Monthly Notices
of the
ROYAL ASTRONOMICAL SOCIETY

MNRAS **482,** 3356–3372 (2019)
Advance Access publication 2018 October 29

doi:10.1093/mnras/sty2931

Reliable mass calculation in spherical gravitating systems

Foivos I. Diakogiannis,[1,2★] Geraint F. Lewis,[3] Rodrigo A. Ibata,[4] Magda Guglielmo,[3] Mark I. Wilkinson[5] and Chris Power[2]

[1] *Data61, CSIRO, Floreat WA 6014, Australia*
[2] *International Center for Radio Astronomy Research, University of Western Australia, 35 Stirling Highway, Crawley, WA 6009, Australia*
[3] *Sydney Institute for Astronomy, School of Physics, A28, University of Sydney, NSW 2006, Australia*
[4] *Observatoire Astronomique, Université de Strasbourg, CNRS, 11, rue de l Université, F-67000 Strasbourg, France*
[5] *Department of Physics & Astronomy, University of Leicester, Leicester LE1 7RH, UK*

Accepted 2018 October 25. Received 2018 October 4; in original form 2018 April 12

ABSTRACT

We present an innovative approach to the methodology of dynamical modelling, allowing practical reconstruction of the underlying dark matter mass without assuming both the density and anisotropy functions. With this, the mass–anisotropy degeneracy is reduced to simple model inference, incorporating the uncertainties inherent with observational data, statistically circumventing the mass–anisotropy degeneracy in spherical collisionless systems. We also tackle the inadequacy that the Jeans method of moments has on small data sets, with the aid of Generative Adversarial Networks: we leverage the power of artificial intelligence to reconstruct the projected line-of-sight velocity distribution non-parametrically. We show, with realistic numerical simulations of dwarf spheroidal galaxies, that we can distinguish between competing dark matter distributions and recover the anisotropy and mass profile of the system.

Key words: methods: statistical – techniques: radial velocities – galaxies: dwarf – galaxies: kinematics and dynamics – galaxies: statistics.

1 INTRODUCTION

Whilst dark matter represents the dominant mass component of the universe, its true nature remains elusive. Astrophysical probes of the properties of dark matter in large galaxies and galaxy clusters are typically hampered by the complexities of baryonic physics, and the complex coupling of the properties of kinematic tracers and the underlying form of the gravitational potential.

In recent years, considerable focus has been given to dwarf spheroidal galaxies in the local universe. With a stellar mass of $\sim 10^7\ \mathrm{M}_\odot$, these are seen to be both devoid of gas, limiting the impact of baryonic astrophysics, and sufficiently simple to allow the determination of the gravitational potential of the dominant dark matter component from the stellar motions. However, traditional approaches of determining the distribution of dark matter in dSphs are limited by both the influence of the observational uncertainties and the mathematical complexity of deriving the properties of the dark matter.

One such approach, the Schwarzschild (1979) method, attempts to determine the underlying dark matter distribution through the reconstruction of the observed luminosity and kinematic properties of a galaxy using a library of precomputed orbits in trial potentials. Via the appropriate weighting of the components of the library for a particular mass model, the optimal fit to the data can be recovered and the mass determined. However, the computational aspects of the Schwarzschild method makes implementation highly impractical. Building a high-resolution orbit library to survey the likelihood of millions of mass models is currently computationally prohibitive.

Other approaches are based upon the Jeans equation (Binney 1980), which relates the properties of kinematic tracers to the form of the gravitational potential. When applying the Jeans equation, there are two key ingredients, the distribution of dark matter, and a velocity anisotropy, β, which describes the relationship between radial and tangential orbits within the structure. In established approaches, it is typical to assume a functional form for the dark matter distribution, such as a Navarro–Frenk–White (Navarro, Frenk & White 1996) or a Plummer (1911) profile, and a functional form for β, optimizing the parameters of both based upon the observational data. Given the mathematical form of the Jeans equation, however, the resultant determination of the mass depends upon the assumed form for β, with various combinations of the adopted mass profile and β providing equally acceptable fits to the data. Known as the 'mass–anisotropy degeneracy' (hereafter MAD), this is generally accepted as a fundamental limitation of Jeans-based approaches (Merrifield & Kent 1990, see also Read & Steger 2017).

In this contribution, we present a new approach to address the MAD in the Jeans formalism, relying upon a parametrized functional form, known as a B-Spline, to account for the implicit relationship between the dark matter profile and the velocity anisotropy.

★ E-mail: foivos.diakogiannis@data61.csiro.au

Published by Oxford University Press on behalf of the Royal Astronomical Society

In this latest version of the JEANS (Diakogiannis et al. 2017) approach, the *tight*-JEANS (hereafter t-JEANS), we represent both the unknown radial *and* tangential velocity dispersions as B-splines. Then, we allow the data to give them the correct geometric shape. In this way, we avoid having to assume the functional form of all, but one, of the unknown functions used in the modelling process. Then, even with competing dark matter models that have equal numbers of unknown coefficients, we end up with statistical fits of different quality. The key point is that by demanding that these curves be as simple as possible, i.e. that they are represented by a minimal number of variables, competing dark matter density models give different qualitative fits to the data. This eventually allows us to statistically discriminate between competing mass models and thus transform the MAD to a mere model inference problem. For the case of small data sets (of the order of 1000 tracer stars), we use Generative Adversarial Networks (hereafter GANs, Goodfellow et al. 2014) to reconstruct non-parametrically the underlying projected line-of-sight (LOS) velocity distribution. With this, we artificially augment the data to arbitrarily large numbers, and obtain reliable estimates for the moments of the LOS velocity distribution with an unprecedented density of points. The combination of t-JEANS modelling with the GANs for artificial data augmentation is a powerful approach for reliable mass estimates.

In Section 2 we present a short review of the Jeans mass modelling method. In Section 3 we give the details of the data sets we used as well as the pre-processing method we followed. In Section 4 we give a detailed description of the t-JEANS algorithm. In Section 5 we present our findings and in Section 6 we discuss the reasons behind the efficiency of the t-JEANS. Finally in Section 7 we present our concluding remarks.

2 A REVIEW OF THE JEANS MODELLING METHODOLOGY

In this section we present an overview and analysis of the established (Binney & Tremaine 2008) methodology of Jeans modelling. We continue by providing a proof for the uniqueness of the anisotropy profile upon assuming a specific functional form for the mass density profiles of stars, $\rho_\star$, and dark matter (hereafter DM), $\rho_\bullet$.

The Jeans modelling approach subject to the assumption of spherical symmetry is fully contained in the following two equations:

$$-\frac{\mathrm{d}\Phi}{\mathrm{d}r} = \frac{1}{\rho_\star}\frac{\mathrm{d}}{\mathrm{d}r}\left(\rho_\star\sigma_{\rm rr}^2\right) + \frac{2}{r}\beta(r)\sigma_{\rm rr}^2 \tag{1}$$

$$\sigma_{\rm los}^2(R) = \frac{2}{\Sigma_\star(R)}\int_R^{r_{\rm vir}}\left(1-\beta(r)\frac{R^2}{r^2}\right)\frac{r\,\rho_\star\sigma_{\rm rr}^2}{\sqrt{r^2-R^2}}\mathrm{d}r. \tag{2}$$

Here, Φ is the total potential of the system, $\rho_\star$ the stellar tracer density, $\sigma_{\rm rr}^2$ the radial velocity dispersion, $\Sigma_\star$ is the projected tracer surface density, $\sigma_{\rm los}^2$ is the observed LOS velocity dispersion, β is the anisotropy profile defined by $\beta(r) = 1 - \sigma_{\rm tt}^2/(2\sigma_{\rm rr}^2)$, and R and r are, respectively, the projected and 3D distance radii from the centre of the system. Although the integral in equation (2) usually has infinity as its upper bound, here we define $r_{\rm vir}$ as the distance in which the DM mass density, $\rho_\bullet$, profile falls to approximately $\rho_\bullet(r_{\rm vir}) \approx 200\rho_{\rm crit}$. For all practical purposes, this is a useful numerical approximation that does not alter our findings. With the exception of the observed LOS velocity dispersion, $\sigma_{\rm los}^2$, and the projected tracer density profile, $\Sigma_\star$, all remaining functions ($\rho_\star$, $\rho_\bullet$, $\sigma_{\rm rr}^2$, β) are unknown and need to be determined from the data. Therefore, the system is underdetermined.[1] In practice we can make a very good approximation to the functional form of the tracer density profile, $\rho_\star$ given deep photometry of the dSph, and we are thus left with three unknown functions, $\{\rho_\bullet(r), \sigma_{\rm rr}^2(r), \beta(r)\}$ in a system of two equations.

Although there are variations[2] to the general methodology, the common established (Binney & Tremaine 2008) starting point to solving this system of coupled integrodifferential equations with respect to the unknowns $\rho_\bullet(r)$, $\beta(r)$ and $\sigma_{\rm rr}^2(r)$, is to assume parametric functional forms for the DM mass density, $\rho_\bullet$, and the anisotropy profile, $\beta(r)$. In an iterative approach (assuming for simplicity we have full knowledge of the tracer profile, $\rho_\star$), one proposes a set of values for the parameters that define $\rho_\bullet$ and β, then solves the differential equation (1) with respect to[3] $\sigma_{\rm rr}^2$ and substitutes the result in equation (2). The validity of the numerical values of the parameters that define $\rho_\bullet$ and β is tested by comparing the model $\sigma_{\rm los}^2$ with the observables. This iterative process is performed until some convergence criterion is met. The rationale behind this approach is that when we consider parametric forms for $\rho_\bullet$ and β, the system becomes overdetermined (since equations 2 and 1 are evaluated in various distinct locations, r_i, R_j) and thus a solution exists.

It needs to be emphasized though that once we make an assumption for the parametric form of one of the three unknown functions, $\{\rho_\bullet, \sigma_{\rm rr}^2, \beta\}$, the system of two equations with (the remaining) two unknowns is closed. That is, the remaining two functions can be fully determined without the need for their parametric representation. There exist published (Binney & Mamon 1982; Solanes & Salvador-Sole 1990; Dejonghe & Merritt 1992; Mamon & Boué 2010) exact solutions to the system of these equations (termed *inversion* techniques) that make a parametric assumption for only one of the three unknown functions. These prove that it is an unnecessary assumption to assume two of the three unknown functions in parametric form. Usually, assuming more parametric forms than necessary, increases the uncertainty in the model parameters, thus making the distinction between competing mass models even more difficult.[4]

It is insightful to separate the process of solving the system of coupled integrodifferential equations (1) and (2), in two distinct approaches: the exact numerical solution of the equations to perfect noiseless data and the statistical fitting to noisy data. Clearly, all conclusions we can draw from knowledge gained in exact solutions of the system of Jeans equations can be transferred to the case of statistical fitting, while the converse is not always true. In the following, we focus on the exact numerical solution.

[1] One though needs to be precise in the definition of the number of 'unknowns'. Usually, we make assumptions for the functional form of these unknown functions that depend on some parameters. It is the number of these parameters that define the necessary number of equations to close the system. Then, for either exact (numerical solutions) or overdetermined systems (statistical fitting), we evaluate each of the equations (1) and (2) in a set of distinct locations, r_i, R_j that are equal or greater in numbers to the number of unknown parameters.

[2] These include, e.g. using higher moments (Łokas & Mamon 2003) of $\sigma_{\rm los}^2$, or different assumptions on the distribution function of the system, i.e. different penalty functions when comparing the LOS velocity dispersion with observables.

[3] Subject to the boundary condition $\lim_{r\rightarrow r_{\rm vir}}\sigma_{\rm rr}^2 \approx 0$.

[4] In addition, there are well-known methods for solving numerically systems of coupled integrodifferential equations, such as finite differences and finite element methods (Ŝolín 2005; Jalali & Tremaine 2011), wavelets (Bertoluzza et al. 2008), and B-splines discretization (Höllig 2003).

2.1 Uniqueness of the anisotropy profile for a given mass model

In this section, we provide a theorem that upon making an assumption for the functional form of the tracer and DM mass densities, $\rho_\star$, $\rho_\bullet$, and the LOS dispersion, $\sigma^2_{\rm los}$ there exists a unique anisotropy profile, β. For our purposes, we consider we have full knowledge of the above-mentioned functions, $\rho_\star$, $\rho_\bullet$, and $\sigma^2_{\rm los}$. We solve equation (1) with respect to $\beta(r)$ and substitute it under the integral sign of equation (2). Then we end up with a single integrodifferential equation (subject to the virial boundary condition $\lim_{r\to r_{\rm vir}}\sigma^2_{\rm rr}(r)\approx 0$), namely

$$\sigma^2_{\rm los}(R)=\frac{2}{\Sigma_\star(R)}\int_R^{r_{\rm vir}}\left[K_{\rm A}\left(\frac{{\rm d}(\rho_\star\sigma^2_{\rm rr})}{{\rm d}r}+\rho_\star\frac{{\rm d}\,\Phi}{{\rm d}r}\right)\right.\\ \left.+\,K_{\rm B}\rho_\star\sigma^2_{\rm rr}\right]{\rm d}\,r, \tag{3}$$

where $K_{\rm A}$ and $K_{\rm B}$ are kernel functions defined by

$$K_{\rm A}(r,R)=\frac{R^2}{\sqrt{r^2-R^2}},\qquad K_{\rm B}(r,R)=\frac{2r}{\sqrt{r^2-R^2}}.$$

This equation has one unknown, the radial velocity dispersion, $\sigma^2_{\rm rr}$. That is, assuming perfect knowledge of the LOS velocity dispersion profile, $\sigma^2_{\rm los}(R)$, if we could solve this equation for the unknown $\sigma^2_{\rm rr}$ we would obtain for each assumption of a mass model, $\{\rho_\star, \Phi(r)\}$, a radial velocity dispersion, $\sigma^2_{\rm rr}$. The question arises: is the solution with respect to $\sigma^2_{\rm rr}$ unique?

THEOREM 1. *The solution of equation (3) with respect to $\sigma^2_{\rm rr}$ for given $\sigma^2_{\rm los}(R)$, $\rho_\star(\mathrm{r})$ and $\Phi(\mathrm{r})$ profiles, is unique.*

We provide the proof of Theorem 1 in the Appendix. This result, which complements the published inversion techniques, has the following implication: we can assume only one of the three unknown functions and thus reduce *the uncertainty* of the modelling parameters (in comparison with the uncertainty we get by assuming parametric forms for two unknown functions, as is customary). For the case of statistical fitting, we can use hierarchical models (e.g. smoothing splines) of varying complexity, that make model selection possible and this is the key for breaking *statistically* the Jeans degeneracy: the missing ingredient (an additional equation) is replaced by the model selection criterion. Thus, if we allow the total mass, $M(r)$ to vary, i.e. if we assume a different mass model for the same $\sigma^2_{\rm los}$ profile, then the anisotropy profile will generally be different. However there is one important constraint we need to consider, namely, the projected virial theorem (discussed in detail in Section 4.5.1). The projected virial theorem does not depend on the anisotropy, which implies that not all mass profiles are consistent with the projected kinetic energy evaluated from $\sigma^2_{\rm los}$. However, the projected virial theorem, on its own, is not sufficient to break the degeneracy (it is a single scalar equation, therefore the total number of equations is still less than the unknowns). It can only further reduce the feasible solution space of where the $M(r)$ function resides. We will discuss this further in Section 6. In Section 5.1 we provide numerical examples of the uniqueness of the kinematic profile for an assumed mass density.

