Resolving the mass–anisotropy degeneracy of the spherically symmetric Jeans equation – I. Theoretical foundation

Foivos I. Diakogiannis,1⋆ Geraint F. Lewis1 and Rodrigo A. Ibata2

1Sydney Institute for Astronomy, School of Physics, A28, University of Sydney, NSW 2006, Australia
2Observatoire Astronomique, Université de Strasbourg, CNRS, 11, rue de l Université, F-67000 Strasbourg, France

Accepted 2014 June 9. Received 2014 June 7; in original form 2014 January 14

ABSTRACT
A widely employed method for estimating the mass of stellar systems with apparent spherical symmetry is dynamical modelling using the spherically symmetric Jeans equation. Unfortunately, this approach suffers from a degeneracy between the assumed mass density and the second-order velocity moments. This degeneracy can lead to significantly different predictions for the mass content of the system under investigation, and thus poses a barrier for accurate estimates of the dark matter content of astrophysical systems. In a series of papers, we describe an algorithm that removes this degeneracy and allows for unbiased mass estimates of systems of constant or variable mass-to-light ratio. The present contribution sets the theoretical foundation of the method that reconstructs a unique kinematic profile for some assumed free functional form of the mass density. The essence of our method lies in using flexible B-spline functions for the representation of the radial velocity dispersion in the spherically symmetric Jeans equation. We demonstrate our algorithm through an application to synthetic data for the case of an isotropic King model with fixed mass-to-light ratio, recovering excellent fits of theoretical functions to observables and a unique solution. The mass–anisotropy degeneracy is removed to the extent that, for an assumed functional form of the potential and mass density pair (Φ, ρ), and a given set of line-of-sight velocity dispersion σ02 observables, we recover a unique profile for σ2 and σ4. Our algorithm is simple, easy to apply and provides an efficient means to reconstruct the kinematic profile.

Key words: methods: miscellaneous – methods: statistical – galaxies: general – galaxies: kinematics and dynamics – galaxies: statistics.

1 INTRODUCTION

The spherically symmetric Jeans equation (hereafter SSJE) is an important tool for the estimation of the mass content of stellar structures that exhibit spherical symmetry. It has been used widely (see Binney & Tremaine 2008) for the dynamical modelling of globular clusters, dwarf spheroidal and elliptical galaxies with nearly spherical shape. However, a problem with this approach is that there exists a degeneracy between the assumed mass density and the velocity distribution of the system, which can lead to erroneous mass estimates. Describing the mass content of a stellar system accurately is crucial for identifying dark matter (hereafter DM) structures and to test the standard Λ cold dark matter model. Therefore, it would be important if this degeneracy could be completely removed.

There has been extensive effort (e.g. Binney & Mamon 1982; Tonry 1983; Merritt 1987; Merrifield & Kent 1990; Dejonghe & Merritt 1992; Lokas 2002, and others) to resolve this problem in recent years with significant, yet not complete, success. There are two main approaches in attacking the problem. One approach is to assume a functional form for the mass density and then try to recover the correct second-order velocity moments. The other is to define a class of distribution functions f(E, L) and try to infer qualitative and quantitative results for the velocity distribution of actual stellar systems through the use of second, fourth or higher velocity moments of the observables. It should be stated that both approaches try to estimate a unique kinematic profile for a given mass density. Both have advantages and disadvantages. The first has the advantage that we can make a good prediction of the functional form of the mass density from observed brightness distributions. However, this method is limited by the use of only the second velocity moments, thus it cannot account for the general velocity distribution. The second approach can, in principle, estimate the full distribution function (hereafter DF). However, we do not have a direct comparison of f(E, L) with E, L observables to be certain of our assumption on the functional form of f(E, L). Thus, it is possible that it introduces a bias in the derived measures.

© 2014 The Authors

Published by Oxford University Press on behalf of the Royal Astronomical Society
In this paper, we focus on the first approach. The seminal work of Binney & Mamon (1982, see also Tonry 1983) presented an algorithm that, for self-consistent systems and an assumed mass density, yields a unique constant mass-to-light ratio $\gamma$ and second radial $\sigma_{rr}^2$ and tangential $\sigma_{tt}^2$ velocity moments. Although this method is elegant, significant and very useful, it presents some difficulties and has limitations. The major limitation, as the authors point out, is that it cannot account for a variable mass-to-light ratio, i.e. when a separate DM component is present the method is not applicable. Another difficulty is that the accuracy of the algorithm was demonstrated using synthetic data that had very small errors ($\leq 3$ per cent of the actual values), which is rarely realistic in practise. Furthermore, one needs to define first a fitted profile to the line-of-sight velocity dispersion $\sigma_{los}^2$ and then use this as a theoretical function to infer the functional form of the anisotropy $\beta$. This $\sigma_{los}^2$ fit will always have some uncertainty due to errors in the observables. The authors demonstrate that this uncertainty does not affect their qualitative results, i.e. they can still distinguish between radially or tangentially biased profiles. Unfortunately quantitatively, with such a procedure, there is error propagation that degrades the quality of the estimates, particularly if there are large uncertainties in the data. It was argued by van der Marel (1994) that this method requires knowledge of the profile of the projected velocity dispersions corrected for the effects of seeing and spatial binning and that these corrections can be exceedingly difficult to make, especially near the centre of the system under study.

Ibata et al. (2013) presented an algorithm for the evaluation of the mass content of a system with variable mass-to-light ratio and a varying anisotropy profile, i.e. the method can account for a separate DM component. This method uses splines to define the radial $\Delta \sigma_{rr}^2(r)$ and tangential $\Delta \sigma_{tt}^2(r)$ velocity dispersions, as well as the mass density $\Delta \rho(r)$ in a dense set of radial positions $r_i$. The individual value of each profile at each position $r_i$ is treated as a free parameter and is estimated through a Markov Chain Monte Carlo (MCMC) scheme subject to some physically plausible constraints. This method, although efficient, uses a large number of free parameters ($\sim 389$) and is computationally expensive, thus making model comparison through Bayesian model inference a very difficult task.

Our work focuses on the task of determining unique second-order velocity moments and accurate mass estimates performed using the SSJE. In this paper, we develop the basic mathematical framework of our algorithm. Thus, we limit the application of our method to a simple example of a system with a fixed mass-to-light ratio. In Diakogiannis, Lewis & Ibata (2014, hereafter Paper II), we expand the theoretical model and validate our method by giving a detailed analysis of applications to various systems with constant and variable mass-to-light ratio. In the current approach, the only assumption we make is the functional form of the mass density $\rho(r)$ of the system. From this, facilitating comparison with observables, we recover the mass content and a unique kinematic profile of the stellar system. Then, the correct mass model hypothesis can be inferred through Bayesian inference methods. Our method is valid even in the case where there are two separate components, e.g. stars and DM (Paper II). It is simple, easy to apply and computationally inexpensive. The key idea behind our method is this: the line-of-sight velocity dispersion, $\sigma_{los}^2$, depends on both the radial, $\sigma_{rr}^2$, and tangential, $\sigma_{tt}^2$, velocity dispersions; since we do not know the functional form of the kinematic quantities $\sigma_{rr}^2(r)$ or $\sigma_{tt}^2(r)$, we can use the SSJE to eliminate the tangential component, $\sigma_{tt}^2(r)$, dependence from $\sigma_{los}^2$ and approximate $\sigma_{tt}^2(r)$ with a smooth Computer Aided Geometric Design (CAGD) curve. Comparison of $\sigma_{los}^2$ with line-of-sight velocity dispersion observables gives $\sigma_{rr}^2(r)$ both the correct geometric shape and estimates of its numerical value. This avoids any bias in the mass estimates from assumption of a specific anisotropy profile. The CAGD tools we use are B-spline functions. Once $\sigma_{los}^2(r)$ is known, we can always use the SSJE to estimate the tangential velocity dispersion, $\sigma_{tt}^2(r)$, thus recover, within uncertainties, the anisotropy profile.

