MNRAS **000**, 1–4 (2022) Preprint 21 September 2022 Compiled using MNRAS LATEX style file v3.0

Bayesian Analysis of Gas-free Galaxies in Simulations I: Reviving Stochasticity

Alexander Rawlings,[1]★ Others,[1]
[1] *Department of Physics, Gustaf Hällströmin katu 2, FI-00014, University of Helsinki, Finland*

Accepted XXX. Received YYY; in original form ZZZ

ABSTRACT
We present the Bayesian Analysis of Gas-free Galaxies in Simulations (BAGGinS) project, show that stochasticity is ubiquitous with collisionless merger sims.

Key words: keyword1 – keyword2 – keyword3

1 INTRODUCTION

Something cool

Introduce core sersic profile here.

$$\Sigma(R) = \Sigma' \left[1 + \left(\frac{r_\mathrm{b}}{R}\right)^\alpha\right]^{\gamma/\alpha} \exp\left[-b_n \left(\frac{R^\alpha + r_\mathrm{b}^\alpha}{R_\mathrm{e}^\alpha}\right)^{1/(\alpha n)}\right], \tag{1}$$

where

$$\Sigma' = \Sigma_\mathrm{b} 2^{-\gamma/\alpha} \exp\left[b_n \left(2^{1/\alpha} r_\mathrm{b}/R_\mathrm{e}\right)\right] \tag{2}$$

2 NUMERICAL SIMULATIONS

Describe the numerical set up, code, etc

3 STATISTICAL ANALYSIS

We perform full Bayesian Hierarchical Modelling for the projected mass density profiles of our simulated galaxies. The modelling is performed using the statistical software Stan[1], which itself implements Hamiltonian Monte Carlo (HMC).

3.1 Hamiltonian Monte Carlo

An excellent overview of HMC may be found in Betancourt (2017), with a brief description provided here. Mention NUTS here somewhere?

In numerically sampling an arbitrary target distribution, we are required to sample from the *typical set*, or those regions of the target distribution which provide the dominant contribution to an expectation value. Simplistically, this may be achieved using any Markov generating process, such as the widely-used Metropolis-Hastings (MH) algorithm, albeit at the expense of inefficient and biased sampling. The MH algorithm generates from a starting position q a proposal for the next position q', with proposals accepted with some probability $Q(q'|q)$, where Q is often chosen to be the normal distribution. Sampling with increased efficiency can be achieved by using information about the geometry of the typical set of the target distribution through a vector field aligned with the typical set, which may be estimated using the gradient of the target distribution. To align the gradient vector field with the typical set vector field, we can introduce auxillary momenta variables p, promoting our D-dimensional target distribution to a projection of a $2D$-dimensional canonical distribution in some phase space:

$$P(q,p) = P(p|q)P(q) = e^{-H(q,p)} \tag{3}$$

where H is an invariant Hamiltonian function. The desired vector field may then be generated from the Hamiltonian H using Hamilton's equations. This formulation also lends itself to improved sampling efficiency by decomposing the Hamiltonian into a *kinetic energy* term and a *potential energy* term, with the latter equivalent to the logarithm of the target distribution.

$$\begin{aligned} H(q,p) &= -\log P(q,p) && (4)\\ &= -\log\left(P(p|q)P(q)\right) && (5)\\ &= -\log P(p|q) - \log P(q) && (6) \end{aligned}$$

As the Hamiltonian is phase space conserving, the Hamilton equations may be accurately integrated using a volume-preserving integration scheme, such as symplectic integrators. Of these integrators, the leapfrog scheme offers both a simple implementation and high accuracy over long integration times.

3.2 Hierarchical Models

In general, a model, whether empirical or physical, is a mathematical description of some observed data involving one or more model parameters. There are a vast number of methods, such as least-squares fitting and χ^2-minimisation, that aim to recover the optimal or 'best-fit' parameters for a model given some data. Hierarchical models extend this concept by generalising parameters as realisations of an underlying multi-dimensional population distribution with some given shape parameters, or *hyperparameters*.

★ E-mail: alexander.rawlings@helsinki.fi
[1] https://mc-stan.org

Given a sample of observations that may be reasonably assumed to be drawn from some larger population, we are able to use Bayesian analysis to recover the optimal latent parameters, constrained by prior information we have on the hyperparameters.