It should be stated that we can choose equally well to assume a functional form for the anisotropy profile, and leave the mass density to be deduced by the data (Mamon & Boué 2010, see also Read & Steger 2017): in this case, the total mass of the system *follows* from the assumptions of the anisotropy model, β, for a given data set of observables. This can be very easily seen from the following: again, assuming perfect knowledge of $\sigma^2_{\rm los}$ profile, once we use a specific functional form for the anisotropy β, the system of equations that describes a stellar dynamical system is

$$\frac{1}{\Sigma_\star}\int_R^{r_{\rm vir}}\left[\rho_\star K_1(r,R)\sigma^2_{\rm rr}+\rho_\star K_2(r,R)\sigma^2_{\rm tt}\right]{\rm d}\,r=\sigma^2_{\rm los}(R) \tag{4}$$

$$\sigma^2_{\rm tt}-2(1-\beta(r))\sigma^2_{\rm rr}=0 \tag{5}$$

$$\frac{1}{\rho_\star}\frac{\rm d}{{\rm d}r}\left(\rho_\star\sigma^2_{\rm rr}\right)+\frac{2\sigma^2_{\rm rr}-\sigma^2_{\rm tt}}{r}=-\frac{GM_{\rm tot}(r)}{r^2}. \tag{6}$$

These are three equations with respect to the three unknowns $\sigma^2_{\rm rr}$, $\sigma^2_{\rm tt}$, and the total mass $M_{\rm tot}$. The system of equations, equations (4) and (5) is complete, i.e. we have a unique solution for $\sigma^2_{\rm rr}$ and $\sigma^2_{\rm tt}$. Then, from the last equation (6) we calculate the total mass, $M_{\rm tot}$, whose value depends solely on the tracer stellar density, $\rho_\star$, and the kinematic profile, $\sigma^2_{\rm rr}$, $\sigma^2_{\rm tt}$. Therefore, when we model by *assuming* a specific anisotropy profile, β, we effectively pre-specify the mass content of the system. In this reasoning, we did not need to adopt any assumptions for the parametric form of the DM mass density.

3 DATA

In this section we provide an overview of the data sets we used to validate our methodology. We describe how we pre-process the data and create validation and test data sets for the t-JEANS solver, as well as how we use GANs to generate large artificial samples of data for the case of small data sets.

We test our algorithm with the GAIA CHALLENGE[5] suite of mock simulations, in particular the spherically symmetric sets. The mock suites provide a snapshot of the full 6D information of the tracer profile, x, y, z, v_x, v_y, v_z. For modelling each of the systems, we used only the projected positions, x, y, and the LOS velocity, v_z. Following the GAIA CHALLENGE guidelines, we used the data that include velocity errors. For each datum we considered that this error is equal to the 2 per cent of the true v_z velocity. Our training data set, $D_{\rm train}$, consists of the values, $\{x, y, v_z\}_j, j=1,\dots,N_\star$, as well as the second-order moments, $\sigma^i_{\rm los}$, of the projected LOS velocity.

As a sample of the various data sets, we chose the PlumCuspOM, PlumCuspIso, PlumCuspTan, and NonPlumCoreOM suites. For the first three we use the suites with 10k targets, and for the last one we use both 10k and 1k data sets. The Plummer-like family of tracer profiles was chosen based on the knowledge that most Stellar profiles observed in nature are cored; the latter three models were considered to be representative by the curator of the GAIA CHALLENGE (Read & Steger 2017). In particular, the NonPlumCoreOM is a notoriously difficult set to model, and this is the reason why for this particular one we also include a set with only 1k targets. Each of these data sets was modelled with assumptions for the stellar and DM profiles only. We model each system with two competing models: one with the true parametric form (with parameters recovered from the fitting process), and one with an incorrect parametric assumption (again the parameters are fitted to the data). We report the combinations of Stellar and DM models we used in Tables 2 and 6. In all model fits, the anisotropy profile is evaluated from the data.

The reference anisotropy profiles that these data sets were created from are two, namely constant and Ossipkov-Merritt (Osipkov

[5] http://astrowiki.ph.surrey.ac.uk/dokuwiki/doku.php

Table 1. Synthetic data sets: parameters α, β, γ are dimensionless numbers. Distance parameters $r_{\star,\bullet}$ are in kpc, while $\rho_{0\bullet}$ in $\mathrm{M_\odot\ pc^{-3}}$.

Data set	$\theta_\star = [r_\star, \alpha_\star, \beta_\star, \gamma_\star]$	$\theta_\bullet = [\rho_{0\bullet}, r_\bullet, \alpha_\bullet, \beta_\bullet, \gamma_\bullet]$	$\beta(r)$ anisotropy
PlumCuspOM (10k)	[0.1,2,5,0.1]	[0.064,1,1,3,1]	β_{OM}, $r_{\mathrm{a}} = 0.1$
PlumCuspIso (10k)	[0.25,2,5,0.1]	[0.064,1,1,3,1]	$\beta_0 = 0.0$
PlumCuspTan (10k)	[0.5,2,5,0.1]	[0.0239, 2, 1, 4, 1]	$\beta_0 = -0.5$
NonPlumCoreOM (1k, 10k)	[0.25,2,5,1]	[0.400,1,1,3,0]	β_{OM}, $r_{\mathrm{a}} = 0.25$

Table 2. Competing mass models for the various 10k data sets. We report the average error on unseen test data, D_{test}. The true models from which the data were produced are with bold fonts. In all cases the test error, χ^2_{test}, selects the correct model.

Data set	Stellar model	DM model	χ^2_{test}
PlumCuspOM, 10k	Plummer	Burkert	223.972
PlumCuspOM, 10k	**gH**	**NFW**	**217.337**
PlumCuspIso, 10k	Plummer	Burkert	207.001
PlumCuspIso, 10k	**gH**	**NFW**	**204.812**
PlumCuspTan, 10k	Plummer	Burkert	192.115
PlumCuspTan, 10k	**gH**	**gH**	**192.017**
NonPlumCoreOM, 10k	Plummer	NFW	348.126
NonPlumCoreOM, 10k	**gH**	**gH**	**340.255**

1979; Merritt 1985):

$$\beta(r) = \begin{cases} \beta_0, & \text{constant} \\ \dfrac{r^2}{r^2 + r_{\mathrm{a}}^{\alpha}}, & \text{OM} \end{cases} . \tag{7}$$

The mass density profile that these data sets follow, both for the stellar and the DM components, is given by a power law from Zhao (1996), for a variety of reference parameters:

$$\rho(r) = \rho_0 \left(\frac{r}{r_s}\right)^{-\gamma} \left[1 + \left(\frac{r}{r_s}\right)^{\alpha}\right]^{(\gamma-\beta)/\alpha} . \tag{8}$$

In addition, the GAIA CHALLENGE data sets make the approximation that the stellar tracer mass is negligible in comparison to the DM mass component. In order to account for this we normalized the total tracer mass to unity, i.e. $M_\star^{\mathrm{tot}} = 1\,\mathrm{M_\odot}$.

We report the reference model parameters for each of the data sets used in Table 1. In Tables 2 and 6 we record the combinations of Stellar and DM mass models we used for the modelling process as well as the test error for the various competing models.

3.1 Data pre-processing

In this section we describe the process that we followed in order to create the LOS velocity dispersion, σ^i_{los}, values, that we use as training data.

Each bin contains $N_{\mathrm{targets}} = \sqrt{N_\star}$. For example, for 10^4 stars, we have $N_{\mathrm{bin}} = 10^2$. This approach gives equal Poisson error (Read & Steger 2017) for each datum. We modelled the distribution of stars, within each i bin, as a Gaussian centred at zero. The likelihood of this model, for each bin i, is

$$P_i(v_j|s, D) = \prod_{j=1}^{N_{\mathrm{targets}}} \frac{\exp\{-v_j^2(s^2 + (\delta v_j)^2)^{-1}/2\}}{\sqrt{2\pi(s^2 + (\delta v_j)^2)}}, \tag{9}$$

where v_j is the value of the LOS velocity of star j in bin i, δv_j is the associated error, and s the standard deviation of the Gaussian distribution. For each bin i, we perform a Markov Chain Monte Carlo (MCMC) process using the likelihood equation (9), to estimate the marginalized distribution of the parameter s. It should be clear that this MCMC process is used only in the data pre-processing stage. It should not be confused with the MCMC we perform later for the estimation of marginalized distributions for the stellar, $\theta_\star$, and DM, $\theta_\bullet$, parameters. The LOS velocity dispersion data values, σ^i_{los}, we use at each location R_i (centre of the ith radial bin), is the mode value of the histogram, $\hat{s}^2 = \sigma^i_{\mathrm{los}}$. The associated error, $\delta\sigma^i_{\mathrm{los}}$, is the 1σ uncertainty of s^2. Thus, $D_{\mathrm{train}} = \{x, y, v_z\}_j \cup \{\sigma^i_{\mathrm{los}}, \delta\sigma^i_{\mathrm{los}}\}, j = 1, \dots, N_\star$ and $i = 1, \dots, N_{\mathrm{bin}}$.

In addition to the above LOS moments, we draw 100 random samples, $\sigma^{ij}_{\mathrm{los}}$ ($j = 1, \dots, 100$), from the marginalized distribution of σ_i^2, that we keep for a validation data set, D_{val}, and 200 random samples that we use for test sets, D_{test}. During the Evolutionary Algorithm (hereafter EA) training, the validation set is used for the selection of the smoothing parameters, θ_{smooth} (Section 4). During the MCMC training phase, instead of using the mode value σ^i_{los} of the LOS dispersion as the moments data, in each iteration of the solver, we select random realizations, $\sigma^i_{\mathrm{los}} = \sigma^{ij}_{\mathrm{los}}$ (random j), from the validation data set, D_{val}. In this way we incorporate the uncertainty of the moments data as prior information to the modelling process. The test set, D_{test}, is used for the model selection between competing models after the EA phase. In Fig. 1 we plot for the case of the PlumCuspTan model the binned σ^i_{los} values (black solid line), as well as the validation values, $\sigma^{ij}_{\mathrm{los}} \in D_{\mathrm{val}}$, for each bin i.

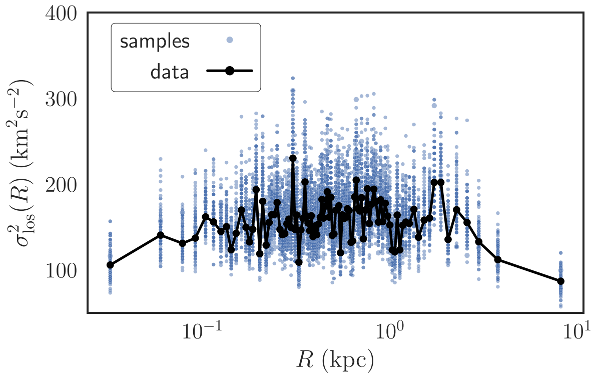

Figure 1. Binned σ^2_{los} (black solid line) data as well as validation σ^2_{los} data set D_{val} for the PlumCuspTan model.

3.2 Data augmentation for small data sets using GANs

In this section we briefly describe the application of GANs for the numerical reconstruction of the 3D projected LOS velocity distribution, $f(x, y, v_{\mathrm{los}})$ from the NonPlumCoreOM 1k data set. Our goal is to give an intuitive understanding behind the reason that this method is so effective and not to detail the GAN methodology (see Goodfellow 2017 for a pedagogical introduction).

A fundamental limitation to the method of moments, in the Jeans framework, is that it requires a wealth of data to be successful. This is because the moments of the data, as a product of the summary information of the underlying distribution, are much fewer in number than the original unbinned data set. This is more evident especially when the original data set is small (from a few hundred to 1k stars) as is often the case in astronomical data sets (e.g. of dSph galaxies). We overcome this difficulty by applying a pre-processing step, where we create synthetic data from a generative model, that resembles the true underlying distribution. That is, we create synthetic data to complement the original data set and thus acquire a large number of LOS velocity moments. We do so only for the 1k NonPlumCoreOM data set (although the method can be applied to the 10k as well for higher quality results). For this task, artificial intelligence actors (GANs) are excellent generative models, since they learn by 'looking' at the real data, i.e. *by example*, and are not bound by assumptions of the mathematical form of the underlying distribution.

The general framework of the GANs consists of a set of two competing artificial neural networks (hereafter ANNs). The first, the Generator (hereafter G), takes as input a vector of random numbers and tries to create fake (synthetic) data whose distribution resembles the distribution of the true training data set. The second, the Discriminator (hereafter D), takes as input, true data, drawn randomly from the training distribution, or fake data, created randomly from G, and tries to predict whether the data that it was given are genuine (real) or fake. During training, the goal of G is to make D perform a mistake, i.e. the goal of G is to generate as authentic looking synthetic data as possible. The goal of D is to discriminate the true data from the fake ones and debunk the efforts of G. This framework is a minimax two-player game. During training both players become proficient in their task. When this process reaches equilibrium, G is a faithful approximator of the true underlying distribution of the training data set. This method is unsupervised training that in practice means there is no upper bound on the quality of the data approximation.

This method has been applied successfully, with impressive results, in artificial intelligence generative tasks, such as the creation of high quality images (Karras et al. 2017), for the creation of synthetic MRI scans for enhanced deep neural network training (Shin et al. 2018), for motion transfer in videos (Chan et al. 2018) and many more cases where the data distribution is anything but 'easy' to express mathematically (if not impossible).

For our particular needs we construct a PYTORCH (Paszke et al. 2017) implementation of Wasserstein GANs with gradient penalty (hereafter WGAN-GP, Gulrajani et al. 2017). We chose WGAN-GP because it is one of the most reliable GAN frameworks for stability in training. The architectures we used for the G and D ANNs are summarized in Table 3. The input to the generator is a random 10D multinomial distribution, $z \sim \mathcal{N}(0, 1)^{10} \in \Re^{10}$. In Table 4 we detail the hyper parameter values we used during GANs training. In addition, in order to avoid overfitting the NonPlumCoreOM 1k data set, we augmented the data with random rotations on the x, y plane and reflections with respect to x and y axis. In particular we followed the transformations $(x, y, v_{\rm los}) \rightarrow (-x, y, -v_{\rm los})$ and $(x, y, v_{\rm los}) \rightarrow (x, -y, -v_{\rm los})$. For zero mean $v_{\rm los}$ stellar systems, these reflections are like observing the target from the opposite direction of the initial observer: clearly the physics of the system should not change. This type of information should be viewed as 'prior knowledge encoding' of the modelling process with neural networks.