The structure of our paper is the following: in Section 2, we describe the degeneracy of the SSJE in a detailed mathematical formulation. In Section 3, we give an extended presentation of smoothing B-spline CAGD curves and functions, how we combine them with the SSJE and the dynamical mass model we use. In Section 4, we describe the statistical inference methods. In Section 5, we present a simple example. In this, we reconstruct fully the mass content of the system and the kinematic profile, using synthetic data of brightness and line-of-sight velocity dispersion $\sigma_{los}^2$. In Section 6, we discuss various aspects of our method, and we comment on the optimum smoothing problem of the B-spline representation. Finally in Section 7, we conclude our work.

2 JEANS DEGENERACY IN DETAIL

Consider a self-gravitating stellar system in dynamical equilibrium. Under the SSJE framework, this system is described through the mass density $\rho(r)$, the potential $\Phi(r)$ and the second velocity moments $\sigma_{rr}^2(r)$ and $\sigma_{tt}^2(r)$. The SSJE is customarily written in the form

$$-\frac{d\Phi}{dr} = \frac{1}{\rho} \frac{d(\rho \sigma_{rr}^2)}{dr} + \frac{2\beta}{r} \sigma_{tt}^2,$$  \hspace{1cm} (1)

where

$$\beta = 1 - \frac{\sigma_{tt}^2}{2\sigma_{rr}^2} \hspace{1cm} (2)$$

is the Binney anisotropy parameter (Binney & Mamon 1982, see also Binney & Tremaine 2008). The connection with observables is performed through the line-of-sight velocity dispersion, namely

$$\sigma_{los}^2(R) = \frac{2}{\Sigma(R)} \int_{R}^{R_t} \left(1 - \beta(r) \frac{r^2}{R^2}\right) \frac{r \rho \sigma_{rr}^2}{\sqrt{r^2 - R^2}} \, dr,$$  \hspace{1cm} (3)

where $r_t$ is the tidal radius of the physical system. Note that $\beta$ is multiplied with $\sigma_{rr}^2$, and this increases the complexity of the set of equations (1) and (3).

The traditional approach of using the SSJE for dynamical modelling is to assume a mass density $\rho(r)$ and a functional form for the $\beta(r)$ anisotropy profile. Then, one evaluates $\sigma_{los}^2(r)$ from equation (1), substitutes into equation (3) and compares with observables. For an assumed mass density, any $\beta(r)$ functional form defines a severe restriction on the system and inserts bias in the mass estimates. Choosing different $\beta(r)$ functions in general can result in significantly different results for both the mass estimates and the kinematic profile of the system (Merritt 1987).

As mentioned in the introduction, this is the problem we are going to resolve: for an assumed mass density $\rho(r)$, we will recover the unique kinematic profile as it is described through the second moments of radial $\sigma_{rr}^2$ and tangential $\sigma_{tt}^2$ velocities. We must emphasize that this does not remove the degeneracy on the assumption of the mass density, i.e. a different assumption on $\rho(r)$ will in general

---

1 For self-consistent systems, potential $\Phi(r)$ and mass density $\rho(r)$ are related through Poisson’s equation.

2 Here, we consider that in a spherical coordinate system ($r$, $\theta$, $\phi$), the tangential velocity dispersion is defined as $\sigma_{tt}^2 = \sigma_{\theta\theta}^2 + \sigma_{\phi\phi}^2$.  

MNRAS 443, 598–609 (2014)
lead to a different kinematic profile \( \sigma_2^2(r) \) and \( \sigma_3^2(r) \) that still reproduces the observables. However, again this will be unique for the given \( \rho(r) \).

3 Mathematical Formulation

Since B-spline functions are not widely used in the astronomical community, we will give a short description of them. In this section, we will introduce B-spline curves and functions and describe in detail how we use B-spline functions in the SSJE. We will also give definitions for the mass density \( \rho(r) \) of the dynamical models we use. The standard reference for B-spline functions is De Boor (1978). For practical applications, the interested reader will find great help in books of CAGD, such as Rogers (2001) and Farin (2002). All the above references provide information on available libraries for B-splines in FORTRAN and C programming languages. For our needs, we used the GNU Scientific Library that has an implementation of B-spline bases.

In short, a B-spline function \( f(x) \) is a linear combination of some constant coefficients \( a_i \) with some polynomial functions \( B_{i,k}(x) \) (B-spline basis functions) of a given degree \((k-1)\), i.e. \( f(x) = \sum a_i B_{i,k}(x) \). These polynomial functions \( B_{i,k}(x) \) are smooth and consist of polynomial pieces joined together in a special way. We will start with the definition of B-spline basis functions and then proceed to B-spline curves and functions.

3.1 B-spline basis

Let \( k \) be a positive integer and \( \xi_i \), represent a non-decreasing sequence of \( m+1 \) real numbers, \( \xi_0 \leq \xi_1 \leq \ldots \leq \xi_m \). We will refer to this sequence as the knot sequence. Each of these \( \xi_i \) are called knots. The integer \( k \) is called the degree of the B-spline basis and should not be confused with the degree of the polynomial pieces (degree = \( k-1 \)). We say that a knot \( \xi_i \) has multiplicity \( p \) if it appears \( p \) times in the knot sequence \((p \leq k)\).

The elements \( B_{i,k}(x) \) of a B-spline basis of order 1 (polynomial degree = 0) are defined through the formula

\[
B_{i,1}(x) = \begin{cases} 
1, & \text{if } \xi_i \leq x < \xi_{i+1} \\
0, & \text{otherwise}
\end{cases}
\] (4)

A B-spline basis of order \( k \) is defined for all real numbers \( x \) through the Cox–de Boor recursive algorithm

\[
B_{i,k}(x) = \omega_{i,k} B_{i,k-1}(x) + (1 - \omega_{i+1,k}) B_{i+1,k-1}(x),
\] (5)

where

\[
\omega_{i,k} = \begin{cases} 
\frac{x-\xi_i}{\xi_{i+k-1}-\xi_i}, & \text{if } \xi_{i+k-1} \neq \xi_i \\
0, & \text{otherwise}
\end{cases}
\] (6)

Thus, B-spline basis functions \( B_{i,k}(x) \) are polynomials of degree \( k-1 \). In this definition, we follow the convention that whenever division by zero appears we treat the whole fraction as zero, i.e. \( 0/0 = 0 \).

We list here some important properties of the B-spline basis which are related to our needs for the development of our method. This is not a complete list. In our effort to emphasize the importance of these properties in applications, we shall frequently refer to the coefficients \( a_i \) of a B-spline function \( f(x) = \sum a_i B_{i,k}(x) \), despite the fact that we formally define these functions in a later subsection.

3 There are also some excellent online notes by C. K. Shene http://www.cs.mtu.edu/~shene/COURSES/cs3621/NOTES/

Figure 1. Top panel: B-spline basis functions of order \( k = 4 \) (polynomial degree = 3) for the uniform knot sequence \([0, 0, 0, 0, 1/3, 2/3, 1, 1, 1, 1]\). Bottom panel: a B-spline representation of a function \( f(x) = \sum a_i B_{i,k}(x) \) for the set of coefficients \( a_i = \{0, 1, 2, 1.25, 4.0, 0.5\} \) on the same knot sequence. The B-spline basis functions multiplied with the corresponding coefficient are also plotted. The control polygon is the dashed line, and the black squares are the positions of the control points. Due to the multiplicity of the first and last knots, the function \( f(x) \) attains the values of the first \( a_1 = 0 \) and last \( a_6 = 0.5 \) coefficients.

(i) B-spline basis functions \( B_{i,k}(x) \) are linearly independent.

(ii) \( B_{i,k}(x) \) is a degree \( k-1 \) polynomial in \( x \). This is a restriction on the differentiability of the functions we are going to consider later.

(iii) Each basis function \( B_{i,k}(x) \geq 0 \) for any \( x \). Then, any change in sign of a B-spline function results from a change in the sign of the coefficients \( a_i \). This is a very important property, since if we have a positive function (such as \( \sigma_3^2 \)) that we wish to expand in a B-spline basis, then by demanding the coefficients \( a_i \) of this expansion to be positive, we guarantee this restriction.

