Draft version August 2, 2015
Preprint typeset using LaTeX style emulateapj v. 12/16/11

ACTION-BASED DYNAMICAL MODELS FOR THE MILKY WAY DISK WITH *ROADMAPPING* AND OUR IMPERFECT KNOWLEDGE OF THE "REAL WORLD"

W. Trick[1,2], J. Bovy[3,4], and H.-W. Rix[1]

Draft version August 2, 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 mono-abundance populations (*MAP*) in the Galactic disk (Bovy & Rix 2013; Binney & McMillan 2011; Binney 2012). *RoadMapping* is a full likelihood analysis that models the observed positions and velocities of *MAP* stars with an equilibrium, three-integral quasi-isothermal distribution function (qDF) in an axisymmetric potential. In preparation for the application to the large data sets of modern Galactic surveys like Gaia, we create and analyze a large suite of mock data sets and develop qualitative "rules of thumb" which characteristics and limitations of data, model and machinery affect constraints on the potential and qDF 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 seems to be no stars that are on manifestly more diagnostic orbits. [TO DO: Write about results on measurement errors] We investigate how small deviations of the stars' distribution from the assumed qDF affect the modelling: Having less stars at small Galactocentric radii R and with lower velocities as predicted by the qDF recovers the potential more reliably than having too many stars at small R and with higher velocities. The deviations of the true orbit distribution from the qDF introduced by binning stars into *MAP*s in the ([Fe/H],[α/Fe]) plane and due to abundance errors do not matter much, as long as the qDF parameters of two neighbouring *MAP*s 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 is the fundamental tool to infer the gravitational potential of the Milky Way from the positions and motions of its stars (Rix & Bovy 2013; Binney 2011b) [TO DO]. The observational information on the phase-space coordinates of stars are currently growing at a rapid pace, and will be taken to a whole new level by the upcoming Gaia data. Yet, rigorous and practical modelling tools that turn this information 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: Modelling tools for the MW: a) Made-to-measure: De Lorenzi et al. (2007)(based on Syer & Tremaine (1996) , best application to to bulge Bissantz et al. (2004), Hunt & Kawata (2014) also have a tool for Gaia at hand, b) Streams: Johnston et al. (1999) c) Action-based distribution function modelling Sanders & Binney (2015) Piffl et al. (2014) d) torus modelling McMillan & Binney 2003, models galaxy determining the potetial e) Jeans modelling Büdenbender et al. (2015) Loebman et al. (2012)]

Accurately determining the Galactic gravitational potential is fundamental for understanding its dark matter and baryonic structure [REF]. Accurately determining the stellar-population dependent orbit distribution function is a fundamental contraint on the Galaxy's formation history.

Open questions about the MW's potential and structure, on which future modelling attempts will hopefully give more definite answers are: What is the local dark matter density (Zhang et al. 2013; Bovy & Tremaine 2012)? Is the Milky Way's dark matter halo flattened ([REF])? Is the MW disk maximal (Sackett 1997) and, to be able to disentangle halo and disk contribution (Dehnen & Binney 1998), what is the disk's overall mass

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???]
[4] Hubble fellow

scale length (Bovy & Rix 2013)?

Open questions about the star's distribution within the MW, which dynamical modelling can help to constrain, are: How are stellar kinematics and their chemical abundances are related (Sanders & Binney (2015),[REF])? In particular, does the disk have a thin/thick disk dichotomy (Gilmore & Reid (1983)) or is it a continuum of many exponential disks (Bovy et al. (2012d))? How does radial migration affect the orbit distribution (Sellwood & Binney 2002; Roškar et al. 2008a,b; Schönrich & Binney 2008; **?**) [TO DO: These are References from Rix & Bovy 2013 - should I use all of them?]?
To address these questions, observed stellar positions and motions need to be turned into full orbits - which stresses again the importance of having a reliable model for the MW's gravitational potential.

In the era of big Galactic surveys all of this could soon be within our reach. Not only will there be full 6D stellar phase-space coordinates for a thousand million of stars measured by Gaia to unprecedented precision by the end of 2016. But already with existing surveys (e.g., SEGUE (Beers et al. 2006), RAVE (Steinmetz et al. 2006), LAMOST (Newberg et al. 2012), APOGEE (Majewski 2012), Gaia-ESO (Gilmore et al. 2012), GALAH (Freeman 2012) [TO DO: I just copied this from Melissas Cannon paper. Should I reference all of them??? Not in reference list yet.]) and sophisticated machine-learning tools (e.g. *The Cannon* by Ness et al. (2015)) to combine them, we will soon have huge data sets at our disposal.
In this work we present a rigorous, robust and reliable dynamical modelling machinery, strongly building on previous work by Binney & McMillan (2011); Binney (2012); Bovy & Rix (2013); Bovy (2015) and explicitly developed to exploit and deal with these large data sets in the future.
There is a variety of practical approaches to dynamical modelling of discrete collisionless tracers (such as the stars in the Milky Way) [REF]. Most of them – explicitly or implicitly – describe 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 [Binney 2011abcdefg???].

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). **?** 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 [Fe/H] and [α/Fe] (Bovy et al. (2012b), Bovy et al. (2012c), Bovy et al. (2012d)).

Bovy & Rix (2013) implemented a 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 Gdwarf (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 *MAP*s 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 (~ 2 kpc) and on the disk to dark matter ratio at the Solar radius [TO DO: quote number???].

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 modelling machinery by Bovy & Rix (2013), making extensive use of the *galpy* python package (Bovy (2015)). *RoadMapping* relaxes some of the restraining assumptions Bovy & Rix (2013) had to made, 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. What is it about the data, the model and the machinery itself, that limits our recovery of the true gravitational potential?

In the light of Gaia we explicitely analyze how well the modelling machinery behaves in the limit of large data. For a huge number of stars three statistical aspects become important, that are hidden behind Poisson noise for smaller data sets: (i) We have to make sure that our modelling is an un-biased and asymptotically normal estimator (§3.1). (ii) Numerical inaccuracies in the actual modelling machinery start to matter and need to be avoided (§**??**). (iii) Parameter estimates become so precise, that we start to be able to distinguish between similar models. We therefore want more flexibility and more free fit parameters in the potential and DF model. The modelling machinery itself needs to be flexible and fast in effectively finding the best fit parameters for a large set of parameters. The improvements made to the machinery used in Bovy & Rix (2013) are presented in §2.6.

