# Astrostatistics within PCMI

## http://pcmi.cnrs.fr

**SOC:** Jérôme Pety (IRAM & LERMA, pety@iram.fr), François Boulanger (LERMA), Pierre Gratier (LAB), Franck Le Petit (LERMA), Marc-Antoine Miville-Deschenes (CEA).

**Participants:** Erwan Allys (ENS), Olivier Berné (IRAP), Jérôme Bobin (CEA), Sylvain Bontemps (LAB), Sandrine Bottinelli (IRAP), Emeric Bron (LERMA), Cédric Colling (CEA), Frédéric Galliano (CEA), Maryvonne Gerin (LERMA), Isabelle Grenier (CEA), Cécile Gry (LAM), Antoine Gusdorf (ENS), Marc Huertas (LERMA), Annie Hughes (IRAP), Théo Joubaud (CEA), Pouria Khalaj (IPAG), Bilal Ladjelate (IRAM), Rosine Lallement (GEPI), Antigone Lambert-Huyghe (CEA), David Languignon (LERMA), Thibaut Le Bertre (LERMA), François Levrier (ENS), Douglas Marshall (CEA), Estelle Moraux (IPAG), Frédérique Motte (IPAG), Alejandra Recio-Blanco (OCA), Arabindo Roy (LAB), Jeremy Scholtys (CEA), Charlotte Vastel (IRAP), Sixin Zhang (ENS)

28 August 2018
Version 1.0

**Abstract**

Datasets produced by telescopes, numerical models and simulations become extremely large. Data mining in these large volumes of data and comparison between observations and models require the development of new methods, often based on statistical approaches. This in turn requires building bridges between the PCMI community and the applied mathematics / statisticians communities. One outcome of this workshop is that several independent works used similar underlying ideas, showing the need for more interactions in order to compare methodologies in details. The PCMI community should organize a serie of workshops dedicated to these subjects in the next 5 years, and partnerships between the PCMI community and statisticians will be supported.

# 1 Summary

In recent years, the PCMI community has contributed to different major achievements. Thanks to Planck, we now have an all-sky survey of the dust polarization. The successive date releases of Gaia offers the possibility to constrain the ISM 3D structure over distances of a few kpc. The advent of wide bandwidth receiver in radio-astronomy turns almost every observation into a line survey. This leads to the acquisition of blind line surveys over tens of objects of the same category (PDRs, dense cores, young stellar object, ...) as well as the first systematic observations of large fraction of local GMCs over the 3mm atmospheric window. The gain in sensitivity of radio-interferometers now enables to systematically detect the GMCs in the normal ("main-sequence") nearby galaxies. This opens the possibility to study environmental effects (e.g., metallicity, galaxy dynamic, ...) on the properties of the ISM.

In all these examples, observatories routinely deliver high sensitivity observations at high angular resolution over wide fields and wide bandwidths. At the same time, simulation and modeling capabilities yield datasets comparable in size and complexity. These datasets contain much more information than traditional analyses can easily extract. The PCMI community thus enters the big data era that will enable it to restate his main science questions (structure of the ISM and its link with turbulence, star and planet formation, chemical complexity and heritage, ...) in statistical terms. A first step to reach this goal was to organize a two-days workshop that mostly gathers members of the PCMI community. The goals were to make a census of the statistical methods used in the community, to identify statistical methods that could be useful, and to discuss how to build bridges between statisticians and members of the PCMI community who (wish to) use statistical methods. Indeed, while the PCMI community knows the cutting-edge science questions, statisticians have the expertise to devise the best statistical method to address specific questions. PCMI will encourage the organization of workshops that gather both communities in order to prepare the best application to the funding calls that CNRS puts forward.