In this work, we have 36 merger configuration families, with each family containing ten perturbed realisations of the same initial state. As we only slightly perturb the SMBH phase space coordinates, each *child* simulation can be thought of as a sampled observation of the general galaxy-merger configuration for its corresponding family. Thus, we are able to employ hierarchical modelling to recover latent parameter distributions for various observable quantities of the merger remnant.

Robust Bayesian analysis relies on a well-informed choice of the (hyper-) parameter prior distributions. Broadly speaking, there are three categories of prior distributions used:

(i) *Uninformative* priors: no prior information is given to the prior distribution, and thus all physically-allowed values are equally likely. A uniform distribution over the parameter domain is used.
(ii) *Weakly-informative* priors: a small amount of prior information is encoded into the prior distribution, however the distribution is chosen to have a large variance. This generally ensures parameter values close to the expected order of magnitude are favoured over values that are 'insensibly' extreme. An example is a normal distribution $\mathcal{N}(\mu = 0, \sigma^2 = 100)$ for a parameter than is expected to be in the range $[-1, 1]$.
(iii) *Strongly-informative* priors: a large amount of prior information is encoded into the prior distribution, with small variance. This can lead to severely-biased parameter estimates from Bayesian analysis due to excessive 'weight' given to the prior distribution over the data. An example is a normal distribution $\mathcal{N}(\mu = 0, \sigma^2 = 0.5)$ for a parameter than is expected to be in the range $[-1, 1]$.

In this work, we opt for weakly-informative priors: from the literature, we know the approximate range of 'sensible' values that our model parameters may take, but ensure that the prior distribution variance is wide so as to not excessively weight our priors relative to our data.

Additionally, we must also consider the shape of each parameter's prior distribution, as parameters are generally constrained by some physical condition. For example, a normalising radius a cannot take a value less than 0, so a normal distribution, which has an unbounded domain, is an inappropriate choice for the prior distribution of this variable. Choosing a log-normal or gamma distribution with support on $[0, \infty)$ would represent a physically-motivated choice for the prior distribution here.

3.3 Predictive Checks

Check names: forward fold, push forward To ensure consistency between the (hyper-) parameter prior distributions and the observed data, we perform both prior-predictive and posterior-predictive checks of our modelling process.

Prior predictive checks involve sampling model parameters directly from the model prior distributions, and ensuring that the observed dependent variable can be reasonably explained by those prior distributions given the observed independent variable as input (Gabry et al. 2019). If a model is described by a set of parameters θ which are drawn from a prior distribution P, then the model parameters are sampled according to $\theta^{\rm sim} \sim P(\theta)$. Simulated data $y^{\rm sim}$ may then be sampled from the sampling distribution Q given the set $\theta^{\rm sim}$, so that $y^{\rm sim} \sim Q(y|\theta^{\rm sim})$.

Posterior predictive checks are similar to prior predictive checks, except that the posterior distributions that are recovered from the Bayesian analysis are sampled instead of the prior distributions of the model, and are thus conditioned on the data (Gelman et al. 2015). For some data dependent y, new dependent data $y^{\rm rep}$ *using the original independent data* is predicted following:

$$P(y^{\rm rep}|y) = \int_\Omega P(y^{\rm rep}|\theta)P(\theta|y)\mathrm{d}\theta \tag{7}$$

Ideally, the the observed dataset y should appear somehow typical of those datasets $r^{\rm rep}$ that are predicted. In this work, we visually classify a 'good' fit as one where the data falls within the 50% highest density interval (HDI, also referred to as the posterior density region). The α% HDI region is defined (e.g. Gelman et al. 2015) as the region containing:

(i) $100(1-\alpha)$% of the probability, and
(ii) the density within the HDI region is always greater than the density outside the region.

The HDI differs from a regular confidence interval in that a confidence interval does not impose condition (ii), making it unsuitable for describing distributions other than those that are unimodal and symmetric.

In addition to visual inspection of the posterior predictive distribution, we wish to quantitatively assess the statistical ability of our model to capture the key features we are interested in. Critically, the exact form of such a test will depend on which features we define as being 'important' for the model to capture. Typically though, a sense of how well a posterior distribution captures the first and second moments of a distribution (i.e., the mean and variance) are metrics we will be guided by.