(iv) For a given knot sequence \( \xi_0, \ldots, \xi_m \), there exist \( n \) B-spline basis functions \( B_{i,k}(x) \), of order \( k \), where \( n = m + 1 - k \). If we wish to use a given polynomial order B-spline basis, and a given number of coefficients \( a_i \), the number of knot points is uniquely determined.

(v) On any point \( x \in [\xi_i, \xi_{i+1}] \) at most \( k \) basis functions are nonzero. Then, for a B-spline function \( f(x) = \sum a_i B_{i,k}(x) \) for a given \( x \in [\xi_i, \xi_{i+k}] \), only a subset of all coefficients \( a_i \) will contribute to the value of \( f(x) \). We shall refer to this property as the local modification scheme of B-splines.
vector is \( \dim(\xi) = 10 \), thus according to our definition \( m = 9 \). Then, there exist \( n = m + 1 - k = 6 \) linearly independent bases of order \( k = 4 \).

### 3.2 B-spline curves

A B-spline curve in two-dimensional space is the linear combination of some \( n \) constant vector coefficients\(^4\) \( c_i \) with the B-spline basis functions \( B_{i,k}(x) \)

\[
P(x) = \sum_{i=1}^{n} c_i B_{i,k}(x),
\]

\( P(x) \) is the position vector that traces the curve parametrized by \( x \).

The position vectors \( e_i \) are called control points. They define the \( n \) vertices of an open polygon which is called the control polygon. This control polygon defines the shape of the B-spline curve. By adjusting the control points, the curve acquires a different geometric shape.

We state without proof two very important properties of B-spline curves.

(i) An important class of B-spline curves is the one for which the first \( \xi_0 \) and last \( \xi_m \) knots have multiplicity \( p \) equal to the order \( k \) of the B-spline curve. It can be proved then, that the B-spline curve passes from the first \( e_1 = P(\xi_0) \) and last \( e_n = P(\xi_m) \) points of the control polygon. This is crucial for our subsequent analysis since, if from some physical considerations we know the boundary conditions at the beginning or the end of a curve, then we know the coordinates of the first \( e_1 \) or last \( e_n \) control points. This property in combination with the smooth behaviour of B-spline curves proves to be a severe restriction on our models. All curves we are going to consider have multiplicity \( p = k \) in the first and last knots.

(ii) Having defined a knot vector \( \xi \) and knowing the control points \( e_i \), then these completely determine the tangent curve \( T(x) \) of \( P(x) \). This is simply

\[
T(x) = \frac{dP(x)}{dx} = \sum_{i=1}^{n} c_i \frac{dB_{i,k}(x)}{dx}.
\]

Since the basis functions \( B_{i,k}(x) \) are known polynomial functions, so are their derivatives. Thus, the control points \( e_i \) define the curve and all of its derivatives. This is a remarkable property for our needs in dynamical analysis. Each time we encounter an unknown function that participates in some differential equation, then by using a B-spline representation of the function we no longer need to solve the differential equation. Instead, we simply need to calculate the unknown coefficients \( a_i \) through some algebraic process.\(^5\) We shall see later that this property removes the complexity in the SSJE of having to calculate \( \sigma^2_n(R) \).

### 3.3 B-spline functions

A B-spline function is the linear combination of some constant coefficients \( a_i \) with the B-spline basis functions \( B_{i,k}(x) \)

\[
f(x) = \sum_{i=1}^{n} a_i B_{i,k}(x).
\]

\( a_i \) are the control points and \( B_{i,k}(x) \) are the B-spline basis functions.

\( \xi_i \) are the knot sequence, \( n \) is the number of knots, and \( k \) is the order of the B-spline basis.

The properties of B-spline curves are transferred also to B-spline functions.

(i) For a B-spline function \( f(x) = \sum_{i=1}^{n} a_i B_{i,k}(x) \) defined on some knot vector, if the multiplicity of the first \( \xi_0 \) and last \( \xi_m \) knot is equal to the B-spline basis order \( k \) then \( f(\xi_0) = a_1 \) and \( f(\xi_m) = a_n \).

(ii) For a given knot sequence \( \xi_0, \ldots, \xi_m \), the constant coefficients \( a_i \) uniquely determine the function \( f(x) \) and all of its derivatives.

An example of a B-spline function is given in the bottom panel of Fig. 1. We define our function with the use of the B-spline basis functions \( B_{i,k}(x) \) that are on the top panel (order \( k = 4 \), knot sequence \( \xi = [0, 0, 0, 0, 1/3, 2/3, 1, 1, 1, 1] \)) and the set of coefficients \( a_i = \{0, 1, 2, 1.25, 4.0, 0.5\} \); we plot the function \( f(x) = \sum_{i=1}^{n} a_i B_{i,k}(x) \), the weighted B-spline basis functions \( a_i B_{i,k}(x) \) as well as the control polygon of the B-spline curve \( P(x) = (x, f(x)) \). The coordinates of the control points are given by \( e_i = \xi^*_i \), where \( \xi^*_i \) are called Greville abscissae and are not to be confused with the knot points \( \xi_i \). These are defined as the mean position of \( k - 1 \) consecutive knots \( \xi_i \):

\[
\xi^*_i = \frac{1}{k-1} (\xi_i + \xi_{i+1} + \cdots + \xi_{i+k-2}).
\]

see Farin (2002) for details.

B-spline curves and functions are used extensively in CAGD and in statistical modelling of data, whenever a smoothing model function is needed. The quality of the resulting fit depends on the order \( k \) of the spline, on the distribution of knot points, and on the number of coefficients. There is no optimum choice since all of the above parameters depend on our data. We need to use model comparison for the best choices of order \( k \), knot distribution and number of knot points. In general, a bad choice of all the above parameters can result in overfitting or underfitting to the data. Bayesian inference solves partially this problem by finding the model that has the optimum knot order \( k \) and number of coefficients \( a_i \). Again, there still remains the problem of optimum smoothing, since it may be the case that we have data with large errors that result in unphysical oscillatory behaviour in the functions we represent with B-spline bases. We give a solution to this in Paper II by introducing a smoothing penalty that uses information of the smoothness from ideal theoretical models.

Our goal is to use a B-spline function representation for the radial velocity dispersion \( \sigma^2_n(r) \)

\[
\sigma^2_n(r) = \sum_{i=1}^{n} a_i B_{i,k}(x).
\]

Doing so, we recover the values of the coefficients \( a_i \) from comparison with observational values of the line-of-sight velocity dispersion \( \sigma^2_0(R) \), thus determining the anisotropy of the system in a unique way.

### 3.4 Choice of knot sequence

The distribution of knot points \( \xi_i \) is one way to affect the geometric shape of a curve described by a B-spline function\(^7\) \( P(x) = (x, f(x)) \). For our purposes, we want to approximate a physical quantity, i.e. \( \sigma^2_n(r) \), with a B-spline representation. This approximation is better if we have more knot points distributed around regions

\( k \) is the order of the B-spline basis.

\( \xi_i \) are the knot sequence, \( n \) is the number of knots, and \( k \) is the order of the B-spline basis.

\( a_i \) are the control points and \( B_{i,k}(x) \) are the B-spline basis functions.

\( \xi_i \) are the knot sequence, \( n \) is the number of knots, and \( k \) is the order of the B-spline basis.

\( a_i \) are the control points and \( B_{i,k}(x) \) are the B-spline basis functions.

\( \xi_i \) are the knot sequence, \( n \) is the number of knots, and \( k \) is the order of the B-spline basis.

\( a_i \) are the control points and \( B_{i,k}(x) \) are the B-spline basis functions.

\( \xi_i \) are the knot sequence, \( n \) is the number of knots, and \( k \) is the order of the B-spline basis.

\( a_i \) are the control points and \( B_{i,k}(x) \) are the B-spline basis functions.

\( \xi_i \) are the knot sequence, \( n \) is the number of knots, and \( k \) is the order of the B-spline basis.