Different characteristics of the data might influence the success of the parameter recovery. (i) In an era where we can choose data from different MW surveys, it might be worth to explore if different regions within the MW

(i.e. differently shaped or positioned survey volumes) are especially diagnostic to recover the potential (§**??**). (ii) What happens if our knowledge about the selection function, specifically the completeness of the data set within the survey volume, is not perfect (§3.3)? (iii) How to account for measurement errors in the modelling (§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 we hope to recover the true potential, when our potential and DF models deviate from the true potential and DF. For the DF we specifically investigate two of our assumptions in §3.5: First, what would happen if the stars within *MAP*s do intrinsically not follow a single qDF as assumed by Ting et al. (2013); Bovy & Rix (2013). Second, and assuming *MAP*s do indeed follow the qDF, what would be the effect of pollution of *MAP*s through stars from neighbouring *MAP*s in the ([Fe/H],[α/Fe]) plane due to too big abundance errors or bin sizes.
And last but not least we test in §**??** how well the modelling works, if our assumed potential family deviaties from the true potential.

For all of these 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 §2.1-2.3 following the procedure in §2.4, analysed according to the description of the machinery in §2.5-2.6 and the results are presented in §3 and discussed in §4.

The strongest assumption that goes into this kind of dynamical modelling might be the idealization of the Galaxy to be axi-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.

2. DYNAMICAL MODELLING

2.1. *Actions and Potential Models*

Actions.— Orbits in axisymmetric potentials are best described and fully specified by the three actions J_R, J_z and $J_\phi = L_z$. They are integrals of motion and generally defined as

$$J_i = \frac{1}{2\pi} \int_{\text{orbit}} p_i(t)\, \mathrm{d}x_i(t) \tag{1}$$

and 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 [REF]. 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). Even though actions are the optimal choice as orbit labels and arguments for stellar distribution functions, their computation is very expensive.

Action calculation.— The action calculation depends on the choice of potential in which the star moves: The spherical isochrone (Binney & Tremaine 2008) is the only potential for which Eq. (1) takes an analytic form. For axisymmetric 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 Eq. (1) is integrated up to infinity. 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).[5]

Potential models.— In our modelling we assume a family of parametrized potentials with a fixed number of free parameters. We use different kinds of potentials: Besides the Milky Way like potential from Bovy & Rix (2013) ("MW13-Pot") with bulge, disk and halo, we also extensively use 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. In addition we use the 2-component Kuzmin-Kutuzov Stäckel potential by Batsleer & Dejonghe (1994) ("KKS-Pot"), which displays a disk and halo structure and also provides exact actions. Table 1 summarizes all reference potentials together used in this work with their free parameters p_Φ. The density distribution of these potentials is illustrated in Fig. 1.

2.2. *Distribution Function*

Distribution Function.— Motivated by the findings of Bovy et al. (2012b,c,d) and Ting et al. (2013) about the simple phase-space structure of *MAP*s , and following Bovy & Rix (2013) and their successful application, we also assume that each *MAP* follows a single qDF of the form given by Binney & McMillan (2011). This qDF is a function of the actions $\boldsymbol{J} = (J_R, J_z, L_z)$ and has the form

$$\begin{aligned} &\mathrm{qDF}(\boldsymbol{J} \mid p_{\mathrm{DF}}) \\ &= f_{\sigma_R}(J_R, L_z \mid p_{\mathrm{DF}}) \times f_{\sigma_z}(J_z, L_z \mid p_{\mathrm{DF}}) \end{aligned} \tag{2}$$

with

$$\begin{aligned} f_{\sigma_R}(J_R, L_z \mid p_{\mathrm{DF}}) = n \times \frac{\Omega}{\pi \sigma_R^2(R_g)\kappa} \exp\left(-\frac{\kappa J_R}{\sigma_R^2(R_g)}\right) \\ \times [1 + \tanh(L_z/L_0)] \end{aligned} \tag{3}$$

$$f_{\sigma_z}(J_z, L_z \mid p_{\mathrm{DF}}) = \frac{\nu}{2\pi\sigma_z^2(R_g)} \exp\left(-\frac{\nu J_z}{\sigma_z^2(R_g)}\right) \tag{4}$$

$$\tag{5}$$

Here $R_g \equiv R_g(L_z)$ and $\Omega \equiv \Omega(L_z)$ are the (guidigcenter) 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). 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 random small

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

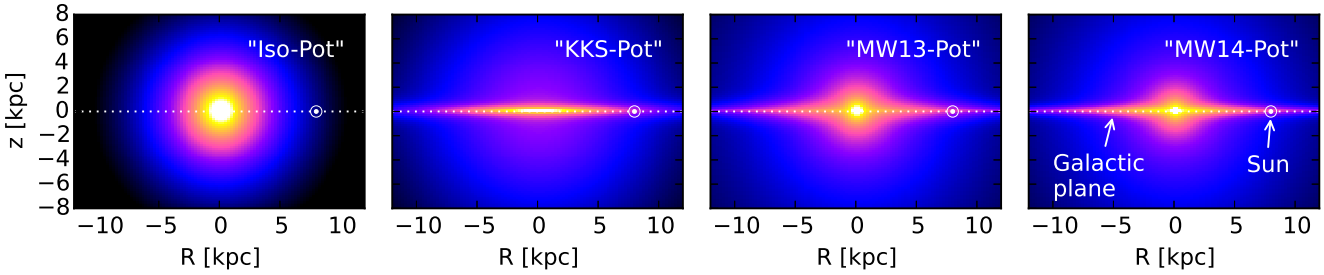

Figure 1. Density distribution of the four reference galaxy potentials in Table 1, for illustration purposes. These potentials are used throughout this work for mock data creation and potential recovery. [TO DO: Halo sichtbarer machen, evtl. mit isodensity contours]

value ($L_0 = 10 \times R_\odot/8 \times v_{\rm circ}(R_\odot)/220$).
For this qDF to be able to incorporate the findings by Bovy et al. 2012??? about the phase-space structure of *MAP*s summarized in §1, we set the functions n, σ_R and σ_z, which indirectly set the stellar number density and radial and vertical velocity dispersion profiles,

$$n(R_g \mid p_{\rm DF}) \propto \exp\left(-\frac{R_g}{h_R}\right) \tag{6}$$

$$\sigma_R(R_g \mid p_{\rm DF}) = \sigma_{R,0} \times \exp\left(-\frac{R_g - R_\odot}{h_{\sigma_R}}\right) \tag{7}$$

$$\sigma_z(R_g \mid p_{\rm DF}) = \sigma_{z,0} \times \exp\left(-\frac{R_g - R_\odot}{h_{\sigma_z}}\right). \tag{8}$$

The qDF for each *MAP* has therefore a set of five free parameters $p_{\rm DF}$: 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_{σ_R} and h_{σ_z}, that describe the radial decrease of the velocity dispersion. The *MAP*s we use for illustration through out this work are summarized in Table 2.

