Summary, Conclusion and Outlook

> "Je crois qu'on obtiendra des résultats étonnants. C'est justement pour cela que je ne puis rien vous en dire; car si je les prévoyais, que leur resterait-il d'étonnant?"
> — Henri Poincaré (1900)

Summary

The main subject of this thesis is the process of data assimilation for the analysis of the cosmological large-scale structure. It aims at finding the best realizations of a physical model of structure formation in light of the data, while fully accounting for all uncertainties inherent to the inference problem. To this end, the BORG (Bayesian Origin Reconstruction from Galaxies) algorithm derives the initial conditions and produces physical large-scale structure reconstructions, by assimilating survey data into a cosmological model. It is an inference engine that allows the simultaneous analysis of the morphology and the formation history of the cosmic web, a subject introduced in this work that we refer to as chrono-cosmography. The present thesis contributed to the establishment of physical large-scale structure inference as a functional and effective tool for the analysis of real survey data.

We started by a review of the standard picture of LSS formation, discussing the gravitational self-evolution of the dark matter fluid and introducing cosmological perturbation theory (chapter 1). We then examined the accuracy of Lagrangian perturbation theory, a tool widely applied in LSS data analysis and also a key ingredient of the BORG algorithm (chapter 2). We characterized the approximation error in particle realizations produced by LPT at first or second order instead of fully non-linear gravity. In particular, we analyzed the one-, two- and three-point statistics of the density field, examined the displacement field and compared the volume of different cosmic web elements. In spite of visual similarities, we found that LSS realizations produced by LPT and by N-body simulations can drastically differ in some regimes, an effect of general interest for data analysis.

Since unique recovery of signals from data subject to observational effects (incomplete sky coverage, selection effects, biases and noise) is not possible, BORG uses a Bayesian approach to quantify uncertainties. We discussed the fundamental concepts, mathematical framework and computer implementation of Bayesian probability theory (chapter 3). Building upon these notions, we introduced physical large-scale structure inference with the BORG algorithm (chapter 4). We exposed the data model and described the numerical sampler, based on the Hamiltonian Monte Carlo technique. Our approach allows the exploration of the posterior distribution, which lives in an extremely high-dimensional parameter space (usually of the order of 10 million free parameters), in computationally reasonable times. As a result, it provides a sampled representation of the large-scale structure inferred from the data, in the form of four-dimensional cosmographic maps of the matter distribution. We applied the BORG algorithm to the Sloan Digital Sky Survey main galaxy sample data and presented our analysis (chapter 5). Our results constitute accurate three-dimensional reconstructions of the present density and velocity fields, but also of the initial conditions and of the formation history of the large-scale structures in the observed domain.

The full data-assimilation problem is deeply non-linear, implying that probability distributions for observed cosmic fields are far from a multivariate Gaussian. We discussed the challenges associated with a fully non-linear description of late-time structure formation. We proposed a fast method to improve the correspondence between density fields in approximate models and in full numerical simulations (chapter 6). The technique relies on remapping the one-point distribution of approximate fields using information extracted from simulations, and allows to extend the validity of LPT beyond shell-crossing, in the mildly non-linear regime. We introduced the concept of non-linear filtering of BORG samples (chapter 7). This procedure improves constrained realizations by augmenting them with physically plausible information at small scales. We checked the accuracy of the fast COLA scheme as a non-linear filter versus GADGET-2, and used it to produce a large ensemble of non-linear BORG-COLA samples for subsequent use.

We finally made use of our results for cosmic web analysis. With the VIDE toolkit, we produced and analyzed constrained catalogs of cosmic voids in the Sloan volume. In doing so, we showed that the inference of voids at the level of the dark matter field, deeper than with the galaxies, is achievable, and that suitable inference technology is capable of tapping a mine of information even in existing surveys. In particular, we found at least one order of magnitude more voids at all scales considered, between 5 and 40 Mpc/h. As a consequence, our method yields a drastic reduction of statistical uncertainty for the determination of void properties and carries a vast potential for their use as cosmological probes. We presented a probabilistic analysis of the dynamic cosmic web, dissected into voids, sheets, filaments, and clusters, on the basis of the tidal field (chapter 9). We examined the history and characterized the information content of our web-type maps. This study demonstrated that our inference framework allows self-consistent propagation of observational uncertainties to cosmic web analysis, and counts among the pioneering steps toward a data-supported connection between cosmology and information theory. Eventually, we introduced a new framework for optimal decision-making based on the web-type posterior probabilities and the strength of data constraints (chapter 10). We obtained efficient statistical summaries of our inference results and outlined more general applications to classification problems in the face of uncertainty.