Table 3. Generator and Discriminator network architectures. We follow PYTORCH semantics to denote the dimensionality and type of the layers and non-linear activations we used. Here `LDIM` = 10 is the dimensionality of the latent space that we sample and feed into the Generator, `DIM` = 512 is the number of features in the linear Layers, and `XDIM` = dim$(x, y, v_{\rm los}) = 3$ is the dimensionality of the projected observation space.

Layer	Generator	Discriminator
1	Linear(LDIM,DIM)	Linear(XDIM,DIM)
Activation	LeakyReLU(α = 0.01)	LeakyReLU(α = 0.01)
2	Linear(DIM, DIM)	Linear(DIM, DIM)
Activation	LeakyReLU(α = 0.01)	LeakyReLU(α = 0.01)
3	Linear(DIM, DIM)	Linear(DIM, DIM)
Activation	LeakyReLU(α = 0.01)	LeakyReLU(α = 0.01)
4	Linear(DIM, XDIM)	Linear(DIM, 1)

Table 4. Training hyperparameters of the GANs system. `NBATCH` is the batch size, `NCRITIC` is the number of training iterations that the D performs for a single G training iteration, `LDIM` is the dimensionality of the input random number to the G. For the gradient descent we used the Adam optimizer (Kingma & Ba 2014). The input data set that the D was trained on was the NonPlumCoreOM 1k.

Parameter	Value
NBATCH	128
NCRITIC	5
LDIM	10
Optimizer	Adam (`lr` = 1e-4,$\beta_1 = 0.5$, $\beta_2 = 0.9$)

In Fig. 3 we plot on the $(R, v_{\rm los})$ plane the synthetic data generated from the GANs against the 1k and 10k NonPlumCoreOM data sets. We generated ~25k synthetic data points by training the Discriminator, D, on the NonPlumCoreOM 1k data set. This resulted in approximately 160 $\sigma^2_{\rm los}$ binned values for the LOS velocity dispersion profile. In Fig. 4 we plot the LOS velocity dispersion profile from the GAN data as well as the true 1k and 10k dispersion profiles. In all panels the reference profile (dashed curve) is overplotted. Clearly, the GAN generated profile is of high quality. In fact, the uncertainty of the data points around the reference profile is smaller than even the case of the original 10k data set. This happens because the GAN system learns more information of the underlying distribution from the NonPlumCoreOM 1k data set than what the moments of the 10k sample can describe. As a result, with higher number of targets (25k) we end up with a LOS velocity dispersion profile of smaller uncertainty than the 10k original data set. A small bias is apparent in the last two $\sigma^2_{\rm los}$ data points, probably because the GANs overfit the outliers at the edges of the radial distance of the 1k data set. This bias may also be due to the system of GANs not having reached the optimum equilibrium when we terminated training. Finally in Fig. 5 we compare the projected density (brightness for $\Upsilon_{\rm V} = 1$) of the tracer population. It should be noted that we did not experiment with new architectures, training schemes, or hyperparameter optimization. We just used the proposed implementation scheme from Gulrajani et al. (2017) for their toy model of 25 2D Gaussian distributions. There is huge scope for improvement and adaptation for individual data sets of this technique for data augmentation in astronomy in various sub-disciplines. Here, we are merely scratching the surface of the potential of this technology.

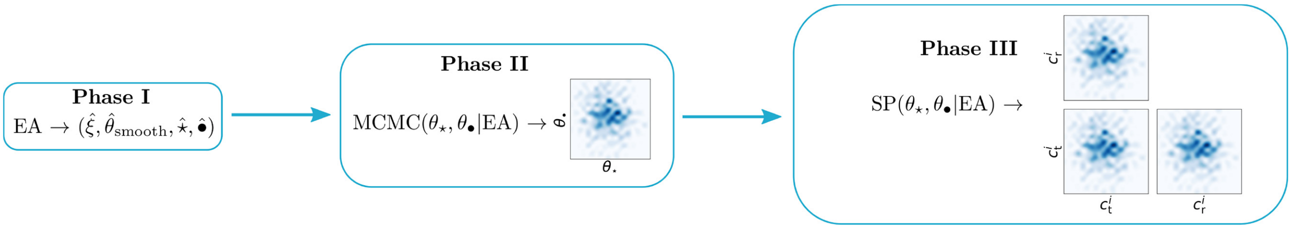
Phase I
EA → $(\hat{\xi}, \hat{\theta}_{\rm smooth}, \hat{\star}, \hat{\bullet})$
Phase II
MCMC$(\theta_\star, \theta_\bullet|{\rm EA}) \to \theta_\bullet$
$\theta_\bullet$
Phase III
SP$(\theta_\star, \theta_\bullet|{\rm EA}) \to$
c_r^i
c_t^i
c_t^i
c_r^i

Figure 2. The t-JEANS algorithm.

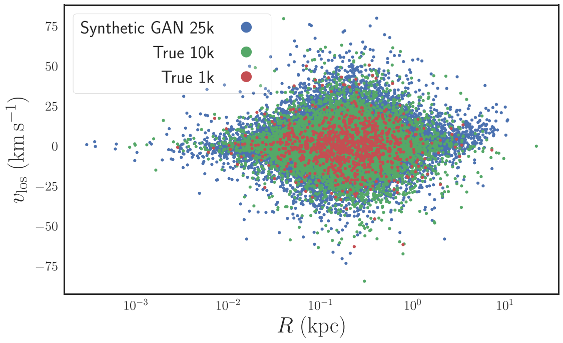

Figure 3. Comparison of NonPlumCoreOM 1k, 10k true data sets against the GAN generated synthetic data on the $(R, v_{\rm los})$ plane.

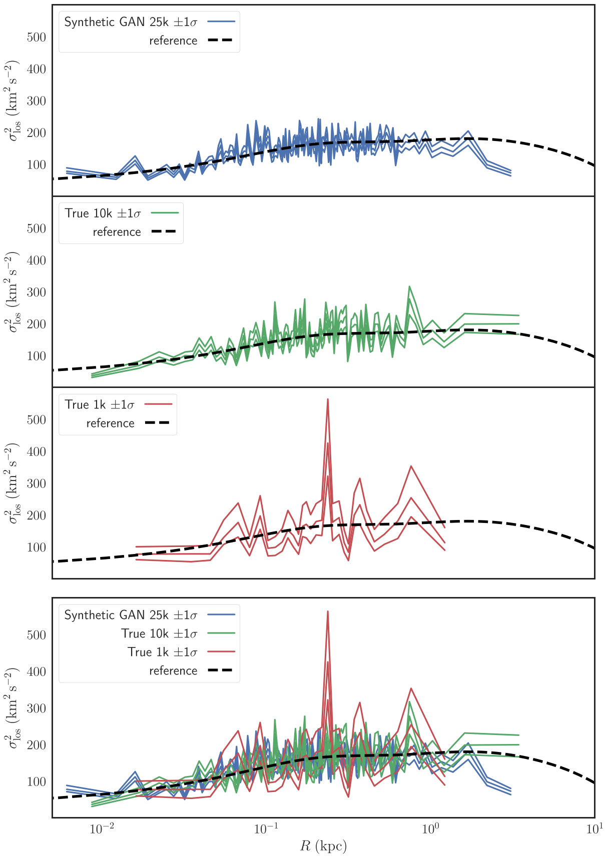

Figure 4. Comparison of NonPlumCoreOM 1k, 10k true dispersion profiles against the GAN generated dispersion profile. Overplotted is the true reference profile. The GAN generated profile was created from information from the NonPlumCoreOM 1k data set only.

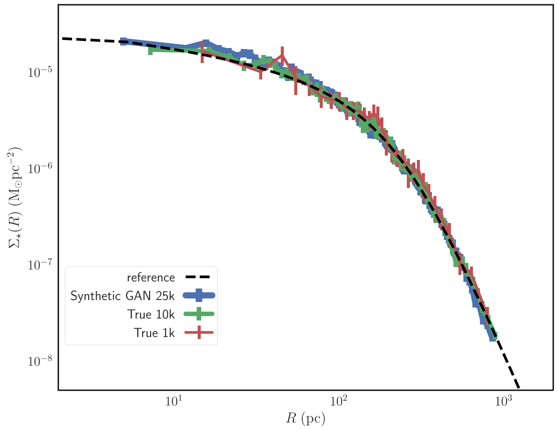

Figure 5. Comparison of NonPlumCoreOM 1k, 10k true projected density profiles against the GAN generated projected density profile. Overplotted is the true reference profile.

4 THE T-JEANS SOLVER

In this section we present an overview of the t-JEANS algorithm we developed for accurate mass estimates in spherically symmetric self-gravitating systems.

The JEANS (Diakogiannis et al. 2017) algorithm is a numerical solver that estimates the mass content and the kinematic profile of spherically symmetric gravitating systems. It models independent of anisotropy, $\beta(r)$, assumptions and it requires parametric functional forms for the mass density profiles. The best mass model is selected with the use of model selection criteria (Diakogiannis et al. 2017) (Akaike Information Criterion, Sugiura 1978; Burnham & Anderson 2002, hereafter AICc). The radial velocity dispersion profile, $\sigma^2_{\rm rr}$, is represented as a ‘free form’ B-spline function, $\sigma^2_{\rm rr}(r) = \sum_i a^i B_i(r)$. The correct kinematic profile is inferred from the data. The solver uses information of brightness and LOS velocity moments, $\sigma^2_{\rm los}$, to estimate marginalized distributions of the mass model parameters as well as the coefficients, a^i, that describe the radial dispersion profile, $\sigma^2_{\rm rr}$.

The algorithm consists of three distinct phases. In the first phase it evaluates the *simplest* kinematic profile that gives a satisfactory representation[6] to the data, as well as the most probable mass model. This is achieved with the use of evolutionary optimization and quadratic programming. In the second phase, JEANS evaluates the optimum smoothing parameters from ideal theoretical models. Finally, in phase three the algorithm performs MCMC inference, for the determination of marginalized distributions of the model parameters.

[6] That is, the best B-spline basis, $B_i(x)$, according to the bias-variance trade-off (Hastie, Tibshirani & Friedman 2001).

The new version t-JEANS is significantly modified compared to the previously published version (Diakogiannis et al. 2017). In the first phase we again evaluate the optimum B-spline basis, as well as the statistically most favoured mass model. We introduce a new quadratic programming formalism – the Dynamic Moments Solver (hereafter DMS) – for the numerical solution of the system of coupled integrodifferential equations (1) and (2). In the latest version of the JEANS we expand both the radial, σ_{rr}^2, and tangential, σ_{tt}^2, profiles in a B-spline basis of the order of $k = 4$ (degree $= 3$),[7] i.e. $\sigma_{\mathrm{rr}}^2(r) = c_{\mathrm{r}}^i B_i(r)$, $\sigma_{\mathrm{tt}}^2(r) = c_{\mathrm{t}}^i B_i(r)$. This allows us to treat the Jeans equation as a local, r_i, constraint in the quadratic optimization problem of estimating the velocity moments, σ_{rr}^2, σ_{tt}^2. In combination with the local support of B-spline functions, this translates to more equations for the unknown coefficients c_{r}^i, c_{t}^i that further reduce the feasible solution space. In comparison with the old version of the JEANS, by solving the Jeans equation (16) with respect to σ_{tt}^2 and substituting under the integral sign of the σ_{los}^2 definition (equation 14), we loose the local equations that c_{r}^i and c_{t}^i coefficients participate after the last datum. By keeping the Jeans equation as a constraint we can evaluate equations for c_{r}^i and c_{t}^i in all space $r \in [0, r_{\mathrm{vir}}]$. This has a direct positive impact on the quality of the recovered anisotropy profile, $\beta(r)$.

In a similar fashion to the first version of the JEANS, we do not invert the dynamical equations, thus we avoid the problem of having to integrate/differentiate noisy numerical functions. We also include additional global and local constraints that guarantee that the kinematic profiles lead to physically acceptable solutions ($\sigma_{\mathrm{rr}}^2, \sigma_{\mathrm{tt}}^2 \geq 0,\ \forall r \in [0, r_{\mathrm{vir}})$). The fitness function is modified in order to include information from the full LOS kinematics. The optimum smoothing parameters are now evaluated directly from the data according to the best bias-variance tradeoff using a validation data set, D_{val}. The model selection is performed using a hold out test data set, D_{test}. In phase two we perform MCMC inference for the unknown stellar and DM mass model parameters, $\theta_\star$, $\theta_\bullet$. In this phase, the kinematic profile is treated as a nuisance parameter. Finally in the third phase, we perform stochastic programming (SP) in order to determine confidence intervals for the velocity dispersion profiles, $\sigma_{\mathrm{rr}}^2, \sigma_{\mathrm{tt}}^2, \sigma_{\mathrm{los}}^2$.

In more detail (Fig. 2), the distinct phases of the t-JEANS are the following:

(1) An evolutionary optimization (EA) phase. In this phase we determine: (i) the simplest (best) B-spline basis[8] for the representation of the unknown radial, σ_{rr}^2, and tangential, σ_{tt}^2 velocity dispersions, (ii) the best candidate mass models and, (iii), the best smoothing[9] parameters, $\theta_{\mathrm{smooth}} = \{\lambda_1, \beta_1, \lambda_2, \beta_2\}$. For the evaluation of the smoothing parameters we use a validation data set, D_{val}, created from random sampling from the LOS σ_{los}^i marginalized distributions (see Fig. 1). The optimum smoothing parameters are the ones that minimize the validation error for all random samples, σ_{los}^i. We give more details of this process in the section where we describe the fitness function. For the model selection, we use a 'hold-out' LOS moments test data set, D_{test} (Section 3.1), and we perform model selection (Section 4.8) based on the out-of-sample prediction error (*generalization test error*). This approach gives more robust model selection (in comparison with predictive information criteria; Gelman, Hwang & Vehtari 2014), since it heavily penalizes models that do not generalize well on unseen data. It should be stressed, however, that the efficiency of the model selection process depends crucially on the number of available data points.

(2) A MCMC analysis, keeping the B-spline basis and the smoothing parameters fixed, for the best mass model. In this scheme, the radial and tangential coefficients, c_{r}^i, c_{t}^i are treated as nuisance parameters: they are estimated at each iteration from the DMS. This phase produces marginalized distributions of the parameters of stellar, $\theta_\star$, and DM, $\theta_\bullet$, mass densities, $\{\rho_\star, \rho_\bullet\}$.

(3) A stochastic programming (SP) phase, where the $\theta_\star$, $\theta_\bullet$ parameters are used iteratively in the DMS. This produces marginalized distributions for the radial and tangential coefficients, c_{r}^i, c_{t}^i, *subject to local and global dynamical constraints*. This last phase gives the required uncertainty of LOS and radial and tangential velocity dispersions.