Tracer Density. — One crucial point in our dynamical modelling technique (§???), as well as in creating mock data (§2.4), is to calculate the (axisymmetric) spatial tracer density $\rho_{\rm DF}(\boldsymbol{x} \mid p_\Phi, p_{\rm DF})$ 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 $N_{\rm velocity}$-th order Gauss-Legendre quadrature for each integral:

$$\begin{aligned}
&\rho_{\rm DF}(R, |z| \mid p_\Phi, p_{\rm DF}) \\
&= \int_{-\infty}^{\infty} {\rm qDF}(\boldsymbol{J}[R, z, \boldsymbol{v} \mid p_\Phi] \mid p_{\rm DF})\, {\rm d}^3\boldsymbol{v} \qquad (9)\\
&\approx \int_{-N_{\rm sigma}\sigma_R(R|p_{\rm DF})}^{N_{\rm sigma}\sigma_R(R|p_{\rm DF})} \int_{-N_{\rm sigma}\sigma_z(R|p_{\rm DF})}^{N_{\rm sigma}\sigma_z(R|p_{\rm DF})} \int_0^{1.5 v_{\rm circ}(R_\odot)} \\
&\qquad {\rm qDF}(J[R, z, \boldsymbol{v} \mid p_\Phi] \mid p_{\rm DF})\, {\rm d}v_T\, {\rm d}v_z\, {\rm d}v_R, \qquad (10)
\end{aligned}$$

where $\sigma_R(R \mid p_{\rm DF})$ and $\sigma_z(R \mid p_{\rm DF})$ are given by eq. (7) and (8) and the integration ranges are motivated by Fig. 2. For a given p_Φ and $p_{\rm DF}$ we explicitly calculate the density on $N_{\rm spatial} \times N_{\rm spatial}$ 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 $z > 0$. The total number of actions that need to be calculated to set up the density interpolation grid is $N^2_{\rm spatial} \cdot N^3_{\rm velocity}$. Fig. ??? shows the importance of choosing $N_{\rm spatial}$, $N_{\rm velocity}$ and $N_{\rm sigma}$ sufficiently large in order to get the density with an acceptable numerical accuracy.

2.3. *Selection Function*

Galactic Coordinate System. — Our modelling takes place in the Galactocentric rest-frame with cylindrical coordinates $\boldsymbol{x} \equiv (R, \phi, z)$ and corresponding velocity components $\boldsymbol{v} \equiv (v_R, v_\phi, v_z)$. If the stellar phase-space data is given in observed coordinates, position $\tilde{\boldsymbol{x}} \equiv (\alpha, \delta, m - M)$ in right ascension α, declination δ and distance modulus $(m - M)$, and velocity $\tilde{\boldsymbol{v}} \equiv (\mu_\alpha, \mu_\delta, v_{\rm los})$ as proper motions $\boldsymbol{\mu} = (\mu_\alpha, \mu_\delta)$ [TO DO: cos somwhere???] and line-of-sight velocity $v_{\rm los}$, the data $(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{v}})$ has to be converted first into the Galactocentric rest-frame coordinates $(\boldsymbol{x}, \boldsymbol{v})$ using the sun's position and velocity. For simplicity we assume for the sun

$$(R_\odot, \phi_\odot, z_\odot) = (8\ {\rm kpc}, 0°, 0\ {\rm kpc})$$
$$(v_{R,\odot}, v_{T,\odot}, v_{z,\odot}) = (0, 230, 0)\ {\rm km\ s^{-1}}.$$

Selection Function. — A survey's selection function can be understood as a subvolume in the space of observables: e.g. position on the plane of the sky (limited by the pointing of the survey), 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).
Within the framework of this paper, using only mock data for testing purposes, we ignore target cuts in colors and metallicity and simply use spatial selection functions, which we define as

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

It's value describes the probability to observe a star at $\boldsymbol{x}$.
For the observed volume we use simple geometrical shapes: Either a sphere of radius $r_{\rm max}$ with the sun at its center, or a "wedge", which we define as the angular segment of an cylindrical annuli, i.e. the volume with $R \in [R_{\rm min}, R_{\rm max}], \phi \in [\phi_{\rm min}, \phi_{\rm max}], z \in [z_{\rm min}, z_{\rm max}]$ 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.
The completeness is, in our framework, a function of position with $0 \leq {\rm completeness}(\boldsymbol{x}) \leq 1$ everywhere inside the observed volume. It could be understood as a position-dependent detection probability. Unless explicitly stated

otherwise, we use everywhere

$$\text{completeness}(\boldsymbol{x}) = 1.$$

2.4. *Mock Data*

One goal of this work is to test how the loss of information in the process of measuring stellar phase-space coordinates can affect the outcome of the modelling. To 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. In further steps we can manipulate and modify these mock data sets to mimick observational effects.
The distribution function is given in terms of actions and angles. The transformation $(\boldsymbol{J}_i, \boldsymbol{\theta}_i) \longrightarrow (\boldsymbol{x}_i, \boldsymbol{v}_i)$ is however difficult to perform and computationally much more expensive than the transformation $(\boldsymbol{x}_i, \boldsymbol{v}_i) \longrightarrow (\boldsymbol{J}_i, \boldsymbol{\theta}_i)$. We propose 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.

Preparation: Tracer density.— We first setup the interpolation grid for the tracer density $\rho(R, |z| \mid p_\Phi, p_{\rm DF})$ generated by the given qDF and according to §2.2 and Eq. 10. For the creation of the mock data we use $N_{\rm spatial} = 20$, $N_{\rm velocity} = 40$ and $N_{\rm sigma} = 5$.

