

ACTION-BASED DYNAMICAL MODELLING OF THE MILKY WAY DISK WITH *ROADMAPPING* AND OUR IMPERFECT KNOWLEDGE OF THE “REAL WORLD”

WILMA H. TRICK^{1,2}, JO BOVY³, AND HANS-WALTER RIX¹

Draft version September 18, 2015

ABSTRACT

We present *RoadMapping*, a dynamical modelling machinery that aims to recover the Milky Way’s (MW) gravitational potential and the orbit distribution of stellar populations in the Galactic disk. *RoadMapping* is a full likelihood analysis that models the observed positions and velocities of stars with an equilibrium, three-integral distribution function (DF) in an axisymmetric potential. In preparation for the application to the large data sets of modern surveys like Gaia, we create and analyze a large suite of mock data sets and develop qualitative “rules of thumb” for which characteristics and limitations of data, model and machinery affect constraints on the potential and DF most. We find that, while the precision of the recovery increases with the number of stars, the numerical accuracy of the likelihood normalisation becomes increasingly important and dominates the computational efforts. The modelling has to account for the survey’s selection function, but *RoadMapping* seems to be very robust against small misjudgments of the data completeness. Large radial and vertical coverage of the survey volume gives in general the tightest constraints. But no observation volume of special shape or position and stellar population should be clearly preferred, as there seem to be no stars that are on manifestly more diagnostic orbits. We propose a simple approximation to include measurement errors at comparably low computational cost that works well if the distance error is $\lesssim 10\%$. The model parameter recovery is also still possible, if the proper motion errors are known to within 10% and are $\lesssim 2 \text{ mas yr}^{-1}$. We also investigate how small deviations of the stars’ distribution from the assumed DF influence the modelling: An over-abundance of high velocity stars affects the potential recovery more strongly than an under-estimation of the DF’s low-velocity domain. Selecting stellar populations according to mono-abundance bins of finite size can give reliable modelling results, as long as the DF parameters of two neighbouring bins do not vary more than 20% [TO DO: CKECK]. As the modelling has to assume a parametric form for the gravitational potential, deviations from the true potential have to be expected. We find, that in the axisymmetric case we can still hope to find a potential that is indeed a reliable best fit within the limitations of the assumed potential. Overall *RoadMapping* works as a reliable and unbiased estimator, and is robust against small deviations between model and the “real world”.

Keywords: Galaxy: disk — Galaxy: fundamental parameters — Galaxy: kinematics and dynamics — Galaxy: structure

1. INTRODUCTION

Stellar dynamical modelling can be employed to infer the Milky Way’s gravitational potential from the positions and motions of individual stars (Binney & Tremaine 2008; Binney 2011; Rix & Bovy 2013). Observational information on the 6D phase-space coordinates of stars is currently growing at a rapid pace, and will be taken to a whole new level in number and precision by the upcoming data from the Gaia mission (Perryman et al. 2001). Yet, rigorous and practical modelling tools that turn position-velocity data of individual stars into constraints both on the gravitational potential and on the distribution function (DF) of stellar orbits, are scarce (Rix & Bovy 2013) [TO DO: more references] [TO DO: References that explain that the modelling is scarce, or previous modelling approaches??] [TO DO: Hans-Walter suggested a Sanders & Binney ref-

erence, but I’m still not sure to what kind of paper: modelling approach or review of scarce modelling tools...]

The Galactic gravitational potential is fundamental for understanding the Milky Way’s dark matter and baryonic structure (Rix & Bovy 2013; McMillan 2012; Strigari 2013; Read 2014) and the stellar-population dependent orbit distribution function is a basic constraint on the Galaxy’s formation history (Binney 2013; Rix & Bovy 2013; Sanders & Binney 2015) [TO DO: more references].

There is a variety of practical approaches to dynamical modelling of discrete collisionless tracers, such as the stars in the Milky Way (e.g. Jeans modelling: Kuijken & Gilmore (1989), Bovy & Tremaine (2012), Garbari et al. (2012), Zhang et al. (2013), Büdenbender et al. (2015); action-based DF modelling: Bovy & Rix (2013), Piffl et al. (2014), Sanders & Binney (2015); torus modelling: McMillan & Binney (2012, 2013); Made-to-measure modelling: Syer & Tremaine (1996), De Lorenzi et al. (2007) or Hunt & Kawata (2014). Most of them – explicitly or implicitly – describe

Electronic address: trick@mpia.de

¹ Max-Planck-Institut für Astronomie, Königstuhl 17, D-69117 Heidelberg, Germany

² Correspondence should be addressed to trick@mpia.de.

³ University of Toronto [TO DO: What is Jo’s current address??]

the stellar distribution through a distribution function.

Actions are good ways to describe orbits, because they are canonical variables with their corresponding angles, have immediate physical meaning, and obey adiabatic invariance (McMillan & Binney 2008; Binney 2010; Binney & McMillan 2011; Binney 2011). Recently, Binney (2012) and Bovy & Rix (2013) [TO DO: are these the correct references??] proposed to combine parametrized axisymmetric potentials with DF’s that are simple analytic functions of the three orbital actions to model discrete data. Binney (2010) and Binney & McMillan (2011) had proposed a set of simple action-based (quasi-isothermal) distribution functions (qDF). Ting et al. (2013) and Bovy & Rix (2013) showed that these qDF’s may be good descriptions of the Galactic disk, when one only considers so-called mono-abundance populations (MAP), i.e. sub-sets of stars with similar $[\text{Fe}/\text{H}]$ and $[\alpha/\text{Fe}]$ (Bovy et al. 2012b,c,d).

Bovy & Rix (2013) implemented a rigorous modelling approach that put action-based DF modelling of the Galactic disk in an axisymmetric potential in practice. Given an assumed potential and an assumed DF, they directly calculated the likelihood of the observed (\vec{x}, \vec{v}) for each sub-set of MAP among SEGUE G-dwarf stars (Yanny et al. 2009). This modelling also accounted for the complex, but known selection function of the kinematic tracers. For each MAP, the modelling resulted in a constraint of its DF, and an independent constraint on the gravitational potential, which members of all MAPs feel the same way.

Taken as an ensemble, the individual MAP models constrained the disk surface mass density over a wide range of radii ($\sim 4 - 9$ kpc), and proved a powerful constraint on the disk mass scale length and on the disk-to-dark-matter ratio at the Solar radius.

Yet, these recent models still leave us poorly prepared with the wealth and quality of the existing and upcoming data sets. This is because Bovy & Rix (2013) made a number of quite severe and idealizing assumptions about the potential, the DF and the knowledge of observational effects (such as the selection function). All these idealizations are likely to translate into systematic error on the inferred potential or DF, well above the formal error bars of the upcoming data sets.

In this work we present *RoadMapping* (“Recovery of the Orbit Action Distribution of Mono-Abundance Populations and Potential INference for our Galaxy”) - an improved and refined version of the original dynamical modelling machinery by Bovy & Rix (2013), making extensive use of the *galpy* Python package (Bovy 2015) and the *Stäckel Fudge* for fast action calculations by Binney (2012). *RoadMapping* is robust and well-tested and explicitly developed to exploit and deal with the large data sets of the future. *RoadMapping* explores and relaxes some of the restraining assumptions that Bovy & Rix (2013) made and is more flexible and more adept in dealing with large data sets. In this paper we set out to explore the robustness of *RoadMapping* against the breakdowns of some of the most important

assumptions of DF-based dynamical modelling. Our goal is to examine which aspects of the data, the model and the machinery itself limit our recovery of the true gravitational potential.

In the light of the imminent Gaia data, we analyze how well *RoadMapping* behaves in the limit of large data. For a huge number of stars three aspects become important, that may be hidden behind Poisson noise for smaller data sets: (i) We have to make sure that *RoadMapping* is an unbiased estimator (Section 3.1). (ii) Numerical inaccuracies in the actual modelling machinery must not be an important source of systematics (Section 2.6). (iii) As parameter estimates become much more precise (Section 3.1, we need more flexibility in the potential and DF model. The modelling machinery therefore has to be effective in finding the best fit parameters for a large set of free model parameters. The improvements made in *RoadMapping* as compared to the machinery used in Bovy & Rix (2013) are presented in Section 2.7.

We also explore how different aspects of the observational experiment design impact the parameter recovery. (i) In an era where we can choose data from different MW surveys, it might be worth to explore the importance of the survey volume geometry, size and shape, and if different regions within the MW might be especially diagnostic to constrain the potential (Section 3.2). (ii) What if our knowledge of the sample selection function is imperfect, and potentially biased (Section 3.3)? (iii) How to best account for individual measurement errors in the modelling (Section 3.4)?

One of the strongest assumptions is to restrict the dynamical modelling to a certain family of parametrized models. We investigate how well we can hope to recover the true potential, when our potential and DF models do not encompass the true potential and DF. First, we examine in Section ?? what would happen if the stars within MAPs do intrinsically not follow a single qDF as assumed by Ting et al. (2013) and Bovy & Rix (2013). Second, we test in Section ?? how well the modelling works, if our assumed potential family deviates from the true potential.

The strongest assumption that goes into this kind of dynamical modelling might be the idealization of the Galaxy to be axis-symmetric and being in steady state. We do not investigate this within the scope of this paper but strongly suggest a systematic investigation of this for future work.

For all of the above aspects we show some plausible and illustrative examples on the basis of investigating mock data. The mock data is generated from galaxy models presented in Sections 2.1-2.4 following the procedure in Section 2.5, analysed according to the description of the *RoadMapping* machinery in Sections 2.6-2.7 and the results are presented in Section 3 and discussed in Section ??.

[TO DO: Comment from Hans-Walter: Make sure, any topic/issue appears only once] [TO DO: Is now one quarter shorter than before. But maybe shorten it even more...] [TO DO: Comment from Hans-Walter: Make

clear "new in this paper", "general background", "exactly as in BR13"]

2. DYNAMICAL MODELLING

[TO DO: HW: In this section you have to indicate somehow, where you recapitulate BR13 and what is added new. "as in BR13", "beyond BR13"]

In this section we summarize the basic elements of *RoadMapping*, the dynamical modelling machinery presented in this work, which in many respects follows Bovy & Rix (2013).

2.1. Coordinate System

Our modelling takes place in the Galactocentric rest-frame with cylindrical coordinates $\mathbf{x} \equiv (R, \phi, z)$ and corresponding velocity components $\mathbf{v} \equiv (v_R, v_\phi, v_z)$. If the stellar phase-space data is given in observed heliocentric coordinates, position $\tilde{\mathbf{x}} \equiv (\text{RA}, \text{DEC}, m - M)$ in right ascension RA, declination DEC and distance modulus $(m - M)$ as proxy for the distance from the sun, and velocity $\tilde{\mathbf{v}} \equiv (\mu_{\text{RA}}, \mu_{\text{DEC}}, v_{\text{los}})$ as proper motions $\boldsymbol{\mu} = (\mu_{\text{RA}}, \mu_{\text{DEC}})$ [TO DO: cos somewhere??] in both RA and DEC direction and line-of-sight velocity v_{los} , the data $(\tilde{\mathbf{x}}, \tilde{\mathbf{v}})$ has to be converted first into the Galactocentric rest-frame coordinates (\mathbf{x}, \mathbf{v}) using the sun's position and velocity. We assume for the sun

$$(R_\odot, \phi_\odot, z_\odot) = (8 \text{ kpc}, 0^\circ, 0 \text{ kpc})$$

$$(v_{R,\odot}, v_{T,\odot}, v_{z,\odot}) = (0, 230, 0) \text{ km s}^{-1}.$$

2.2. Actions and Potential Models

Orbits in axisymmetric potentials are best described and fully specified by the three actions $\mathbf{J} \equiv (J_R, J_z, J_\phi = L_z)$, defined as

$$J_i = \frac{1}{2\pi} \oint_{\text{orbit}} p_i dx_i, \quad (1)$$

and which depend on the potential via the connection between position x_i and momentum p_i along the orbit. Actions have a clear physical meaning: They quantify the amount of oscillation in each coordinate direction of the full orbit [TO DO: REF: HW suggested Binney & Tremaine (2008), but I can't find a corresponding statement in the book]. The position of a star along the orbit is denoted by a set of angles, which form together with the angles a set of canonical conjugate phase-space coordinates (Binney & Tremaine 2008, §3.5.1).