4.1 Mass models

For our modelling purposes we used the following candidate mass models:

$$\rho(r) = \begin{cases} \dfrac{\rho_0}{[1+(r/r_{\mathrm{s}})^2]^{5/2}} & \text{Plummer} \\ \dfrac{\rho_0}{(1+r/r_{\mathrm{s}})[1+(r/r_{\mathrm{s}})^2]} & \text{Burkert} \\ \dfrac{r_{\mathrm{s}}^3 \rho_0}{r(r^2+r_{\mathrm{s}}^2)^2} & \text{NFW} \\ \text{Eq (8)} & \text{generalized Hernquist} \end{cases} . \tag{10}$$

We model each data set with two different mass model assumptions, the correct one and an incorrect one. Our goal is to demonstrate that given sufficient data it is possible, in principle, to statistically infer the most probable model using model selection criteria.

4.2 Dynamic moments solver

In this section we describe the mathematical representation of the problem, i.e. the dynamic equations that enable us to recover the radial and tangential velocity moments, from knowledge of the LOS velocity dispersion, σ_{los}^2, the tracer, $\rho_\star$, and the DM, $\rho_\bullet$, mass densities. The DMS solves the system of coupled integrodifferential equations (1 and 2) by discretizing the solution space using B-splines. This is achieved by expanding the unknown radial, σ_{rr}^2, and tangential, σ_{tt}^2, velocity moments in a B-spline basis[10]

$$\sigma_{\mathrm{rr}}^2(r) = c_{\mathrm{r}}^i B_i(r) \tag{11}$$

$$\sigma_{\mathrm{tt}}^2(r) = c_{\mathrm{t}}^i B_i(r). \tag{12}$$

The DMS takes as input the knots, ξ_i, the stellar parameters, $\theta_\star$, the DM parameters, $\theta_\bullet$, and the smoothing penalty variables, $\theta_{\mathrm{smooth}} = \{\lambda_1, \beta_1, \lambda_2, \beta_2\}$ and gives as output the coefficients c_{r}^i, c_{t}^i that fully describe the radial and tangential velocity moments. Using the approximation equations (11) and (12), the task is transformed to a convex optimization problem (quadratic programming). The software library we use in t-JEANS for the quadratic optimization is IBM's CPLEX.[11]

[7] The lower the degree of the B-spline basis, the smaller the condition number of the system of equations.

[8] Equivalently, the knots ξ_i that define the simplest basis.

[9] The description of each of the four smoothing parameters, θ_{smooth}, is given in Section 4.6.

[10] We use Einstein summation convention, where double repeated indices indicate summation. E.g. $\sigma_{\mathrm{rr}}^2(r) = a^i B_i(r) \equiv \sum_{i=1}^{n_{\mathrm{basis}}} a^i B_i$.

[11] Free academic license.

Table 5. Summary of the DMS solver (DMS($\theta|D$) → ($c_{\rm r}$, $c_{\rm t}$)) in the JEANS modelling approach. **Assumptions:** (i) Spherical symmetry, (ii) virial equilibrium, (iii) parametric form for the stellar and DM mass profiles. **Input:** $\theta \equiv \{\theta_\star, \theta_\bullet, \xi, \lambda_{1,2}, \beta_{1,2}\}$ and training data, $D = \{R_i, \sigma^i_{\rm los}, \delta\sigma^i_{\rm los}\}$. **Output:** $\sigma^2_{\rm tt} = c^i_{\rm r} B_i$, $\sigma^2_{\rm rr} = c^i_{\rm t} B_i$.

		Mathematical formula
Objective function		$\min \mathcal{F} = \sum_i (\sigma^2_{\rm los}(R_i) - \boldsymbol{\sigma}^i_{\rm los})^2 + \lambda_1[\sum_{i=1}^{n_{\rm coeff}-1} \beta_1(\Delta^1 c^i_{\rm r})^2 + (1-\beta_1)(\Delta^2 c^i_{\rm r})^2]$ $+\lambda_2[\sum_{i=1}^{n_{\rm coeff}-2} \beta_2(\Delta^1 c^i_{\rm t})^2 + (1-\beta_2)(\Delta^2 c^i_{\rm t})^2]$
Model function		$\sigma^2_{\rm los}(R) = c^i_{\rm r}{\rm I}^{\rm r}_i(R) + c^i_{\rm t}{\rm I}^{\rm t}_i(R)$
Local constraints		
	Jeans:	$-\rho_\star \frac{{\rm d}\,\Phi}{{\rm d}\,r} = \left(\frac{{\rm d}(\rho_\star B_i)}{{\rm d}x} + \frac{2\rho_\star B_i}{x}\right) c^i_{\rm r} + \frac{\rho_\star B_i}{x} c^i_{\rm t}$
	sign:	$c^j_{\rm r} B_j(x) \geq 0, \quad c^j_{\rm t} B_j(x) \geq 0$
	boundary:	$\sigma^2_{\rm rr}(0) = \sigma^2_{\rm tt}(0)/2$ $\sigma^2_{\rm rr}(r_{\rm vir}) = \sigma^2_{\rm tt}(r_{\rm vir}) = 0$
Global constraints		
	projected virial:	$2(K^{\rm Rlos}_i c^i_{\rm r} + K^{\rm Tlos}_i c^i_{\rm t}) + W = 0$

For clarity in notation, it is convenient to represent the DMS as a function:

$$\mathrm{DMS}(\theta|D_{\rm train}) \rightarrow (c^i_{\rm r}, c^i_{\rm t}),$$

where $\theta \equiv \{\boldsymbol{\xi}, \theta_\star, \theta_\bullet, \theta_{\rm smooth}\}$ are the parameters that define the B-spline basis, the tracer, and DM profiles, as well as the smoothing penalty regularization. The goal of the DMS is to minimize the training error of the LOS velocity dispersion:

$$\chi^2_{\rm train} = \sum_i^{N_{\rm bins}} \left(\frac{\sigma^2_{\rm los}(R_i) - \sigma^i_{\rm los}}{\delta\sigma^i_{\rm los}}\right)^2 \qquad (13)$$

subject to various local and global dynamic equations (constraints). We separate these constraints into local, boundary, and global constraints. In addition we will impose some regularization conditions (smoothing) in the minimization process, in order to reduce the condition number of the linear system and avoid oscillatory solutions. We formally define the objective function of the DMS in Section 4.6. In Table 5 we summarize the system of equations and the objective function that fully describe the DMS. We proceed by stating exactly the mathematical equations we use in the t-JEANS.

The LOS velocity dispersion under the B-spline approximation of the velocity moments is given by

$$\sigma^2_{\rm los} = \frac{1}{\Sigma_\star(R)}\left(\int_R^{r_{\rm vir}} \rho_\star K_1 \sigma^2_{\rm rr}\,{\rm d}\,r + \int_R^{r_{\rm vir}} \rho_\star K_2 \sigma^2_{\rm tt}\,{\rm d}\,r\right), \qquad (14)$$

where

$$\Sigma_\star(R) = \int_R^{r_{\rm vir}} \rho_\star K_3\,{\rm d}\,x$$

is the projected tracer mass density and

$$K_1(r, R) = \frac{2(r^2 - R^2)}{r\sqrt{r^2 - R^2}}$$

$$K_2(r, R) = \frac{R^2}{r\sqrt{r^2 - R^2}}$$

$$K_3(r, R) = \frac{2r}{\sqrt{r^2 - R^2}}$$

are kernel functions. Applying the B-spline approximation (equations 11 and 12) and defining

$${\rm I}^{\rm r}_i(R) = \frac{1}{\Sigma_\star}\int_R^{r_{\rm vir}} \rho_\star K_1(r, R) B_i(r)\,{\rm d}\,r$$

$${\rm I}^{\rm t}_i(R) = \frac{1}{\Sigma_\star}\int_R^{r_{\rm vir}} \rho_\star K_2(r, R) B_i(r)\,{\rm d}\,r,$$

the linearized LOS velocity dispersion takes the form

$$\sigma^2_{\rm los}(R) = c^i_{\rm r}{\rm I}^{\rm r}_i(R) + c^i_{\rm t}{\rm I}^{\rm t}_i(R). \qquad (15)$$

This is the model function that we compare with observables, subject to physical constraints. It is linear with respect to the unknown coefficients, $c^i_{\rm r}$, $c^i_{\rm t}$, something that simplifies the solution and allows for convex optimization.

4.3 Local constraints

These constraints are termed local, because they are valid in the whole extent of the system, $r \in [0, r_{\rm vir}]$. We evaluate these at the positions of the Greville abscissae of the B-spline basis.

4.3.1 Jeans constraints

The spherically symmetric Jeans equation (SSJE) is

$$-\,\rho_\star\frac{{\rm d}\,\Phi}{{\rm d}\,r} = \frac{{\rm d}(\rho_\star\sigma^2_{\rm rr})}{{\rm d}\,r} + \rho_\star\frac{(2\sigma^2_{\rm rr} - \sigma^2_{\rm tt})}{r}. \qquad (16)$$

The linearized form of SSJE that results from the B-spline approximation is

$$-\,\rho_\star\frac{{\rm d}\,\Phi}{{\rm d}\,r} = \left(\frac{{\rm d}(\rho_\star B_i)}{{\rm d}\,r} + \frac{2\rho_\star B_i}{r}\right)c^i_{\rm r} + \frac{\rho_\star B_i}{r}c^i_{\rm t}. \qquad (17)$$

4.3.2 Sign constraints

We demand the velocity moments to be positive in all solution space:

$$\sigma^2_{\rm rr}(r) \geq 0$$

$$\sigma^2_{\rm tt}(r) \geq 0.$$

In terms of the kinematic coefficients:

$$c_{\mathrm{r}}^{j} B_j(r) \geq 0 \tag{18}$$

$$c_{\mathrm{t}}^{j} B_j(r) \geq 0. \tag{19}$$

4.4 Boundary constraints

These apply at the origin and at the virial radius of the system.

$$\sigma_{\mathrm{rr}}^2(0) = \sigma_{\mathrm{tt}}^2(0)/2$$
$$\sigma_{\mathrm{rr}}^2(r_{\mathrm{vir}}) = \sigma_{\mathrm{tt}}^2(r_{\mathrm{vir}}) = 0.$$

The reasoning for the $\sigma_{\mathrm{rr}}^2(0) = \sigma_{\mathrm{tt}}^2(0)/2$ boundary condition is the following: we expect that all tangential motions at the limit $r \to 0$ become radial. That is, if we draw the tangent line to a circle of radius r, as the radius approaches zero, the tangent line approaches the origin $r = 0$ of the coordinate system. In the limiting case where $r \to 0$ the tangent line passes from the origin (it is actually a degenerate case: all directions are equivalent). In this respect it is our understanding that in this limit the tangential and radial motions are indistinguishable. This is why we expect that their dispersions will be equal at $r \to 0$. With regards to the second boundary constraint, it is proven (Dejonghe & Merritt 1992) that for a self-consistent system in virial equilibrium the radial and tangential velocity dispersions vanish in the limit of the virial radius.

4.5 Global constraints

In this category fall constraints of local functions are integrated over all space.

4.5.1 *Projected virial theorem*

The virial theorem states (Binney & Tremaine 2008; Merritt 2013) that if K is the total kinetic energy of a system, and $\mathcal{W}$ its total potential energy, then for a system in dynamic equilibrium:

$$2K + \mathcal{W} = 0. \tag{20}$$

For a spherically symmetric system,

$$\mathcal{W} = 4\pi \int_0^{r_{\mathrm{vir}}} \rho_\star \left(-\frac{\mathrm{d}\,\Phi}{\mathrm{d}\,r}\right) r^3\,\mathrm{d}\,r.$$

The total kinetic energy of a system, defined via the LOS velocity dispersion is

$$K^{\mathrm{los}} = \frac{3}{2}\int_{R=0}^{r_{\mathrm{vir}}} \mathrm{d}\,R\, 2\pi R \Sigma_\star(R)\, \sigma_{\mathrm{los}}^2(R).$$

Substituting $\sigma_{\mathrm{los}}^2(R)$ from equation (15), we have

$$K^{\mathrm{los}} = c_{\mathrm{r}}^{i} K_i^{\mathrm{Rlos}} + c_{\mathrm{t}}^{i} K_i^{\mathrm{Tlos}}, \tag{21}$$

where

$$K_i^{\mathrm{Rlos}} = 3\pi \int_{R=0}^{r_{\mathrm{vir}}} R\Sigma_\star(R)\mathrm{I}_i^{\mathrm{r}}(R)\,\mathrm{d}R$$
$$K_i^{\mathrm{Tlos}} = 3\pi \int_{R=0}^{r_{\mathrm{vir}}} R\Sigma_\star(R)\mathrm{I}_i^{\mathrm{t}}(R)\,\mathrm{d}R.$$

Substituting in equation (20) yields

$$2(K_i^{\mathrm{Rlos}} c_{\mathrm{r}}^{i} + K_i^{\mathrm{Tlos}} c_{\mathrm{t}}^{i}) + \mathcal{W} = 0. \tag{22}$$

This is an additional constraint on the $c_{\mathrm{r}}^{j}, c_{\mathrm{t}}^{j}$ coefficients. From the perspective of linear/quadratic programming algorithmic structure, equation (22) is a hyperplane equation with respect to the unknown coefficients, $c_{\mathrm{t}}^{j}, c_{\mathrm{r}}^{j}$, that further reduces feasible solution space.

The projected virial theorem is also a hard bound on the value of the total gravitational energy of the stellar and dark matter ($\star\bullet$) interaction. Furthermore, it is clear that since the K^{los} value is independent of the anisotropy profile (i.e. it is an observational fact), then it is impossible to have only the stellar component with some peculiar anisotropy profile to represent the observables. In other words, the total gravitational energy of the system is fixed from the total kinetic energy as this is estimated from the LOS dispersion, σ_{los}^2. That is, the constraint of virial equilibrium does not allow one to vary the anisotropy profile β to fit any desired mass profile.

4.6 Objective function

The DMS objective function that relates observables, σ_{los}^2, with the model function (equation 15) is given by

$$\begin{aligned}\mathcal{F} = &\sum_i \left(\frac{\sigma_{\mathrm{los}}^2(R_i) - \sigma_{\mathrm{los}}^i}{\delta\sigma_{\mathrm{los}}^i}\right)^2 \\ &+ \lambda_1 \left[\beta_1 \sum_{i=1}^{n_{\mathrm{coeff}}-1} (\Delta^1 c_{\mathrm{r}}^i)^2 + (1-\beta_1) \sum_{i=1}^{n_{\mathrm{coeff}}-2} (\Delta^2 c_{\mathrm{r}}^i)^2\right] \\ &+ \lambda_2 \left[\beta_2 \sum_{i=1}^{n_{\mathrm{coeff}}-1} (\Delta^1 c_{\mathrm{t}}^i)^2 + (1-\beta_2) \sum_{i=1}^{n_{\mathrm{coeff}}-2} (\Delta^2 c_{\mathrm{t}}^i)^2\right],\end{aligned} \tag{23}$$

where the difference operators $\Delta^{1,\,2}$ are defined by

$$\Delta^1 c_{\mathrm{r,t}}^i = c_{\mathrm{r,t}}^{i+1} - c_{\mathrm{r,t}}^i$$
$$\Delta^2 c_{\mathrm{r,t}}^i = \Delta^1(\Delta^1 c_{\mathrm{r,t}}^i) = c_{\mathrm{r,t}}^{i+2} - 2c_{\mathrm{r,t}}^{i+1} c_{\mathrm{r,t}}^i + c_{\mathrm{r,t}}^i.$$

The coefficients, $\lambda_{1,2}$ regulate the amount of smoothing penalty on each of the velocity dispersions. The coefficients $\beta_{1,2}$ regulate the relative contribution of the first and second derivative penalties for each velocity dispersion. This smoothing penalty is efficient and very fast to evaluate in comparison with previous efforts (Diakogiannis, Lewis & Ibata 2014b; Diakogiannis et al. 2017). It is the same penalty used in the P-splines (Eilers, Rijnmond & Marx 1996) formulation in statistical smoothing.