ISM objects (GMCs, filaments, cores, ...) are usually extracted through the analysis of the spatial distribution of the emission of one or a few tracers. These objects are then classified in pre or proto-stellar cores, young stellar objects, jets, proto-planetary disks, etc... based on their continuum spectral energy distribution or on their molecular content. The increase of the data quantity opens the possibility to use automatic methods such as clustering. Two challenges have been identified to ensure clear success. First, the definition of ISM objects is often elusive, while their automatic extraction requires to explicitly define their exact properties. Without this first step, different methods will provide different set of objects under the same name just because they are tailored to different (often implicit) definitions. This is difficult because the ISM is a continuous medium while observations are often noise-limited. For example, the CO(1-0) line emission is traditionally used to identify GMCs. However, the increase of sensitivity showed that this emission is more extended than previously thought and often connect different GMCs, impairing usual methods to define GMCs. This raises the question of what defines a GMC in the underlying physics, and thus whether it is still possible to classify the emission of 12CO(1-0) in meaningful objects. For instance, is there a particular scale at which the physics change (Are GMCs virialized objects compared to the unbound surrounding atomic diffuse gas?) and how can it be measured? This example also illustrates the second challenge: The fact that observations deliver an instantaneous snapshot of a continuous, opened

medium evolving with time. While it is probably possible to classify this snapshot, it remains to show that the found classes are meaningful from the science viewpoint.

In statistics, the characterization of the properties of objects, and thus the ability to classify them, is divided into two others categories. On one hand, unsupervised classification takes a data set and try to cluster it depending on some of its generic statistical properties (for instance, the fact that the probably distribution function of one of its physical variable has several peaks indicating that some values are privileged). This enables to discover unexpected results but this clustering somehow becomes arbitrary for transition objects. On the other hand, the fit of parametrized models enables accurate predictions but may become biased if the model does not correctly represent the data.

With the advent of large observational datasets, automating the extraction of parameters (for instance, column densities) becomes an important issue. The current consensus is that the Bayesian approach (potentially hierarchical Bayesian approach) is the best one because it correctly constrain the uncertainties of correlated quantities. However, this method is computationally intensive. An alternative is to train a neuronal network on modeled data before applying it on true data to extract the required parameters. The advantage is to have a fast estimation. The challenge is to get reliable estimations of the uncertainties and of the biases. For instance, when a neuronal network is used on a dataset for which it has not been trained, it may silently bias the results.

More generally, machine learning algorithms require to be trained on large datasets. This suggests that a large effort of analysis is still required on a fair amount of data using classical approaches in order to reach the point where machine learning techniques get profitable. Noise and systematics must also be adequately sampled in order that machine learning techniques works correctly. An alternative to the use of already analyzed data sets is to produce massive grid of models that will cover the space of parameters to be searched. It is thus important to start producing realistic grid of models for the main objects we study in the PCMI community: PDRs, shocks, dense cores, proto-planetary disks, etc... This is all the more difficult that not only the physics (including radiative transfer) and chemistry but also the source structure must be correctly described and varied. The reward of such an effort is that it will then become possible to gain physical and chemical insights by analyzing the grids of models themselves: For instance, what are the best tracers of a given parameter such as ionization, density, or even evolutionary state (chemical clocks) in a more distant future? Why these are the best tracers?

The study of the turbulent nature of the ISM and its link to the star formation efficiency has the additional specificity that turbulence is a inherently random process where gravity, cooling/heating, magnetic field produce structures at different scales. We thus need to deal with a non-linear, multi-scale random process. Progresses require an in-depth extensive comparison of observations with 3D simulations. This effort is challenging because the observations deliver convoluted information (indirect tracer of the bulk of the mass, integrated over the line of sight, intensity instead of density or temperature, one velocity component instead of a vector, polarization fraction instead of magnetic field, ...), which are difficult to reproduce exactly in 3D simulations of the ISM. Two main directions are currently explored to bridge this gap. First, powerful statistical diagnostics can answer physical questions important for the simulators: For instance, the amount of momentum that is injected in solenoidal vs compressive mode of turbulence. Indeed, energy injected in vortices will be more efficiency to counter-act gravity. Second, other studies try to ask whether it is possible to describe the ISM random process with only 20 parameters (as in cosmology), and how applicable this description would be (to what mass/volume of the ISM)? In other words, the comparison with simulations should be done in the space (e.g., decomposition on Gaussian, Fourier transform, wavelet transform, etc...) where the statistical properties can be described most efficiently (in a sparse way). The fact that structures (filaments, dense cores, etc...) exist in the medium also indicates that the non-Gaussianity of the random processes (e.g., dense structures filling a small amount of the volume) play a key role and that this statistical description must go to higher order that the usual correlation analysis. One open question is how this description could be linked to the one based on extraction of objects (filaments, cores, ...), and thus the star formation.