Even though actions are excellent orbit labels and arguments for stellar distribution functions, their computation is typically very expensive and depends on the choice of potential in which the star moves. The spherical isochrone potential (Henon 1959) is the only [TO DO: Jo suggested "most general Galactic" instead of "only", but the isochrone is actually not Galactic... Ask him.] potential for which Equation 1 takes an analytic form (Binney & Tremaine 2008, §3.5.2). For Stäckel potentials actions can be calculated exactly by the (numerical) evaluation of a single integral. In all other potentials numerically calculated actions will always be approximations, unless Equation 1 is integrated along the whole (often not periodic) orbit. A computational

fast way to get actions for arbitrary axisymmetric potentials is the *Stäckel fudge* by Binney (2012), which locally approximates the potential by a Stäckel potential. To speed up the calculation even more, an interpolation grid for J_R and J_z in energy E , angular momentum L_z and [TO DO: what else??] can be build out of these Stäckel fudge actions, as described in Bovy (2015).⁴

For the gravitational potential in our modelling we assume a family of parametrized potential models with a fixed number of free parameters. We use different kinds of potentials: The Milky Way like potential from Bovy & Rix (2013) (MW13-Pot) with bulge, disk and halo; the spherical isochrone potential (Iso-Pot) in our test suites to make use of the analytic (and therefore exact and fast) way to calculate actions; and the 2-component Kuzmin-Kutuzov Stäckel potential (Batsleer & Dejonghe 1994; KKS-Pot), which displays a disk and halo structure and also provides exact actions. Table ?? summarizes all reference potentials together used in this work with their free parameters p_Φ . The density distribution of these potentials is illustrated in Figure 1.

2.3. Stellar Distribution Functions

Throughout, we assume that the orbits of each MAP can be described by a single qDF of the form given by Binney & McMillan (2011). This is motivated by the findings of Bovy et al. (2012b,c,d) and Ting et al. (2013) about the simple phase-space structure of MAPs, and following Bovy & Rix (2013) and their successful application. This qDF has the form

$$\text{qDF}(\mathbf{J} | p_{\text{DF}}) = f_{\sigma_R}(J_R, L_z | p_{\text{DF}}) \times f_{\sigma_z}(J_z, L_z | p_{\text{DF}}) \quad (2)$$

with

$$f_{\sigma_R}(J_R, L_z | p_{\text{DF}}) = n \times \frac{\Omega}{\pi \sigma_R^2(R_g) \kappa} \exp\left(-\frac{\kappa J_R}{\sigma_R^2(R_g)}\right) \times [1 + \tanh(L_z/L_0)] \quad (3)$$

$$f_{\sigma_z}(J_z, L_z | p_{\text{DF}}) = \frac{\nu}{2\pi \sigma_z^2(R_g)} \exp\left(-\frac{\nu J_z}{\sigma_z^2(R_g)}\right). \quad (4)$$

Here $R_g \equiv R_g(L_z)$ and $\Omega \equiv \Omega(L_z)$ are the (guiding-center) radius and the circular frequency of the circular orbit with angular momentum L_z in a given potential. $\kappa \equiv \kappa(L_z)$ and $\nu \equiv \nu(L_z)$ are the radial/epicycle (κ) and vertical (ν) frequencies with which the star would oscillate around the circular orbit in R - and z -direction when slightly perturbed (Binney & Tremaine 2008, §3.2.3) [TO DO: ask someone, if I'm messing up different definitions of κ]. The term $[1 + \tanh(L_z/L_0)]$ suppresses counter-rotation for orbits in the disk with $L \gg L_0$ which we set to a small value ($L_0 = 10 \times R_\odot/8 \times v_{\text{circ}}(R_\odot)/220$ [TO DO: Jo said, galpy default is 10 km/s kpc. But I got the value actually from the code...]). To match the observed properties of MAPs (see

⁴ [TO DO: Write which numerical accuracy I needed for the grid, as the default values were not good enough.]

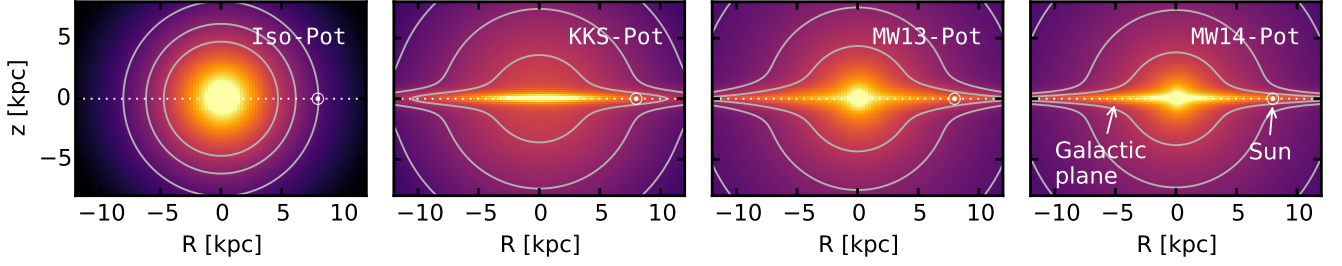


Figure 1. Density distribution of the four reference galaxy potentials in Table ??, for illustration purposes. These potentials are used throughout this work for mock data creation and potential recovery. [TO DO: Potential and/or population names in typewriter]

Bovy et al. 2012b,c,d), we chose the functional forms

$$n(R_g | p_{\text{DF}}) \propto \exp\left(-\frac{R_g}{h_R}\right) \quad (5)$$

$$\sigma_R(R_g | p_{\text{DF}}) = \sigma_{R,0} \times \exp\left(-\frac{R_g - R_\odot}{h_{\sigma,R}}\right) \quad (6)$$

$$\sigma_z(R_g | p_{\text{DF}}) = \sigma_{z,0} \times \exp\left(-\frac{R_g - R_\odot}{h_{\sigma,z}}\right), \quad (7)$$

which indirectly set the stellar number density and radial and vertical velocity dispersion profiles. The qDF for each MAP has therefore a set of five free parameters p_{DF} : the density scale length of the tracers h_R , the radial and vertical velocity dispersion at the solar position R_\odot , $\sigma_{R,0}$ and $\sigma_{z,0}$, and the scale lengths $h_{\sigma,R}$ and $h_{\sigma,z}$, that describe the radial decrease of the velocity dispersion. Throughout this work we use for illustration purposes a few example stellar populations, each following a single qDF, whose parameters are given in Table ?. Most tests use the `hot` and `cool` qDFs from Table ?, which correspond to kinematically hot and cool populations, respectively.

One crucial point in our dynamical modelling technique (§2.6), as well as in creating mock data (§2.5), is to calculate the (axisymmetric) spatial tracer density $\rho_{\text{DF}}(\mathbf{x} | p_\Phi, p_{\text{DF}})$ for a given qDF and potential. We do this by integrating the qDF at a given (R, z) over all three velocity components, using a N_v -th order Gauss-Legendre quadrature for each integral:

$$\rho_{\text{DF}}(R, |z| | p_\Phi, p_{\text{DF}}) = \int_{-\infty}^{\infty} \text{qDF}(\mathbf{J}[R, z, \mathbf{v} | p_\Phi] | p_{\text{DF}}) d^3\mathbf{v} \quad (8)$$

$$\approx \int_{-n_\sigma \sigma_R(R | p_{\text{DF}})}^{n_\sigma \sigma_R(R | p_{\text{DF}})} \int_{-n_\sigma \sigma_z(R | p_{\text{DF}})}^{n_\sigma \sigma_z(R | p_{\text{DF}})} \int_0^{1.5v_{\text{circ}}(R_\odot)} \text{qDF}(\mathbf{J}[R, z, \mathbf{v} | p_\Phi] | p_{\text{DF}}) dv_T dv_z dv_R, \quad (9)$$

where $\sigma_R(R | p_{\text{DF}})$ and $\sigma_z(R | p_{\text{DF}})$ are given by Equations 6 and 7 and the integration ranges are motivated by Figure 2. The integration range $[0, 1.5v_{\text{circ}}(R_\odot)]$ over v_T is in general sufficient (only for observation volumes at smaller Galactocentric radii with larger velocities this upper limit needs to be increased). For a given p_Φ and p_{DF} we explicitly calculate the density on $N_x \times N_x$ regular grid points in the (R, z) plane; in between grid points the density is evaluated with a bivariate spline interpolation.

The grid is chosen to cover the extent of the observations (for $|z| \leq 0$, because the model is symmetric in z by construction). The total number of actions that need to be calculated to set up the density interpolation grid is $N_x^2 \cdot N_v^3$. §2.6 and Figure ?? show the importance of choosing N_x , N_v and n_σ sufficiently large in order to get the density with an acceptable numerical accuracy [TO DO: Jo thinks that this statement is difficult to understand here, because you have not yet talked about the normalization].

2.4. Selection Functions

Any survey's selection function can be understood as defining an effective sample subvolume in the space of observables: e.g. position on the plane of the sky (the survey area), distance from the sun (limited by the brightness of the stars and the sensitivity of the detector), colors and metallicity of the stars (limited by survey mode and targeting).

We simply use spatial selection functions, which describe the probability to observe a star at \mathbf{x} ,

$$\text{sf}(\mathbf{x}) \equiv \begin{cases} \text{completeness}(\mathbf{x}) & \text{if } \mathbf{x} \text{ within observed volume} \\ 0 & \text{outside.} \end{cases}$$

For the observed volume we use simple geometrical shapes. Either a sphere of radius r_{max} with the sun at its center, or an angular segment of an cylindrical annulus (`wedge`), i.e. the volume with $R \in [R_{\text{min}}, R_{\text{max}}]$, $\phi \in [\phi_{\text{min}}, \phi_{\text{max}}]$, $z \in [z_{\text{min}}, z_{\text{max}}]$ within the model galaxy. The sharp outer cut of the survey volume could be understood as the detection limit in apparent brightness in the case, where all stars have the same luminosity. Here $0 \leq \text{completeness}(\mathbf{x}) \leq 1$ everywhere inside the observed volume, so it can be understood as a position-dependent detection probability. Unless explicitly stated otherwise, we simplify to $\text{completeness}(\mathbf{x}) = 1$.

2.5. Mock Data

We will rely on mock data as input to explore the limitations of the modelling. We investigate this, we assume first that our measured stars do indeed come from our assumed families of potentials and distribution functions and draw mock data from a given true distribution. Subsequently, we manipulate and modify these mock data sets to mimic observational effects.