Step 1: Drawing positions from the selection function.— To get positions $\boldsymbol{x}_i$ for our mock data stars, we first sample random positions (R_i, z_i, ϕ_i) uniformly from the observed volume. Then we apply a rejection Monte Carlo method to these positions using the pre-calculated $\rho_{\rm DF}(R, |z| \mid p_\Phi, p_{\rm DF})$. In an optional third step, if we want to apply a non-uniform selection function, $\text{sf}(\boldsymbol{x}) \neq$ const. within the observed volume, we use the rejection method a second time. The sample then follows

$$\boldsymbol{x}_i \longrightarrow p(\boldsymbol{x}) \propto \rho_{\rm DF}(R, z \mid p_\Phi, p_{\rm DF}) \times \text{sf}(\boldsymbol{x}).$$

Step 2: Drawing velocities according to the distribution function.— The velocities are independent of the selection function and observed volume. For each of the positions (R_i, z_i) we now sample velocities directly from the $\text{qDF}(R_i, z_i, \boldsymbol{v} \mid p_{Phi}, p_{\rm DF})$ 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 set according to the required:

$$(\boldsymbol{x}_i, \boldsymbol{v}_i) \longrightarrow p(\boldsymbol{x}, \boldsymbol{v}) \propto \text{qDF}(\boldsymbol{x}, \boldsymbol{v} \mid p_\Phi, p_{\rm DF}) \times \text{sf}(\boldsymbol{x}).$$

Example:— Fig. 2 shows examples of mock data sets in configuration space $(\boldsymbol{x}, \boldsymbol{v})$ and action space. The qDF represents realistic stellar distributions in position-velocity 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 2). The distribution of tangential velocities v_T is skewed because of asymmetric drift [TO DO: Find out, if we need an explanation for asymmetric drift here]
The distribution in action space demonstrates the intuitive physical meaning of actions: The stars of the "cool" *MAP* 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" *MAP* 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 $L_z \sim R v_{\rm circ}$ and the different radial extent of both volumes. The volume above the plane contains more stars with higher J_z, because stars with small J_z can't 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 Fig. 2. [TO DO: Part of this could also be mentioned in the figure caption.]

Introducing measurement errors.— If we want to add measurement errors to the mock data, we need to apply two modifications to the above procedure.
First, measurement errors are best described in the phase-space of observables. We use the heliocentric coordinate system right ascension and declination (α, δ) and distance modulus $(m - M)$ as proxy for the distance from the sun, the proper motion in both α and δ direction (μ_α, μ_δ) and the line-of-sight velocity $v_{\rm los}$. For the conversion between these observables and the Galactocentric cylindrical coordinate system in which the analysis takes place, we need the position and velocity of the sun, which we set for simplicity in this study to be $(R_\odot, z_\odot) = (8, 0)$ kpc and $(v_R, v_T, v_z) = (0, 230, 0)$ km s^{-1}. We assume Gaussian measurement errors in the observables $\tilde{\boldsymbol{x}} = (\alpha, \delta, (m - M))$, $\tilde{\boldsymbol{v}} = (\mu_\alpha, \mu_\delta, v_{\rm los})$.
Second, in the case of distance errors, stars can virtually scatter in and out of the observed volume. To account for this, we first draw "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. [TO DO: Can I say it like this???] We then sample velocities (given the "true" positions of the stars) as described above and perturb them according to the measurement errors as well.

2.5. *Likelihood*

Form of the likelihood.— As data we use the positions and velocities of stars coming from a given *MAP* and survey selection function $\text{sf}(\boldsymbol{x})$,

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

The model that we fit to the data is a parametrized potential and a single qDF with a given number of fixed

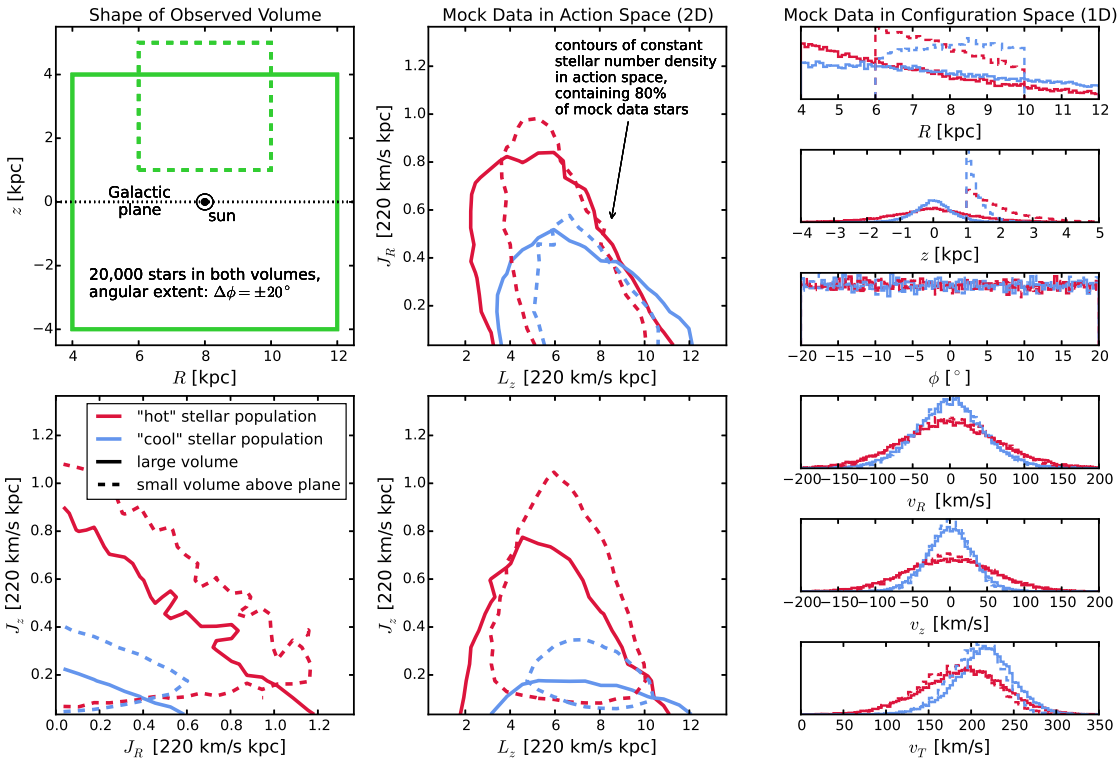

Figure 2. Distribution of mock data in action space (2D iso-density contours enclosing 80% of the stars, in the two central and the lower left panel) and configuration space (1D histograms in right panels), depending on shape and position of the survey observation volume and temperature of the stellar population. The parameters of the mock data model is given as Test ① in Table 3. 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 centered on the Galactic plane (solid lines) and a smaller one above the plane (dashed lines).

and free parameters,

$$p_M = \{p_{\rm DF}, p_\Phi\},$$

We fit the qDF parameters (see §2.2) with a logarithmically flat prior, i.e. flat priors in

$$\begin{aligned} p_{\rm DF} := \{ & \ln\left(h_R/8{\rm kpc}\right), \\ & \ln\left(\sigma_R/220{\rm km\ s^{-1}}\right), \ln\left(\sigma_z/220{\rm km\ s^{-1}}\right), \\ & \ln\left(h_{\sigma_R}/8{\rm kpc}\right), \ln\left(h_{\sigma_z}/8{\rm kpc}\right) \ \}. \end{aligned}$$

