methods explore the parameter space through a sequence of ballistic moves whose direction is changed depending on the local geometry of the parameter space. Doing so, they are able to minimize the randomness needed to achieve a complete exploration, reduced to the stochasticity of the direction change, while maximizing their speed. No accept-reject scheme is required as correctness is ensured by the choice of the successive directions, which leads also to relieve the necessity of a fine tuning of the sampling parameters. Therefore they appear now as serious candidates for tackling this type of complex statistical inference problems and are under growing attention both in statistics and physics [10-14]. The implementation of a PDMC scheme can however be more involved than the ones of simpler MCMC schemes and requires a careful mathematical analysis of the posterior distribution to get access to its full efficiency. Investigating the smoothness and convexity properties of the posterior distribution, by giving information on how the posterior distribution behaves, could indeed lead to simplifications in the computations of the time of the direction changes, yielding important reduction in the computational complexity if a bound on the infinitesimal probability decrease can be obtained. However, today, the regularity and convexity properties of the posterior distributions are poorly known, although preliminary numerical investigations hint at a smooth behaviour. A better understanding of the properties conserved by the Hamiltonian flow of the physical forward evolution and how to exploit them algorithmically could lead to major progress, beyond the issue of a PDMC implementation, as, conversely, this theoretical analysis could shed a new light on the physical argument used in the BORG framework. Finally, the introduction of a PDMC scheme in the BORG framework would constitute their first real case HPC implementation and should promote their practical use in other scientific fields dealing with large-scale stochastic simulations, such as climate science and high-energy physics, as well as reveal potential future challenges and improvements.

References:
[1] J. Jasche and B. D. Wandelt. MNRAS, 432(2):894–913, 04 (2013).
[2] J. Jasche and G. Lavaux. A&A, 625, A64 (2019).
[3] LSST Science Collaboration (Abell, P. A., et al.) arXiv:0912.0201 [4] Y.-Z. Ma and D. Scott, Phys. Rev. D, 93, 083510 (2016).
[5] B. M. Schaefer, arXiv:1701.04469.
[6] C. P. Robert and G. Casella. Monte Carlo Statistical Methods. (2nd ed.). Springer (2004).
[7] S. Duane, A. D. Kennedy, B. J. Pendleton, D. Roweth, Phys. Letters B 195, 2 (1987).
[8] R. M. Neal, Bayesian Learning for Neural Networks. Springer-Verlag New-York (1996).
[9] M. Michel, S. C. Kapfer, and W. Krauth. J. Chem. Phys. 140:054116 (2014).
[10] M. Michel, A. Durmus and S. Sénécal. JCGS, 29(4), 689 (2020).
[11] A. Bouchard-Côté, S. J. Vollmer, and A. Doucet. Journal of the American Statistical Association,113 (522) 855 (2018).
[12] T. A. Kampmann, H. H. Boltz, and J. Kierfeld. Journal of Computational Physics, 281:864 (2015). [13] A. Durmus, A. Guillin, and P. Monmarché. To appear in Annales Institut Henri Poincaré (P/S), (2021).
[14] J. Bierkens, P. Fearnhead and G. Roberts. The Annals of Statistics, 47(3):1288–1320, (2019).

Directeur : Pr Yanick HEURTEAUX
tél. 04.73.40.50.64
E-mail : yanick.heurteaux@uca.fr

Secrétariat : Lucie NAUD
tél. 04.73.40.53.76
E-mail : edsf.drv@uca.fr