CASBI: Chemical Abbundance Simulation Based Inference

Giuseppe Viterbo

June 11, 2024

Contents

Chapter 1

Abstract

Galaxies evolve trough merging events and destroy lower-mass systems over their lifetimes. The contribution that those lower-mass system brings to the modern picture has been freezes in stellar halos by the long orbital timescales, making the relicts of this objects retain part of their initial progenitor orbit. But dynamical information is not enough to disentangle these components, and the complementary chemical information helps to characterized these building blocks. In fact, merging events tend to quench the star formation rate of these objects, making the chemical abundance plane (iron abundance against α element abundance) a distinct imprint that retains information on the conditions of formation of their stars, like the total mass and the age of the system until the merging event. These theoretical background allows us to attempt to decompose the stellar halo into its components, unraveling the merging history. In the modern era of large N-body galaxy simulations, we recast this problem into an SBI pipeline to recover the properties of this building block, (e.g. total stellar mass, infall time, ...) using the chemical abundance plane as observables. We therefore present CASBI (Chemical Abundance Simulation Based Inferece), a python package to recover the posterior probability of properties of building blocks of Milky Way like galaxy's halo. Moreover, CASBI incorporate conditional neural network architectures as generator to obtain observables from parameters, smoothly interpolating on region of the parameters space that weren't fully covered during the N-body simulations.

Chapter 2

Previous work

2.1 The reconstruction of the Assembly history of the Milky Way

Inferring the assembly history of the Milky Way is a challenging task, even in the era of the astrometric Gaia mission and its 6 dimensional phase space data, and the complementary chemical information obtained from the wide-field spectroscopic programs such as the GALAH survey (De Silva et al. 2015), the H3 survey (Conroy et al. 2019b), APOGEE (Majewski et al. 2017), RAVE (Steinmetz et al. 2006), SEGUE (Yanny et al. 2009), and LAMOST (Cui et al. 2012). The dynamical times of the accreted objects are far longer than the age of the host galaxy, allowing the phase space to retain part of the information on the original orbit parameters. On the other hand, the chemical space is dependent on the star formation history, in particular type II SNe produce α-elements and iron with a almost constant ratio, while type Ia SNe produce more efficiently iron. Another factor that governs the chemical space is the total mass of the galaxy, since the more massive galaxies are more capable to resist the expulsion of metals due to feedback mechanism. The crossmatch between Gaia and spectroscopic data allowed for the discovery of the "Gaia-Sausage-Enceladus" (GSE) Belokurov et al. 2018; Helmi et al. 2018, a massive accretion event whose remnant now dominates the observation of the inner stellar halo of our Galaxy. Robustly identify distinct structure is challenging, and disentangle the components in fully phase mix situations is nearly impossible. In order to characterize the assembly history [1] propose to use the "CARDs", the chemical abundance ratio distributions of the stars, obtained from the FIRE-2 zoom-in cosmological simulations of MW-mass galaxies (Wetzel et al. 2016). Although similar to CASBI in how to leverage N-body simulations, this method do not recovers posteriors for the parameters of the accreted objects, but rather considers the host halo as a linear combinations of templates CARDs, treating each coefficient as the fraction of mass contribution from the accretion event, and tries to recover those coefficients by maximize a loss that compares the observed CARDs with the combination of the templates. This method and CASBI share two more similarities: 1. Both of these methods are meant to be used on simulations, and the integrations of observational data is not yet implemented, even though theoretically possible. 2. Both rely on the assumption that the chemical space of accreted and isolated dwarf galaxies is very similar, due to ram pressure quenching the star formation history of the accreted object and hence 'freezing' these abundance ratios at the infall time. This assumption is investigated further in section 6 of [1]. Another approach is presented in [2], which takes advantage of the mass-metallicity relation to decompose the metallicity distribution functions (MDF) of the host galaxy as a mixture

of accreted halo's MDF, assumed gaussian for each of these building blocks. The objective of this work is to obtain a posterior distribution for the number of galaxies in each luminosity bin, which can be considered a proxy for the star mass. In CASBI we adopt the same superimposition philosophy of the components contribution, but we do not assume neither a prefix or an analytical form for the joint distribution of the chemical abundance, relaxing these assumption and relying only on the available samples from the N-body simulations.

Chapter 3

CASBI

3.1 SBI

CASBI is a Simulation Based Inference (SBI) package to recover the properties of building blocks of Milky Way like galaxy's halo from observations of the chemical abundance plane. The SBI framework has existed along side the more traditional likelihood based inference methods for quite some years already, and has its root in the Approximate Bayes Computation (Rubin 1984), and it has been used in a variety of fields, from cosmology to particle physics. The main difference between SBI and likelihood based methods, like MCMC, is that the former do not require the likelihood function to be known, but rather rely on a simulator to generate synthetic data $\mathbf{x}$ once the input parameters $\boldsymbol{\theta}$ are passed to it, and the inference pipeline is trained based on data-parameters pairs $(\mathbf{x}, \boldsymbol{\theta})$.