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

The likelihood of the data given the model $\mathscr{L} = (D \mid p_M)$ is the product of the probabilities for each star to move in the potential with p_Φ, being within the survey's selection function and it's orbit to be drawn from the qDF with $p_{\rm DF}$, i.e.

$$\begin{aligned} & \mathscr{L}(p_M \mid D) \\ & \equiv \prod_i^N P(\boldsymbol{x}_i, \boldsymbol{v}_i \mid p_M) \\ & = \prod_i^N \frac{1}{(r_o v_o)^3} \cdot \frac{{\rm qDF}(\boldsymbol{J}_i \mid p_M) \cdot {\rm sf}(\boldsymbol{x}_i)}{\int {\rm d}^3x\, {\rm d}^3v\ {\rm qDF}(\boldsymbol{J} \mid p_M) \cdot {\rm sf}(\boldsymbol{x})} \\ & \propto \prod_i^N \frac{1}{(r_o v_o)^3} \cdot \frac{{\rm qDF}(\boldsymbol{J}_i \mid p_M)}{\int {\rm d}^3x\ \rho_{\rm DF}(R, |z| \mid p_M) \cdot {\rm sf}(\boldsymbol{x})}, \end{aligned} \quad (11)$$

where N is the number of stars in the data set D. In the last step we used eq. (9). The factor $\prod_i {\rm sf}(\boldsymbol{x}_i)$ is independent of the model parameters, so we simply evaluate Eq. (11) in the likelihood calculation. We find the best set of model parameters by maximising the likelihood.

A word on units.— We evaluate the likelihood in a scale-free potential within a Galactocentric coordinate system which is defined as $v_{\rm circ}(R = 1) = 1$. The circular velocity at the sun's radius, $v_{\rm circ}(R_\odot = 8.{\rm kpc}) \sim 230{\rm km\ s^{-1}}$, determines the total mass amplitude of the galaxy potential. In the modelling all data and model parameters are re-scaled to spatial units of $r_o := R_\odot$ or velocitiy units of $v_o := v_{\rm circ}(R_\odot)$. The prefactor $1/\left(r_o v_o\right)^3$ in eq. (11) makes sure that the likelihood has the correct units to satisfy:

$$\int P(\boldsymbol{x}, \boldsymbol{v} \mid p_M)\, {\rm d}^3x\, {\rm d}^3v \propto 1$$

Including this prefactor is crucial when $v_{\rm circ}(R_\odot)$ is a free fitting parameter.

Numerical accuracy in calculating the likelihood.— The normalisation in Eq. (11) is a measure for the total number of tracers inside the survey volume,

$$M_{\rm tot} \equiv \int {\rm d}^3x\ \rho_{\rm DF}(R, |z| \mid p_m odel) \cdot {\rm sf}(\boldsymbol{x}). \quad (12)$$

In the case of an axisymmetric galaxy model and ${\rm sf}(\boldsymbol{x}) = 1$ 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} (see Eq. (10) and surrounding text) over the survey volume times the observation volume's geometric angular contribution at each (R, z). We perform this integral as a Gauss Legendre quadrature of order 40 in each R and z direction. Unfortunately the evaluation of the likelihood for only one set of model parameters is computationally expensive. The computation speed is set by the number of action calculations required, i.e. the number of stars and the numerical accuracy of the integrals in Eq. ??? needed for the normalisation, which requires $N^2_{\rm spatial} \times N^3_{\rm velocity}$ action calculations. The accuracy has to be chosen high enough, such that a resulting numerical error

$$\delta_{M_{tot}} \equiv \frac{M_{\rm tot}(N_{\rm spatial}, N_{\rm velocity}, N_{\rm sigma}) - M_{\rm tot,true}}{M_{\rm tot,true}} \quad (13)$$

does not dominate the likelihood, i.e.

$$\begin{aligned}&\log \mathscr{L}(p_M \mid D)\\ =&\sum_i^N \log \mathrm{qDF}(\boldsymbol{J_i} \mid p_M) - 3N \log(r_o v_o)\\ &-N\log(M_{\rm tot,true}) - N\log(1+\delta_{M_{tot}}),\end{aligned} \quad (14)$$

with

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

In other words, this error is only small enough, if it does not affect the comparison of two adjacent models whose likelihoods differ, to be clearly distinguishable, by a factor of 10. Otherwise numerical inaccuracies could lead to systematic biases in the potential and DF fitting. For data sets as large as N = 20,000 stars in one *MAP* , which in the age of GAIA could very well be the case [TO DO: Really???], we would need a numerical accuracy of 0.005% in the normalisation. Fig. 3 demonstrates that the numerical accuracy we use in the analysis, $N_{\rm spatial} = 16$, $N_{\rm velocity} = 24$ and $N_{sigma} = 5$, does satisfy this requirement.

Dealing with measurement errors.— We assume Gaussian errors in the observable space $\boldsymbol{y}_i \equiv (\tilde{\boldsymbol{x}}_i, \tilde{\boldsymbol{v}}_i) = (\alpha, \delta, (m-M), \mu_\alpha, \mu_\delta, v_{\rm los})$,

$$\begin{aligned}N[\boldsymbol{y}_i, \sigma_{y,i}](\boldsymbol{y'}) &= N[\boldsymbol{y'}, \sigma_{y,i}](\boldsymbol{y}_i)\\ &\equiv \prod_k \frac{1}{\sqrt{2\pi\sigma^2_{y,k}}} \exp\left(-\frac{(y_{i,k}-y'_k)^2}{2\sigma^2_{y,k}}\right),\end{aligned}$$

where $y_{i,k}$ are the coordinate components of $\boldsymbol{y}_i$. Observed stars follow the (quasi-isothermal) distribution function ($\mathrm{DF}(\boldsymbol{y}) \equiv \mathrm{qDF}(\boldsymbol{J}[\boldsymbol{y} \mid p_\Phi] \mid p_{\rm DF}]$) for short), convolved with the error distribution $N[0, \sigma_y](\boldsymbol{y})$. The selection function $\mathrm{sf}(\boldsymbol{y})$ acts on the space of (error affected) observables. Then the probability of one star coming from potential p_Φ, distribution function $p_{\rm DF}$ and being affected by the measurement errors $\sigma_{\boldsymbol{y}}$ becomes

$$\begin{aligned}&\tilde{P}(\boldsymbol{y}_i \mid p_\Phi, p_{\rm DF}, \sigma_{\boldsymbol{y},i})\\ \equiv&\frac{\mathrm{sf}(\boldsymbol{y}_i)\cdot \int \mathrm{d}^6y'\ \mathrm{DF}(\boldsymbol{y'})\cdot N[\boldsymbol{y}_i, \sigma_{\boldsymbol{y},i}](\boldsymbol{y'})}{\int \mathrm{d}^6y\ \mathrm{DF}(\boldsymbol{y})\cdot \int \mathrm{d}^6y'\ \mathrm{sf}(\boldsymbol{y'})\cdot N[\boldsymbol{y}, \sigma_{\boldsymbol{y},i}](\boldsymbol{y'})}.\end{aligned}$$