The distribution function is given in terms of actions and angles. The transformation $(\mathbf{J}_i, \boldsymbol{\theta}_i) \rightarrow (\mathbf{x}_i, \mathbf{v}_i)$ is however difficult to perform and computationally much more

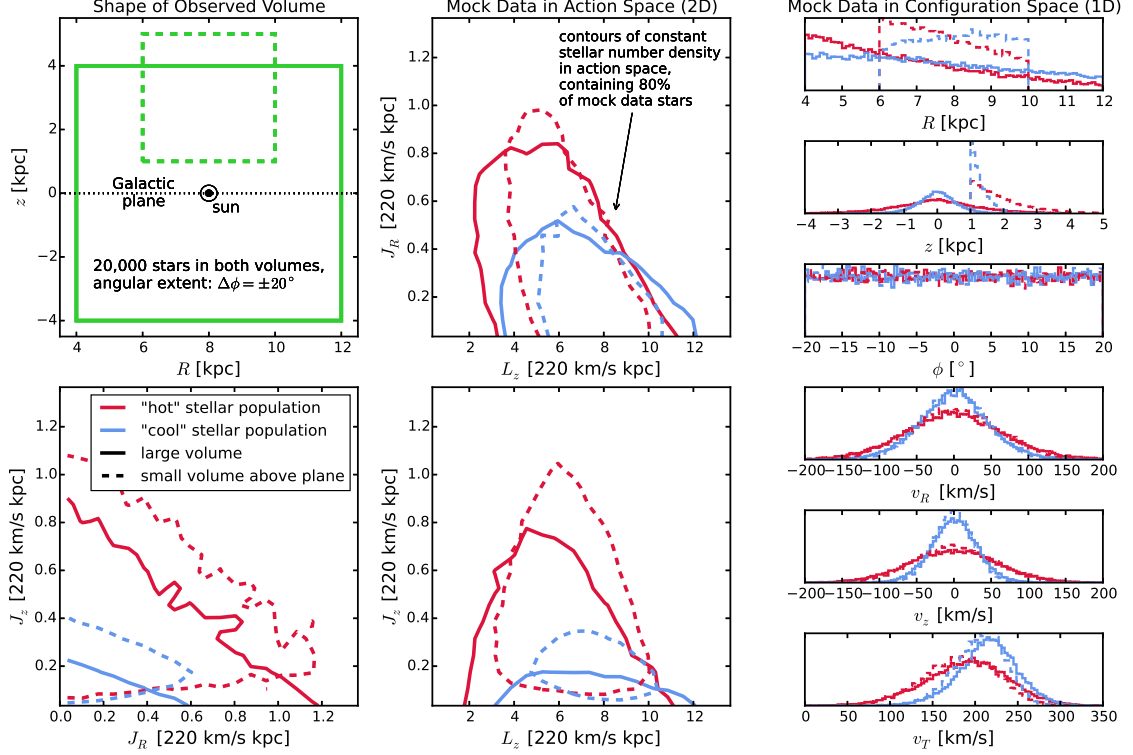


Figure 2. Distribution of mock data in action space (2D iso-density contours, enclosing 80% of the stars, the two central and the lower left panel) and configuration space (1D histograms, right panels), depending on shape and position of the survey observation volume and temperature of the stellar population. The parameters of the mock data model is given as Test ?? in Table ?. In the upper left panel we demonstrate the shape of the two different wedge-like observation volumes within which we were creating each a hot (red) and cool (blue) mock data set: a large volume centred on the Galactic plane (solid lines) and a smaller one above the plane (dashed lines). The distribution in action space visualizes how orbits with different actions also reach into different regions within the Galaxy. The 1D histograms on the right illustrate that qDFs generate realistic stellar distributions in galactocentric coordinates $(R, z, \phi, v_R, v_z, v_T)$. [TO DO: fancybox Legend] [TO DO: Potential and/or population names in typewriter font] [TO DO: Jo suggests to make two or three separate figures out of this. I'm not yet convinced, as I think it is nice and tidy like this.]

expensive than the transformation $(\mathbf{x}_i, \mathbf{v}_i) \rightarrow (\mathbf{J}_i, \boldsymbol{\theta}_i)$. We employ a fast and simple two-step method for drawing mock data from an action distribution function, which also accounts effectively for a given survey selection function.

In the first step we draw positions \mathbf{x}_i for our mock data stars from the selection function and tracer density. We start by setting up the interpolation grid for the tracer density $\rho(R, |z| | p_\Phi, p_{\text{DF}})$ generated by the given qDF and according to §2.3 and Equation 9. For the creation of the mock data we use $N_x = 20$, $N_v = 40$ and $n_\sigma = 5$. Next, we sample random positions (R_i, z_i, ϕ_i) uniformly within the entire observable volume. Then we apply a rejection Monte Carlo method to these positions using the pre-calculated $\rho_{\text{DF}}(R, |z| | p_\Phi, p_{\text{DF}})$. To apply a non-uniform selection function, $\text{sf}(\mathbf{x}) \neq \text{const.}$ within the observed volume, we use the rejection method a second time. The resulting sample then follows $\mathbf{x}_i \rightarrow p(\mathbf{x}) \propto \rho_{\text{DF}}(R, |z| | p_\Phi, p_{\text{DF}}) \times \text{sf}(\mathbf{x})$.

In the second step we draw velocities according to the distribution function. The velocities are independent of the selection function within the observed volume. For each of the positions (R_i, z_i) we sample velocities directly from the qDF $(R_i, z_i, \mathbf{v} | p_\Phi, p_{\text{DF}})$ using a rejection method. To reduce the number of rejected velocities, we use a Gaussian in velocity space as an envelope function,

from which we first randomly sample velocities and then apply the rejection method to shape the Gaussian velocity distribution towards the velocity distribution predicted by the qDF. We now have a mock data satisfying $(\mathbf{x}_i, \mathbf{v}_i) \rightarrow p(\mathbf{x}, \mathbf{v}) \propto \text{qDF}(\mathbf{x}, \mathbf{v} | p_\Phi, p_{\text{DF}}) \times \text{sf}(\mathbf{x})$.

Figure 2 shows examples of mock data sets in configuration space (\mathbf{x}, \mathbf{v}) and action space. The mock data from the qDF lead to the expected distributions in configuration space: More stars are found at smaller R and $|z|$, and are distributed uniformly in ϕ according to our assumption of axisymmetry. The distribution in radial and vertical velocities, v_R and v_z , is approximately Gaussian with the (total projected) velocity dispersion being $\sim \sigma_{R,0}$ and $\sim \sigma_{z,0}$ (see Table ?). The distribution of tangential velocities v_T is skewed because of asymmetric drift. The distribution in action space illustrates the intuitive physical meaning of actions: The stars of the cool population have in general lower radial and vertical actions, as they are on more circular orbits. The different relative distributions of the radial and vertical actions J_R and J_z of the hot and cool population is due to them having different velocity anisotropy $\sigma_{R,0}/\sigma_{z,0}$. The different ranges of angular momentum L_z in the two volumes reflect $L_z \sim Rv_{\text{circ}}$ and the different radial extent of both volumes. The volume above the plane contains stars with higher J_z , because stars with small J_z cannot reach that far above the

plane. Circular orbits with $J_R = 0$ and $J_z = 0$ can only be observed in the Galactic mid-plane. An orbit with L_z much smaller or larger than $L_z(R_\odot)$ can only reach into a volume located around R_\odot , if it is more eccentric and has therefore larger J_R . This together with the effect of asymmetric drift can be seen in the asymmetric distribution of J_R in the top central panel of Figure 2.

If we want to add measurement errors to the mock data, we need to apply the following modifications to the above procedure. First, measurement errors are best described in heliocentric observables (see Section 2.1), we therefore assume and apply Gaussian errors to the *true* phase-space coordinates $\tilde{\mathbf{x}} = (\text{RA}, \text{DEC}, (m - M)), \tilde{\mathbf{v}} = (\mu_{\text{RA}}, \mu_{\text{DEC}}, v_{\text{los}})$, where we have taken $(m - M)$ as a proxy for distance. Second, in the case of distance errors, stars can virtually scatter in and out of the observed volume. To account for this, we draw the *true* positions from a volume that is larger than the actual observation volume, perturb the stars positions according to the distance errors and then reject all stars that lie now outside of the observed volume. This procedure mirrors the Poisson scatter around the detection threshold for stars whose distances are determined from the apparent brightness and the distance modulus. We then sample velocities (given the *true* positions of the stars) as described above and perturb them according to the measurement errors as well.

2.6. Data Likelihood

As data we consider here the positions and velocities of stars coming from a given MAP and survey selection function $\text{sf}(\mathbf{x})$,

$$D = \{\mathbf{x}_i, \mathbf{v}_i \mid (\text{star } i \text{ belonging to same MAP}) \wedge (\text{sf}(\mathbf{x}_i) > 0)\}.$$

The model that we fit is specified by a number of fixed and free parameters,

$$p_M = \{p_{\text{DF}}, p_\Phi\}.$$

For the qDF parameters (see Section 2.3) we assume a prior that is flat in

$$p_{\text{DF}} := \{\ln h_R, \ln \sigma_{R,0}, \ln \sigma_{z,0}, \ln h_{\sigma,R}, \ln h_{\sigma,z}\}. \quad (10)$$

The orbit of the i -th star in a potential with p_Φ is labeled by the actions $\mathbf{J}_i := \mathbf{J}[\mathbf{x}_i, \mathbf{v}_i \mid p_\Phi]$ and the qDF evaluated for the i -th star is then $\text{qDF}(\mathbf{J}_i \mid p_M) := \text{qDF}(\mathbf{J}[\mathbf{x}_i, \mathbf{v}_i \mid p_\Phi] \mid p_{\text{DF}})$.

The likelihood of the data given the model is

$$\begin{aligned} \mathcal{L}(D \mid p_M) &\equiv \prod_i^N p(\mathbf{x}_i, \mathbf{v}_i \mid p_M) \\ &= \prod_i^N \frac{\text{qDF}(\mathbf{J}_i \mid p_M) \cdot \text{sf}(\mathbf{x}_i)}{\int d^3x d^3v \text{qDF}(\mathbf{J} \mid p_M) \cdot \text{sf}(\mathbf{x})} \\ &\propto \prod_i^N \frac{\text{qDF}(\mathbf{J}_i \mid p_M)}{\int d^3x \rho_{\text{DF}}(R, |z| \mid p_M) \cdot \text{sf}(\mathbf{x})}, \end{aligned} \quad (11)$$

where N is the number of stars in the data set D , and in

the last step we used Equation 9. The factor $\prod_i \text{sf}(\mathbf{x}_i)$ is independent of the model parameters so we treat it as unimportant proportionality factor in the likelihood calculation. We find the best set of model parameters by maximizing the posterior probability distribution $\text{pdf}(p_M \mid D)$, which is according to Bayes' theorem proportional the likelihood $\mathcal{L}(D \mid p_M)$ times the prior. We assume flat priors in both p_Φ and p_{DF} (see Equation 10) through out this work, then pdf and likelihood can and will be used interchangeably for the remainder of the work.

The normalisation in Equation 11 is a measure for the total number of tracers inside the survey volume,

$$M_{\text{tot}} \equiv \int d^3x \rho_{\text{DF}}(R, |z| \mid p_M) \cdot \text{sf}(\mathbf{x}). \quad (12)$$

In the case of an axisymmetric galaxy model and $\text{sf}(\mathbf{x}) = 1$ everywhere inside the observed volume (i.e. a complete sample as assumed in most tests in this work), the normalisation is essentially a two-dimensional integral in R and z of the interpolated tracer density ρ_{DF} in Equation 9 over the differential survey volume, i.e. $\frac{\partial M_{\text{tot}}}{\partial \phi}(R, z) = \int dR dz \rho_{\text{DF}} \times \frac{\partial V}{\partial \phi}$ [TO DO: missing factor of R ???]. We perform this integral as a Gauss Legendre quadrature of order 40 in each R and z direction. The angular integral, i.e. $M_{\text{tot}} = \int R d\phi \frac{\partial M_{\text{tot}}}{\partial \phi}$, can be solved analytically.