Conclusion and Outlook

What does chrono-cosmography predict about the Universe?

The aim of physical large-scale structure inference is to provide a cosmographic description of a subvolume of the observable Universe, as well as a probabilistic characterization of uncertainties. In this fashion, we have access to a wealth of information on the present and past of this region. This material can be used in a variety of subsequent astrophysical and cosmological analyses, many of which can be already conducted based on the results obtained during this PhD project.

A first class of possible projects build upon the inference of the initial conditions, in which gravitational non-Gaussianity is largely suppressed.

Genesis and growth of the cosmic web. As we have shown, Bayesian large-scale structure inference paves the path toward a high-fidelity description of the complex web-like patterns in cosmic structure. A natural follow-up project is to exploit the richness of the information inferred by BORG, including quantities so far only accessible in simulations, to build and compare classifications of the cosmic web. These will use, for example: the Lagrangian displacement field (DIVA, Lavaux & Wandelt, 2010) or the stretchings and foldings of the dark matter phase-space sheet (ORIGAMI, Falck, Neyrinck & Szalay, 2012).

Primordial non-Gaussianity and inflation. The BORG algorithm has the potential to accurately characterize the statistics of primordial seeds. In particular, estimators of bispectra (Schmittfull, Baldauf & Seljak, 2015) or based on the phases of inferred fields (Obreschkow *et al.*, 2013; Wolstenhulme, Bonvin & Obreschkow, 2015) – for which no prior information is assumed – may detect signatures of primordial non-Gaussianity or constrain models of inflation. While it will be extremely challenging for LSS measurements to improve upon CMB constraints on inflation, the theoretically interesting threshold for many models involving deviations from Gaussian initial conditions has not yet been reached. Hence, it is crucial to develop new methods that will extract this information from the LSS.

Galaxies within the large-scale structure. Physical properties of galaxies (luminosity, color, spin, morphological type, etc.) are known to be correlated with their large-scale environment (see e.g. Lee & Lee, 2008; Park, Kim & Park, 2010; Eardley *et al.*, 2015). Cosmic web analyses enabled by BORG straightforwardly allow to test models of galaxy formation and evolution within their time-varying environment. The resulting information could then be used in future large-scale structure inference procedures, to obtain more refined information on the matter density field traced by these galaxies.

BORG provides a complete description of the gravitational dynamics of the volume of interest. A second class of projects is to use this information and see what it implies for other cosmological observables. In this regard, we use BORG as a forecast-generating machine, whose predictions can be tested with complementary observations in the actual sky.

Effects of the inhomogeneous large-scale structure on photons. Inferred information permits to produce various prediction templates for cross-correlations with other cosmological data sets. In particular, it is possible to predict the effects of inhomogeneities in the LSS on photon properties and geodesics, given galaxy observations:

deviations in redshift (in the radial direction) and weak gravitational lensing effects (in the angular directions). In a similar fashion, dynamic information such as velocity fields and evolution of the gravitational potential can be used to enhance the detectability of secondary effects expected in the cosmic microwave background, such as the kinetic Sunyaev-Zel'dovich effect, the integrated Sachs-Wolfe effect, and the non-linear Rees-Sciama effect.

Cosmological parameters, baryon acoustic oscillations and dynamic dark energy. The incorporation of a physical model in the likelihood provides a natural way to infer cosmological parameters from observations. The work presented in this thesis is also expected to provide an alternative way to reconstruct the baryon acoustic oscillation signal (Padmanabhan *et al.*, 2012) and to infer the equation of state of a possible dark energy component. This approach will yield a more precise picture of the expansion history of the Universe and help to understand the origin of cosmic acceleration.

How do we include more aspects in the data model?

Contrary to traditional approaches, which apply various cosmological tests to data separately and combine constraints in a suboptimal fashion, the approach presented in this thesis automatically and fully self-consistently performs a joint analysis of all aspects. It models their interdependence and accounts for the ways in which different observables can mutually enhance one another. The joint analysis of all phenomena can be used to perform consistency tests of the standard cosmological model and has the potential to rule out some of its possible extensions.