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

$$\begin{aligned}&\tilde{P}(\boldsymbol{y}_i \mid p_\Phi, p_{\rm DF}, \sigma_{\boldsymbol{y},i})\\ \approx&\frac{\mathrm{sf}(\boldsymbol{x}_i)}{\int \mathrm{d}^6y\ \mathrm{DF}(\boldsymbol{y})\cdot \mathrm{sf}(\boldsymbol{x})}\cdot\frac{1}{N_{\rm error}}\sum_n^{N_{\rm error}} \mathrm{DF}(\boldsymbol{x}_i, \boldsymbol{v}[\boldsymbol{y}'_{i,n}])\end{aligned}$$

with

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

In doing so, we ignore errors in the star's position $\boldsymbol{x}_i$ altogether. This simplifies the normalisation drastically and makes it independent of measurement errors, including the velocity errors. Distance errors however are included, but only implicitly in the convolution over the stars' velocity errors in the Galactocentric restframe. We calculate the convolution using Monte Carlo integration with $N_{\rm error}$ samples drawn from the full error Gaussian in observable space, $y'_{i,n}$.

2.6. *Fitting Procedure*

We search the $(p_\Phi, p_{\rm DF})$ parameter space for the maximum of the likelihood in Eq. (11) using a 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 (or rather the posterior probability distribution) using a Monte-Carlo Markov Chain (MCMC) approach.

Fitting Step 1: Nested-grid search.— The $(p_\Phi, p_{\rm DF})$ 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 $M = (p_\Phi, p_{\rm DF})$, we set up a sufficiently large initial grid with 3^N regular grid points.[6]
- *Evaluation.* We evaluate the likelihood at each grid-point. Because of the many conputationally expensive $\boldsymbol{x}, \boldsymbol{v} \xrightarrow{p_\Phi} \boldsymbol{J}$ transformations that have to be performed for each new set of p_Φ parameters, an outer loop iterates over the p_Φ parameters and precalculates the actions, while an inner loop evaluates the likelihood Eq. (11) for all qDF parameters $p_{\rm DF}$ with the actions in the given potential and (analogously to Fig. 9 in Bovy & Rix (2013)).
- *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

[6] To get a better feeling where in parameter space the true $p_{\rm DF}$ parameters lie, we fit eq. (???) directly to the data. This gives a very good inital guess for $\sigma_{R,0}$ and $\sigma_{z,0}$. To improve the estimate for h_R, we fit eq. (???) only to stars within a thin wedge around $(R = 0, z = 0)$ and then apply the relation in fig. 5 in Bovy & Rix (2013) between the stars' measured scale length $h_R^{\rm out}$ and the qDF tracer scale length $h_R^{\rm in} = h_R$.

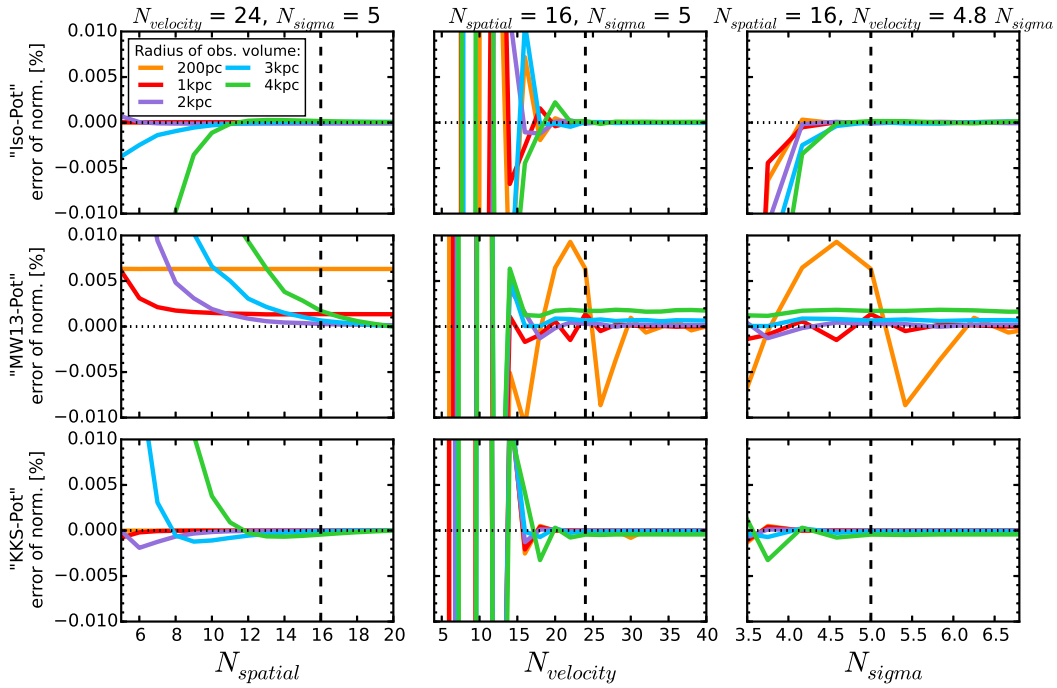

Figure 3. Relative error of the likelihood normalization in Eq. (13) depending on the accuracy of the density calculation in Eq. (10) (and sourrounding text). The different colors represent calculations for different radii of the spherical observation volume around the sun, as indicated in the legend. $N_{\rm spatial}$ is the number of regular grid points in each R and $z > 0$ within the observed volume on which the tracer density is evaluated according to Eq. (10). At each (R, z) a Gauss-Legendre integration of order $N_{\rm velocity}$ is performed over an integration range of $\pm N_{\rm spatial}$ times the dispersion in v_R and v_z and $[0, 1.5 v_{\rm circ}(R_\odot)]$ [TO DO: update if required] in v_T. To integrate the interpolated density over the observed volume to arrive at the likelihood normalization in Eq. (12), we perform a 40th-order Gauss-Legendre integration in each R and z direction. We compare the convergence of the normalisation for the "hot" qDF in three potentials, "Iso-Pot", "MW13-Pot" and "KKS-Pot" (see also test ⑨ in Table 3 for all other model details). In each column of plots we keep two of the accuracy parameters fixed (indicated on top), while the third parameter is varied. (Caption continues on next page.)