4.7 Fitness function

The EA phase of the t-JEANS solver evaluates the simplest B-spline basis that best represents the observables. This is a nested optimization: the EA parameters consist of the stellar, $\theta_\star$, the DM, $\theta_\bullet$, and the smoothing penalty variables, $\theta_{\mathrm{smooth}} = \{\lambda_1, \beta_1, \lambda_2, \beta_2\}$. Once these parameters, $\theta = \{\theta_\star, \theta_\bullet, \theta_{\mathrm{smooth}}\}$, are proposed, then the problem is a quadratic programming optimization problem, with respect to the $c_{\mathrm{r}}^{i}, c_{\mathrm{t}}^{i}$ unknown constants. The optimal variables, $\hat{c}_{\mathrm{r}}^{\,i}, \hat{c}_{\mathrm{t}}^{\,i}$ for the proposed θ parameters are evaluated with the DMS. The evaluation of the model though, takes into account information from both the DMS and the full kinematics. For the full kinematics, we use definitions (Mamon, Biviano & Boué 2013) based on assumptions of a Gaussian distribution function for the velocities (in 3D space), truncated at the escape velocity of the system.

The fitness function is defined with the usage of model selection criteria (BIC, AICc) and the following penalty functions:

$$\mathrm{AICc} = 2\sum_i \frac{1}{2}\left(\frac{\hat{\sigma}^2_{\mathrm{los}}(R_i) - \sigma^i_{\mathrm{los}}}{\delta\sigma^i_{\mathrm{los}}}\right)^2 + 2n + \frac{2n(n+1)}{N_{\mathrm{data}} - n - 1}$$

$$\chi_{\mathrm{smooth}} = \frac{1}{N_{\mathrm{sample}}}\sum_{j=1}^{N_{\mathrm{sample}}}\sum_{i=1}^{N_{\mathrm{bins}}}\left(\frac{\hat{\sigma}^2_{\mathrm{los}}(R_i) - \sigma^{ij}_{\mathrm{los}}}{\delta\sigma^i_{\mathrm{los}}}\right)^2$$

$$\chi_{\mathrm{virial}} = \left|\frac{2K^{\mathrm{los}}}{W} - \frac{W}{2K^{\mathrm{los}}}\right|$$

$$\mathrm{BIC} = -2\sum_{i=1}^{N_{\mathrm{batch}}} \log q(R_i, v^i_{\mathrm{los}})) + n\log(N_{\mathrm{batch}}),$$

where

$$q(R, v_{\mathrm{los}}) = \frac{2\pi R}{M^{\star}_{\mathrm{tot}}} g(R, v_{\mathrm{los}}) \tag{24}$$

$$g(R, v_{\mathrm{los}}) = \int_R^{r_{\mathrm{vir}}} \frac{r\rho_{\star}(r)}{\sqrt{r^2 - R^2}} h(v_{\mathrm{los}}|R, r)\mathrm{d}r \tag{25}$$

$$h(v_{\mathrm{los}}|R, r) = \frac{\exp[-\frac{v^2_{\mathrm{los}}}{2\sigma^2_z(R,r)}]}{\sqrt{2\pi\sigma^2_z(R, r)}\,\mathrm{erf}\{v_{\mathrm{esc}}(R)/\sqrt{2\sigma^2_z(R, r)}\}} \tag{26}$$

$$\sigma^2_z(R, r) = \sigma^2_{\mathrm{rr}}(1 - (R/r)^2) + \sigma^2_{\mathrm{tt}}(R/r)^2/2 \tag{27}$$

are the full kinematics definitions from the MAMPOSSt (Mamon et al. 2013) algorithm. The projected virial theorem is satisfied by the quadratic programming solver, within some numerical tolerance. We found that we got slightly faster convergence by also penalizing this explicitly in the EA solver, with the χ_{virial} term.

The values $\hat{\sigma}^2_{\mathrm{los}}(R_i)$ are the solutions from the DMS for the given input parameters $\theta = \{\boldsymbol{\xi}, \theta_{\star}, \theta_{\bullet}, \lambda_{1,2}, \beta_{1,2}\}$. The data values, σ^i_{los} and $\sigma^{ij}_{\mathrm{los}}$ are produced from a binning scheme as described in Section 3.1. We remind the reader that the values $\sigma^{ij}_{\mathrm{los}} \in D_{\mathrm{val}}$ are j sampled values for each bin i. They are used as a validation set for determining the smoothing parameters, $\theta_{\mathrm{smooth}} = \{\lambda_1, \beta_1, \lambda_2, \beta_2\}$. N_{batch} is the size of a random sample (without replacement) of full kinematics data of stars from the population. We use $N_{\mathrm{batch}} = 1000$: this provides a good approximation to the full kinematics likelihood and allows for faster convergence.

The fitness function, treated as a maximization problem, is the product of four components, namely

$$f_{\mathrm{DMS}} = \frac{1}{1 + \mathrm{AICc}}$$

$$f_{\mathrm{full\,kin}} = \frac{1}{1 + \mathrm{BIC}}$$

$$f_{\mathrm{smooth}} = \frac{1}{1 + \chi_{\mathrm{smooth}}}$$

$$f_{\mathrm{vir}} = \frac{1}{1 + \chi_{\mathrm{vir}}}.$$

Then, the fitness function is

$$f_{\mathrm{tot}}(\theta) = f_{\mathrm{DMS}}\ f_{\mathrm{full\,kin}}\ f_{\mathrm{smooth}}\ f_{\mathrm{vir}}. \tag{28}$$

4.8 Model selection

Model selection takes place in two distinct processes inside the t-JEANS. Once we select a set of tracer and DM mass densities, we use the EA in order to find the simplest B-Spline basis for the radial and tangential velocity dispersions. This task is a hierarchical model selection problem (where the various competing models are the ones that have different number and locations of knots, but the same mass density parametric form). For this task, AICc or BIC, based on the training error measure (likelihood) prove to be good choices.

However, when one needs to compare competing mass models that were trained in distinct EA phases, it is best to use out-of-sample data and test how well the model generalizes on unseen (during training) data (Section 3). Once the EA phase is complete, for competing mass models, we evaluate the best model using the hold-out LOS moments test set in a cross-validation manner. The average error on unseen moments test data that we use is

$$\chi^2_{\mathrm{test}} = \frac{1}{2N_{\mathrm{test}}}\sum_{j=1}^{N_{\mathrm{test}}}\sum_{i=1}^{N_{\mathrm{bin}}}\left(\frac{\hat{\sigma}^2_{\mathrm{los}}(R_i) - \tilde{\sigma}^{ij}_{\mathrm{los}}}{\delta\sigma^i_{\mathrm{los}}}\right)^2, \tag{29}$$

where $\tilde{\sigma}^{ij}_{\mathrm{los}}$ are out of sample test data, created for each bin i by random sampling from the marginalized distribution of the data pre-processing MCMC chains. The model with the smallest test error is selected as the best candidate. In our experiments this method has proven to be more robust than predictive training error methods (e.g. AICc), which can have bias from overfitting (Gelman et al. 2014).

4.9 Likelihood function

In Phase II of the t-JEANS we perform an MCMC exploration using the following likelihood (Ibata et al. 2013; Mamon et al. 2013; Diakogiannis, Lewis & Ibata 2014a; Diakogiannis et al. 2014b; Diakogiannis et al. 2017):

$$\mathcal{L} = \left[\prod_{j=1}^{N^{\star}_{\mathrm{bin}}} \frac{\exp\{-\frac{(\Sigma_{\star}(R_j) - \Sigma^j)^2}{2(\delta\Sigma^j_{\star})^2}\}}{\sqrt{2\pi(\delta\Sigma^j_{\star})^2}}\right] \times \left[\prod_{i=1}^{N_{\mathrm{bin}}} \frac{\exp\left\{-\frac{(\hat{\sigma}^2_{\mathrm{los}}(R_i) - \tilde{\sigma}^i_{\mathrm{los}})^2}{2\delta(\sigma^i_{\mathrm{los}})^2}\right\}}{\sqrt{2\pi\delta(\sigma^i_{\mathrm{los}})^2}}\right] \tag{30}$$

$$\times\lambda_1 e^{-\lambda_1 W_1}\lambda_2 e^{-\lambda_2 W_2}\left(\prod_{j=1}^{N_{\mathrm{batch}}} q(R_j, v^j_{\mathrm{los}})\right), \tag{31}$$

where $\Sigma_{\star}(R_j)$ is the projected tracer density at location R_j, $\Sigma^j_{\star}$, and $\delta\Sigma^j_{\star}$ the observed projected mass density and its uncertainty, $\tilde{\sigma}^i_{\mathrm{los}}$ is a random sampled value (at each iteration, we use values from D_{val}) from the ith MCMC binned histograms and W_1, W_2 are given by

$$W_1 = \beta_1 \sum_{i=1}^{n_{\mathrm{coeff}}-1} (\Delta^1 c^i_{\mathrm{r}})^2 + (1 - \beta_1)\sum_{i=1}^{n_{\mathrm{coeff}}-2} (\Delta^2 c^i_{\mathrm{r}})^2 \tag{32}$$

$$W_2 = \sum_{i=1}^{n_{\mathrm{coeff}}-1} \beta_1(\Delta^1 c^i_{\mathrm{t}})^2 + (1 - \beta_1)\sum_{i=1}^{n_{\mathrm{coeff}}-2} (\Delta^2 c^i_{\mathrm{t}})^2. \tag{33}$$

The B-spline knots, and the coefficients $\lambda_{1,2}$, $\beta_{1,2}$ are kept fixed to the values of the best EA solution. The full kinematics likelihood is calculated on each iteration on a random sample (without replacement) of $N_{\mathrm{batch}} = 1000$ stars. This is sufficient for the algorithm to converge in an excellent trade-off between computational efficiency

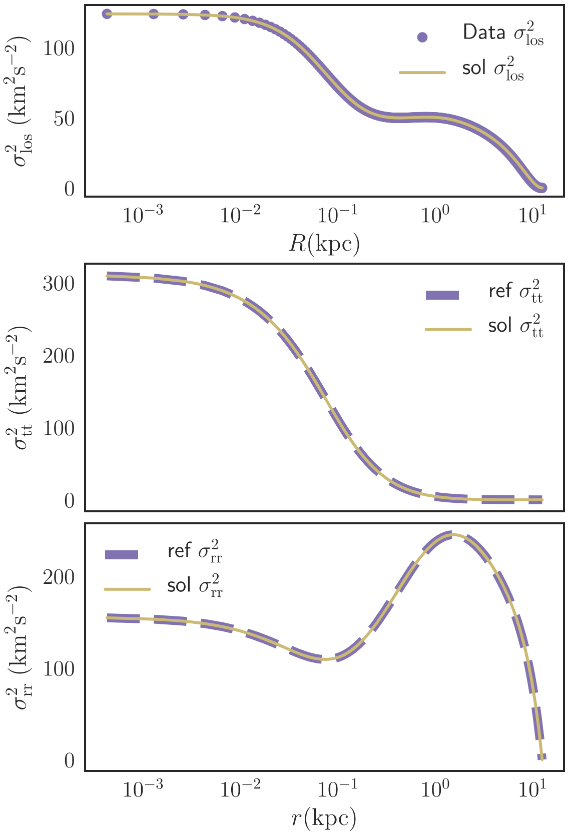

Figure 6. Exact numerical solution of the system of Jeans equations for the PlumCuspOM model using the Dynamic Moments Solver (DMS, see Table 5). In order to obtain this solution we assumed perfect knowledge of the $\sigma^2_{\rm los}$, $\rho_\star$ and $\rho_\bullet$ profiles.

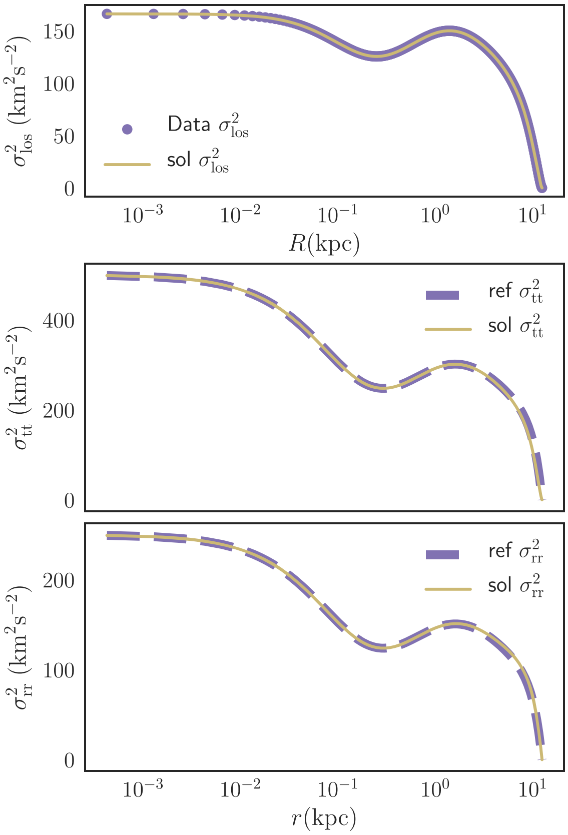

Figure 7. As Fig. 6 for the PlumCuspIso reference profile.

and parameter constraints. We use random samples, $\tilde{\sigma}^i_{\rm los}$, as data in each MCMC iteration in order to avoid overoptimistic constraints for the marginalized distributions of parameters. In this way we incorporate the uncertainty of the binned LOS dispersion values in the marginalized distributions of the $\theta_\star$, $\theta_\bullet$ parameters.