\( a_i \) are the control points and \( B_{i,k}(x) \) are the B-spline basis functions.

\( \xi_i \) are the knot sequence, \( n \) is the number of knots, and \( k \) is the order of the B-spline basis.

\( a_i \) are the control points and \( B_{i,k}(x) \) are the B-spline basis functions.
where our function has greater curvature. If we have no information on where this region might be, we use a uniform distribution of knot points.

For a King mass model, we know that a critical distance from the cluster centre is the King core radius \(r_i\). It is around this point where the several functions of the model appear to have increased curvature. Therefore, we have the option of using a Gaussian knot distribution with mean \(\mu = r_i\) and variance \(\sigma = r_i/\gamma\), where \(r_i\) is the tidal radius of our system. The parameter \(\gamma\) regulates how close to the mean the distribution of points will be. A large value of \(\gamma\) concentrates points around \(r_i\). A value \(\gamma = 1\) results in an approximately uniform distribution in the interval \(\xi \in [0, r_i]\).

Let \(\alpha\) be a uniform sequence of numbers in the interval \([0, 1]\). For this sequence, the following equation gives rise to a Gaussian distribution of points around the mean \(\mu\) with variance \(\sigma\):

\[
\xi(\alpha) = \mu + \sqrt{2\sigma} \text{erf}^{-1} \left[ \frac{\mu - \alpha}{\sqrt{2\sigma}} \left( 1 + \sqrt{\frac{1}{\gamma}} \right) \right].
\]

Equation (10) is produced with the same methodology we use when we wish to create a random Gaussian number \(\xi \in [0, r_i]\) from a uniform random \(\alpha \in [0, 1]\).

In Fig. 2, we plot several possible choices of knot distributions. Specifically, the left-hand panel demonstrates a Gaussian distribution of points around mean \(\mu = 25\) with coefficient \(\gamma = 5\). The middle panel has a uniform distribution of knots, while the right-hand panel is an exponential knot distribution with the majority of knots concentrated exponentially close to the origin.

### 3.5 The SSJE

In this section, we describe how we combine the SSJE with the line-of-sight velocity dispersion \(\sigma_{los}^2\) in order to facilitate comparison with observables. For the case of our Galaxy, where typically, one only has radial velocities, see Appendix A.

In order to apply our method, we write equations (1) and (3) in what we believe to be a much simpler form in terms of \(\sigma^2_\mu\) and \(\sigma^2_\phi\). Furthermore, we simplify the notation by setting

\[
\psi = \sigma^2_\mu(r)
\]

\[
\phi = \sigma^2_\phi(r).
\]

Then the SSJE and \(\sigma^2_{los}(R)\) in the \(\psi, \phi\) representation are

\[
- \frac{d\Phi}{dr} = \frac{d\psi}{dr} + \left( \frac{1}{\rho} \frac{d\rho}{dr} + \frac{2}{r} \right) \psi - \frac{1}{r} \phi
\]

\[
\sigma^2_{los} = \frac{1}{\Sigma(R)} \int_R^{\infty} \rho \left[ 2\psi(r^2 - R^2) + \phi R^2 \right] \, dr.
\]

As we shall see in Section 3.6, the tidal radius \(r_i\) of the system is defined through Poisson’s equation from \(\rho(r)\) and \(\Phi(r)\) and does not depend on the kinematic quantities \(\sigma^2_{\mu}\) or \(\sigma^2_{\phi}\). The problem with equations (13) and (14) is that both functions \(\psi\) and \(\phi\) are unknown, and cannot be deduced from the mass density \(\rho(r)\) or the potential \(\Phi(r)\) of the system. Moreover, \(\psi\) participates also with its first derivative, making the problem even more complex.

We are going to consider the expansion of \(\psi(r)\) in a B-spline basis function of order \(k\). That is,

\[
\psi(r) = \sum_{i=1}^{N_{coeff}} a_i B_{i,k}(r),
\]

where \(B_{i,k}(r)\) are known B-spline basis functions. Then the derivative of this function is merely

\[
\psi^{(1)}(r) = \sum_{i=1}^{N_{coeff}} a_i B_{i,k}^{(1)}(r),
\]

where \(\psi^{(1)}(r) = d\psi/dr\) and \(B_{i,k}^{(1)}(r) = dB_{i,k}(r)/dr\). That is, the derivative of \(\psi\) depends on the same unknown coefficients \(a_i\) but is expanded in a new set of basis functions \(B_{i,k}^{(1)}(r)\). This removes the complexity of not knowing the derivative of \(\psi(r)\). Substituting \(\phi\) from equation (13) in the integrand of \(\sigma^2_{los}\) (equation 14) yields

\[
\sigma^2_{los} = \frac{1}{\Sigma(R)} \int_R^{\infty} \left( 2\rho r + \rho^{(1)}(r^2 - R^2)^{1/2} \psi + \rho R^2 \psi^{(1)} \right) \, dr + \frac{1}{\Sigma(R)} \int_R^{\infty} \frac{\rho R^2}{\sqrt{r^2 - R^2}} \frac{d\Phi}{dr} \, dr.
\]

where \(\rho^{(1)} = d\rho(r)/dr\). Now the line-of-sight velocity dispersion depends on the mass density of the system, the potential and the unknown function \(\psi\) along with its first derivative \(\psi^{(1)}\). Using the basis expansion (equations 15 and 16) yields

\[
\sigma^2_{los} = \sum_{i=1}^{N_{coeff}} a_i \frac{1}{\Sigma(R)} \int_R^{\infty} \left( 2\rho r + \rho^{(1)}(r^2 - R^2)^{1/2} B_{i,k}(r) + \rho R^2 B_{i,k}^{(1)}(r) \right) \, dr + \frac{1}{\Sigma(R)} \int_R^{\infty} \frac{\rho R^2}{\sqrt{r^2 - R^2}} \frac{d\Phi}{dr} \, dr.
\]

We define the following functions:

\[
I_i(R) = \frac{1}{\Sigma(R)} \int_R^{\infty} \left( 2\rho r + \rho^{(1)}(r^2 - R^2)^{1/2} + \rho R^2 B_{i,k}^{(1)}(r) \right) \, dr,
\]

\[
C(R) = \frac{1}{\Sigma(R)} \int_R^{\infty} \frac{\rho(r)R^2}{\sqrt{r^2 - R^2}} \frac{d\Phi(r)}{dr}. 
\]

Then, the value of \(\sigma^2_{los}(R)\) is given by

\[
\sigma^2_{los}(R) = \sum_{i=1}^{N_{coeff}} a_i I_i(R) + C(R).
\]
model. That is, although we cannot know the velocity profile of the
cluster from its mass density \( \rho(r) \) or the potential, we may allow this
to be deduced from the observables. Knowledge of \( a_i \) is equivalent
to knowledge of \( \sigma^2_{\theta} \equiv \psi \) and \( \sigma^2_{\phi} \equiv \phi \).

The coefficients \( a_i \) cannot take arbitrary values. One restriction
to be applied is that both \( \psi \equiv \sigma^2_{\theta} \) and \( \phi \) functions (equation 13) are
positive. Moreover, the function \( \psi \), must equal zero at \( r = r_i \), the
tidal radius of the system, and we assume that \( \phi \) and \( \sigma^2_{\phi} \) are also
zero at \( r = r_i \). This last condition, combined with the smoothness of
B-spline functions, imposes a severe restriction on the possible
cases, \( \sigma \) is known.

Therefore, we do not include \( \psi \) a y C in which we demand
\( \sigma \) is equivalent \( \sigma \) as defined from the DF (equation 22).

\[ r = E \sigma^2 \]

\[ r \geq \psi \]

\[ w \rightarrow \phi \]

\( |w| \propto \) is the radial component of the velocity in spherical coordinates
\( v \rho \) where \( \tilde{\sigma} = \sigma \).

\[ \rho = \rho (r, w(r)) \bigg|_{r=r_i}, r_c = \left( \frac{9 \sigma^2}{4 \pi G \rho_0} \right)^{1/2}. \]

Then for the full description of a King model, we use the following
set of parameters \((w_0, \rho_0, r_c)\). Using the transformed potential \( w \),
the Poisson equation is most conveniently written

\[ \nabla^2 w(r) = - \frac{\rho(r, w)}{\rho_0}. \]  

(24)

The steps followed for a full evaluation of a King model are the following.