Figure 3. (Continued.) We calculate the true normalization with high accuracy as $M_{\rm tot,true} \approx M_{\rm tot}(N_{\rm spatial} = 20, N_{\rm velocity} = 56, N_{\rm sigma} = 7)$. The dashed lines indicate the accuracy used in our analyses: it is better than 0.002% for all three potential types. 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 isochrone and the Staeckel potential exactly, we have to resort to an approximate action calculation for the MW-like potential (see §???). [TO DO: Try to redo yellow curve in MW. Weird, that it does not depend on $N_{spatial}$.???]

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 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 $N_j \times N_{\rm error}$ [TO DO: explain $N_{\rm error}$???] stellar data actions $\boldsymbol{J}_i$. To calculate the normalisation in Eq. (11), $N^2_{\rm spatial} \times N^3_{\rm velocity}$ actions $\boldsymbol{J}_n$ are needed. Formally the spatial coordinates at which the $\boldsymbol{J}_n$ are calculated depend on the $p_{\rm DF}$ parameters via the integration ranges in eq. (10). To relax this dependence we instead use the same velocity integration limits in the likelihood calculations for all $p_{\rm DF}$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 Eq. (???), 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 $p_{\rm DF}$. Choosing a fiducial qDF that is very off from the true qDF can however lead to large biases. The optimal values for the fiducial qDF are the (yet unknown) best fit $p_{\rm DF}$ parameters. We take care of this by setting, in each iteration step of the nested-grid search, the fiducial qDF simply to the $p_{\rm DF}$ 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.

- *Speed Limitations.* Overall the computation speed of this nested-grid approach is dominated (in descending order of importance) by a) the complexity of potential and action calculation, b) the number $N_j \times N_{\rm error} + N^2_{\rm spatial} \times N^3_{\rm velocity}$ of actions to calculate, i.e. the number of stars, error samples and nu-

merical 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 it's 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 exponentially with number of free parameters N. For a large number of free parameters ($N > 4$) a Monte Carlo Markov Chain (MCMC) approach might sample the likelihood (or rather the posterior probability distribution, which is the likelihood times some priors, see §????) much faster. We use *emcee* by Foreman-Mackey et al. (2013) and release the walkers very close to the likelihood 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 Eq. (10) the current qDF parameters as each values can be used as integration ranges. 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.

3. RESULTS

We are now in a position to explore the questions about the ultimate limitations of action based modelling, posed in the introduction:

- Can we still retrieve unbiased model parameter estimates p_M in the limit of large sample sizes?
- What role does the survey volume and geometry play, at given sample size?
- What if our knowledge of the sample selection function is imperfect, and potentially biased?
- How do the parameter estimates deteriorate if the individual errors on the phase-space coordinates become significant?

But we also consider the more fundamental limitations:

- What if the observed stars are not extacly drawn from the family of model distribution functions?
- What happens to the estimate of the potential and the DF, if the actual potential is not contained in the family of model potentials?

We do not explore the breakdown of the assumption that the system is axisymmetric and in steady state. Except of the test suite on measurement errors in §3.4, we assume that the phase-space errors are negligible.

3.1. *Model parameter estimates in the limit of large data sets*

The individual *MAP* in Bovy & Rix (2013) contained typically between 100 and 800 objects, so that each *MAP* implied a quite broad *pdf* for the model parameters $p_M = \{p_\Phi, p_{\rm DF}\}$. Here we explore what happens in the limit of very much larger samples for each *MAP* , say 20,000 objects. As outlined in §2.5 the immediate consequence of larger samples is given by the likelihood normalization requirement, $\log(1 + rel.error) \leq 1/N_{sample}$, (see Eq. 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 $1/N_{\rm sample}$. Further, we must verify that any bias in the *pdf* expectation value is far less than SE, even for quite large samples.

Using sets of mock data, created according to §2.4 and with our fiducial model for p_M in Table 3, Tests ②, ③ and ⑩, we verified that *RoadMapping* satisfies all these conditions and expectations. Fig. 4 illustrates the joint *pdf*'s of all p_M. This figure illustrates that the *pdf*'s are multivariate Gaussians that project 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 with the mock data were drawn. Figure5 then illustrates that the *pdf* width, SE, indeed scales as $1/N_{\rm sample}$. Fig.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.

3.2. *The Role of the Survey Volume Geometry*

Beyond the sample size, the survey volume *per se* must play a role; clearly, even a vast and perfect data set of stars within 100 pc of the Sun, has limited power to tell us about the potential at very different R. Intuitively, having dynamical tracers over a wide range in R suggests to allow tighter constraints on the radial dependence of the potential. To this end, we devise two suites of mock data sets:
The first one draws mock data from the same p_M (see Test ④ in Table 3), but from four different volume wedges (see §2.3), illustrated in the right upper panel of fig. 7. To make the parameter inference comparison very differential, 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 edges). We perform the same test for three different potential models ("Iso-Pot", "MW14-Pot" and "KKS-Pot").
The results of this first mock suite are shown in Fig. 7 for all potential fit parameters. The mock suite was created to investigate the influence of *position and shape* of the survey volume within the Galaxy, together with the *choice of potential model*.
The second suite of mock data sets was already introduced in §3.1 (see also Test ③), where mock data sets were drawn from five spherical volumes around the sun