However, many relevant aspects are still absent of the current BORG data model: a fully non-linear treatment of gravitational structure formation, redshift-space distortions, lightcone effects, photometric redshifts uncertainty, density-dependent selection effects, scale-dependent and stochastic galaxy bias or predictions of non-standard cosmologies. The joint analysis of other probes of the LSS (CMB lensing, weak lensing shear maps, etc.) should also be addressed in our framework, via a joint likelihood or sequential data assimilation. The inclusion of these aspects in the LSS data model involves conceptual, but also technical challenges. Bayesian large-scale structure inference is highly computationally expensive, to the degree that it touches the border of what is currently possible.

In the author's opinion, future progress will not only depend on adequate approximations, but also on the development of new methodological ways to implement sampling. In comparison to the state-of-the-art Hamiltonian Monte Carlo algorithm, efficient, advanced non-linear data assimilation techniques will have to allow a much cheaper statistical inference (presumably by several orders of magnitude), which will open the way for the inclusion of more physical effects.

What can ultimately be learned from the large-scale structure?

How deterministic is the formation of structure in the Universe? In other words, at what scale does one-to-one mapping from initial to final conditions (valid at large, linear and weakly non-linear scales) break down? Since state-of-the-art simulations are still very far from resolving all relevant physical processes, the issue of the scale at which structure formation is non-deterministic is not yet considered crucial for numerical modeling. However, a theoretical understanding of this question would be of central interest in the context of an Effective Field Theory of the LSS (Baumann *et al.*, 2012; Carrasco, Hertzberg & Senatore, 2012; Senatore & Zaldarriaga, 2015). A quantification of the information content of primordial patches that collapse to form structures (such as the Milky Way) characterizes the LSS in a fundamental way and contains a wealth of information on the properties of matter at the highest energies, far beyond the reach of particle colliders.

High amount of primordial information contained in small-scale structures would for example disfavor warm dark matter or any mechanism suppressing small-scale density fluctuations, and could even require further initial degrees of freedom such as isocurvature perturbations. Issues related to the information content of primordial patches have been recently speculatively examined by Neyrinck (2015), who proposed, as a thought experiment, a test about the scale at which structure formation is deterministic. Unfortunately, simulations do not provide much insight into the questions of where the information content is, and how to optimally extract it from the data.

Building upon the inference of initial conditions from which the LSS originates, the first steps toward a practical implementation of a scale-dependent test of determinism in structure formation can be taken. This task will involve the careful definition of a measure of complexity in Lagrangian patches inferred by BORG and of a means to compare initial and final information. It will also require careful analysis of information propagation via Lagrangian transport within a fully probabilistic approach (see sections 5.3.3 and 9.5.1; figures 5.8, 9.3 and 9.6 for a preparatory discussion). Further investigation will have to consider information sinks such as baryonic processes and black hole formation, and information sources that broadcast non-primordial randomness at large scales, such as supernovae, active galactic nuclei and unstable astrophysical phenomena. The link between astrophysical, thermodynamic entropy, as well as statistical, information-theoretic entropy will also have to be clarified.

In the last few years, ESA's Planck mission confirmed our picture of the evolution of the homogeneous Universe to spectacular accuracy and provided the highest precision probe to date of the physical origin of cosmic structure. Challenges for accurate cosmology now arise from studying the inhomogeneous cosmic structure. This research will provide an exceptionally detailed characterization of the cosmic web underlying the observed galaxy distribution, extract information about the nature of dark energy, and furnish unprecedentedly accurate information on the initial conditions from which structure appeared in the Universe. Progress will not only depend on our ability to handle ever larger data sets: crucial to the longer-term aims is developing efficient tools for assimilating data into the forecasts of a physical model and quantifying the information content uniquely encoded in the primordial large-scale structure. Only through such a quantitative statistical approach can we expand our understanding of the dynamic Universe and make significant progress on the age-old puzzles of cosmic beginning and ultimate fate of the Universe. I am confident that the methods and results described in this thesis, counting among the first steps towards precision chrono-cosmography, will contribute to this endeavor.