4.10 Stochastic programming

Once we have clean (after burn-in phase) MCMC chains of the $\theta_\star$, $\theta_\bullet$ parameters, we estimate distributions of $c_{\rm r}$, $c_{\rm t}$ by applying the DMS solver iteratively to each pair of MCMC values $\theta^j_\star$, $\theta^j_\bullet$. Here, the index j indicates the jth MCMC chain. For this computation we keep the smoothing penalty parameters, as well as the B-spline basis, fixed to the best EA values. In functional form

$$\mathrm{PDF}(c^i_{\rm r}, c^i_{\rm t}) = \mathrm{DMS}\left(\theta^j_\star, \theta^j_\bullet \sim \mathrm{PDF}(\theta_\star, \theta_\bullet)|D\right), \qquad (34)$$

where the symbol $\theta^j_\star$, $\theta^j_\bullet \sim \mathrm{PDF}(\theta_\star, \theta_\bullet)$ denotes that $\theta^j_\star$, $\theta^j_\bullet$ are sampled at random from their marginalized distribution $\mathrm{PDF}(\theta_\star, \theta_\bullet)$ (estimated from the MCMC chains). Finally, from the marginalized distributions of $c_{\rm r}$, $c_{\rm t}$, $\theta_\star$, $\theta_\bullet$, we can estimate 1σ uncertainty intervals for the velocity moments and the various mass model functions.

5 RESULTS

In this section we summarize our findings for both the exact solution of the Jeans system of equations and the statistical fitting of the GAIA CHALLENGE data set.

5.1 Exact solutions

In Figs 6 and 7 we plot the exact solutions of the system of the Jeans equations (Table 5) using the DMS solver, for the case of the PlumCuspOM and PlumCuspIso reference profiles. Our aim here is to provide numerical 'proof of concept' examples of Theorem 1. That is, by assuming full knowledge of the LOS velocity dispersion profile, the tracer $\rho_\star$ and $\rho_\bullet$ mass densities, we recover a unique kinematic profile as this is described by the second-order radial, $\sigma^2_{\rm rr}$, and tangential, $\sigma^2_{\rm tt}$, velocity moments. In this approach we are not using smoothing penalty coefficients ($\lambda_{1,\,2} = 0$ in the objective function equation 23) since we have a wealth of data points. For the exact solution we use a large number of B-Spline basis, $\dim\{B_i(r)\} \sim 150$. For each of the two figures, from top to bottom panels: data $\sigma^2_{\rm los}$ and recovered solution, reference and recovered tangential velocity dispersion ($\sigma^2_{\rm tt}$) and reference and recovered radial velocity dispersion profile ($\sigma^2_{\rm rr}$).

5.2 Statistical fitting

5.2.1 10k data sets

Our results are summarized in Figs 8–11 and Table 2. We fully recover the mass content and the anisotropy profiles in a representative sample of synthetic data sets from the GAIA CHALLENGE[12] suite of mock simulations. In Fig. 8 we plot the best-fitting model, as well as the 1σ uncertainty interval for a data set with Plummer-like tracer profile, a Cuspy DM halo, and Ossipkov-Merritt (Osipkov 1979;

[12] http://astrowiki.ph.surrey.ac.uk/dokuwiki/doku.php

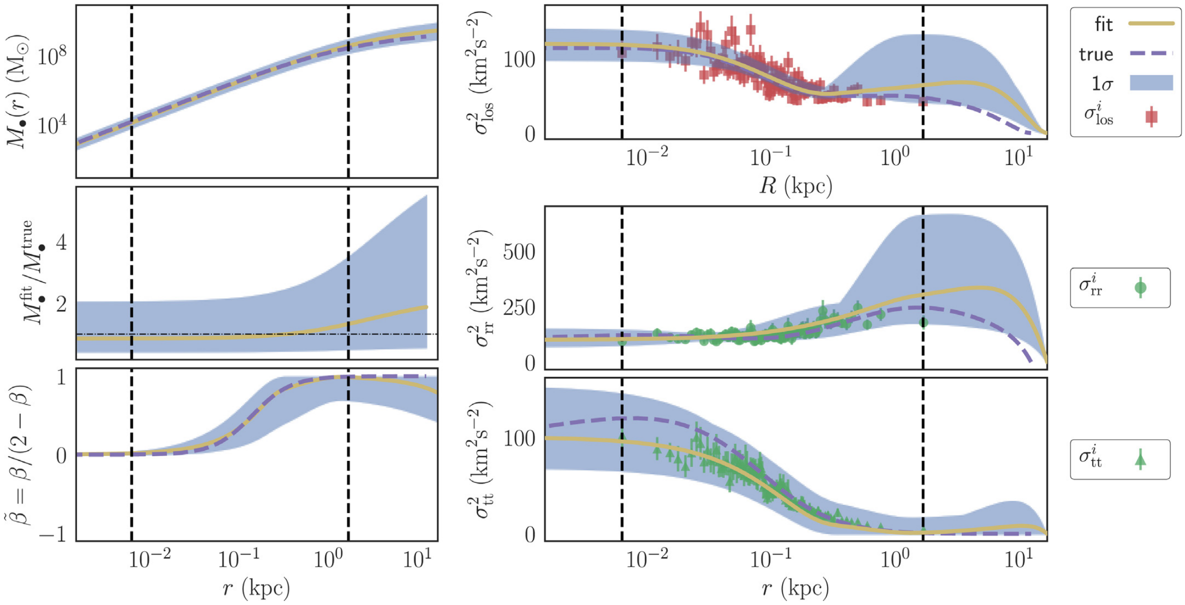

Figure 8. Fit and true reference profiles for the GAIA Challenge data set PlumCuspOM, for 10k targets. In the left-hand panels we plot (top) the DM mass, (middle) the ratio of the estimated mass over the true mass, and (bottom) the anisotropy profile. In the right-hand panels we plot (top) the fit to the LOS observables, (middle and bottom) the recovered radial and tangential profiles. Overplotted are the true $\sigma^i_{\rm rr}$, $\sigma^i_{\rm tt}$ dispersions, as estimated from the data; these were not used in the fitting process. The blue region corresponds to 1σ uncertainty for all of the quantities.

Merritt 1985) velocity anisotropy profile (PlumCuspOM), for 10k stars. In all panels, the vertical dashed lines designate the values of the first and last datum. The reliable region for making predictions is within these lines. Everything outside this region is extrapolation and cannot be trusted. Left panels, from top to bottom: estimated DM mass, the ratio of the fitted to the true DM mass, and the normalized (Read & Steger 2017) velocity anisotropy. Right panels, from top to bottom: LOS velocity dispersion fit and the data we used. The radial (middle), $\sigma^2_{\rm rr}$, and tangential (bottom), $\sigma^2_{\rm tt}$, velocity moments. The data in the middle and bottom panels were not used in the fitting process. They are produced from the true 3D kinematic information and are shown for comparison with the fitted models.

Figs 9–11 are as Fig. 8 for the GAIA CHALLENGE data sets with: a Plummer-like tracer profile with cuspy dark matter halo and isotropic velocity anisotropy (PlumCuspIso), a Plummer-like tracer profile with cuspy DM halo and tangential velocity anisotropy (PlumCuspTan) and a cuspy like (non-Plummer) tracer profile with a cored halo, and Ossipkov-Merritt (Osipkov 1979; Merritt 1985) velocity anisotropy profile (NonPlumCoreOM) data sets. In all four cases, our algorithm selects the correct model and reconstructs robustly the mass content and kinematic profile of the underlying stellar distributions from LOS data only.

In Table 2 we report the results of the mass model selection during Phase I of the t-JEANS. The model selection is performed using the average test error on unseen (during training) data (equation 29) as it has proven to be a more robust discriminator (in comparison with AICc or BIC). We perform model selection after Phase I, in order to reduce computation time. In general, better discrimination results between competing models can be achieved by performing the MCMC process (Phase II) for both competing models and then evaluating the test error, $\chi^2_{\rm test}$ (equation 29). In Table 2 we report the average error, $\chi^2_{\rm test}$, on unseen test data, $D_{\rm train}$ (Section 3). In all four cases, the t-JEANS finds the true underlying models from which the synthetic data were created. In Figs 8–11 we plot the best candidate models as these were selected from the t-JEANS. The plotted results were obtained after Phase III of the t-JEANS. In all cases, our algorithm achieves excellent performance and reconstructs the true underlying profiles.

5.2.2 *The 1k NonPlumCoreOM data set*

In this section we discuss our findings for the 1k NonPlumCoreOM data set as well as the efficiency of the GANs for synthetic data generation. The latter is judged by the quality of the fits.

The NonPlumCoreOM 1k data set, besides being a very difficult data set due to its strong radial anisotropy profile (Read & Steger 2017), also presents a challenge for all Jeans moments based solvers due to its small number of data. Binning 1k data, we end up with as few as 30 binned LOS velocity dispersion values. For a small model, with only three knots for the definition of the B-spline basis, we end up with 5 ($\sigma^2_{\rm rr}$) + 5 ($\sigma^2_{\rm tt}$)+ 4(smoothing penalty)+2 (DM) + 2 (Stellar) = 18 unknown parameters. In addition, the uncertainty of the $\sigma^2_{\rm los}$ binned values is much larger, as is evident from Fig. 4.

For these reasons, we fit both the true 1k profile, as well as the augmented GAN profile. For the case of the 1k data set, the test error based on the sampled MCMC values of the $\sigma^2_{\rm los}$ bins fails to recover the correct model. The augmented GAN data set (approximately 160 binned values) selects the correct model, thus underlining the importance of this data augmentation approach. We summarize the results of the model selection, during the EA phase, in Table 6.

In Fig. 12 we present the fit to the GAN generated data. In the top right-hand panel the $\sigma^2_{\rm los}$ data values are the ones created from the 25k GAN generated synthetic data. In the middle and bottom right-hand panels, the $\sigma^2_{\rm rr}$ and $\sigma^2_{\rm tt}$ data values were not used in the fit. They were estimated from the true NonPlumCoreOM 10k data set and they are placed there for reference only. We used these because the GAN data do not have the full 3D information for us to create these data values for reference. The recovery of the data set is much better than what we would get by using only the NonPlumCoreOM 1k data set. The recovered profile is of lower uncertainty than the

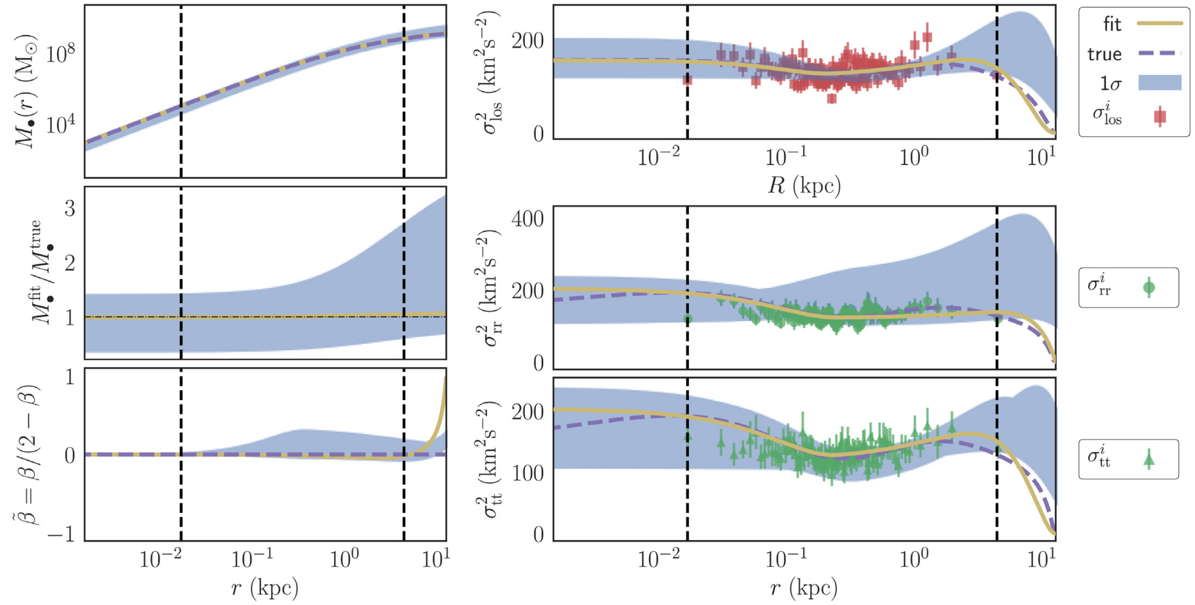

Figure 9. Same as Fig. 8 for the PlumCuspIso 10k data set.

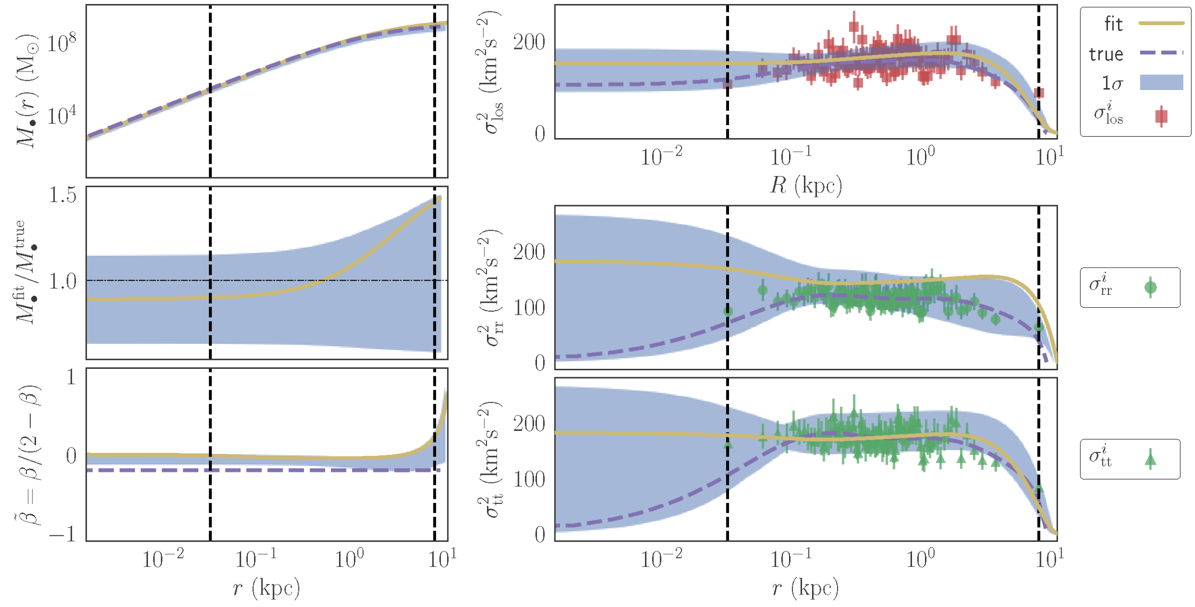

Figure 10. Same as Fig. 8 for the PlumCuspTan 10k data set.