It turns out that the sufficiently accurate evaluation of the likelihood is computationally expensive, even for only one set of model parameters. This expense is dominated by the number of action calculations required, which in turn depends on the number of stars in the sample and the numerical accuracy of the integrals in Equation 9 needed for the normalisation, which requires $N_x^2 \times N_v^3$ action calculations. The accuracy has to be chosen high enough, such that a resulting numerical error

$$\delta_{M_{\text{tot}}} \equiv \frac{M_{\text{tot,approx}}(N_x, N_v, N_\sigma) - M_{\text{tot}}}{M_{\text{tot}}} \quad (13)$$

[TO DO: make sure every M_{tot} is replaced by M_{tot}] does not dominate the likelihood, i.e.

$$\begin{aligned} &\log \mathcal{L}(p_M \mid D) \\ &= \sum_i^N \log \text{qDF}(\mathbf{J}_i \mid p_M) - 3N \log(r_o v_o) \\ &\quad - N \log(M_{\text{tot}}) - N \log(1 + \delta_{M_{\text{tot}}}), \end{aligned} \quad (14)$$

with

$$N \log(1 + \delta_{M_{\text{tot}}}) \lesssim 1.$$

In other words, this error is only small enough if it does not affect the comparison of two adjacent models whose log-likelihoods differ, to be clearly distinguishable, by 1. Otherwise numerical inaccuracies could lead to systematic biases in the potential and DF fitting. For data sets as large as $N = 20,000$ stars, which in the age of Gaia could very well be the case [TO DO: Really???], one needs a numerical accuracy of 0.005% in the normalisation. Figure ?? demonstrates that the numerical accuracy we use in the analysis, $N_x = 16$,

$N_v = 24$ and $N_{\text{sigma}} = 5$, does satisfy this requirement.

[TO DO: Comment by HW: What I am missing in this Section is any distinction of what aspects are "new" (not addressed in existing papers) and what is recapitulated to be coherent.]

If the data is affected by measurement errors, they have to be incorporated in the likelihood. We assume Gaussian errors in the observable space $\mathbf{y} \equiv (\tilde{\mathbf{x}}, \tilde{\mathbf{v}}) = (\text{RA}, \text{DEC}, (m - M), \mu_{\text{RA}}, \mu_{\text{DEC}}, v_{\text{los}})$, i.e. the i -th star's observed $\mathbf{y}_i \sim N[\mathbf{y}'_i, \delta\mathbf{y}_i](\mathbf{y}) = N[\mathbf{y}, \delta\mathbf{y}_i](\mathbf{y}'_i)$ [TO DO: Talk to HW about best notation.], with \mathbf{y}'_i being the true position and velocity of the star. Stars follow the (quasi-isothermal) distribution function ($\text{DF}(\mathbf{y}') \equiv \text{qDF}(\mathbf{J}[\mathbf{y}' | p_\Phi] | p_{\text{DF}})$ for short), convolved with the error distribution $N[0, \delta\mathbf{y}](\mathbf{y}')$ [TO DO: CHECK AGAIN]. The selection function $\text{sf}(\mathbf{y})$ acts on the space of (error affected) observables. Then the probability of one star becomes

$$\begin{aligned} \tilde{p}(\mathbf{y}_i | p_\Phi, p_{\text{DF}}, \delta\mathbf{y}_i) \\ \equiv \frac{\text{sf}(\mathbf{y}_i) \cdot \int d^6\mathbf{y}' \text{DF}(\mathbf{y}') \cdot N[\mathbf{y}_i, \delta\mathbf{y}_i](\mathbf{y}')}{\int d^6\mathbf{y}' \text{DF}(\mathbf{y}') \cdot \int d^6\mathbf{y} \text{sf}(\mathbf{y}) \cdot N[\mathbf{y}', \delta\mathbf{y}_i](\mathbf{y})}. \end{aligned}$$

In the case of errors in distance or position, the evaluation of this is computational expensive - especially if the stars have heteroscedastic errors $\delta\mathbf{y}_i$, for which the normalisation would have to be calculated for each star separately. In practice we apply the following approximation,

$$\begin{aligned} \tilde{p}(\mathbf{y}_i | p_\Phi, p_{\text{DF}}, \delta\mathbf{y}_i) \\ \approx \frac{\text{sf}(\mathbf{x}_i)}{\int d^6\mathbf{y}' \text{DF}(\mathbf{y}') \cdot \text{sf}(\mathbf{x}')} \cdot \frac{1}{N_{\text{error}}} \sum_n^{N_{\text{error}}} \text{DF}(\mathbf{x}_i, \mathbf{v}[\mathbf{y}'_{i,n}]) \end{aligned}$$

with

$$\mathbf{y}'_{i,n} \sim N[\mathbf{y}_i, \delta\mathbf{y}_i](\mathbf{y}')$$

In doing so, we ignore errors in the star's position \mathbf{x}_i [TO DO: something is not clear to HW here] altogether. This simplifies the normalisation drastically and makes it independent of measurement errors, including the velocity errors. Distance errors however are included [TO DO: something is not clear to HW here], but only implicitly in the convolution over the stars' velocity errors in the Galactocentric rest frame. We calculate the convolution using Monte Carlo integration with N_{error} samples drawn from the full error Gaussian in observable space, $\mathbf{y}'_{i,n}$.

2.7. Fitting Procedure

To search the (p_Φ, p_{DF}) parameter space for the maximum of the likelihood in Equation 11, we go beyond the fixed grid search by Bovy & Rix (2013) and employ an effective two-step procedure: The first step finds the approximate peak and width of the likelihood using a nested-grid search, while the second step samples the shape of the likelihood using a Monte-Carlo Markov Chain (MCMC) approach.

Fitting Step 1: Nested-grid search. — The (p_Φ, p_{DF}) parameter space can be high-dimensional. To effectively

minimizing the number of likelihood evaluations before finding its peak, we use a nested-grid approach:

- *Initialization.* For N free model parameters $M = (p_\Phi, p_{\text{DF}})$, we set up a sufficiently large initial grid with 3^N regular grid points.
- *Evaluation.* We evaluate the likelihood at each grid-point similar to Bovy & Rix (2013) (their Figure 9): Because of the many computationally expensive $\mathbf{x}, \mathbf{v} \xrightarrow{p_\Phi} \mathbf{J}$ transformations that have to be performed for each new set of p_Φ parameters, an outer loop iterates over the p_Φ parameters and pre-calculates the actions, while an inner loop evaluates the likelihood Equation 11 for all qDF parameters p_{DF} with the actions in the given potential.
- *Iteration.* To find from the very sparse 3^N likelihood grid a new grid, that is more centered on the likelihood and has a width of order of the width of the likelihood, we proceed as follows: For each of the model parameter in M we marginalize the likelihood by summing over the grid. If the resulting 3 points all lie within 4σ of a Gaussian, we fit a Gaussian to the 3 points and determine a new 4σ fitting range. Otherwise the boundaries of the grid point with the highest likelihood becomes the new fitting range. We proceed with iteratively evaluating the likelihood on finer and finer grids, until we have found a 4-sigma fit range in each of the model parameter dimensions.
- *The fiducial qDF.* For the above strategy to work properly, the action pre-calculations have to be independent of the choice of qDF parameters. This is clearly the case for the $N_j \times N_{\text{error}}$ stellar data actions \mathbf{J}_i . To calculate the normalisation in Equation 11, $N_x^2 \times N_v^3$ actions \mathbf{J}_n are needed. Formally the spatial coordinates at which the \mathbf{J}_n are calculated depend on the p_{DF} parameters via the integration ranges in Equation 9. To relax this dependence we instead use the same velocity integration limits in the likelihood calculations for all p_{DF} s in a given potential. This set of parameters, that sets the velocity integration range globally, $(\sigma_{R,0}, \sigma_{z,0}, h_{\sigma,R}, h_{\sigma,z})$ in Equation 6 and 7, is referred to as the *fiducial* qDF. Using the same integration range in the density calculation for all qDFs at a given p_Φ makes the normalisation vary smoothly with different p_{DF} . Choosing a fiducial qDF that is very different from the true qDF can however lead to large biases. The optimal values for the fiducial qDF are the (yet unknown) best fit p_{DF} parameters. We take care of this by setting, in each iteration step of the nested-grid search, the fiducial qDF simply to the p_{DF} parameters of the central grid point. As the nested-grid search approaches the best fit values, the fiducial qDF approaches automatically the optimal values as well. This is another advantage of the nested-grid search, because the result will not be biased by a poor choice of the fiducial qDF.
- *Computational expense.* Overall the computation speed of this nested-grid approach is dominated (in

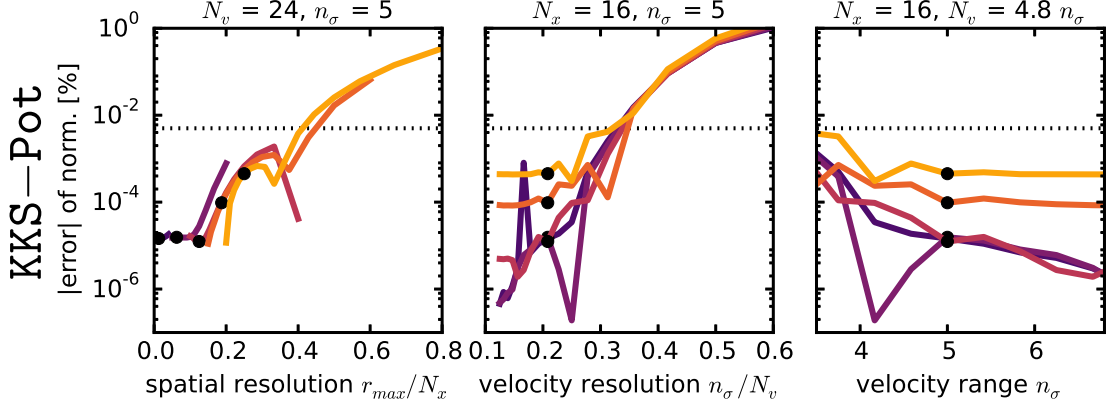


Figure 3. Relative error δM_{tot} of the likelihood normalization M_{tot} in Equation 13 depending on the accuracy of the grid-based density calculation in Equation 9 (and surrounding text). We show how δM_{tot} varies with the spatial resolution (first column), velocity resolution (second column) and velocity integration range (third column) for two different potentials (KKS-Pot in the first row and MW13-Pot in the second row) and five different spherical observation volumes with radius r_{max} (color coded according to the legend). (Test ?? in Table ?? summarizes all model parameters.) N_x is the number of spatial grid points in $R \in R_{\odot} \text{ kpc} \pm r_{\text{max}}$ and $|z| \in [0, r_{\text{max}}]$ on which the density is evaluated according to Equation 9. The spatial resolution in z is therefore r_{max}/N_x and $2r_{\text{max}}/N_x$ in R . This choice is reasonable because the density is symmetric in z and varies less in R than in z , because the tracer scale length of the disk is much larger than its scale height. At each (R, z) of the grid a Gauss-Legendre integration of order N_v is performed over an integration range of $\pm n_{\sigma}$ times the velocity dispersion in v_R and v_z and $[0, 1.5v_{\text{circ}}(R_{\odot})]$ in v_T . n_{σ}/N_v is therefore a proxy for the velocity resolution of the grid. (We vary N_x , N_v and n_{σ} separately and keep the other two fixed at the values indicated above the columns.) To arrive at the approximation $M_{\text{tot,approx}}$ for M_{tot} in Equation 12, we perform a 40th-order Gauss-Legendre integration in each R and z direction of the interpolated density over the observed volume. We calculate the “true” normalization with high accuracy as $M_{\text{tot}} \approx M_{\text{tot,approx}}(N_x = 20, N_v = 56, N_{\sigma} = 7)$. The black dots indicate the accuracy used in our analyses: It is better than 0.002%. Only for the smallest volume in the MW13-Pot (yellow line) the error is only $\sim 0.005\%$. This could be due to the fact, that, while we have analytical formulas to calculate the actions for the Staechel potential KKS-Pot exactly, we have to resort to an approximate action calculation for the MW-like potential MW13-Pot (see Section 2.2). [TO DO: Write $[\delta M_{\text{tot}}]$ on y-axis] [TO DO: Remove MW13-Pot completely from this plot, caption and test table] [TO DO: Caption too long] [TO DO: Rewrite caption, text and table, I changed the plot]

descending order of importance) by a) the complexity of potential and action calculation, b) the number $N_j \times N_{\text{error}} + N_x^2 \times N_v^3$ of actions to calculate, i.e. the number of stars, error samples and numerical accuracy of the normalisation calculations, c) the number of different potentials to investigate (i.e. the number of free potential parameters and number of grid points in each dimension) and d) the number of qDFs to investigate. The latter is also non-negligible, because for such a large number of actions the number of qDF-function evaluations also take some time.