(i) Does the posterior predictive distribution capture the *location* of the data generating process? If our model predicts data $y^{\rm sim}$ significantly offset from our observed data y, then the model is not very useful in predicting future observations generated from the same data generating process as y.
(ii) Does the posterior predictive distribution capture the *scale* of the data generating process? What does this imply??

To answer these questions, we define a test statistic T from which we determine a Bayes-equivalent of the frequentist p-value (Gelman et al. 2015).

$$p_{\rm Bayes} = \Pr\left(T(y^{\rm rep}, \theta) \geq T(y, \theta)|y\right) \tag{8}$$

$$= \iint I_{T(y^{\rm rep},\theta)\geq T(y,\theta)|y} p\left(y^{\rm rep}|\theta\right) p\left(\theta|y\right) \mathrm{d}y^{\rm rep}\mathrm{d}\theta \tag{9}$$

Here, I is the indicator function.

3.4 Assessing the HMC

Probably cite Betancourt (2017) here, maybe others. Is this section necessary, or saying 'default stan exit conditions' sufficient? Don't want to just be rewording stan docs, but also want to be transparent
.

In determining whether or not the multi-dimensional distribution described by a hierarchical model has been well sampled, a number of checks are performed by Stan, which we describe briefly here.

- sampler tree depth
- divergences
- potential energy

Table 1. The parameters for each hyperparameter distribution for the core-Sérsic model, where each distribution is a gamma distribution.

Hyperparameter	α	β
$r_{\mathrm{b},\alpha}$	3.0	2.0
$r_{\mathrm{b},\beta}$	10.0	8.0
$R_{\mathrm{e},\alpha}$	30.0	2.0
$R_{\mathrm{e},\beta}$	40.0	2.0
$\tilde{\Sigma}_{\mathrm{b},\alpha}$	20.0	4.0
$\tilde{\Sigma}_{\mathrm{b},\beta}$	20.0	2.0
γ_α	1.0	2.0
γ_β	4.0	2.0
n_α	16.0	2.0
n_β	4.0	2.0

- effective sample size
- split R-hat values

Figures to do

- Example of forward-modelled semimajor axis and eccentricity
- Example of forward-modelled density profile
- Example of forward-modelled velocity dispersion profile
- comparison latent parameters to SAMI (n, Re) and Thomas (rb)

and maybe another one of BH merger remnant properties.

WHAT ARE THE MAIN OBSERVATIONS WE SEE and WHAT CONCLUSIONS CAN WE DRAW?

(i) stability of hardening rate -> semimajor axis can be robustly predicted

(ii) wide variance of eccentricity -> N-body models offer no quantitative prediction on likelihood of merger, can only comment on what merger looks like *IF* it happens

(iii) variance in observables: trends between observable properties and what observers can actually make use of (don't have this yet), e.g. signatures of SMBH scouring in velocity profile?

ENSURE CONSISTENCY WITH SIGMA AND I. AS WE ARE USING MASS DENSITY, SHOULD BE SIGMA, NOT I, OR ASSUME A M/L RATIO OF 1. NEED TO GO THROUGH AND ENSURE THIS

3.5 Projected Mass Density Profiles

We use hierarchical modelling to recover the population parameter distributions of the core-Sérsic model Equation 1 for each simulation family. We enforce a sharp transition from the core to outer galaxy regions by setting α in Equation 1 to 10.0. A probabilistic graphical model detailing the connection between hyperparameters, latent parameters, and observables is given in Figure 1.

3.5.1 *Observable Quantities*

The hierarchical model takes three observable quantities, denoted by hat-variables in red circles in Figure 1. These variables are the radius $\hat{R}$, and the mean projected mass density $\hat{\Sigma}$, and the standard deviation of the projected mass density $\hat{\sigma}_\Sigma$ measured in the radial bins centred on $\hat{R}$. For each *child* simulation, we choose to 'observe' the first snapshot where the binary semimajor axis is *at most* 10 pc. In the event that there is no snapshot satisfying the above requirement (due to, for example, rapid merging of the SMBH binary in a time period shorter than the snapshot output frequency), the *child* simulation is not included in the analysis. We enforce a minimum of six *child* simulations per family for that family to be analysed with our hierarchical modelling technique. For those *child* simulations which pass our selection criteria, we take 30 random projections of the merger remnant, and determine the projected mass density in 50 radial bins evenly-spaced in logarithm space, with a minimum value of 0.2 kpc and a maximum value of 20.0 kpc. This allows us to build up a distribution of projected mass densities per radial bin, from which $\hat{\Sigma}$ and $\hat{\sigma}_\Sigma$ are determined.