One difficulty of the comparison between simulations and observations is that parameters are well controlled in simulations while observations are often made of mixtures of conditions, all the more that surveys start to cover large regions. This problem is even more acute in nearby galaxy studies where one resolution element is of the order of 40 pc.

In summary, with this new era of Big Data, it is clear that the PCMI community must acquire the expertise on statistical methods to exploit large datasets. Several teams have already begun to explore these methods for various topics. First feedbacks show that the learning time to find the right method for a given problem, to acquire expertise on them, and to obtain reliable results is long. To be efficient, collaborations with statisticians must be developed. Several main actions are required to progress significantly. First, the community needs to precise the definition of the ISM objects (filaments, GMCs, ...). Second, the community needs to produce high quality grid of models. Third, the community is encouraged to organize workshops with statisticians in order to identify the best statistical methods that will enable us to answer our science questions.

# 2 Program

**May 31st 2018** 6 hours of discussion in 3 sessions.

**09h00-10h00** Welcome coffee.

**10h00-10h25** J.Bobin, Component separation on Planck data.

**10h25-10h50** S.Zhang, Statistical model of Non-Gaussian Process with Wavelet Scattering Moments.

**10h50-12h00** Statistical description of the ISM structure. Session 1.

- Jérôme Pety, The dawn of the big data era in radioastronomy.
- Rosine Lallement, ISM 3D.
- Antoine Gusdorf, The interstellar and stellar content of supernova remnants.
- Erwan Allys, Comparing MHD simulations and data.

**12h00-13h30** Lunch.

**13h30-14h25** Statistical description of the ISM structure. Session 2.

- Sylvain Bontemps, The GENESIS project.
- Arabindo Roy, Probing ISM Physics with Delta-Variance.
- Jan Orkisz, Turbulence modes and star formation efficiency.

**14h25-15h30** Modeling data. Session 1.

- Franck Le Petit, ISMDB: Comparison Models-Observations.
- Emeric Bron, Exploiting massive grids of models.
- Frédéric Galliano, Hierarchical Bayesian Inference for dust Emission.

**15h30-16h00** Pause.

**16h00-18h00** Summary and next steps. Session 1.

**June 1st 2018** 4.5 hours of discussion in 3 sessions, plus 1 session to recapitulate and to deal with delays.

**09h00-10h00** Modeling data. Session 2.

- Charlotte Vastel, Automating the analysis of line surveys.
- Sandrine Bottinelli, Analysis of line surveys and grids of models with CASSIS.

**10h00-10h25** M.Huertas, Applications of classification techniques to galaxies.

**10h25-10h50** A.Recio-Blanco, Decision trees in Galactic archeology.

**10h50-11h20** Pause.

**11h20-13h00** Clustering observations.

- Annie Hughes, Characterising Cold Gas in Nearby Galaxies.
- Olivier Berné, Factorial models for unmixing and fusion of hyper spectral data.
- Antoine Marchal, Extracting the multi-phase structure of the ISM from hyper-spectral data.
- Jean-François Robitaille, Multi-scale structure segmentation.
- Jan Orkisz, Filamentary network in Orion B.
- Maryvonne Gerin, The ORION-B project: An example of a multiple line imaging survey.

**13h00-14h00** Lunch.

**14h00-15h30** Summary and next steps. Session 2.

**15h30-16h00** Pause.

**16h00-17h00** Managing delays.