(i) Assign initial values to parameters \((w_0, \rho_0, r_c)\).

(ii) Subject to the initial conditions \( w(r = 0) = w_0 \) and

\[ \frac{dw}{dr} \bigg|_{r=r_i} = 0, \]  

solve Poisson’s equation numerically, to thus obtain \( w(r) \).  

(iii) The mass density \( \rho(r) \) is fully determined upon knowledge
of \( w(r) \).

In the following, we are going to use only the King mass density,
and pretend that we do not know the kinematic quantities \( \sigma^2_{\theta} \)
and \( \sigma^2_{\phi} \) as defined from the DF (equation 22).

4. Statistical Analysis

In this section, we will be using standard Bayesian approaches
model fitting. The reader is directed to standard texts such
as Hastie, Tibshirani & Friedman (2001), Sivia & Skilling (2006)
and Gregory (2010) for further details.

4.1 Likelihood Function

Let \( \theta \) represent the vector of parameters needed to fully describe
a given assumed physical model. These will be the set of defining
parameters of the dynamical model, and the coefficients \( a_i \)
of the B-spline representation of \( \psi \equiv \sigma^2_{\theta} \), i.e. \( \theta = (w_0, \rho_0, \psi, \sigma^2_{\phi}, \sigma^2_{\theta})\).

In this paper, we consider for simplicity an example
of B-spline functions, imposes a severe restriction on the possible
values of \( a_i \). The result is well-defined curves with small error bars.

We will see later that closer to \( r_i \) the variance of the
coefficients becomes small.

From equations (19) and (20), we see that \( C(r_i) = I_i(r_i) = 0 \) by
definition, since the lower and upper limits of the integrals coincide.

Then, we impose the ad hoc restriction that \( \sigma^2_{\theta} \) \( \rightarrow \) 0 as \( r \rightarrow r_i \).

This is easily achieved by adding an artificial data point very close
to \( r_i \) in which we demand \( \sigma^2_{\theta} \approx 0 \) within some very small error.

3.6 Dynamical models

In the following sections, we will reconstruct from synthetic data
the kinematic profile of a stellar system in equilibrium, i.e. \( \sigma^2_{\phi} \)
and \( \sigma^2_{\theta} \) (once \( \sigma^2_{\phi} \) is known, \( \sigma^2_{\theta} \) can be found from the SSJE). We will
assume that the stellar mass content of this system is described by
a King model mass density \( \rho(r) \). In the current contribution, this is
the only mass density we are going to consider. For systems that
contain also a DM component, see Paper II.

For a full description of King models, the reader should consult
King (1966) and Binney & Tremaine (2008). Here, we give for
reference the functional forms we used. A King model is defined through its DF

\[ f(\mathcal{E}) = \begin{cases} \frac{1}{2} \left( e^{-\frac{\mathcal{E}}{\sigma^2}} - 1 \right) & \mathcal{E} < 0 \\ 0 & \mathcal{E} \geq 0 \end{cases}, \]  

(22)

where \( f_0 \) and \( \sigma \) are parameters to be determined from Bayesian
likelihood methods.

Let \( r_i \) denote the tidal radius of the system, i.e. a position beyond
which the mass density and all physical quantities of the system vanish. If \( \Phi(r) \) is the potential, by making use of an arbitrary additive
constant to its definition, we may define as a new potential the
difference:

\[ \Psi = \Phi(r) - \Phi(r_i); \text{now } \Psi \text{ vanishes at the tidal radius.} \]

Furthermore, in order to simplify our calculations, we introduce the
transformation:

\[ w = -\Psi(r)/\sigma^2. \]  

Then,

\[ \mathcal{E} = v_\phi^2 + v_\theta^2 - 2\sigma^2 w(r). \]  

(23)

The mass density of the system \( \rho(r, w) \) can be calculated analytically
with the use of Computer Algebra Systems (e.g. Maxima, Mathematica, Maple), as functions of \( r \) and ‘potential’ \( w(r) \):

\[ \rho(r, w) = 4\pi \rho_0 \int_{v_\phi = 0}^{\sqrt{2\sigma^2 w - v_\phi^2}} \int_{v_\theta = 0}^{\sqrt{2\sigma^2 w - v_\phi^2}} f(\mathcal{E}, L) v_\phi dv_\phi dv_\theta. \]

\( v_\phi \) is the radial component of the velocity in spherical coordinates
\( (v_\phi, v_\theta, v_\phi) \) and \( v^2 = v_\phi^2 + v_\theta^2 \). A model is fully described once we
assign values to its defining parameters and know the functional
form of the ‘potential’ \( w(r) \). The latter is achieved by solving
Poisson’s equation numerically. To do this, we require two additional
assumptions at \( r = 0 \): an initial value for the potential \( w_0 \) and
the equilibrium condition \( \frac{dw}{dr} \bigg|_{r=0} = 0 \).
when $\theta \in \Delta \theta_j$ and 0 otherwise. $N_{\text{data}}$ represents the total number of parameters and $\Delta \theta_j$ the range of possible values for parameter $i$. $\mathcal{L}(D|\theta)$ is the likelihood model.

Our likelihood model must take into account both the brightness and kinematic data. Since these two data sets are mutually independent it follows that

$$\mathcal{L}(D|\theta) = \mathcal{L}(D|\theta)\mathcal{L}_K|\theta) \equiv \mathcal{L}_B|\mathcal{L}_K. \quad (27)$$

For $\mathcal{L}_B$ and $\mathcal{L}_K$, we choose standard Gaussian distributions, i.e.

$$\mathcal{L}_B = \prod_{i=1}^{N_{\text{data}}} \frac{1}{\sqrt{2\pi \delta J_i}} \exp \left(-\frac{(J_i - \Sigma(R)_i)^2}{2(\delta J_i)^2}\right) \quad (28)$$

$$\mathcal{L}_K = \prod_{i=1}^{N_{\text{data}}} \frac{1}{\sqrt{2\pi \delta d_i}} \exp \left(-\frac{(d_i - \sigma_{\text{los}}(R)_i)^2}{2(\delta d_i)^2}\right). \quad (29)$$

$J_i$ is our brightness data values, $\delta J_i$ the error in each value. $d_i$ is the line-of-sight $\sigma_{\text{los}}(R)_i$ data value at position $R_i, \delta d_i$ the corresponding error. In the example presented in Section 5, the brightness $J_i$ and line-of-sight velocity dispersion observables $d_i$ are evaluated on the same positions, $R_i$, however this need not be the case and this does not affect the efficiency of the method.

In order to estimate the highest likelihood values of the parameters $\theta$, we employed an MCMC algorithm, namely a stretch move as described in Goodman & Weare (2010). This method has the advantage of exploring the parameter space efficiently, and the fitted parameters generally do not get stuck around local maxima of the likelihood function. This is an important feature since if there is a degeneracy in pair of $(\rho, \beta)$ values, then we must recover multimodal distributions for the parameters $a_i$. Our MCMC walks were run for sufficient autocorrelation time, so as to ensure that the distributions of parameters were stabilized around certain values.

Two important remarks need to be made here: due to the complexity of the problem, if we increase the number of $a_i$ coefficients to more than 15, the autocorrelation time\(^9\) becomes very large. There were cases in our initial trial runs where we needed to run our MCMC for up to $10^7$ points, because the chains were converging very slowly. The behaviour of the chains is different than in standard parameter estimation of functions. Specifically, the values tend to concentrate at some region quite fast, and then this whole region oscillates slowly until it is eventually stabilized. In general, for our models, we run our MCMC for approximately $2-4 \times 10^6$ points, and this was sufficient. However, we did not need to use more than $7-12$ unknown $a_i$ coefficients.