Fitting Step 2: MCMC.— After the nested-grid search is converged, the grid is centered at the peak of the likelihood and its extent contains the 4σ confidence interval. To actually sample the full shape of the likelihood, we could do a grid search with much finer grid spacing (e.g. $K = 11$ in each dimension). The number of grid points scales as a power of the free parameters N . For a large number of free parameters ($N > 4$) a Monte Carlo Markov Chain (MCMC) approach might sample the *pdf* (which is here equivalent to the likelihood, see §2.6) much faster. We use *emcee* by Foreman-Mackey et al. (2013) and release the walkers very close to the *pdf* peak found by the nested-grid search, which will assure fast convergence in much less than K^N likelihood evaluations. For a sufficiently high numerical accuracy in calculating the integrals in Equation 9 the current qDF at each walker position can be used as the fiducial qDF. To get reasonable results also for slightly lower accuracy, a single fiducial qDF can be used for all likelihood evaluations within the MCMC as well. As fiducial qDF we use the qDF parameters of the likelihood peak, found by the

nested-grid search.

[TO DO: Make consistent use of *pdf* and likelihood.]

3. RESULTS

We are now in a position to explore the limitations of action based modelling posed in the introduction: (i) unbiased estimates; (ii) survey volume; (iii) imperfect selection function; (iv) measurement errors; (v) actual DF or (vi) Potential not spanned by the space of models. We do not explore the breakdown of the assumption that the system is axisymmetric and in steady state. With the exception of the test suite on measurement errors in §3.4, we assume that the phase-space errors are negligible. All tests are also summarized in Table ??.

3.1. Model Parameter Estimates in the Limit of Large Data Sets

The individual MAPs in Bovy & Rix (2013) contained typically between 100 and 800 objects, so that each MAP implied a quite broad *pdf* for the model parameters $p_M = \{p_{\Phi}, p_{\text{DF}}\}$. Here we explore what happens in the limit of much larger samples for each MAP, say 20,000 objects. As outlined in §2.6 the immediate consequence of larger samples is given by the likelihood normalization requirement, $\log(1 + \delta M_{\text{tot}}) \leq 1/N_{\text{sample}}$, (see Equation 14)), which is the modelling aspect that drives the computing time. This issues aside, we would, however, expect that in the limit of large data sets with vanishing measurement errors the *pdf*s of the p_M become Gaussian, with a *pdf* width (i.e. standard error SE of the Gaussian) that scales as $1/\sqrt{N_{\text{sample}}}$. Further, we must verify that any bias in the *pdf* expectation value is considerably less than the error (SE), even for quite large samples.

Using sets of mock data, created according to §2.5 and the fiducial model for p_M (see Table ??, Tests ??, ??, and ??), we verified that *RoadMapping* satisfies all these conditions and expectations. Figure 4 illustrates the joint *pdf*s of all p_M . This figure illustrates that the *pdf* is a multivariate Gaussian that projects into Gaussians when considering the marginalized *pdf* for all the individual p_M . Note that some of the parameters are quite covariant, but the level of their actual covariance depends on the choice of the p_M from which the mock data were drawn. Figure 5 then illustrates that the *pdf* width, SE, indeed scales as $1/\sqrt{N_{\text{sample}}}$. Figure 6 illustrates even more that *RoadMapping* satisfies the central limit theorem. The average parameter estimates from many mock samples with identical underlying p_M are very close to the input p_M , and the distribution of the actual parameter estimates are a Gaussian around it.

[TO DO: I sometimes talk about pdf, sometimes about likelihood. We should make this consistent everywhere. I would use *pdf* everywhere, but I sometimes reference the likelihood equation. How should I write it in this case?]

3.2. The Role of the Survey Volume Geometry

To explore the role of the survey volume (see Section 1) at given sample size, we devise two suites of mock data sets:

The first suite draws mock data from the same p_M , *two different potentials* (Iso-Pot and MW13-Pot, see Test ?? in Table ??), and volume wedges (see Section 2.4) at *different positions within the Galaxy*, illustrated in the right upper panel of Figure 7. To isolate the role of the survey volume geometry, the mock data sets are equally large (20,000) in all cases, and are drawn from identical total survey volumes (4.5 kpc^3 , achieved by adjusting the angular width of the wedges). The results are shown in Figure 7.

The second suite of mock data sets was already introduced in Section 3.1 (see also Test ??), where mock data sets were drawn from five spherical volumes around the sun with different maximum radius, for *two different stellar populations*. The results of this second suite are shown in Figure 6 and demonstrate the effect of the *size of the survey volume*.

Figures 6 and 7 illustrate the ability of *RoadMapping* to constrain model parameters, with the standard error of the *pdf* as measure of the precision on the x -axis. Figure 6 demonstrates that, given a choice of qDF, a larger volume always results in tighter constraints. There is no obvious trend that a hotter or cooler MAP will always give better results [TO DO: Comment from HW: The question of whether a hotter or a colder population gives tighter constraints is an important question, but it seems buried here in a section that is dedicated to another matter, namely the question of volume ... It's OK to leave it here, but somewhere we need to say clearly: whether the population is hot or cold does not make a big and generic difference...]; it depends on the survey volume and the model parameter in question. In Figure 7 the wedges all have the same volume and all give results of similar precision. Minor differences, e.g. with the Iso-Pot potential being less constrained in the wedge with large vertical, but small radial extent, are a special property of the considered potential and parameters, and not a

global property of the corresponding survey volume. In the case of an axisymmetric model galaxy, the extent in ϕ direction is not expected to matter. Overall radial extent and vertical extent seem therefore to be equally important to constrain the potential. In addition Figure 7 implies that for these cases volumes offsets in the radial or vertical direction have at most a modest impact - even in case of the very large sample size at hand.

While it appears that the argument for significant radial and vertical extent is generic, we have not done a full exploration of all combinations of p_M and volumina.

3.3. Impact of Misjudging the Completeness of the Data Set

The completeness function (see Section 2.4) depends on the characteristics and mode of the survey. It can be very complex and is therefore sometimes not perfectly known. We investigate how much the recovery of the potential can be affected by imperfect knowledge of the selection function. We do this by creating mock data with varying incompleteness (within a maximal survey volume), while assuming constant completeness in the analysis. The mock data comes from a sphere around the sun with an incompleteness function that drops linearly with distance r from the sun (see Test ??, Example 1, in Table ?? and Figure 8). This captures the relevant case of stars being less likely to be observed (than assumed) the further away they are (e.g. due to unknown dust obscuration). We demonstrate that the potential recovery with *RoadMapping* is very robust against somewhat wrong assumptions about the radial completeness of the data (see Figure 9). Apparently, much information about the potential comes from the rotation curve measurements in the plane, which is not affected by the incompleteness of the sample. In Appendix .1 we also show that the robustness is somewhat less striking but still persists for small misjudgments of the incompleteness in vertical direction, parallel to the disk plane (Figures 13 and 14). This could model the effect of wrong corrections for interstellar extinction in the plane. We also investigate in Appendix .1 if indeed most of the information is stored in the rotation curve [TO DO: Comment by HW: I don't have an immediate solution for this, but again, it seems the interesting question of "how much of the information is in the rotation curve" is 'hidden' in the section on selection functions...]. For this we use the same mock data sets as analysed in Figures 9 and 14, but without including the tangential velocities in the modelling (by marginalizing the likelihood over v_T). In this case the potential is much less tightly constrained, even for 20,000 stars. For only small deviations of true and assumed completeness ($\lesssim 10\%$) we can however still incorporate the true potential in our fitting result (see Figure 15).

[TO DO: Mention in text or caption how the panels looked that I removed.]

3.4. Dealing with Measurement Errors and their Effect on the Parameter Recovery

Convergence of the error integral. — In §2.6 we introduced how we convolve the model probability with the measurement errors. In the absence of distance errors the accuracy of the parameter recovery is limited by an insufficient MC sampling of the convolution integral in Equation 15. Test ?? in Table ?? and Figure 10 investigate

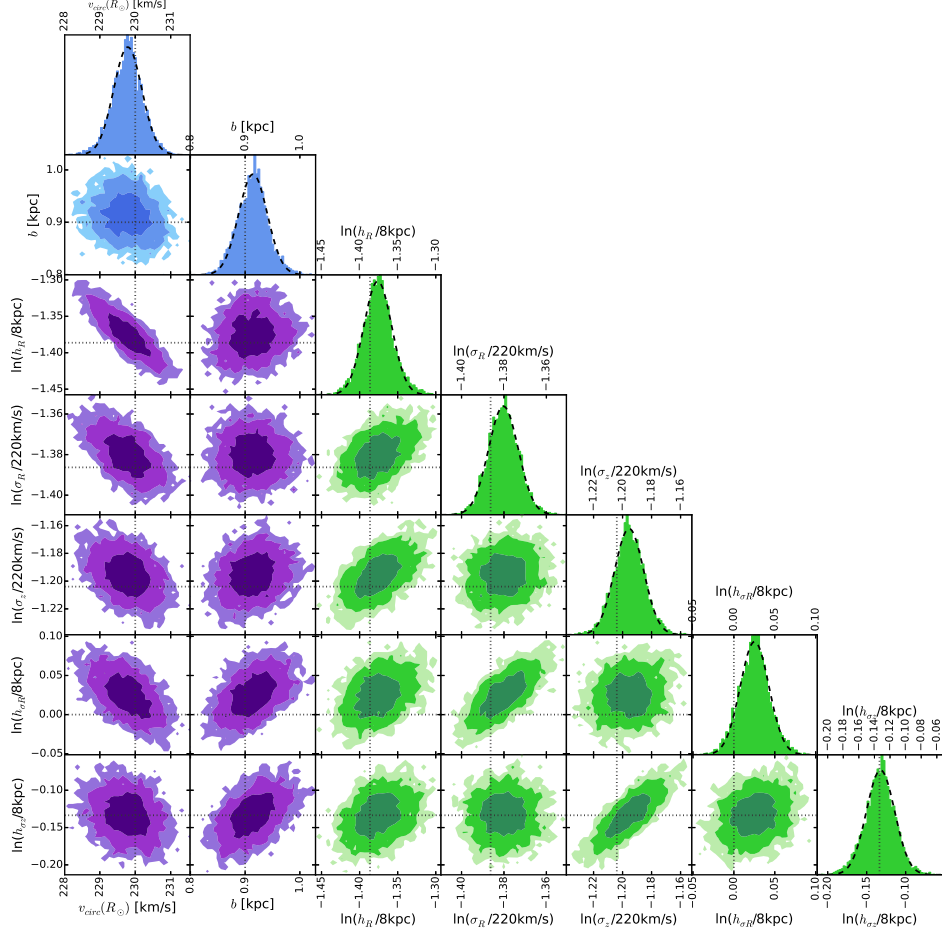


Figure 4. The pdf (proportional to the likelihood in Equation [TO DO]) in the parameter space $p_M = \{p_\Phi, p_{DF}\}$ for one example mock data set created according to Test ?? in Table ?. Blue indicates the pdf for the potential parameters, green the qDF parameters. The true parameters are marked by dotted lines. The dark, medium and bright contours in the 2D distributions represent 1, 2 and 3 sigma confidence regions [TO DO: HW: "likelihood vs. pdf - This is where this matters: is this a confidence on the data or on the parameters?" Don't understand, what he means...], respectively, and show weak or moderate covariances. This analysis was picked among five similar analyses, to have all 1 sigma contours encompass the input values [TO DO: Jo didn't understand this sentence]. The pdf here was sampled using MCMC (with flat priors in p_Φ and $\ln(p_{DF})$ to turn the likelihood in Equation 11 into a full pdf). Because only 10,000 MCMC samples were used to create the histograms shown, the 2D distribution has noisy contours. The dashed lines in the 1D distributions are Gaussian fits to the histogram of MCMC samples. This demonstrates very well that for such a large number of stars, the pdf approaches the shape of a multi-variate Gaussian, as expected from the central limit theorem [TO DO: Jo wrote, that he is not sure if the central limit theorem is directly relevant here]. [TO DO: rename $h_{\sigma R}$ to $h_{\sigma, R}$, σ_R to $\sigma_{R,0}$ and analogous for z]

how many MC samples are needed, given the size of the velocity error, for the integral to be accurate within certain limits:

For each $\delta\mu \in [2, 3, 4, 5]\text{mas yr}^{-1}$ we set up four mock data sets and evaluated the likelihood for different N_{error} . We used $N_{\text{conv}} := 800$ and 1200 MC samples to calculate the numerically converged likelihood for proper motion errors $\delta\mu \leq 3\text{mas yr}^{-1}$ and $\delta\mu > 3\text{mas yr}^{-1}$, respectively (see left panels in Figure 11). We determined the mean bias

$$\text{BIAS}(N_{\text{error}}, \delta\mu) \equiv \frac{1}{4} \sum_{j=1}^4 [\langle p_i \rangle(N_{\text{error}}, \delta\mu)]_j - [\langle p_i \rangle(N_{\text{conv}}, \delta\mu)]_j$$

where $[\langle p_i \rangle(N_{\text{error}}, \delta\mu)]_j$ is the best estimate for the i -th model parameter $p_i \in p_M$ from the analysis of the j -th mock data realisation with $\delta\mu$ using N_{error} MC samples. From this we then generated the curves $N_{\text{error},i}(\delta v_{\text{max}}, \text{BIAS})$ in Figure 10 by linear interpolation, that show how many MC samples are needed for

parameter p_i given a velocity error and a systematic bias in units of the standard error (SE) of the estimate. The proper motion error $\delta\mu$ translates to a velocity error according to

$$\delta v_{\text{max}}[\text{km s}^{-1}] \equiv 4.74047 \cdot r_{\text{max}}[\text{kpc}] \cdot \delta\mu[\text{mas yr}^{-1}], \quad (16)$$

where r_{max} is the maximum distance of stars. We find in Figure 10 the relation

$$N_{\text{error},i}(\delta v_{\text{max}}, \text{BIAS}) \propto (\delta v_{\text{max}})^2.$$

Figure 10 also demonstrates that different model parameters do not have the same sensitivity to the numerical inaccuracies introduced by insufficient sampling. [TO DO: we haven't tested yet, if this plot depends on hotness of population and / or number of stars. But if it takes forever to actually do the calculations, I guess, we just leave it like this.]

Testing the error convolved likelihood approximation. — In absence of distance (modulus) errors our approximation

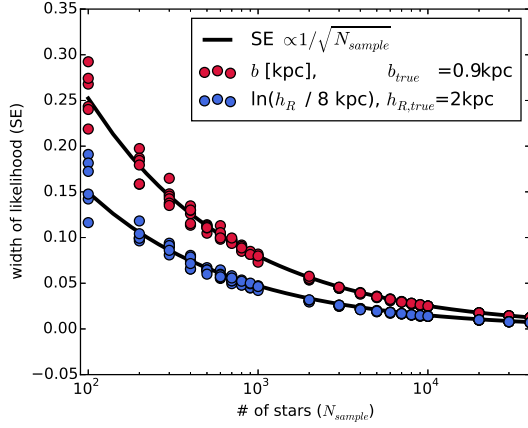


Figure 5. The width of the *pdf* for two fit parameters found from analyses of 132 mock data sets vs. the number of stars in each data set. The mock data was created in the Iso-Pot potential and all model parameters are given as Test ?? in Table ???. The *pdf* (using the likelihood in Equation 11 [TO DO: CHECK]) was evaluated and then a Gaussian was fitted to the marginalized *pdf* of each free fit parameter. The standard error (SE) of these best fit Gaussians is shown for the potential parameter b in kpc (red dots) and for the qDF parameter $\ln(h_R/8\text{kpc})$ in dimensionless units (blue). The black lines are fits of the functional form $\text{SE}(N_{\text{sample}}) \propto 1/\sqrt{N_{\text{sample}}}$ to the data points of both shown parameters. As can be seen, for large data samples the width of the *pdf* behaves as expected and scales with $1/\sqrt{N_{\text{sample}}}$ as predicted by the central limit theorem. [TO DO: fancybox Legend] [TO DO: write pdf instead of likelihood on y-axis] [TO DO: axis labels similar to MC_vs_error plot.] [TO DO: Different colors (blue and green), and different markers (no black rings).]

for the likelihood, which is the model probability convolved with the measurement errors in Equation 15, is equal to the true likelihood. In case there are distance modulus errors, this likelihood links the range of possible velocities (specified by the measurement errors in line-of-sight velocity, proper motion and distance modulus) to a fixed but slightly wrong position, as we ignore the distance error in the position. As the link between position and velocity provides the information about the potential, this will lead to systematic biases in the parameter recovery the larger the distance error becomes. In Test ?? in Table ?? and Figure 11 we investigate the capabilities of Equation 15 with and without distance modulus errors.

The left column of panels in Figure 11 shows how well the approximation works in the absence of distance errors. There seemed to be no biases in the parameter recovery, independent of the size of the proper motion error. (We note that there could be a tiny bias $\ll 1$ SE in the $\ln(\sigma_{z,0}/200 \text{ km s}^{-1})$ qDF parameter, most likely due to insufficient numerical accuracy, but all the other model parameters, also those not shown in the figure, are very well behaved.) Overall the standard errors on the recovered parameters are quite small (a few percent at most for 10,000 stars), which demonstrates that, if we perfectly knew the measurement errors, we still could get very precise constraints on the potential. The constraints also get tighter the smaller the proper motion error becomes. We found that for $\delta\mu = 1 \text{ mas yr}^{-1}$ the precision of the recovered parameters reduce by \sim half compared to $\delta\mu = 5 \text{ mas yr}^{-1}$.

The right column of panels in Figure 11 demonstrates

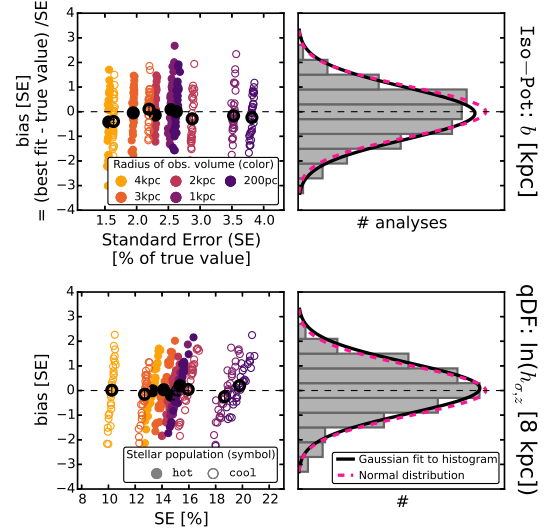


Figure 6. (Un-)bias of the parameter estimates: According to the central limit theorem the best fit values for a large number of data sets, each containing a large number of stars, will follow the Normal distribution. To test this, we create 320 mock data sets, which come from two different stellar populations and five spherical observation volumes (see legends). All model parameters are summarized in Table ?? as Test ??. Bias and relative standard error (SE) are derived from the marginalized *pdf* for one potential parameter (isochrone scale length b in first row) and one qDF parameter ($h_{\sigma,z}$ in second row). The second column displays a histogram of the 320 offsets. As it closely follows a Normal distribution, our modelling method is therefore well-behaved and unbiased. For the 32 analyses belonging to one model we also determine the mean offset and SE, which are overplotted in black in the first two columns (with $1/\sqrt{32}$ as error). [TO DO: Is the scatter of the black symbols too large??? Is the reason for this numerical inaccuracies???] [TO DO: Change test table accordingly, isochrone with $b = 1.5$ is not used anymore] [TO DO: Caption is too long. Make shorter.] [TO DO: r_{max} instead of radius in legend] [TO DO: Leerzeichen fehlt in y-achsenbeschriftung]

the failure of our adopted likelihood approximation in the case of large distance modulus errors. The larger the $\delta(m - M)$, the wronger the recovered parameters become: The systematic biases can get many SEs large. We find however that in case of $\delta(m - M) \leq 0.2 \text{ mag}$ (if also $\delta\mu \leq 2 \text{ mas yr}^{-1}$ and a maximum distance of $r_{\text{max}} = 3 \text{ kpc}$, see Test ?? in Table ??) the parameters can still be recovered within 2 SEs. For most model parameters (except $\ln(\sigma_{z,0}/200 \text{ km s}^{-1})$, as shown in the figure, and $\ln(h_R/8 \text{ kpc})$) even $\delta(m - M) \leq 0.3 \text{ mag}$ still gives biases smaller than 2 SEs. This corresponds to a relative distance error of $\sim 10\%$. This encourages us that for smaller distance modulus errors we really could use our likelihood approximation in Equation 15, which is computationally cheaper than a proper treatment, also on real data sets.

Underestimation of the proper motion error. — We found that in case we perfectly knew the measurement errors (and the distance error is negligible), the convolution of the model probability with the measurement errors gives precise and accurate constraints on the model parameters - even if the error itself is quite large. Now we investigate what would happen if the quoted measurement errors, e.g. the proper motion errors, were actually smaller than the true errors. Figure 12 shows the case for two different stellar populations and an error under-

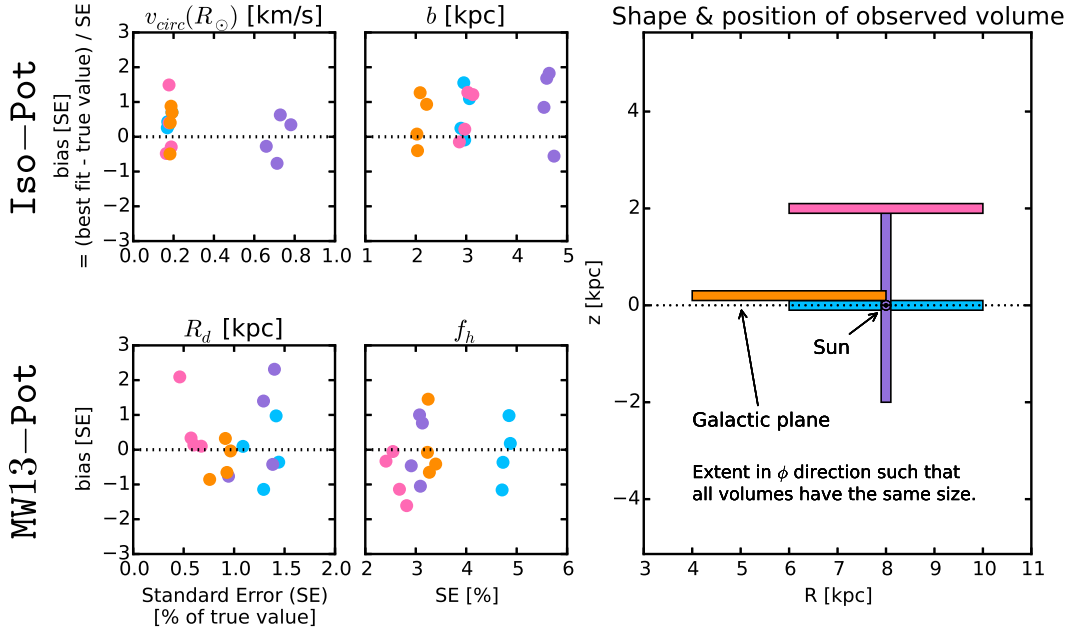


Figure 7. Bias vs. standard error in recovering the potential parameters for mock data stars drawn from four different test observation volumes within the Galaxy (illustrated in the upper right panel) and two different potentials (Iso-Pot and MW13-Pot from Table ??). Standard error and offset were determined as in Figure 6. Per volume and potential we analyse four different mock data realisations; all model parameters are given as Test ?? in Table ?. The colour-coding represents the different wedge-shaped observation volumes. The angular extent of each wedge-shaped observation volume was adapted such that all have the volume of 4.5 kpc^3 , even though their extent in (R, z) is different. Overall there is no clear trend, that an observation volume around the sun, above the disk or at smaller Galactocentric radii should give remarkably better constraints on the potential than the other volumes. [TO DO: Write in Plot "... that all wedges have the same volume".]