one with the NonPlumCoreOM 10k data set, especially close to the outer regions of the data. That is, the GAN generated data set *gives a better fit than the original True NonPlumCoreOM 10k data set* (note that a different range is displayed on the vertical axis in all right-hand panels of Figs 11 and 12). This can be quantified, as can be seen in Fig. 13: in the left-hand panel we plot the true $\sigma^i_{\rm rr}$ profile as this is estimated from the 10k NonPlumCoreOM data set, as well as the highest likelihood fitted profiles, for the GAN data and the 10k NonPlumCoreOM data sets. In the right-hand panel, we do the same for the tangential dispersion, $\sigma^2_{\rm tt}$. In order to quantify the quality of the fits in the unseen latent space of radial and tangential dispersions we estimate the mean square error for the radial and tangential profiles, between the best-fitted profiles and the data:

$$\chi^2_{{\rm rr},10k} = \frac{1}{N_{\rm bin}} \sum_{i=1}^{N_{\rm bin}} \left([\sigma^2_{\rm rr\,10k}(r_i) - \sigma^i_{\rm r}]/\delta(\sigma^i_{\rm r})\right)^2$$

$$\chi^2_{\rm rr,GAN} = \frac{1}{N_{\rm bin}} \sum_{i=1}^{N_{\rm bin}} \left([\sigma^2_{\rm rr\,GAN}(r_i) - \sigma^i_{\rm r}]/\delta(\sigma^i_{\rm r}\right)^2 .$$

And similarly for the tangential profile, $\sigma^2_{\rm tt}$. We find for the ratios

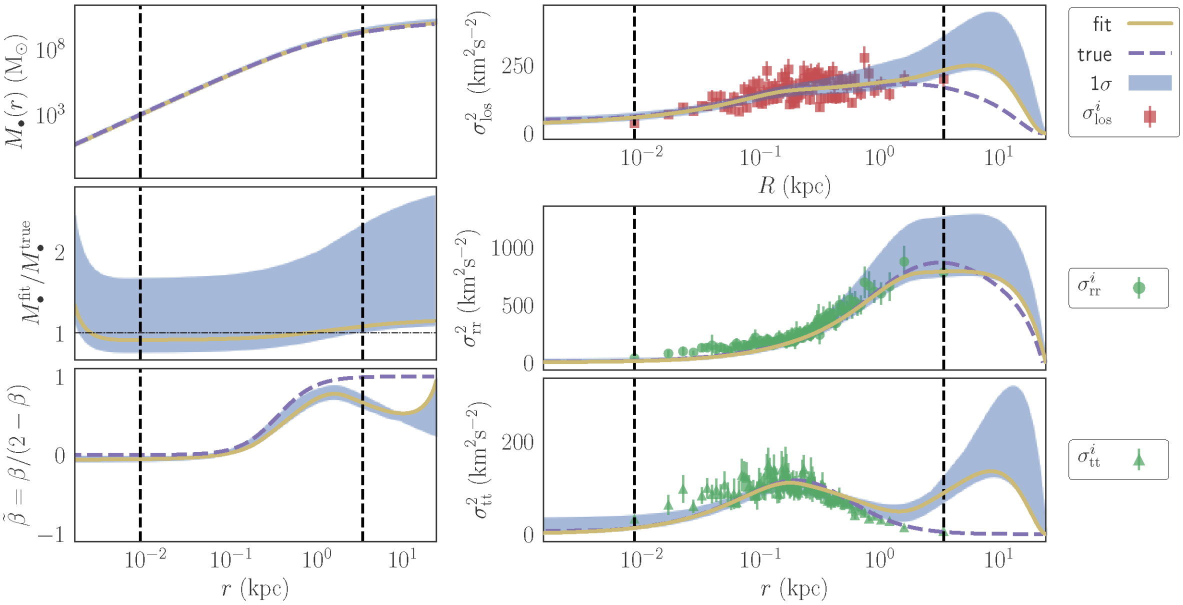

Figure 11. Same as Fig. 8 for the NonPluCoreOM 10k data set.

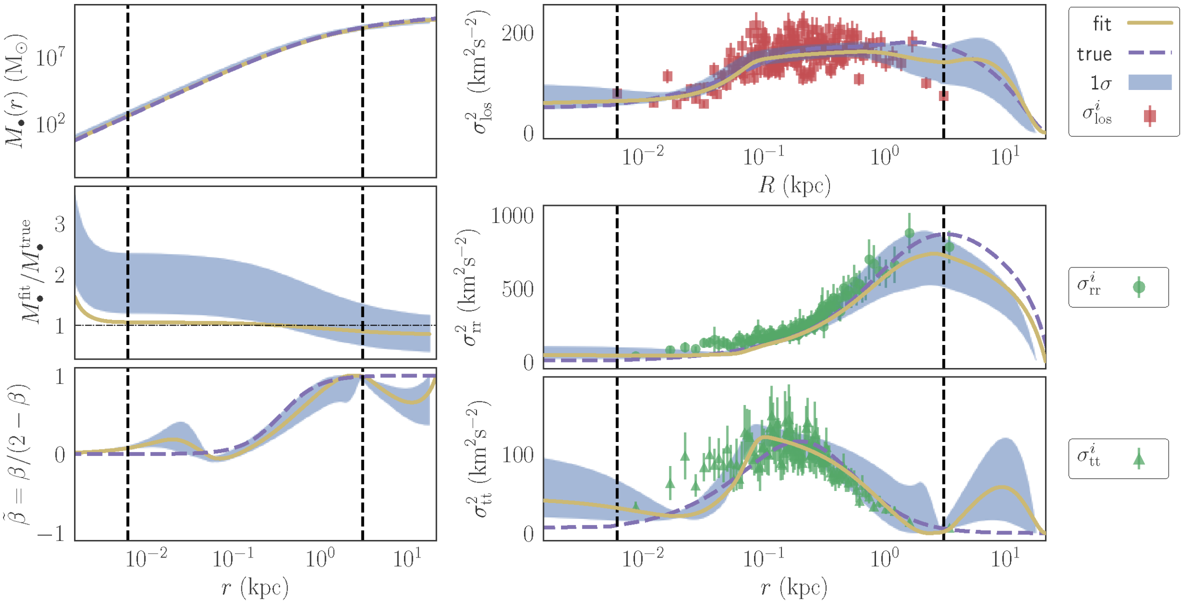

Figure 12. Same as Fig. 8 for the NonPluCoreOM 1k data set. In the top right-hand panel the $\sigma^2_{\rm los}$ data are from the GAN synthetic data generator. In the two bottom right-hand panels the $\sigma^2_{\rm rr}$ and $\sigma^2_{\rm tt}$ moments are from the true NonPlumCoreOM 10k data set (we do not have 3D motions for the GAN generated data).

$$\chi^2_{{\rm rr},10k}/\chi^2_{\rm rr,GAN} = 0.99, \quad \chi^2_{{\rm tt},10k}/\chi^2_{\rm tt,GAN} = 54.26.$$

Therefore, the quality of the $\sigma^2_{\rm rr}$ fit is similar if we train t-JEANS with either the 10k data set, or the GAN generated synthetic data. However, the quality of the $\sigma^2_{\rm tt}$ fit is much worse when t-JEANS is trained with moments from the true 10k data set. This should not come as a surprise. What this means, is that from the 1k of data, the GAN system manages to recover more information than what is hidden in the moments of a 10k data set. Then, with ∼160 binned $\sigma^2_{\rm los}$ data points, it passed more information to the t-JEANS solver, than the moments of the 10k data set can.

The principal criticism that is levelled at the Jeans approach is that one may find solutions to the Jeans equations that require a distribution function that is not positive at all phase-space locations, and is hence unphysical. However, one can always check that the results of our algorithm give a positive DF by testing the solution with a *single* Schwarzschild model. Since the solutions presented above recover the correct input dynamical models from the GAIA CHALLENGE, this step is not necessary here.

6 DISCUSSION

We suspect that the astronomy community's definition of the Jeans degeneracy would be: *many choices of functional forms for $M(< r)$ and $\beta(r)$ (or equivalently $\sigma^2_{\rm rr}$) result in a $\sigma^2_{\rm los}$ profile that is arbitrarily close to the data. Therefore it is not possible to derive a unique mass and anisotropy profile.* This is indeed the case, the system of

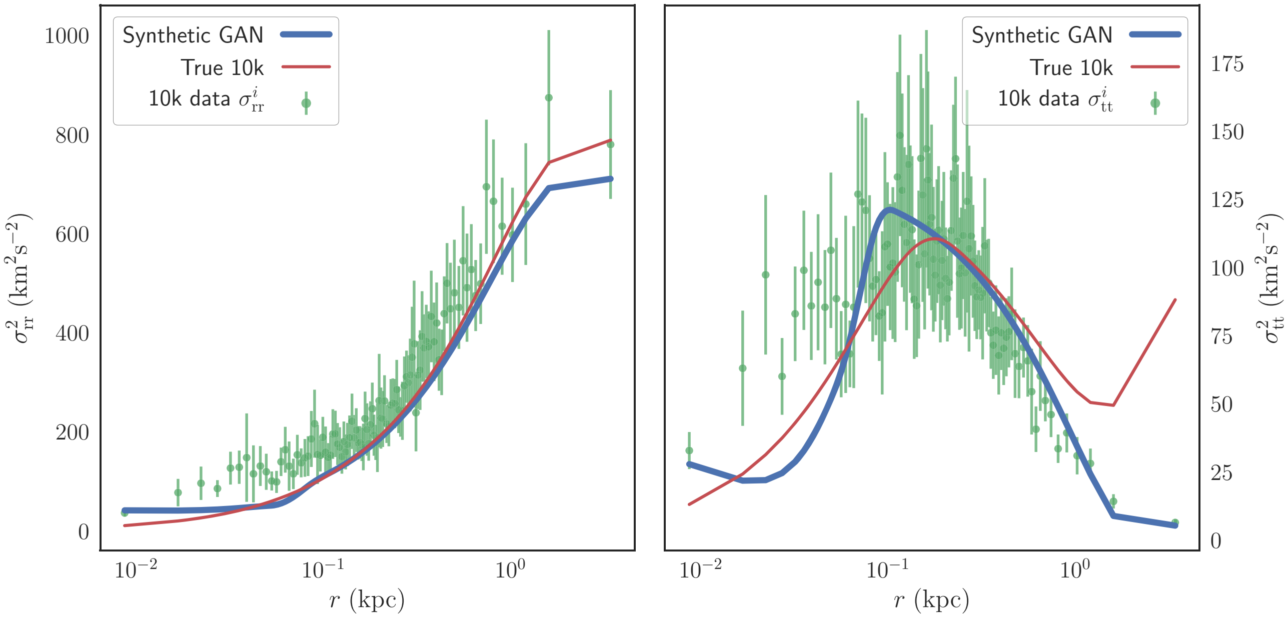

Figure 13. Difference in fitted profiles in latent profiles $\sigma^2_{\rm rr}$, $\sigma^2_{\rm tt}$ for the models trained on the GAN generated synthetic data set and the 10k data set, for the NonPlumCoreOM model.

Table 6. Competing mass models for the NonPlumCoreOM 1k data set, with (GAN) and without (1k) data augmentation. We report the average error on unseen test data, $D_{\rm test}$. The true models from which the data were produced are with bold fonts. The lower test error is also designated with a bold font.

Data set	Stellar	DM	$\chi^2_{\rm test}$
NonPlumCoreOM, 1k	Plummer	NFW	**67.4542**
NonPlumCoreOM, 1k	**gH**	**gH**	69.8274
NonPlumCoreOM, GAN	Plummer	NFW	497.068
NonPlumCoreOM, GAN	**gH**	**gH**	**492.709**

equations is not closed (we need additional constraints that we do not have). However, when it comes to statistical model selection the situation is different. We can provide the additional necessary condition that closes the system of equations by selecting the ‘simplest’ solution that describes well the observable data. The point of emphasis above in bold, that the profile should be *arbitrarily close* to the data, resembles a χ^2 ‘selection’ criterion, which is, however, not a proper model selection method. The key point in t-JEANS to *statistically* break the degeneracy is the realization that we can use hierarchical[13] models that eventually result in different quantitative fits to the data (i.e. different test error). In other words, different assumptions of functional forms for mass, $M(r)$, and anisotropy, β, are no longer quantitatively equivalent.

A special note needs to be made about the fact that the notion of the mass anisotropy degeneracy, when it comes to statistical fitting, is reinforced by the fact that for the majority of stellar systems, the observables are few in number. This makes model selection even more difficult and sustains the belief that, given the availability of data, it is not always possible to discriminate between competing mass models. This is more evident for moment-based mass estimators that rely on summary statistics of the initial data set. The modern semi-supervised machine learning techniques that are actively being developed by the community, such as the GANs for synthetic data generation, are a remedy to this problem.

Some of the main differences of the t-JEANS that allow more efficient treatment of the degeneracy problem, assuming sufficient available data, in comparison with other approaches are

(1) We do not assume two unknown parametric functional forms for both the DM mass density profile, $\rho_\bullet(r)$, and the anisotropy profile, $\beta(r)$. This reduces the uncertainty of the parameters and allows for more robust model selection.

(2) With our choice of hierarchical parametric models (B-splines) for $\sigma^2_{\rm rr}$ and $\sigma^2_{\rm tt}$ we can better statistically discriminate between competing models. This is achieved because hierarchical models find a trade-off between test and train error, and are thus more resilient to over-fitting.

(3) We incorporate a set of physically plausible constraints (Section 4.2) that further reduces the feasible solution space and pushes to the limit the model selection process.

(4) For the moments solver (DMS, Section 4.2), we are not using a single $\sigma^2_{\rm los}$ value for each bin. In contrast, we are using the full MCMC chains to get additional information from the binning scheme. This allows for the estimation of train, validation and test error, as it is used in modern machine learning supervised training techniques.

Although we have not performed a detailed numerical comparison by switching on and off all the constraints we used, we have the following understanding of the effect of each as well as our modelling approach:

(1) The choice of the assumed DM mass model: As with all model selection processes, our effort relies on the assumption that,

[13] With the term hierarchical we mean models that result from the same general equation, but with possibly different complexity. Examples of hierarchical models are a Fourier expansion of a function: $f(x) = \sum_{i=1}^{n} c_n, \cos(nx)$, or a B-spline basis, $f(x) = \sum_{i=1}^{n} c_n B_n(x)$. As n increases we get models of increasing complexity that are derived from the same general equation.

if we try a large set of competing mass models, then one (or some) of them will not be very far from the truth. Then our best solution should approximate reality at a satisfactory level. Our contribution is demonstrating that with the use of a hierarchical basis, satisfactory model selection is possible. Obviously, if our mass model assumptions are away from the truth, we expect that the kinematic fits will also be away from the true anisotropy profile. In our numerical experiments, even with different mass model assumptions, the kinematic profiles tend to be similar. However we cannot conclude, due to the limited number of mass models and data sets we tried, that this is a general feature. In addition, we cannot quantify the ‘anisotropy similarity’ in terms of similarity between competing DM mass models. This is something that requires further investigation. We also note that we have found that the choice of tracer profile affects the derived anisotropy profile significantly.