### 4.2 Bayesian model selection

In our present description, we use Bayesian model selection (Gelman et al. 2003; Gregory 2010) and Nested Sampling (Skilling 2004, hereafter JS04), a method for estimating the evidence for a given likelihood model. For completeness, we give a short introduction to these methods.

Let $M_i$ represent each of the models used in our analysis (e.g. $M_i \equiv \text{King mass density}$, with specific number $n$ of coefficients $a_i$, order of B-spline $k$ and knot distribution). Furthermore, let $I = M_1 + \ldots + M_n$ represent our hypothesis, that at least one of the models is correct. Summation indicates logical ‘or’. Let $\theta$ represent the total number of parameters for each model and $D$ our data set. According to Bayes theorem, the probability of the model parameters $\theta$ given the data set of values is

$$p(\theta|D, M_i, I) = \frac{p(\theta|M_i, I)\mathcal{L}(D|\theta, M_i, I)}{p(D|M_i, I)}. \quad (30)$$

$p(\theta|M_i, I)$ is the prior information on the parameters, $\mathcal{L}(D|\theta, M_i, I)$ is the likelihood as defined in equation (27) and $p(D|M_i, I)$ is the normalization constant for the model $M_i$ under consideration. This constant plays an important role for model selection. Marginalizing over all parameters, for the set of competing hypothesis, the probability of a model given the data is

$$p(M_i|D, I) = \int p(M_i|\theta, M_i, I)p(D|\theta, M_i, I)d\theta. \quad (31)$$

Our level of ignorance of model choice suggests that $p(M_i|D, I) = p(M_i|I)$ for any $i, j$ combination (all models are equiprobable). Hence, the relative ratio of probabilities of two models is

$$p(M_i|D, I) = \frac{p(M_i|I)p(D|M_i, I)}{p(M_j|I)p(D|M_j, I)} = \frac{p(D|M_i, I)}{p(D|M_j, I)} = O_{ij}. \quad (32)$$

$O_{ij}$ is defined as the odds ratio, and it quantifies the comparison of two competing models for the description of observables. $p(D|I)$ is the normalization constant that does not participate in our calculations each time we compute the relative ratio of two models. A measure for model selection is given by Jeffreys table (Table 1). It quantifies the relative ratio of probabilities $p(M_i)/p(M_j)$ of two competing models. See Jeffreys (1961) and Gelman et al. (2003) for further details.

**Table 1. Jeffreys table.**

| $p(M_1)/p(M_2)$ | Strength of evidence |
|-----------------|----------------------|
| $<0$            | Negative (supports M2) |
| 0–1.16          | Barely worth mentioning |
| 1.16–2.3        | Positive |
| 2.3–4.6         | Strong |
| $>4.6$          | Very strong – decisive |

Nested Sampling, introduced by JS04, is an algorithm for the estimation of the normalization parameter $p(D|M_i, I)$. Following his terminology, the evidence $Z_i$ of model $M_i$ is given by

$$Z_i = p(D|M_i, I) = \int p(\theta|M_i, I)\mathcal{L}(D|\theta, M_i, I)d\theta, \quad (33)$$

and corresponds to the normalization constant $p(D|M_i, I)$. Making use of the prior mass $dX = p(\theta|M_i, I)d\theta$, an effective parameter transformation from $\dim(\theta) = n$ to $\dim(X) = 1$, the above integral is simplified:

$$Z_i = \int_0^1 \tilde{\mathcal{L}}(D|X)dX. \quad (34)$$

In order to estimate this quantity and perform model selection, we use MULTINEST (Feroz & Hobson 2008; Feroz, Hobson & Bridges 2009). This algorithm is designed for effective calculation of Bayesian evidence based on Skilling’s algorithm. It gives consistent results even in the case of multimodal likelihood functions.

### 5 Example: Isotropic System with King Mass Density

In this section, we are going to reconstruct the kinematic profile of an isotropic King model ($\beta(r) = 0$). For the notation of the total
Marginalized distributions of mass model parameters for the isotropic system with King mass density and 14 data points, as estimated from our algorithm. In this case, we kept fixed the mass-to-light ratio. The ratio corresponds to $n = 5 + 1$ coefficients $a_i$. The red dashed lines correspond to the reference values $\{w_0^{\text{ref}}, \rho_0^{\text{ref}}, r_0^{\text{ref}}\} = \{5, 200, 5\}$ from which synthetic data were created. The reference values of the mass model are well within the boundaries of the estimated values.

Number $n$ of unknown coefficients $a_i$, we use the following scheme: based on the restriction that all the quantities that describe the cluster must be zero at the tidal radius, the final coefficient will be $a_5 = 0$. This extra coefficient does not go into the likelihood analysis, hence we break the total number of coefficients to the sum of unknowns plus one which represents this last coefficient. We use this notation in the figure captions and in Table 2 where we list the Bayesian evidence.

For each evaluation of our likelihood based on our parameters $\theta = (w_0, \rho_0, r_c, a_1, \ldots, a_{n-1})$, we need to construct a B-spline representation of the radial velocity dispersion $\psi \equiv \sigma_\psi^2 = \sum_i a_i B_i(r)$, $\sum_i a_i = 0$. We keep the order $k$ of the B-spline basis fixed. However, we use an adaptive knot distribution. For each set of proposed parameters that define the dynamical model $(w_0, \rho_0, r_c)$, there exists a unique tidal radius $r_c$. This is defined from the solution of the Poisson equation and depends on the brightness profile. We define the knot sequence $\xi_i$ from a Gaussian distribution of knots, around the mean $\mu = r_c$ for the choice of $\gamma = 3$ in the interval $r \in [0, r_c]$ (Section 3.4). The choice for the value of $\gamma$ was taken after many tests on synthetic data, by comparison of the Bayesian evidence for each fit. We use a mild concentration of knot points around $r_c$ since we know from theory that this is a point of interest in the sense of increased curvature of the corresponding functions. We need in general more knots around regions of increased curvature. We emphasize, however, that we can recover the correct kinematic profile also with a uniform knot distribution.

We use only 14 data points; this small number is used in order to demonstrate the power of the method for realistic applications. Our data consists of synthetic brightness values $J(R) = \Sigma(R)/\Gamma$ and synthetic line-of-sight velocity dispersion values $\sigma_\psi^2$. For simplicity, we assume a fixed mass-to-light ratio $\Gamma = 1$ and the following set of values for the defining parameter of the King reference profile: $\{w_0 = 5, \rho_0 = 200, r_c = 5\}$. Each of the synthetic profile values is constructed by adding a random error to the reference profile, either to the brightness $J$ or to the line-of-sight velocity dispersion $\sigma_\psi^2$. For this example, the error is a random 10% on the actual value of the reference profile. The set of random values is created by the following scheme:

$$J_i = J(R_i) + \delta J_i, \quad \delta J_i = 0.1J(R_i)g_1,$$

$$d_i = \sigma_\psi^2(R_i) + \delta d_i, \quad \delta d_i = 0.1\sigma_\psi^2(R_i)g_2,$$

where $g_1, g_2$ are two distinct Gaussian random numbers of mean zero and dispersion equal to one. $\delta J_i$ is the random error on the brightness value $J_i$, and $\delta d_i$ is the random error of the line-of-sight velocity dispersion $\sigma_\psi^2(R_i)$ at position $R_i$.

The first thing we need to establish is if the mass content of the system is recovered correctly. In order to have the most accurate results, we run MULTINEST for the evaluation of Bayesian evidence for a set of values for the order $k$ of the B-spline basis and the number of coefficients. The values of all these parameters can be seen in Table 2. Optimum choices result for order $k = 4$ and $n = 6 + 1$ coefficients and order $k = 5$ and $n = 5 + 1$ coefficients. Since these values are the same within error estimates, we choose for our fits the model with $k = 5$ and $n = 5 + 1$. For this choice of parameters, we plot the highest likelihood fitting models.