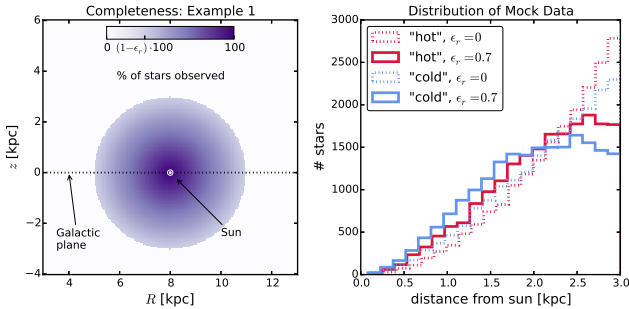


Figure 8. Selection function and mock data distribution for investigating radial incompleteness of the data. All model parameters are summarized as Test ??, Example 1, in Table ?. The survey volume is a sphere around the sun and the percentage of observed stars is decreasing linearly with radius from the sun, as demonstrated in the left panel. How fast this detection/incompleteness rate drops is quantified by the factor ϵ_r . Histograms for four data sets, drawn from two MAPs (hot in red and cool in blue, see Table ??) and with two different ϵ_r , 0 and 0.7, are shown in the right panel for illustration purposes. [TO DO: Potential and/or population names in typewriter font]

estimation of 10% and 50%.

Overall the parameter recovery gets worse the larger the proper motion error and the stronger the underestimation. The relation between the bias due to error misjudgment and the size of the proper motion error seems to be linear.

For the recovery of the isochrone potential scale length b the hotness of the population does not matter (see lower left panel in Figure 12). The circular velocity $v_{\text{circ}}(R_{\odot})$ is, as always, better measured by cooler than by hotter

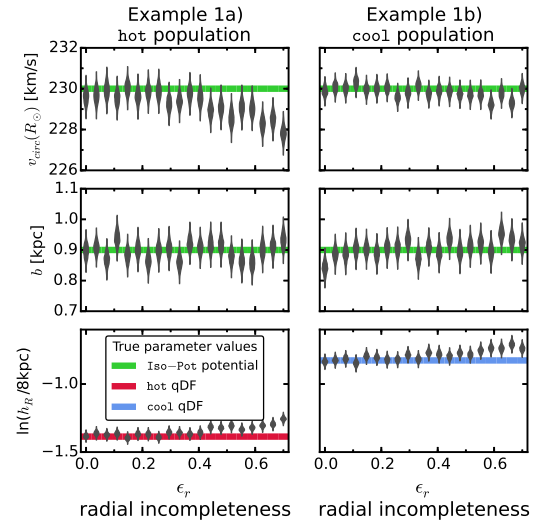


Figure 9. Influence of wrong assumptions about the radial incompleteness of the data on the parameter recovery with RoadMapping. Each mock data set was created having different incompleteness parameters ϵ_r (shown on the x -axis and illustrated in Figure 8) and the model parameters are given as Test ??, Example 1, in Table ?. The analysis however didn't know about the incompleteness and assumed that all data sets had constant completeness within the survey volume ($\epsilon_r = 0$). The marginalized likelihoods from the fits are shown as violins. The green lines mark the true potential parameters (Iso-Pot) and the red and blue lines the true qDF parameters (hot MAP in red and cool MAP in blue), which we tried to recover. The RoadMapping method seems to be very robust against small to intermediate deviations between the true and the assumed data incompleteness.

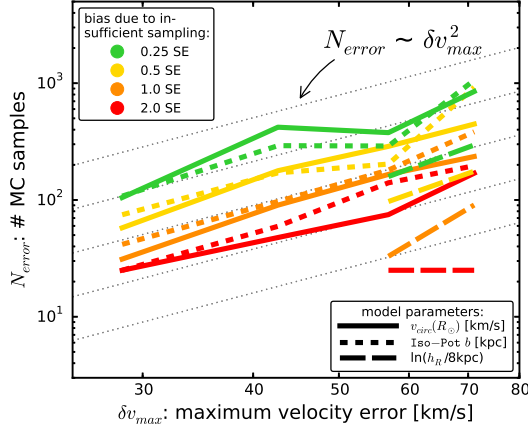


Figure 10. Number of Monte Carlo (MC) samples N_{error} needed for the numerical error convolution in Equation 15, given the maximum velocity error δv_{max} in the sample to reach a given accuracy. An insufficient sampling of the convolution integral leads to systematic biases in the reconstruction of the true model parameters. The size of the bias is color coded as indicated in the legend and is given in units of the standard error (SE). The model parameters, marked by different symbols, have different sensitivities to the numerical inaccuracy of the error convolution, therefore the range in N_{error} for the same given bias. Here we assume that the distance error is zero and the proper motion error $\delta\mu$ translates to a velocity error according to Equation 16 and $\delta v_{\text{los}} \ll \delta v_{\text{max}}$. All model parameters are listed in Table ?? as Test ??. The number of MC samples needed increases with the velocity error as $N_{\text{error}} \propto (\delta v_{\text{max}})^2$, as can be seen especially well in the inset figure for the potential parameter $v_{\text{circ}}(R_{\odot})$. All lines are fits of this functional form to each four points derived for a given model parameter (symbol) and bias (color). The large scatter in the points comes from low number statistics and errors introduced by linear interpolation of the bias vs. N_{error} relation found from the analyses. [TO DO: rename $h_{\sigma R}$ to $h_{\sigma, R}$, σ_R to $\sigma_{R,0}$ and analogous for z] [TO DO: some of the 25 MC sample analyses have to be re-done. (Currently running on cluster.))] [TO DO: Rewrite caption. I changed the whole plot.]

APPENDIX

Influence of wrong assumptions about incompleteness of the data parallel to the Galactic plane

In §3.3 we found a striking robustness of the *RoadMapping* modelling approach against wrong assumptions about the radial incompleteness of the data set. To further test this result, we investigate a different completeness function that drops with distance from the Galactic plane (see Test ??, Example 2, in Table ?? and Figure 13). We get a similar robust behaviour for small deviations, and only slightly less robustness for larger deviations. That an explanation for this robustness could be, that a lot of information about the potential comes from the rotation curve, which is not affected by incompleteness, is demonstrated in Figure 15.

Marginalization over v_T .— The likelihood in Equation 11 is marginalized over the coordinate v_T as follows

$$\begin{aligned} \mathcal{L}(p_M | D)|_{(v_T \text{ marg.})} &= \prod_i^N P_{(v_T \text{ marg.})}(\mathbf{x}_i, v_{R,i}, v_{z,i} | p_M) \\ &\equiv \prod_i^N v_0 \cdot \int_0^{1.5 v_{\text{circ}}(R_{\odot})} dv_T P(\mathbf{x}_i, v_{R,i}, v_T, v_{z,i} | p_M) \end{aligned}$$

where $P(\mathbf{x}, \mathbf{v} | p_M)$ is the same as in Equation 11 and the numerical integral over v_T is performed as a 24th order Gauss-Legendre quadrature. The additional factor of v_0 is needed to get the units of $P_{(v_T \text{ marg.})}(\mathbf{x}_i, v_{R,i}, v_{z,i} | p_M)$ right.

[TO DO: Mention in text or caption how the panels looked that I removed.]

REFERENCES

Batsleer, P., & Dejonghe, H. 1994, A&A [TO DO], 287, 43
 Binney, J. 2010, MNRAS, 401, 2318
 Binney, J., & McMillan, P. 2011, MNRAS, 413, 1889

Binney, J. 2011, Pramana, 77, 39
 Binney, J. 2012, MNRAS, 426, 1324
 Binney, J. 2012, MNRAS, 426, 1328

populations (see upper left panel in Figure 12).

We find that the recovery of the qDF parameters on the other hand is more strongly affected by the misjudgment of the velocity error for *cooler* stellar populations. The measured velocity dispersion is the convolution of the intrinsic dispersion with the measurement errors. If the proper motion error is underestimated, the deconvolved velocity dispersion is larger than the intrinsic velocity dispersion and the relative difference is bigger for a cooler population (see upper right panel for $\sigma_{z,0}$ in Figure 12). The intrinsic velocity dispersion is also cooler at larger radii than at smaller radii, therefore the deconvolved dispersion is overestimated more strongly at large R and the velocity dispersion scale length will be overestimated as well (see lower left panel for $h_{\sigma,z}$ in Figure 12). We get analogous results for the qDF parameters $\sigma_{R,0}$ and $h_{\sigma,R}$. The recovery of the tracer density scale length h_R is not affected by the misjudgment of velocity errors.

The most important and encouraging result from Figure 12 is, that for an underestimation of 10% the bias is still $\lesssim 2$ sigma [TO DO: can I say sigma??] - even for proper motion errors of almost 3 mas/yr.

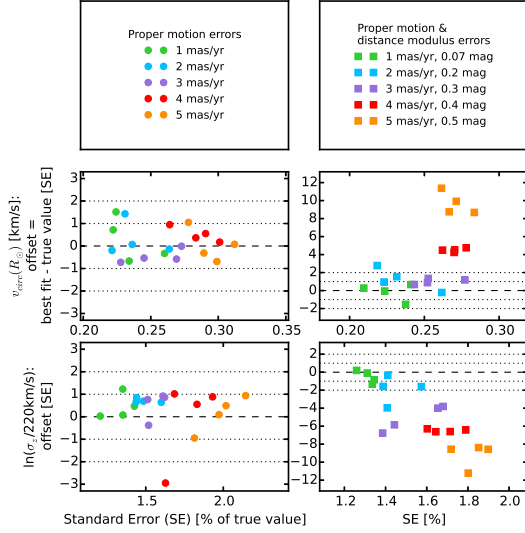


Figure 11. (Caption on next page.)