Recent advance of this technique was made possible by the use of machine learning models to emulate conditional probability distributions, a technique know as Neural Density Estimation (NDE) (Papamakarios 2019). The NDE is achieved by training a Normalizing Flow architecture, a generative model that allows to obtain samples from a complex distribution $p(x)$ by constructing a series of **bijiective** transformations $f^i_{\phi_i}$ that map x to a latent space z that is distributed as a simple distribution, like a Gaussian. Accordingly to [4], in the end the models learns the following schema parameters ϕ_i:

$$p(x) \sim x \overset{f^1_{\phi_1}}{\longleftrightarrow} h_1 \overset{f^2_{\phi_2}}{\longleftrightarrow} h_2 \dots \overset{f^K_{\phi_K}}{\longleftrightarrow} z \sim \mathcal{N}(z; 0, \mathcal{I}), \tag{3.1}$$

by maximizing the negative log likelihood as loss function and using the change of variable formula as follows:

$$\begin{aligned} \log p(x) &= \log p(z) + \log \left| \det \left(\frac{\partial z}{\partial x} \right) \right| \\ &= \log p(z) + \sum_{i=1}^{K} \log \left| \det \left(\frac{\partial h_i}{\partial h_{i-1}} \right) \right|, \end{aligned} \tag{3.2}$$

where the last term is sum of the log determinant of the Jacobian of the transformations $f^i_{\phi_i}$. Once the model is trained it is easy to sample from the distribution $p(x)$ by sampling from the latent space z and applying the inverse transformations $(f^1_{\phi_1})^{-1} \circ \dots \circ (f^K_{\phi_K})^{-1}$. In order to keep the sum of log determinant tractable, the use of *Coupling layers* allows to split the input x along its dimensions and apply a transformation only to a subset of the dimensions, using the other as input for the transformation and keeping it fixed. The subset is than changed at each layers, allowing to have a

permutation invariant transformation. The transformations $f^i_{\phi_i}$ are usually very simple invertible transformation like a translation and a scaling, or splines functions.

Following the discussion presented in [3], in Bayesian analysis we have the choice to approximate either the Posterior, the Likelihood or the Likelihood ratio, and this choice depend mostly on the problem that one wants to solve. In our case, due to the complexity of the Likelihood distribution of the chemical abundance space, we choose to approximate the Posterior distributions, and so we adopted the Neural Posterior Estimate that can be trained using the negative loglikelihood as loss function:

$$\begin{aligned} \mathcal{L}_{NPE}(\boldsymbol{\theta}) &= -\mathbb{E}_{\mathcal{D}_{train}} \log \hat{\mathcal{P}}(\theta_i | x_i) \\ &= -\mathbb{E}_{\mathcal{D}_{train}} \log \left(\frac{p(\theta)}{\tilde{p}(\theta)} q_\omega(\theta_i, x_i) \right), \end{aligned} \tag{3.3}$$

where our Posterior distribution $\hat{\mathcal{P}}(\theta_i|x_i)$ is approximated by the product of the ratio of the prior $p(\theta)$ and proposal distribution $\tilde{p}(\theta)$ and the neural conditional distribution $q_\omega(\theta_i, x_i)$, parametrized by the parameters ω.

Many excellent framework for handling SBI analysis are already available, and CASBI is build on top of the **Ltu-ILI** python package [3]. In particular, CASBI analysis were performed relying on the **sbi** backend [5] to train a *Neural Posterior Estimate* of the parameters' posteriors. The preprocessing of the data is described in Section 3.2, the details of the training of the NPE is described in Section 3.3, and the final package is described in Section 3.4.

3.2 NIHAO

As described in the presentation paper [6], the

3.3 Two step Inference

,

3.4 Python package

Chapter 4

Analysis

4.1 NIHAO UHD

Chapter 5

Conclusion

5.1 Future work

5.1.1 GRUMPY

5.1.2 Number halos has a free parameter: hierarchical sbi

5.1.3 True test: GAIA

Bibliography

[1] Emily C. Cunningham, Robyn E. Sanderson, Kathryn V. Johnston, Nondh Panithanpaisal, Melissa K. Ness, Andrew Wetzel, Sarah R. Loebman, Ivanna Escala, Danny Horta, and Claude-André Faucher-Giguère. Reading the CARDs: The Imprint of Accretion History in the Chemical Abundances of the Milky Way's Stellar Halo. *The Astrophysical Journal*, 934(2):172, August 2022.

[2] Alis J Deason, Sergey E Koposov, Azadeh Fattahi, and Robert J J Grand. Unravelling the mass spectrum of destroyed dwarf galaxies with the metallicity distribution function. *Monthly Notices of the Royal Astronomical Society*, 520(4):6091–6103, February 2023.

[3] Matthew Ho, Deaglan J. Bartlett, Nicolas Chartier, Carolina Cuesta-Lazaro, Simon Ding, Axel Lapel, Pablo Lemos, Christopher C. Lovell, T. Lucas Makinen, Chirag Modi, Viraj Pandya, Shivam Pandey, Lucia A. Perez, Benjamin Wandelt, and Greg L. Bryan. LtU-ILI: An All-in-One Framework for Implicit Inference in Astrophysics and Cosmology, February 2024.

[4] Diederik P. Kingma and Prafulla Dhariwal. Glow: Generative Flow with Invertible 1x1 Convolutions, July 2018.

[5] Alvaro Tejero-Cantero, Jan Boelts, Michael Deistler, Jan-Matthis Lueckmann, Conor Durkan, Pedro J. Gonçalves, David S. Greenberg, and Jakob H. Macke. Sbi: A toolkit for simulation-based inference. *Journal of Open Source Software*, 5(52):2505, August 2020.

[6] Liang Wang, Aaron A. Dutton, Gregory S. Stinson, Andrea V. Macciò, Camilla Penzo, Xi Kang, Ben W. Keller, and James Wadsley. NIHAO project I: Reproducing the inefficiency of galaxy formation across cosmic time with a large sample of cosmological hydrodynamical simulations. *Monthly Notices of the Royal Astronomical Society*, 454(1):83–94, November 2015.