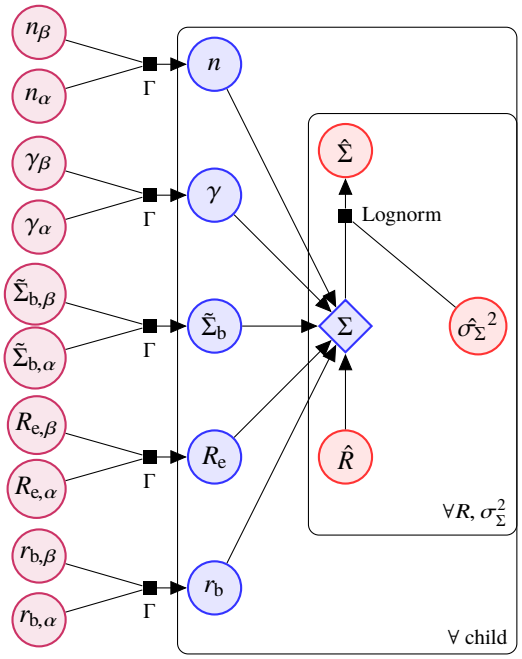

Figure 1. Probabilistic graphical model for the likelihood of the core-Sérsic model. Probabilistic parameters and observable quantities (hatted variables) are shown as circles, whereas deterministic quantities are represented by diamonds. Hyperparameters are represented by purple circles, latent parameters in blue, and observable quantities as red circles. The nodes indicate the distribution used in constructing the prior distributions (for latent parameters) and likelihood function (for $\hat{\Sigma}$). The radial box represents observed and latent variables for each radial bin for a *child* , and the child box represents latent parameters for each *child* simulation in a family.

3.5.2 *Latent Parameters for Observables*

With the omission of the parameter α in the modelling process, there are five latent parameters (r_b, R_e, $\tilde{\Sigma}_\mathrm{b}$, γ, and n, the set of which we denote as κ) that describe each *child* simulation in a family. Here, we define $\tilde{\Sigma}_\mathrm{b} \equiv \Sigma_\mathrm{b}/10^{10}$, where Σ_b is defined in Equation 2. These latent parameters, depicted as non-hatted variables in blue circles in Figure 1, are sampled to obtain a latent estimate on the projected mass density, Σ. As each latent parameter in the set κ is strictly non-negative, we use as prior distributions for each latent parameter a gamma distribution with hyperparameters $(\kappa_\alpha, \kappa_\beta)$, which are themselves drawn from a gamma distribution. To ensure our hyperparameter prior distributions are weakly-informative, the hyperparameters are drawn from the distributions in Table 1 for all simulation families. The sampled hyperparameter distributions are shown as corner plots in Figures 3-5.

We ensure our hyperparameter and latent parameter distributions are representative of our observed data by performing prior predictive checks, an example of which is shown for the A-C-3.0-0.05 family in Figure 2. Critically, both the range of values spanned by the prior distributions and the general shape of the distribution suggests that the data is well represented by the model and the prior distributions on the hyper and latent parameters.

As likelihood function, we use the lognormal distribution Lognorm($\Sigma, \hat{\sigma}^2_\Sigma$) as it a) has support on $[0, \infty)$, and b) produces normally-distributed variates in logarithm space, which is representative of the distribution of $\hat{\Sigma}$ for a given radial bin when we backwards-model and naïvely bin our observations of $\hat{\Sigma}$ for that radial bin.

Write explicitly the posterior distribution we are sampling from.

3.6 Latent Parameter Posterior Distributions

We compute the posterior distribution using Stan, ensuring that the sampler exits without encountering any divergences, and that all R-hat (is a summary stat, but how to draw? Don't want to confuse with observed radius) values are within the range [0.99, 1.01] . We first inspect how the posterior distribution appears visually by computing posterior predictive checks, an example of which is shown in Figure 6 for the family A-C-3.0-0.05. We find that the data is well described by the 50% BCI introduce what BCI is of the posterior distribution, as desired.