(2) The boundary condition at the origin $\sigma_{\rm rr}^2(r=0)=\sigma_{\rm tt}^2(r=0)/2$ can result as a limiting case of the Jeans equation (1), as $r\rightarrow 0$, for non-divergent DM potentials. However, it helps numerically inside the solver to keep it separate. The boundary condition at the virial radius was used mainly for the domain of definition of the B-spline basis (it requires a closed finite interval). As we cannot deduce the profile further than the last datum, this constraint contributes in combination with the projected virial theorem.

(3) The MAMMPOST-style LOSVD helps to constrain more robustly the kinematic profile beyond the half-light radius. It proved helpful in the case of the difficult NonPlumCoreOM data set. In the other three data sets, even without it, the recovered fits were excellent.

(4) The projected virial theorem can alter the solution space significantly, for a given mass model assumption. For example we find that if we run an MCMC exploration with and without it, the parameter chains for the same model converge at different non-overlapping regions. It should also be noted that this is a very difficult constraint to implement numerically in an MCMC scheme, because it is a hard bound and does not allow efficient mixing of the chains. It is possible that there is a connection between the projected virial theorem constraint, and the approach of the virial shape parameters taken by Read & Steger (2017, see also Richardson & Fairbairn 2014), however we have not verified this. It is also interesting to note that despite the fact that the kinematic profile is essentially ‘free’ after the last datum, the virial theorem still helps reducing the feasible solution space.

A special note needs to be made on the particular choice of representation: in t-JEANS we represent the kinematic profile with the variables $\sigma_{\rm rr}^2$ and $\sigma_{\rm tt}^2$ instead of $\sigma_{\rm rr}^2$ and β. This is because in the former representation, with the use of B-splines, we can linearize the system of equations (thereby greatly simplifying the solution). In contrast if we use $\sigma_{\rm rr}^2$ and β, then from equations (1) and (2) it is apparent that due to the product term, $\sigma_{\rm rr}^2\beta$, the system of equations is not linear. It should be emphasized that the choice of representation on its own is not adequate to statistically break the degeneracy. By linearizing the system of equations, however, we gained additional insight to the problem. The linearized equations were the key ingredient that led us to seek additional constraints (e.g. virial theorem) that further reduce the feasible solution space.

Finally, we need to emphasize again the importance of using large data sets for model selection. When these are not available, GANs can be one starting point towards the correct solution. t-JEANS – or any other algorithm – will fail in the absence of sufficient data.

6.1 The case of multiple stellar population dynamics

The linearization of the system of equations in the Jeans formalism yields some useful insights for the case of multiple stellar populations. When it is feasible to separate the stellar population into multiple stellar sub-populations (assuming two for simplicity) that are evolving under the influence of the same DM potential, the system of equations describing the system becomes

$$\begin{aligned}\sigma_{1\rm los}^2 &= \frac{1}{\Sigma_{1\star}(R)}\left(\int_R^{r_{\rm vir}}\rho_{1\star}K_1\sigma_{1\rm rr}^2\,{\rm d}\,r+\int_R^{r_{\rm vir}}\rho_{1\star}K_2\sigma_{1\rm tt}^2\,{\rm d}\,r\right)\\ \sigma_{2\rm los}^2 &= \frac{1}{\Sigma_{2\star}(R)}\left(\int_R^{r_{\rm vir}}\rho_{2\star}K_1\sigma_{2\rm rr}^2\,{\rm d}\,r+\int_R^{r_{\rm vir}}\rho_{2\star}K_2\sigma_{2\rm tt}^2\,{\rm d}\,r\right)\end{aligned}$$

and the corresponding Jeans equations are

$$\begin{aligned}-\,\rho_{1\star}\frac{{\rm d}\,\Phi}{{\rm d}\,r} &= \frac{{\rm d}(\rho_{1\star}\sigma_{1\rm rr}^2)}{{\rm d}\,r}+\rho_{1\star}\frac{(2\sigma_{1\rm rr}^2-\sigma_{1\rm tt}^2)}{r}\\ -\rho_{1\star}\frac{{\rm d}\,\Phi}{{\rm d}\,r} &= \frac{{\rm d}(\rho_{2\star}\sigma_{2\rm rr}^2)}{{\rm d}\,r}+\rho_{2\star}\frac{(2\sigma_{2\rm rr}^2-\sigma_{2\rm tt}^2)}{r}.\end{aligned}$$

This is a set of four equations, with five unknowns (assuming, for simplicity, that the stellar tracer densities, $\rho_{(1,\,2)\star}$, are known), namely, $\sigma_{(1,2)\rm rr,tt}^2$, $\rho_\bullet(r)$. The system of equations is still not closed (in fact, irrespective of the number of sub-populations, we will always have one more unknown function than equations). However, from the insight we get from the linearized equations (say, using B-splines), we understand that if the profiles of the stellar populations are significantly different (i.e. the determinant of the linearized system is not zero), then the solution space is reduced significantly. Depending on the statistical uncertainty of the observables, this may be enough to accurately describe the underlying DM structure. In contrast, if the profiles of the sub-populations are identical, the linear systems are identical (their determinant is zero) and no additional reduction of the feasible solution space is possible. Clearly, the linearization of the equations with the use of B-splines (or other suitable complete bases, e.g. wavelets), besides being a useful numerical scheme, also allows us to gain further insight into the degeneracy problem.

For systems with multiple stellar populations where we are trying to deduce more than one kinematic profile from scarce data, the GAN synthetic data generation can be a game changer for the estimation of the different brightness and LOS velocity dispersion profiles. The reason being that it can construct robust velocity dispersion data with small uncertainties over the extent of the system under investigation.

7 CONCLUSIONS

In this work we describe a new method for reliable mass determination independent of the mass–velocity anisotropy degeneracy. The efficiency of our method is tested on synthetic data from the GAIA CHALLENGE suite of mock simulations. In all cases our algorithm reconstructs accurately the underlying kinematic profile as well as the mass content of the data sets. Our method includes: (i) a new way of solving numerically the Jeans equations, subject to physically plausible local and global constraints, using quadratic programming. (ii) a new way for performing supervised learning in the framework of Jeans mass modelling, using samples from LOS velocity dispersion MCMC chains as ‘unseen’ validation and test data sets. Based on this, we present a new approach in performing regularization and model selection. (iii) The application of GANs for augmenting data sets, thereby making the t-JEANS moments solver method reliable in situations where the available samples possess a relatively small number of stars.

ACKNOWLEDGEMENTS

FID, GFL, and CP acknowledge support from Australian Research Council Discovery Project (DP140100198). FID also thanks the University of Western Australia for support through a Research Collaboration Awards (PG12105204). GFL thanks the European Southern Observatory (ESO) for support as a visiting astronomer and for hosting him in Garching where the final stages of the preparation of this publication were undertaken. RAI gratefully acknowledges support from a ‘Programme National Cosmologie et Galaxies’ grant.

REFERENCES

Bertoluzza S., Falletta S., Russo G., Shu C., 2008, Numerical Solutions of Partial Differential Equations, Advanced Courses in Mathematics - CRM Barcelona. Birkhäuser, Basel, 202
Binney J., 1980, MNRAS, 190, 873
Binney J., Mamon G. A., 1982, MNRAS, 200, 361
Binney J., Tremaine S., 2008, Galactic Dynamics: 2nd edn. Princeton Univ. Press, Princeton, 920
Burnham K. P., Anderson D. R., 2002, Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach. Springer-Verlag, New York, 488
Chan C., Ginosar S., Zhou T., Efros A. A., 2018, preprint (arXiv:e-prints)
Dejonghe H., Merritt D., 1992, ApJ, 391, 531
Diakogiannis F. I., Lewis G. F., Ibata R. A., 2014a, MNRAS, 443, 598
Diakogiannis F. I., Lewis G. F., Ibata R. A., 2014b, MNRAS, 443, 610
Diakogiannis F. I., Lewis G. F., Ibata R. A., Guglielmo M., Kafle P. R., Wilkinson M. I., Power C., 2017, MNRAS, 470, 2034
Eilers P. H. C., Rijnmond D. M., Marx B. D., 1996, Stat. Sci., 11, 89
Gelman A., Hwang J., Vehtari A., 2014, Stat. Comput., 24, 997
Goodfellow I., Pouget-Abadie J., Mirza M., Xu B., Warde-Farley D., Ozair S., Courville A., Bengio Y., 2014, in Ghahramani Z., Welling M., Cortes C., Lawrence N. D., Weinberger K. Q., eds, Advances in Neural Information Processing Systems 27. Curran Associates, Inc.p. 2672
Goodfellow I. J., 2017, CoRR, abs/1701.00160
Gulrajani I., Ahmed F., Arjovsky M., Dumoulin V., Courville A. C., 2017, CoRR, abs/1704.00028
Hastie T., Tibshirani R., Friedman J., 2001, The Elements of Statistical Learning–Springer Series in Statistics. Springer-Verlag, New York
Höllig K., 2003, Finite Element Methods with B-Splines. Frontiers in Applied Mathematics. Society for Industrial and Applied Mathematics, Philadelphia.
Ibata R., Nipoti C., Sollima A., Bellazzini M., Chapman S. C., Dalessandro E., 2013, MNRAS, 428, 3648
Jalali M. A., Tremaine S., 2011, MNRAS, 410, 2003
Karras T., Aila T., Laine S., Lehtinen J., 2017, CoRR, abs/1710.10196
Kingma D. P., Ba J., 2014, CoRR, abs/1412.6980
Łokas E. L., Mamon G. A., 2003, MNRAS, 343, 401
Mamon G. A., Boué G., 2010, MNRAS, 401, 2433
Mamon G. A., Biviano A., Boué G., 2013, MNRAS, 429, 3079
Merrifield M. R., Kent S. M., 1990, AJ, 99, 1548
Merritt D., 1985, AJ, 90, 1027
Merritt D., 2013, Dynamics and Evolution of Galactic Nuclei. Princeton Univ. Press, Princeton.
Navarro J. F., Frenk C. S., White S. D. M., 1996, ApJ, 462, 563
Osipkov L. P., 1979, Soviet Astron. Lett., 5, 42
Paszke A. et al., 2017, NIPS-W
Plummer H. C., 1911, MNRAS, 71, 460
Read J. I., Steger P., 2017, MNRAS, 471, 4541
Richardson T., Fairbairn M., 2014, MNRAS, 441, 1584
Schwarzschild M., 1979, ApJ, 232, 236
Shin H.-C., Tenenholtz N. A., Rogers J. K., Schwarz C. G., Senjem M. L., Gunter J. L., Andriole K., Michalski M., 2018, preprint (arXiv:e-prints)
Solanes J. M., Salvador-Sole E., 1990, A&A, 234, 93
Ŝolín P., 2005, Partial Differential Equations and the Finite Element Method. Pure and Applied Mathematics: A Wiley Series of Texts, Monographs and Tracts. Wiley, Hoboken
Sugiura N., 1978, Commun. Stat., 7, 13
Zhao H., 1996, MNRAS, 278, 488

APPENDIX: PROOF OF THEOREM 1

Let us assume that there exist two radial profiles, $\sigma^2_{1\mathrm{rr}}$ and $\sigma^2_{2\mathrm{rr}}$ that give the same LOS dispersion profile. Then,

$$\sigma^2_{\mathrm{los}}(R) = \frac{2}{\Sigma_\star(R)} \int_R^{r_{\mathrm{vir}}} \left[K_A \left(\frac{\mathrm{d}(\rho_\star \sigma^2_{1\mathrm{rr}})}{\mathrm{d}\,r} + \rho_\star \frac{\mathrm{d}\,\Phi}{\mathrm{d}\,r} \right) + K_B \rho_\star \sigma^2_{1\mathrm{rr}} \right] \mathrm{d}\,r$$

$$\sigma^2_{\mathrm{los}}(R) = \frac{2}{\Sigma_\star(R)} \int_R^{r_{\mathrm{vir}}} \left[K_A \left(\frac{\mathrm{d}(\rho_\star \sigma^2_{2\mathrm{rr}})}{\mathrm{d}\,r} + \rho_\star \frac{\mathrm{d}\,\Phi}{\mathrm{d}\,r} \right) + K_B \rho_\star \sigma^2_{2\mathrm{rr}} \right] \mathrm{d}\,r.$$

Subtracting the above equations yields

$$\int_R^{r_{\mathrm{vir}}} \left[K_A \left(\frac{\mathrm{d}\,\rho_\star \Delta\sigma^2_{\mathrm{rr}}}{\mathrm{d}\,r} \right) + K_B \rho_\star \Delta\sigma^2_{\mathrm{rr}} \right] \mathrm{d}\,r = 0, \tag{A1}$$

where $\Delta\sigma^2_{\mathrm{rr}} = \sigma^2_{2\mathrm{rr}} - \sigma^2_{1\mathrm{rr}}$. In order for this integral to be identically zero for all values of the parameter R, the integrand must be zero, i.e.

$$K_{\mathrm{A}} \left(\frac{\mathrm{d}\,f}{\mathrm{d}\,r} \right) + K_{\mathrm{B}}\, f = 0, \tag{A2}$$

where we have set $f(r) = \rho_\star \Delta\sigma^2_{\mathrm{rr}}$. For the case of $R = 0$ the result is trivial $f = 0$, i.e. $\sigma_{1\mathrm{rr}} = \sigma_{2\mathrm{rr}}$. For the case $r > R > 0$, we manipulate equation (A2):

$$\begin{aligned} &\mathrm{d}\,f/\mathrm{d}r + K_{\mathrm{B}}/K_{\mathrm{A}}\, f = 0 \rightarrow \\ &\mathrm{d}\,f/\mathrm{d}r + \frac{2r}{R^2} f = 0 \rightarrow \\ &\mathrm{d}\,f/f = -2r/R^2 \mathrm{d}r \rightarrow \\ &f = A \exp\{-r^2/R^2\}, \end{aligned}$$

where A is the constant of integration, that will be determined from the virial boundary condition: since the last equation holds for all r, R, it will also hold for $r = r_{\mathrm{vir}}$ and $R = r_{\mathrm{vir}}/2$, where r_{vir} is the virial radius of the system. However for $r = r_{\mathrm{vir}}$, it is

$$\lim_{r\to r_{\mathrm{vir}}} \sigma^2_{\mathrm{rr}}(r) = \lim_{r\to r_{\mathrm{vir}}} \sigma^2_{\mathrm{tt}}(r) = 0.$$

Hence, $\lim_{r\to r_{\mathrm{vir}}} \sigma^2_{1rr}(r) = \lim_{r\to r_{\mathrm{vir}}} \sigma^2_{2rr}(r) = 0$, i.e. $f(r_{\mathrm{vir}}) = 0$. Then $A\exp(-4) = 0$, i.e. $A = 0$, then $f(r) = 0$ and $\sigma_{1\mathrm{rr}} = \sigma_{2\mathrm{rr}}$ for all r.

This proof is also valid for spherically symmetric systems subject to an external gravitational field: in this case as $r \to r_{\mathrm{vir}}$ both the radial and tangential velocity dispersions approach the same constant value (Dejonghe & Merritt 1992), thus again at the virial radius of the system $f \to 0$.

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