In Fig. 3, we plot the histograms of the mass model defining parameters $(w_0, \rho_0, r_c)$. Despite the fact that we allowed completely free freedom in the kinematic profile $(\sigma_\psi^2)$ through the B-spline representation, the recovered parameters of the mass model are distributed around the reference values $(w_0 = 5, \rho_0 = 200$ and $r_c = 5$. This is the most important result: the mass content is completely reconstructed. We note that we could have left the mass-to-light ratio as a free parameter and this would still be recovered (Paper II). In Fig. 4, we plot the density plots of the MCMC walks for these parameters.

Next we need to see how well the kinematic profile, i.e. $\sigma_\psi^2$ and $\sigma_\psi^2$, is approximated. In the top-left panel of Fig. 5, we plot the line-of-sight velocity dispersion $\sigma_\psi^2$ of the synthetic data, the true curve, and the fitted curve. It can be seen that the fit is excellent. In the bottom-left panel, we plot the synthetic brightness data and the highest likelihood fitting profile. The simultaneous fit to brightness and kinematic data is excellent; this is important for cases with real data sets where we need to consider the brightness fit as well.

The top-right panel deserves special attention: we plot again the line-of-sight velocity dispersion and the 1σ uncertainty interval (yellow-shaded region) that corresponds to the uncertainty in the estimates of $a_i$, and does not account for the variance of the mass model parameters. The region close to the origin $(r = 0)$ demonstrates greater variance. This is to be expected, since there the data points end, and the coefficient $a_i = \sigma_\psi^2(0)$ exhibits greater variance. This is a result of the local modification scheme of the B-spline basis functions. Recall that at each point $x \in [\xi_1, \xi_{n+1}]$, only a finite number of B-spline bases $B_i(x)$ are non-zero. For the case of $a_1$, since it is at the beginning of the knot distribution, the only

The prior range for all coefficients $a_i$ that define $\sigma_\psi^2$ was in the range $(0, 50)$.

Once $\sigma_\psi^2$ is known, the tangential $\sigma_\psi^2$ can be evaluated from the SSJE.

### Table 2. Bayesian evidence for isotropic system with King mass model.

| Order | Number of coefficients $n$ | $\ln Z \pm \delta(\ln Z)$ |
|-------|---------------------------|----------------------------|
| $k = 4$ | $n = 5 + 1$ | $-50.15 \pm 0.19$ |
| $k = 4$ | $n = 6 + 1$ | $-41.19 \pm 0.20$ |
| $k = 5$ | $n = 5 + 1$ | $-41.44 \pm 0.20$ |
| $k = 5$ | $n = 6 + 1$ | $-45.06 \pm 0.20$ |

From left to right: first column is the order $k$ of the B-spline representation of $\psi \equiv \sigma_\psi^2$. Second column is the number $n$ of coefficients $a_i$. The third column is the value of Bayesian evidence as estimated from MULTINEST. The highest value of $\ln Z$ corresponds to the most probable model.
Figure 4. Density plots of the MCMC walks for the defining parameters \((w_0, \rho_0, r_c)\) of the King mass model.

Figure 5. Top-left panel: synthetic line-of-sight velocity dispersion data, \(\sigma^2_{\text{los}}\) reference value (red line) and fit (blue line) from the solution of the SSJE using \(n = 5 + 1\) coefficients \(a_i\) for the B-spline representation of \(\sigma^2_{\text{rr}}\). The order of the B-spline basis is \(k = 5\) and the fit is for 14 synthetic data points. Bottom-left panel: fit of the brightness profile of the mass model. Top-right panel: \(\sigma^2_{\text{los}}\) reference value (red line) and fit (blue line). We also plot the functions \(C(R)\) (equation 20) and the weighted \(a_i I_i(R)\) (equation 19) that define the value of \(\sigma^2_{\text{los}}\) (equation 21). Bottom-right panel: the theoretical \(\sigma^2_{\text{rr}}\) from the King profile and from the B-spline representation, as estimated by the MCMC procedure. We also plot the weighted B-spline basis \(a_i B_{n,k}(x)\). The yellow-shaded region in all panels corresponds to 1\(\sigma\) uncertainty intervals of the coefficients \(a_i\) keeping the defining parameters of the mass model fixed to the highest likelihood values.
Resolving the mass–anisotropy degeneracy

Figure 6. Density plots of the MCMC walks for the parameters $a_i$ in the B-splines representation of $\sigma^2_{rr} = \sum_i a_i B_{i,k}(r)$ for the isotropic system with a King mass model. For this fit, we used 14 synthetic data points; the order of the B-spline basis is $k = 5$.

non-zero B-spline basis is $B_{1,5}$. For all other intermediate points, more than one coefficient $a_i$ contributes to the curve estimate, thus the variance is smaller. As we move towards the tidal radius of the system $r_t$, the variance of the fitted values decreases. This results from the smoothness of B-spline functions as well as the boundary condition that the curve $P(r) = (r, \sigma^2_{rr})$ must pass through the point $(r_t, 0)$.

In the same panel, we plot the weighted functions $a_i I_i(r)$ and $C(R)$. $C(R)$ depends only on the mass density of the system (equation 20) and not in any way on the kinematic profile. This function rises up positive and then falls slowly asymptotically. On the other hand, the majority of the weighted functions\(^\text{12}\) acquire negative values for $R > 10$. Then the line-of-sight profile is constructed because the weighted functions $a_i I_i$ are subtracted from $C(R)$. That is, the coefficients $a_i$ are regulated by the fact that they must reduce the value of the sum $\sum_i a_i I_i(R) + C(R)$ in a smooth way until it drops to zero at $r_t$. This means that the mass profile defines the behaviour of the majority of coefficients $a_i$ away from the origin. Again, this results from the smoothness of B-spline functions and the requirement that they must pass through the point $(r_t, 0)$. This result is general, and does not depend on the specific B-spline representation of $\sigma^2_{rr}$.

In the bottom-right panel, we plot the highest likelihood $\sigma^2_{rr}$ and the true value. The fit is again excellent. The yellow-shaded region corresponds to the 1σ uncertainty interval of the $a_i$ coefficients only. In the same panel, we plot the weighted B-spline basis $a_i B_{i,k}(x)$. Their linear combination constructs $\sigma^2_{rr}$. Observe again that due to the smooth behaviour of the B-spline functions, and the requirement that all quantities drop to zero at $r = r_t$, the variance of the fitted values goes to zero as we move away from the system centre.

In Fig. 6, we plot the MCMC density plots for the various parameters $a_i$ of the B-spline representation of $\sigma^2_{rr}$. As the index $i$ increases the coefficients $a_i$ correspond to control points further away from the cluster centre. Their variance, as expected and as it is evident from the density plots, is reduced away from the cluster centre.

\(^{12}\)Recall that $a_i \geq 0$ from the restriction that $\sigma^2_{rr} \geq 0$, since $\sigma^2_{rr} = \sum_i a_i B_{i,k}(r)$ and $B_{i,k}(r) \geq 0$. Actually, this is a stronger constraint than necessary: it is possible to have some $a_i < 0$ and still the radial velocity profile be positive.
6 DISCUSSION

Having established the basic mathematical framework of our method, there remain some more issues to be addressed. An important question is how many coefficients \( a_i \) can we use? These cannot be arbitrary in number, as our choice is limited by the number of available data. This results from the local modification scheme of B-spline functions: if we use a very large number of coefficients, we end up with regions of the domain of definition of the B-spline function between data points, that are regulated by some \( a_i \) coefficients that do not participate in the likelihood function. Bayesian model inference can solve this problem, since it heavily penalizes models with increased complexity (greater number of coefficients).

In fact, this is so important that we cannot apply our method without using model selection, Bayesian or frequentist (e.g. Generalized Cross Validation). It may also be the case that our data are noisy and result in kinematic profiles (\( \sigma_r^2 \) and \( \sigma_\theta^2 \)) that have unphysical variations. This is related to the problem of underfitting, overfitting and optimum smoothing that we further develop in Paper II. There we will address the issue of optimum smoothing, by defining a curvature penalty on the B-spline representation using information of smoothness from ideal theoretical models.