As part of the hierarchical modelling sampling process, we obtain in addition to the posterior distribution on $\Sigma(R)$ the latent parameter distributions from which $\Sigma(R)$ is determined. For the AC progenitor class, we find the core radius to be ~ 1.4 kpc, effective radius ~ 7.2 kpc (recall progenitor A had $R_e \sim 4$ kpc), normalising density of $\sim 3.1 \times 10^9\ M_\odot$ kpc^{-2}, core slope index of ~ 0.05, and Sérsic index of ~ 1.8.

- corner plot description
- comparison between rperi values, maybe new section?

4 DISCUSSION

Discuss said groundbreaking results

5 CONCLUSIONS

Leave the reader with a good feeling

Software: Ketju (Rantala et al. 2017), Gadget-3 (Springel 2005), Stan (Carpenter et al. 2017), NumPy (Harris et al. 2020), SciPy (Virtanen et al. 2020), Matplotlib (Hunter 2007), pygad (Röttgers et al. 2020), Arviz (Kumar et al. 2019).

ACKNOWLEDGEMENTS

P.H.J. acknowledge the support by the European Research Council via ERC Consolidator Grant KETJU (no. 818930).

The numerical simulations used computational resources provided by the CSC – IT Center for Science, Finland.

DATA AVAILABILITY

The inclusion of a Data Availability Statement is a requirement for articles published in MNRAS. Data Availability Statements provide a standardised format for readers to understand the availability of data underlying the research results described in the article. The statement may refer to original data generated in the course of the study or to third-party data analysed in the article. The statement should describe and provide means of access, where possible, by linking to the data or providing the required accession numbers for the relevant databases or DOIs.

REFERENCES

Betancourt M., 2017, arXiv e-prints, p. arXiv:1701.02434
Carpenter B., et al., 2017, Journal of Statistical Software, 76, 1
Gabry J., Simpson D., Vehtari A., Betancourt M., Gelman A., 2019, Journal of the Royal Statistical Society Series A, 182, 389
Gelman A., Carlin J. B., Stern H. S., Dunson D. B., Vehtari A., Rubin D. B., 2015, Bayesian Data Analysis, 3rd ed. edn. Chapman and Hall/CRC
Harris C. R., et al., 2020, Nature, 585, 357
Hunter J. D., 2007, Computing in Science & Engineering, 9, 90
Kumar R., Carroll C., Hartikainen A., Martin O., 2019, Journal of Open Source Software, 4, 1143
Rantala A., Pihajoki P., Johansson P. H., Naab T., Lahén N., Sawala T., 2017, ApJ, 840, 53
Röttgers B., Naab T., Cernetic M., Davé R., Kauffmann G., Borthakur S., Foidl H., 2020, MNRAS, 496, 152
Springel V., 2005, MNRAS, 364, 1105
Virtanen P., et al., 2020, Nature Methods, 17, 261

APPENDIX A: SOME EXTRA MATERIAL

If you want to present additional material which would interrupt the flow of the main paper, it can be placed in an Appendix which appears after the list of references.

This paper has been typeset from a TEX/LATEX file prepared by the author.