Figure 11. Parameter recovery using the approximation for the measurement error convolved likelihood in Equation 15. All model parameters used to create the mock data sets analyzed for this figure are given as Test ?? in Table ???. The mock data sets in the left panels have only errors in line-of-sight velocity and proper motions, while the data sets in the right panels also have distance modulus errors, as indicated in the legends in the first row. The size of the error is color coded. The other panels plot the offset of the recovered model parameter to the true parameter vs. the relative standard error for two of the seven model parameters, the potential parameter $v_{\text{circ}}(R_0)$ and qDF parameter $\sigma_{z,0}$. For data sets with proper motion error errors $\delta(m-M) \leq 3 \text{ mas yr}^{-1}$ Equation 15 was evaluated with $N_{\text{error}} = 800$, for $\delta(m-M) > 3 \text{ mas yr}^{-1}$ we used $N_{\text{error}} = 1200$. In the absence of distance errors Equation 15 gives unbiased results, for $\delta(m-M) \geq 3 \text{ mas yr}^{-1}$ (which corresponds in this test to $\delta v_{\text{max}} \lesssim 43$, see Equation 16) however biases of several sigma [TO DO: can I say sigma????] are introduced as Equation 15 in only an approximation for the true likelihood in this case. [TO DO: rename σ_z to $\sigma_{z,0}$]

Binney, J. 2013, NAR [TO DO: emulateapj doesn't know NAR], 57, 29

Binney, J., & Tremaine, S. 2008, Galactic Dynamics: Second Edition, by James Binney and Scott Tremaine. ISBN 978-0-691-13026-2 (HB). Published by Princeton University Press, Princeton, NJ USA, 2008.

Bovy, J., & Tremaine, S. 2012, ApJ, 756, 89

Bovy, J., Rix, H.-W., & Hogg, D. W. 2012b, ApJ, 751, 131

Bovy, J., Rix, H.-W., Hogg, D. W. et al., 2012c, ApJ, 755, 115

Bovy, J., Rix, H.-W., Liu, C., et al. 2012, ApJ, 753, 148

Bovy, J., & Rix, H.-W. 2013, ApJ, 779, 115

Bovy, J. 2015, ApJS, 216, 29 [TO DO]

Büdenbender, A., van de Ven, G., & Watkins, L. L. 2015, MNRAS, 452, 956

Dehnen, W. 1998, AJ, 115, 2384

De Lorenzi F., Debattista V.P., Gerhard O., Sambhus N. 2007, MNRAS, 376, 7

Famaey, B., & Dejonghe, H. 2003, MNRAS, 340, 752

Foreman-Mackey, D., Hogg, D. W., Lang, D., & Goodman, J. 2013, PASP [TO DO], 125, 306

Garbari, S., Liu, C., Read, J. I., & Lake, G. 2012, MNRAS, 425, 1445

Henon, M. 1959, Annales d'Astrophysique, 22, 126

[TO DO: In which order should I give the references????] [TO DO: replace the references which I typed myself with the ones from ADS.] [TO DO: Check if all references are actually used in paper. ???]

Holmberg, J., Nordström, B., & Andersen, J. 2009, A&A, 501, 941

Hunt, J. A. S., & Kawata, D. 2014, MNRAS, 443, 2112

Jurić, M., Ivezić, Ž., Brooks, A., et al. 2008, ApJ, 673, 864

Kawata, D., Hunt, J. A. S., Grand, R. J. J., Pasetto, S., & Cropper, M. 2014, MNRAS, 443, 2757

Klement, R., Fuchs, B., & Rix, H.-W. 2008, ApJ, 685, 261

Kuijken, K., & Gilmore, G. 1989, MNRAS, 239, 605

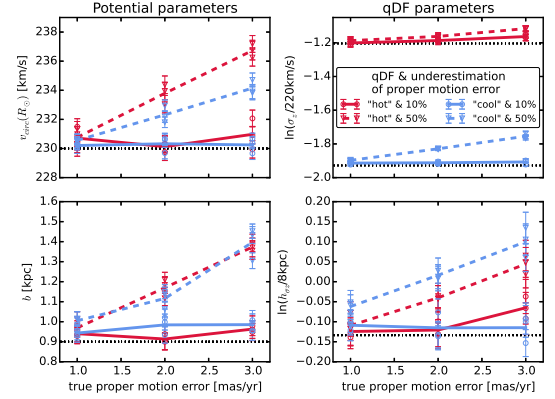


Figure 12. Effect of a systematic underestimation of proper motion errors in the recovery of the model parameters. The true model parameters used to create the mock data are summarized as Test ?? in Table ??, four of them are given on the y -axes and the true values are indicated as black dashed lines. The velocities of the mock data were perturbed according to Gaussian errors in the RA and DEC proper motions as indicated on the x -axis. The circles and triangles are the best fit parameters of several mock data set assuming the proper motion error, with which the model probability was convolved, was underestimated in the analysis by 10% or 50%, respectively. The error bars correspond to 1 sigma [TO DO: Can I say sigma????] confidence. The lines connect the mean of each two data realisations and are just guides to the eyes. [TO DO: rename $h_{\sigma z}$ to $h_{\sigma_{z,0}}$, σ_z to $\sigma_{z,0}$] [TO DO: Potential and/or population names in typewriter font]

McMillan, P. 2011, MNRAS, 414, 2446

McMillan, P. J. 2012, European Physical Journal Web of Conferences, 19, 10002

McMillan, P. J., & Binney, J. J. 2008, MNRAS, 390, 429

McMillan, P. J., & Binney, J. 2012, MNRAS, 419, 2251

McMillan, P. J., & Binney, J. J. 2013, MNRAS, 433, 1411

Navarro, J. F., Helmi, A., & Freeman, K. C. 2004, ApJ, 601, L43

Ness, M., Hogg, D. W., Rix, H.-W. et al., 2015, ApJ, 808, 16

Nordström, B., Mayor, M., Andersen, J., et al. 2004, A&A, 418, 989

Perryman, M. A. C., de Boer, K. S., Gilmore, G., et al. 2001, A&A, 369, 339

Piffl, T., Binney, J., McMillan, P. J., et al. 2014, MNRAS, 445, 3133

Read, J. I. 2014, Journal of Physics G Nuclear Physics, 41, 063101

Rix, H.-W., & Bovy, J. 2013, [TO DO] A& ARv, 21, 61

Sackett, P. 1997, ApJ, 483, 103

Sanders, J. L., & Binney, J. 2015, MNRAS, 449, 3479

Sellwood, J. A. 2010, MNRAS, 409, 145

Steinmetz, M. et al., 2006, AJ, 132, 1645

Strigari, L. E. 2013, Phys. Rep., 531, 1

Syer D., Tremaine S. 1996, MNRAS, 282, 223

Ting, Y.-S., Rix, H.-W., Bovy, J., & van de Ven, G. 2013, MNRAS, 434, 652

Yanny, B., Rockosi, C., Newberg, H. J., et al. 2009, AJ, 137, 4377

Zhang, L., Rix, H.-W., van de Ven, G., et al. 2013, ApJ, 772, 108

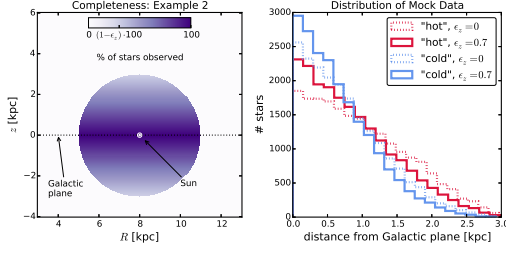


Figure 13. Selection function and mock data distribution for investigating vertical incompleteness of the data. All model parameters are summarized as Test ??, Example 2, in Table ?. The survey volume is a sphere around the sun and the percentage of observed stars is decreasing linearly with distance from the Galactic plane, as demonstrated in the left panel. How fast this detection/incompleteness rate drops is quantized by the factor ϵ_z . Histograms for four data sets, drawn from two MAPs (hot in red and cool in blue, see Table ?) and with two different ϵ_z , 0 and 0.7, are shown in the right panel for illustration purposes. [TO DO: Potential and/or population names in typewriter font]

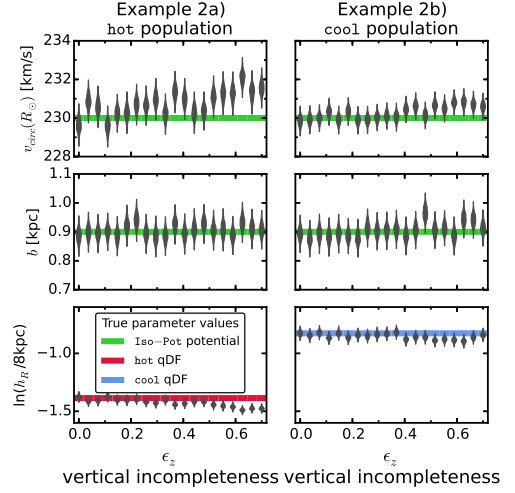


Figure 14. Influence of wrong assumptions about the incompleteness parallel to the Galactic plane of the data on the parameter recovery with *RoadMapping*. Each mock data set was created having different incompleteness parameters ϵ_z (shown on the x -axis and illustrated in Figure 13) and the model parameters are given as Test ??, Example 2, in Table ?. The analysis however didn't know about the incompleteness and assumed that all data sets had constant completeness within the survey volume ($\epsilon_z = 0$). The marginalized likelihoods from the fits are shown as violins. The green lines mark the true potential parameters (Iso-Pot) and the red and blue lines the true qDF parameters (hot MAP in red and cool MAP in blue), which we tried to recover. The *RoadMapping* method seems to be robust against small to intermediate deviations between the true and the assumed vertical data incompleteness, as well as the radial incompleteness in Figure 14.

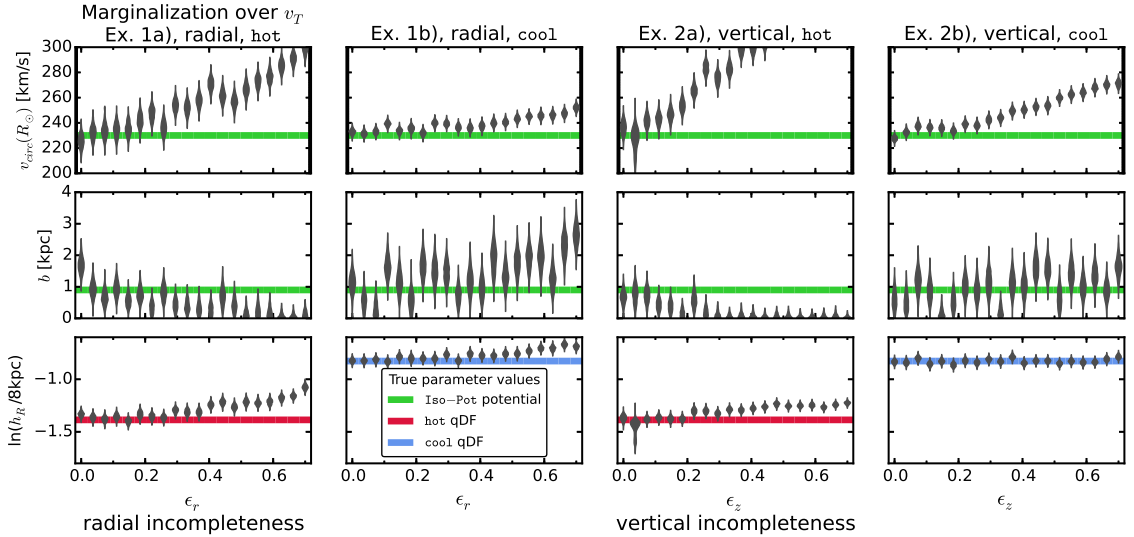


Figure 15. Influence of wrong assumptions about radial and vertical incompleteness on the parameter recovery, when *not* including information about the tangential velocities in the analysis. The mock data sets are the same as in Figure 9 and 14, but this time we did not include the data coordinates v_T in the analysis and therefore marginalized the likelihood over v_T instead (see §.1). This demonstrates that a lot of information about the potential is actually stored in the rotation curve, i.e. $v_T(R)$, which is not affected by removing stars from the data set. But even if we do not include v_T we can still recover the potential within the errors, at least for small ($\epsilon_z \lesssim 10\%$).