# Exchange Visit Grant

## Scientific Report

## Proposal Title : **A Bayesian perspective to estimate parameters in galaxy modeling using hierarchical models**

## Application Reference: **4161**

## 1 Purpose of the visit

The visit was the first one in a series aimed at tackling the problem of modelling the Galaxy from Gaia data. The main objective of the collaboration that starts with this visit is to develop the statistical tools necessary to model the Galactic structure and history from the Gaia data. We aim to do this by creating a hierarchical model that encodes all the prior information that we have about the distribution of Galactic sources in the various parameters, from the top level, down to the direct Gaia observables. In particular, this includes (i) inference of the geometric parameters: shape, size and orientation of the bulge, disks, and halo; (ii) derivation of an intragalactic extinction distribution; and inference of the general laws that govern the star formation: the initial mass function (IMF) and the star formation rate (SFR). In this visit, we start by looking at the possibility to apply hierarchical bayesian models to recover the parameters of the Besanon Galaxy model as simulated by GOG.

The group involved in the development of the modelling tool is composed by: Coryn Bailer-Jones (MPIA), Luis Sarro (UNED), Francesca Figueras (UB), Xavier Lury (UB) and Annie Robin (UFC)

# 2 Description of the progress made during the visit

Obviously, the project objectives are to ambitious to be solved in a three months visit. Therefore, we concentrated on solving a simplified model that serves as a basis for future developments. In order to tackle the aforementioned objectives we proposed four action lines to work with:

- *Selection of effective statistical tools to estimate structural parameters of the galaxy from GAIA observations.*

  I initially proposed a first, simple hierarchical model to link stellar masses, ages and the spatial density function with the initial mass function and stellar formation rate. The model is produced in the form of a directed acyclic graph that encodes the conditional dependencies amongst the variables. This first model was discussed and several modifications applied to it: in particular, it is necessary to include a model of the uncertainties in stellar masses and ages caused by different effects: random (propagated from the Gaia observables to the infered masses and ages) and systematic (specially, the particular choice of the evolutionary tracks and stellar atmosphere libraries). This uncertainties could be included either by using a probabilistic analytic model convolved with the true masses obtained from the Besanon Galaxy model, or by using the posterior probabilities derived by the *Aeneas algorithm* proposed by the DPAC-CU8 team [LBJS+12].

  A second hierarchical model was proposed to tackle the comments/corrections above. At the end of the visit, I have implemented a hierarchical Bayesian model to perform the estimation of the IMF and SFR parameters with both approaches defining two different likelihoods, either with a specific probabilistic model or the probabilities calculated with Aeneas.

- *Parameter estimation of different galactic structures (outer bulge, disks and galactic halo).*

  As stated in the original application, we planned to test the feasibility of this approach by trying to recover the input ingredients of a GOG simulation of a particular realisation of the Besanon Galaxy model. The Barcelona group lead by Francesca Figueras provided various datasets for different Galaxy regions, albeit in different photometric systems than measured by GAIA. During the *Access Catalogue Kick Off Meeting* held in Barcelona (DPAC-CU9) we agree with Annie Robin and Francesca Figueras to focus the initial studies into the thin disk compoenent of the Galaxy and later extend the model to other galactic structures. As a result of the discussions, I modified the spatial density component of the hierarchical model to better describe the simulations.

- *Different measurements of the influence of the extinction model and impact hyperpriors impact in the Bayesian inferences.*

  Given that the current model focuses in the higher levels of the hierarchy, we did not need to include extinction in our considerations. This is nevertheless foreseen for future evolutions of the model

- *Computational tools to sample from the posterior distributions.*

  We used three different approaches/algorithms to obtain an independent and identically distributed sample from the posterior distribution of the parameters. There were three main reasons for using these algorithms: (i) they admit the structure of hierarchical Bayesian modeling, (ii) they are based on different theoretical approaches on sampling the posterior, and (iii), these algorithms can be parallelised for scalability to the Gaia dataset size.

  The three approaches were the `mcmcee` module [FMHLG12] (a Python implementation of the Affine Invariant Markov chain Monte Carlo [GW10] sampler); the `Stan` package [Sta13] which is a C++ implementation of the No-U-Turn sampler [HGss]; and the `Multinest` algorithm [FHB08] which is a variant of the Nested Sampling algorithm [Ski04] that incorporates a splitting criterion in the parameter space based on minimal ellipsoidal boundaries.

# 3 Description of the main results

1. A hierarchical Bayesian model which incorporates likelihoods of measurements potentially affected by different kinds of effects (random and systematic errors): (i) true measurements, mainly simulations from Besanon Galaxy

| wMean | wSD | bMean | Mode | Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
|---|---|---|---|---|---|---|---|---|---|
| 1.313 | 0.045 | 1.313 | 1.321 | 1.087 | 1.270 | 1.316 | 1.314 | 1.357 | 1.543 |
| 2.211 | 0.091 | 2.211 | 2.216 | 1.822 | 2.125 | 2.216 | 2.217 | 2.303 | 2.712 |

Table 1: Summary statistics for the Rstan sampler, with wMean, wSD the weighted mean and standard deviation. The bMean is the batched mean and Min., 1st Qu., 3rd Qu., Max. are minimum, maximum, first and third quantile respectively.

model; (ii) measurements derived from a convolution of a true measurement and a noise model; (iii) measurements based on the posterior probability obtained by Aeneas algorithm.