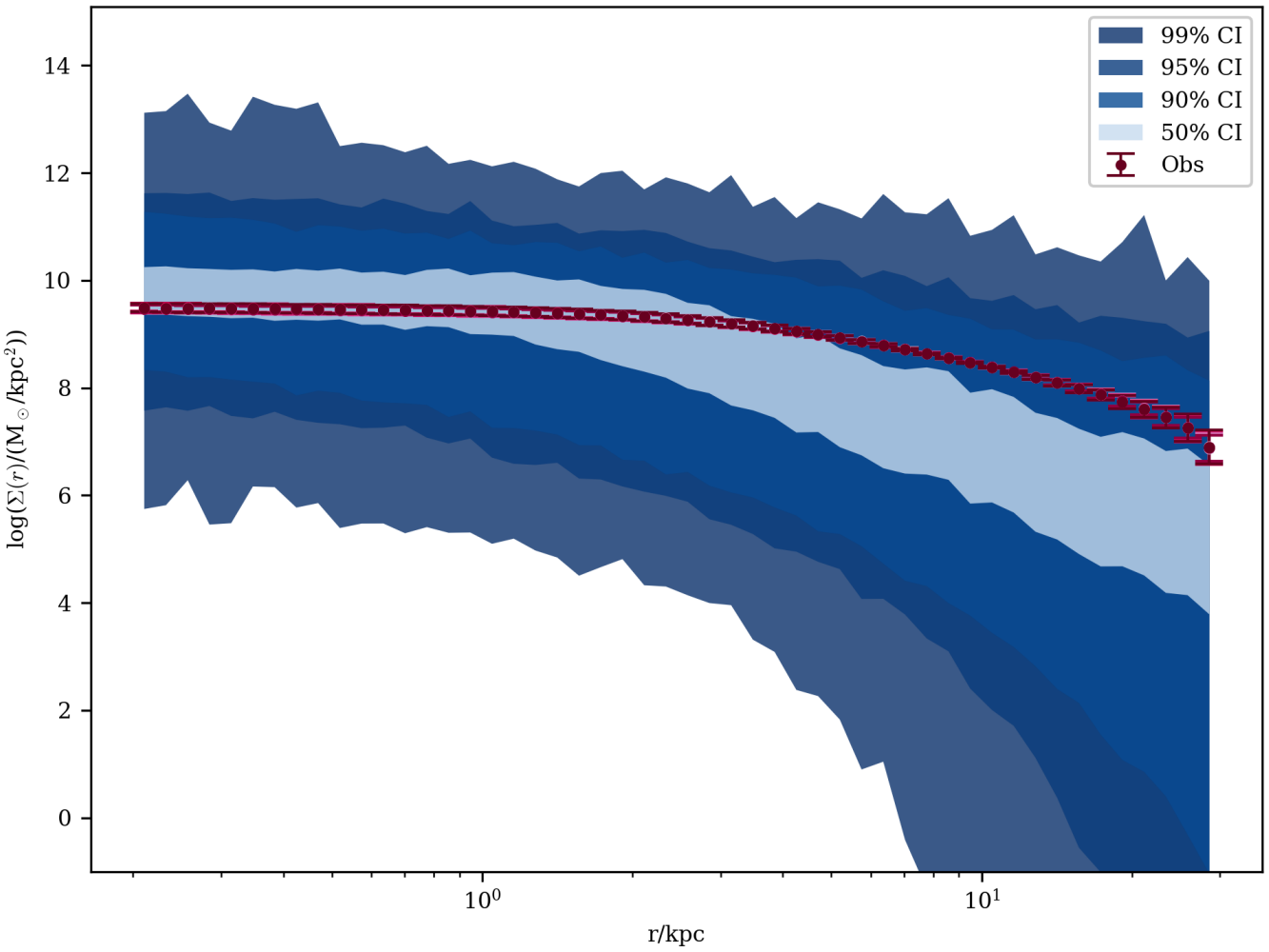

Figure 2. Prior predictive estimate for the core-Sérsic fit. Darker shaded regions correspond to higher Bayesian Credible Intervals (BCIs). The observed data is overlaid. From the prior predictive estimate, it is seen that the model and the assumed parameter hyperdistributions are able to describe the observed data, in that the observed data lies within the range allowed by the prior predictive estimate.