Another important issue is whether the reconstructed kinematic profile is physically acceptable. When we observe a real stellar system, then this possesses a kinematic profile that is a physical realization, i.e. we cannot question whether \( \sigma_r^2 \) and \( \sigma_\theta^2 \) are correct, since this is dictated from nature (although we may question the reliability of the observations!). The question is if the decomposition to \( \sigma_r^2 \) and \( \sigma_\theta^2 \) is physically acceptable for a given mass density \( \rho(r) \). For a general single stellar model with a constant mass-to-light ratio, to the point where our assumption of mass density is a good approximation, then it must be. We expect that if we have a good approximation to the real brightness distribution then the kinematic profile must be also in good proximity with reality. If any of our assumptions is significantly flawed, then we will observe some unrealistic behaviour in \( \sigma_r^2 \) through the use of the SSJE.

Is the kinematic profile that reproduces the line-of-sight velocity dispersion, i.e. \( \sigma_r^2 \) and \( \sigma_\theta^2 \), unique? Let us assume that we know a complete theoretical functional form of \( \rho_{\text{los}} \) and the mass density, \( \rho(r) \), of a self-gravitating system. In principle, it is possible to find \( n \) positions \( R_i \) and form a linear system from equation (21) for the unknowns \( a_i \) in total. Care must be taken, since \( R_i \) positions must span all of the distance \([0, R]\) for the system to have a solution. A natural choice is to use the collocation points (Greville abscissae, see Section 3.3) for the \( R_i \). The matrix \( I(R_i) \) that is produced in this way, is a band matrix due to the local modification scheme of B-spline bases. Then, if this matrix \( I(R_i) \) is invertible, solving for \( a_i \) results in a unique solution to the system. It is important to note here that, a given knot distribution will result a unique solution; however, this solution may be far from optimum. It will trace the correct \( \sigma^2_{\text{los}} \) and \( \sigma^2_r \) profiles, but it may have non-physical variational behaviour. Using an optimization algorithm (e.g. Genetic Algorithm) for the number of coefficients, \( a_i \), positions of knots, \( \xi_i \), and order, \( k \), of the B-spline basis, it is in principle possible to find a solution for the unknowns \( a_i \) and knot distribution that will give in a desirable accuracy the kinematic profile \( \rho_{\text{los}} \), and in consequence, \( \sigma^2_r \). Within this desired accuracy, the kinematic profile will be unique. That is, for a given line-of-sight velocity dispersion \( \sigma^2_{\text{los}}(R) \) and a given potential–mass density pair \( (\Phi, \rho) \), if and only if \( I(R_i) \) is invertible, there exists a unique decomposition to \( \sigma^2_\theta \) and \( \sigma^2_r \). Therefore, there exists a unique anisotropy profile. The fact that the B-spline functions are an approximation to the true function \( \sigma^2_r \) should not worry us. B-spline bases form a set of ‘as complete as possible’ bases for the representation of a function. Actually, in the limit where \( n \rightarrow \infty \), a B-spline representation of a function is its Taylor expansion around some position \( x \) (De Boor 1978). For the case where we have discrete data, we cannot speak of a unique profile. Rather we have a unique family of profiles, within the statistical uncertainty of the model parameters as this is estimated from the MCMC. If the system is degenerate, using a Markov chain exploration of parameter space will in general result in multiple peaks in the marginalized distributions of the \( a_i \) coefficients and the defining parameters of the mass models.

7 CONCLUSIONS

In this paper, we describe the basic mathematical framework for the removal of the mass–anisotropy degeneracy of the SSJE. This is achieved to the level that for an assumed functional form of the pair \( (\Phi(r), \rho(r)) \) of potential and mass density, and a given data set of brightness \( I \) and line-of-sight velocity dispersion observables \( \sigma^2_{\text{los}} \), we reconstruct a unique kinematic profile \( (\sigma_r^2, \sigma_\theta^2) \), within the statistical uncertainties. The uniqueness of this follows from an exploration of parameter space through an MCMC scheme, i.e. we do not present a formal mathematical proof of why this is the case.

Our algorithm combines smoothing B-splines with dynamical equations of physical systems and reconstructs accurately the kinematic profile and the mass content of a stellar system. This is, for a constant or variable mass-to-light ratio \( \Upsilon \). In the current contribution, we present a simple example of an isotropic King profile with fixed mass-to-light ratio. We explore models with variable \( \Upsilon \) in Paper II.

Finally, we note that the idea of combining smoothing splines with equations from dynamics and allowing statistical inference to give the correct shape of unknown functions is quite general. It can be applied to any system in which there are physical quantities of unknown theoretical functional form where we wish to determine their approximate shape. This avoids bias in model parameter estimates and allows for a better understanding of physical models.

ACKNOWLEDGEMENTS

FID acknowledges the University of Sydney International Scholarship for the support of his candidature. GFL acknowledges support from ARC Discovery Project (DP110100678) and Future Fellowship (FT100100268). The authors would like to thank Nick Bate as well as the anonymous referee for useful comments and suggestions on the manuscript.

REFERENCES

Binney J., Mamon G. A., 1982, MNRAS, 200, 361
Binney J., Tremaine S., 2008, Galactic Dynamics, 2nd edn. Princeton Univ. Press, Princeton, NJ
De Boor C., 1978, Applied Mathematical Sciences. Springer-Verlag, Berlin Dejonghe H., Merritt D., 1992, ApJ, 391, 531
Diakogiannis F. I., Lewis G. F., Ibata R. A., 2014, MNRAS, in press (Paper II)
Farin G., 2002, Curves and Surfaces for CAGD: a Practical Guide, 5th edn.
Morgan Kaufmann Publishers Inc., San Francisco
Feroz F., Hobson M. P., 2008, MNRAS, 384, 449
Feroz F., Hobson M. P., Bridges M., 2009, MNRAS, 398, 1601
Gelman A., Carlin J., Stern H., Rubin D., 2003, Chapman & Hall/CRC
Texts in Statistical Science: Bayesian Data Analysis, 2nd edn. Chapman & Hall/CRC, London
APPENDIX A: OUR EXCEPTIONAL MILKY WAY

Relative to much of the halo, our Solar system resides close to the centre of our Galaxy. As a result, many observable stellar velocities are approximately radial velocities $v_r$. We need then to adapt our formalism to the fact that the direct observable quantity is now $\sigma_{rr}^2$. We expand the tangential velocity dispersion in a finite B-spline basis, according to

$$\sigma_{tt}^2 = \phi = \sum_i b_i B_{i,k}(r), \quad (A3)$$

where $b_i$ are constant coefficients to be determined, and define the shape of $\sigma_{tt}^2$. Then, the radial velocity dispersion $\sigma_{rr}^2$ is now expressed in terms of the unknown coefficients $b_i$ and the mass model. Substituting equation (A3) in equation (A2) and collecting terms, in a similar process as in Section 3.5, yields

$$\psi = \sum_{i=1}^n b_i \tilde{I}_i(r) + \tilde{C}(r), \quad (A4)$$

where

$$\tilde{I}_i(r) = \frac{1}{r^2 \rho} \int_0^r r \rho B_{i,k}(r) \, dr \quad (A5)$$

$$\tilde{C}(r) = -\frac{1}{r^2 \rho} \int_0^r r^2 \rho \, \frac{d\Phi}{dr} \, dr. \quad (A6)$$

Comparison of equation (A4) with observables, can result in marginalized distributions for the defining parameters of the mass model as well as the unknown coefficients $b_i$ that define the shape of $\sigma_{tt}^2$. Back substitution into equation (A4) results in the corresponding radial velocity dispersion $\sigma_{rr}^2 \equiv \psi(r)$.

Note that, although in our observables, we do not have information on the tangential velocity distribution, this knowledge is acquired through the use of the SSJE. Indeed, knowing $\sigma_{tt}^2$ or $\sigma_{rr}^2$, we can calculate one from the other. Also, due to the higher number of data points compared to the unknown parameters, equation (A4) can be used for estimation of the kinematic profile and of the mass model.

This paper has been typeset from a TeX/LaTeX file prepared by the author.