2. R code for drawing samples from the posterior distribution of the parameters using Nested sampling, Ellipsoidal Sampling or the Multinest algorithm. This library was made in collaboration with René Andrae (MPIA).

3. A full R implementation of the simplified hierarchical model for parameter estimation. Several test runs were carried out and several statistics were computed to summarize the posterior distributions. Table 1 shows the summary statistics for the best reference run (obtained with *rstan*), and Fig. 1 shows a converged sample for the same sampler, and the $\alpha$ parameter in the IMF defined by Kroupa 2008.

4. Bayesian sensitivity study to analyse the variations in the posterior samples due to different hyper-prior distributions. In a first approximation, we have tried hyper-priors for the IMF parameters based on the jeffreys, uniform and normal distributions. These do not show significant differences in the resulting posteriors, but the grid of hyperparameters was too coarse to draw conclusions from this.

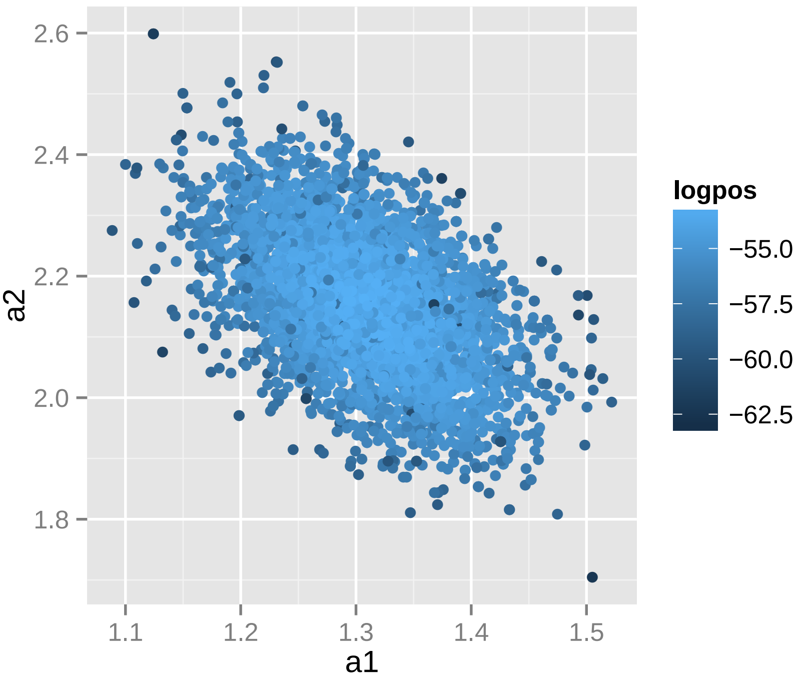

Figure 1: Converged sample for the *rstan* algorithm

## 4 Future collaboration with host institution (if applicable)

## 5 Projected publications/articles resulting or to result from the grant.

- We are currently working on a paper for submission to Annals of Applied Statistics entitled (draft version) *A hierarchical model to unveil the Milky Way history* I. *The Besanon Galaxy model, a proof of concept.* We expect to submit the paper before end of 2013.
- We also plan to publish the R package with our implementation of the Multinest algorithm including Nested Sampling and Ellipsoidal sampling. The release will be accompanied by the submission of a manuscript to the R Journal (peer-reviewed).

## 6 Other comments

I would like to thank to Luis Sarro, Francesca Figueras, Coryn Bailer-Jones and Annie Robin for their enormous patience, good ideas and hard work on this project.

## References

[FHB08] F Feroz, M P Hobson, and M Bridges. MultiNest: an efficient and robust Bayesian inference tool for cosmology and particle physics. *arXiv.org*, (4):1601–1614, September 2008.

[FMHLG12] Daniel Foreman-Mackey, David W Hogg, Dustin Lang, and Jonathan Goodman. emcee: The MCMC Hammer. *arXiv.org*, page 3665, February 2012.

[GW10] Jonathan Goodman and Jonathan Wear. Ensemble samplers with affine invariance. *Communications in Applied Mathematics and Computational Science*, 5(1):65–80, January 2010.

[HGss] Matthew D. Hoffman and Andrew Gelman. The no-U-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, In press.

[LBJS$^+$12] C. Liu, C. A. L. Bailer-Jones, R. Sordo, A. Vallenari, R. Borrachero, X. Luri, and P. Sartoretti. The expected performance of stellar parametrization with gaia spectrophotometry. *Monthly Notices of the Royal Astronomical Society*, 426(3):2463–2482, 2012.

[Ski04] John Skilling. Nested Sampling. In *BAYESIAN INFERENCE AND MAXIMUM ENTROPY METHODS IN SCIENCE AND ENGINEERING: 24th International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering. AIP Conference Proceedings*, pages 395–405. Killaha East, Kenmare, Kerry, Ireland, November 2004.

[Sta13] Stan Development Team. Stan: A c++ library for probability and sampling, version 1.3, 2013.