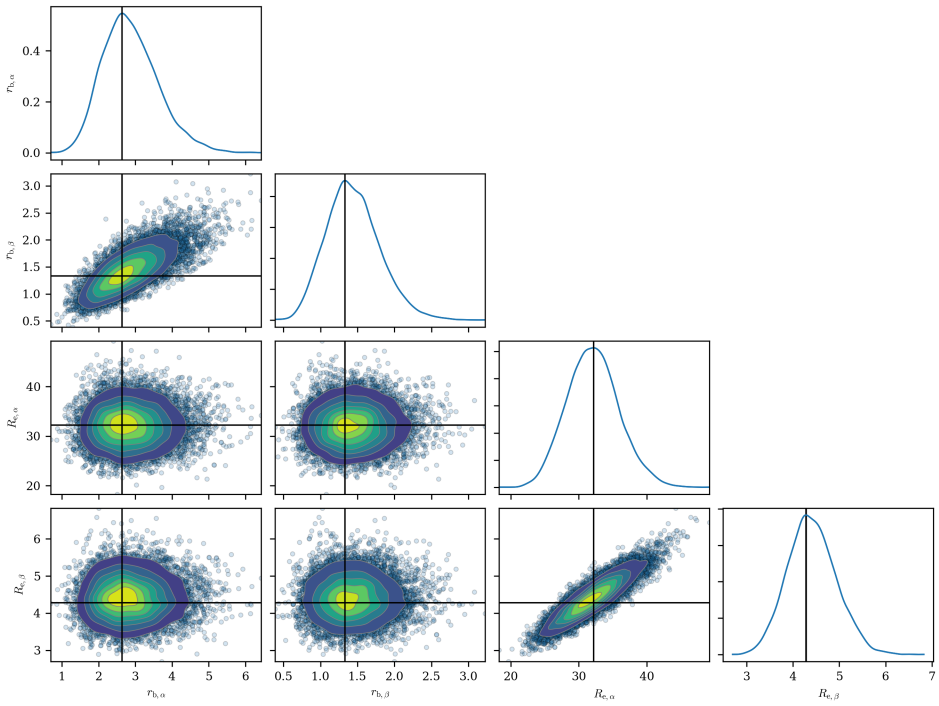

Figure 3. Corner plot of hierarchical model hyperparameters. Each of the hyperdistributions are modelled as Gamma distributions with shape parameters α and β. For example, the hyperdistribution on the core radius r_b is modelled as $r_\mathrm{b} \sim \mathrm{Gamma}(r_{\mathrm{b},\alpha}, r_{\mathrm{b},\beta})$. Critically, the hyperdistributions are unimodal with highest-density regions located some distance from the distribution edges, indicating the HMC sampling has converged well.

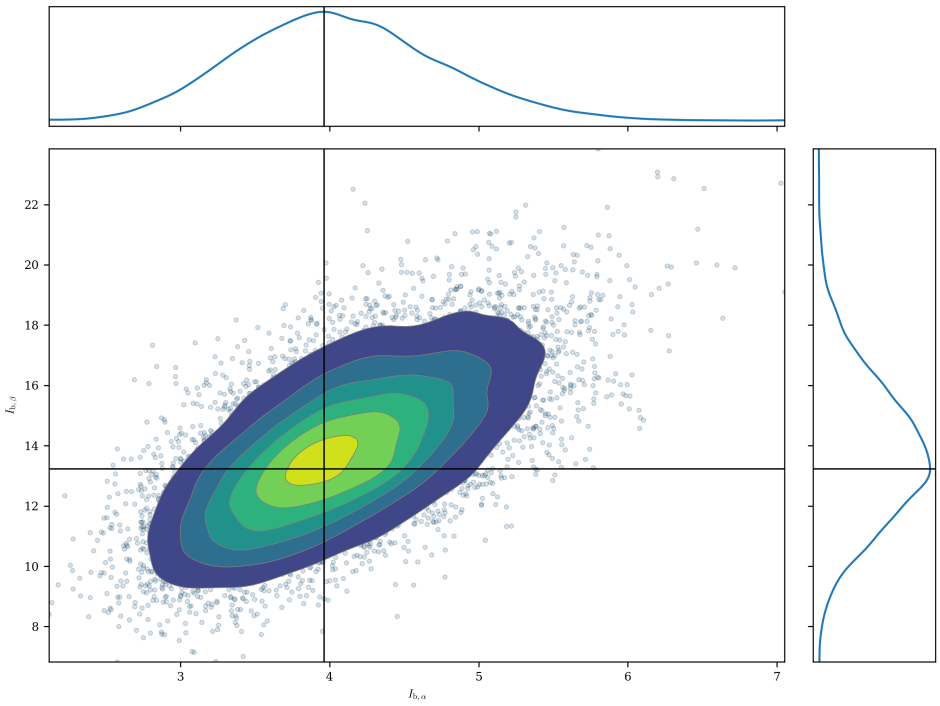

Figure 4. Same as Figure 3

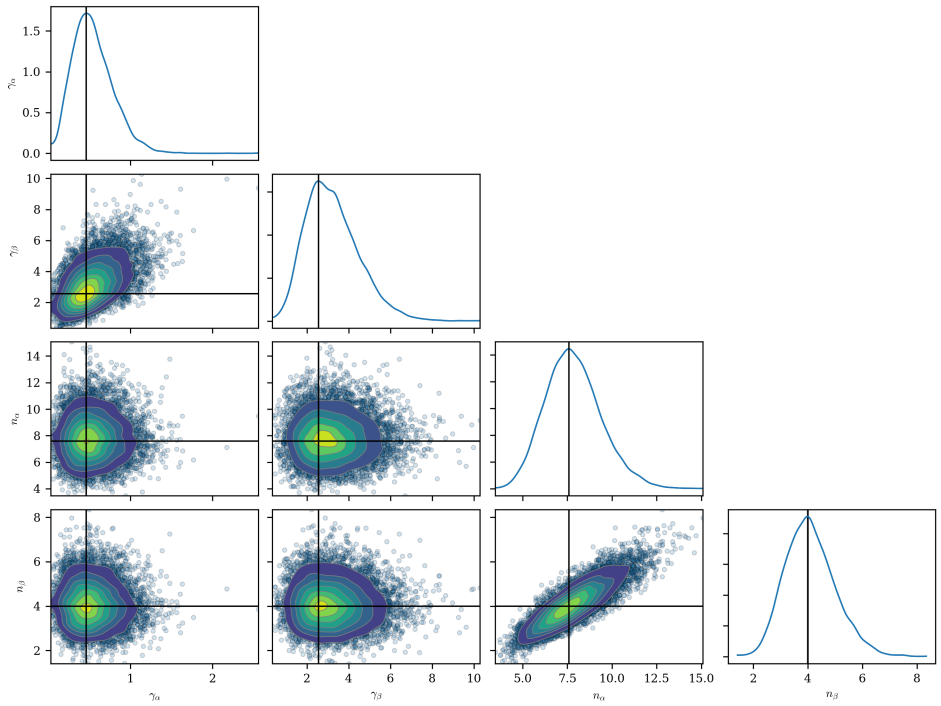

Figure 5. Same as Figure 3

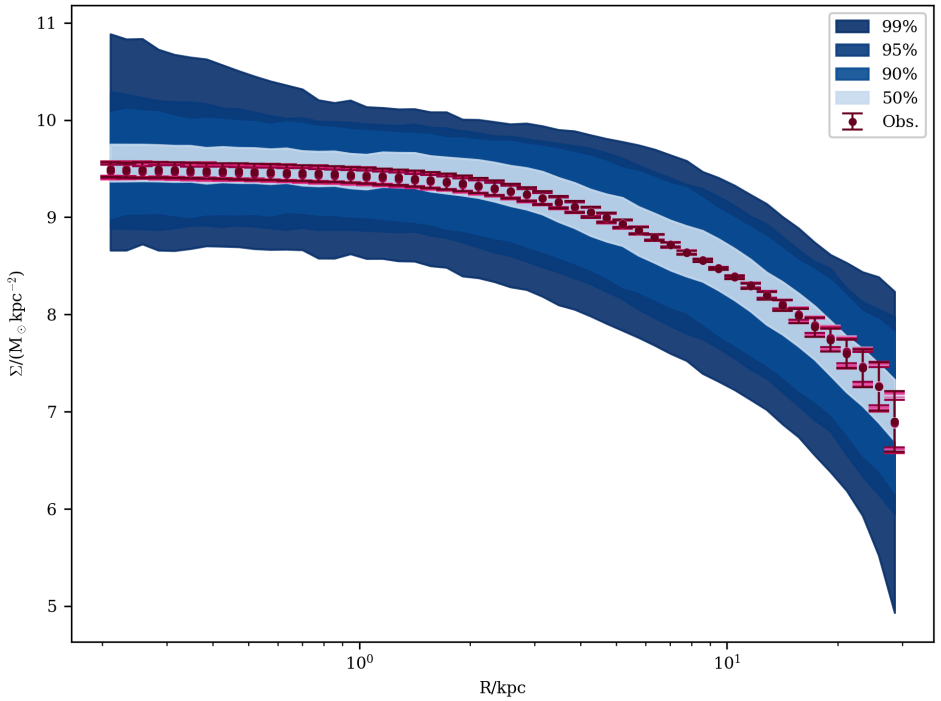

Figure 6. Posterior predictive estimate for the core-Sérsic fit. Darker shaded regions correspond to higher Bayesian Credible Intervals (BCIs). The observed data is overlaid, and seen to lie well within the 50% BCI.