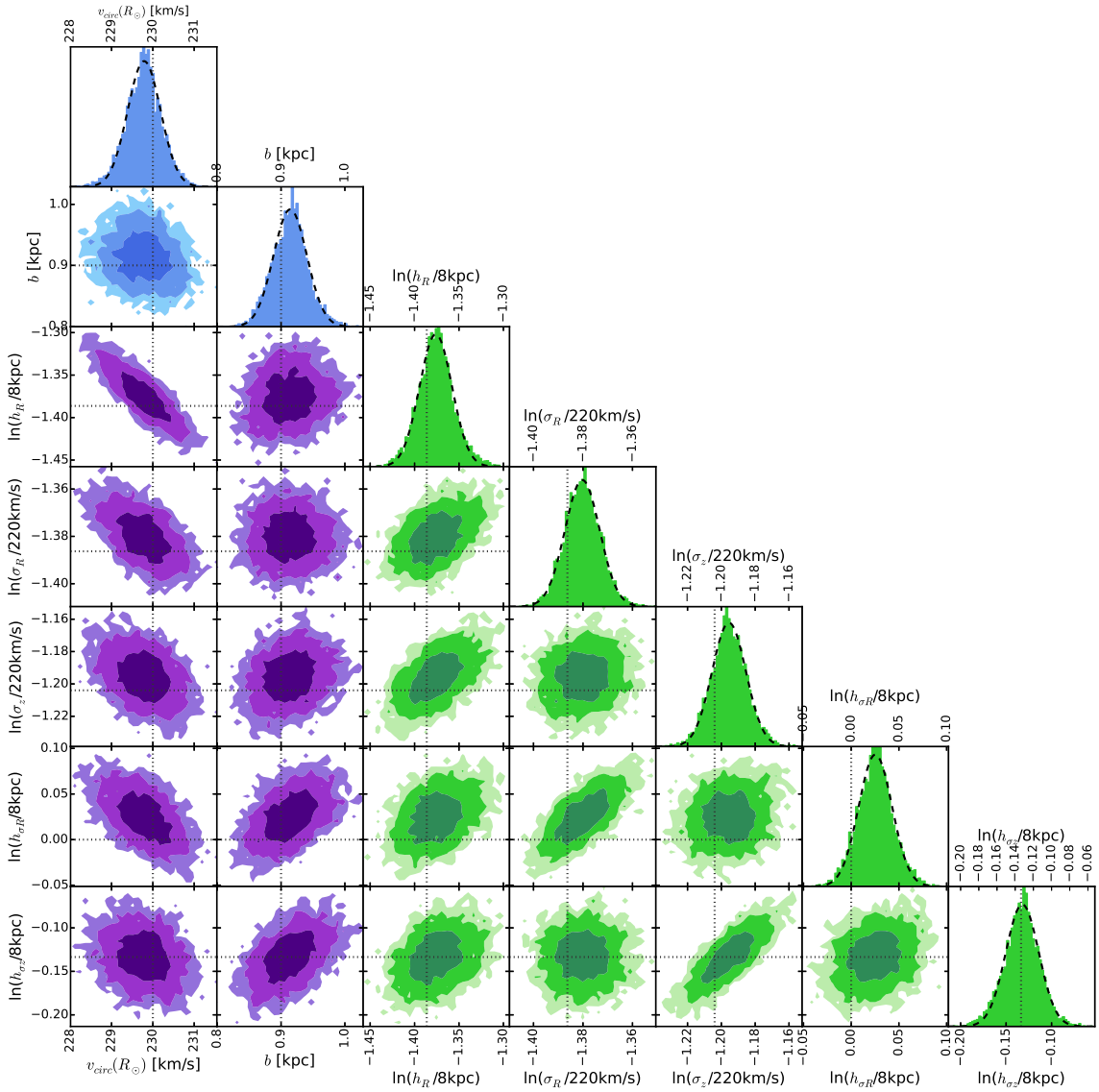

Figure 4. The likelihood in Eq. (11) in the parameter space $p_M = \{p_\Phi, \ln(p_{\rm DF})\}$ for one example mock data set created according to Test 10 in Table 3. Blue indicates the likelihood for the potential parameters, green the qDF parameters. The true parameters are marked by dotted lines. The dark, medium and bright contours in the 2D distributions represent 1, 2 and 3 sigma confidence regions, 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. The likelihood here was sampled using MCMC (with flat priors in p_Φ and $\ln(p_{\rm DF})$ to turn the likelihood 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 likelihood approaches the shape of a multi-variate Gaussian, as expected from the central limit theorem.

with different maximum radius, for two "Iso-Pot"s and two different *MAP*s .

The results of this second suite are shown in Fig. 6 for one potential and one qDF parameter ($\ln(h_{\sigma,z})$) and demonstrate the effect of the *size and extent* of the survey volume together with the *choice of the MAP and model parameters*.

The left panels in Fig. 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. Fig. 6 demonstrates that, given a choice of qDF and potential parameters, a larger volume always results in tighter constraints. There is no obvious trend that a hotter or cooler *MAP* will always give better results; it depends on the survey volume and the model parameter in question. In Fig. 7 the wedges all have the same size and all give results of similar precision. Minor differences, e.g. with the "Iso-Pot" potential being less constraint 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 Fig. 7 implies that for these cases volumes offsets in the radial or vertical direction have at most modest impact - at least at given sample size.

While we believe the argument for significant radial and

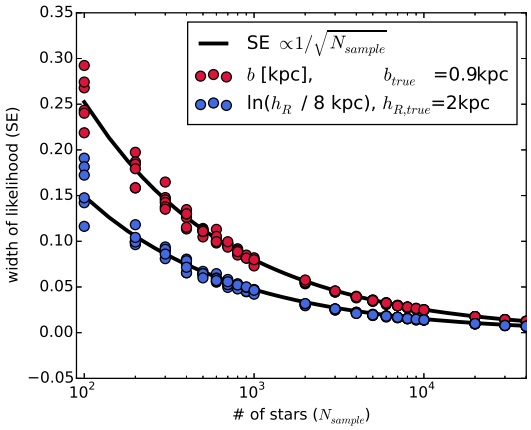

Figure 5. The width of the likelihood for two fit parameters found from analyses of 132 mock data sets vs. the number of stars in each data set. The mock data was created in the "Iso-Pot" potential and all model parameters are given as Test ② in Table **??**. The likelihood in Eq. 11 was evaluated on a grid and a Gaussian was fitted to the marginalized likelihoods 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 likelihood behaves as expected and scales with $1/\sqrt{N_{\text{sample}}}$ as predicted by the central limit theorem.

vertical extent is generic, we have not done a full exploration of all combinations of p_M and volumina.

3.3. *What if our assumptions on the (in-)completeness of the data set are incorrect?*

The selection function of a survey could be described by a spatial survey volume and a completeness function, which determines the fraction of stars observed at a given location within the Galaxy with a given brightness, metallicity etc (see §2.3). The completeness function depends on the characteristics and mode of the survey, can be very complex and is therefore sometimes not perfectly known. We investigate how much an imperfect knowledge of the selection function can affect the recovery of the potential. We model this by creating mock data with varying incompleteness, while assuming constant completeness in the analysis. The mock data comes from a sphere around the sun and an incompleteness function that drops linearly with distance r from the sun (see ⑤, Example 1, in Table **??** and Fig. **??**).
This could be understood as a model for the important effect of stars being less likely to be observed the further away they are. We demonstrate that the potential recovery with *RoadMapping* is very robust against somewhat wrong assumptions about the (in-)completeness of the data (see fig. **??**). A lot of information about the potential comes from the rotation curve measurements in the plane, which is not affected by applying an incompleteness function. In Appendix §A.1 we also show that the robustness is somewhat less striking but still given for small misjudgments of the incompleteness in vertical direction, parallel to the disk plane (fig. **??** and **??**). This could model the effect of wrong corrections for dust obscuration in the plane. We also investigate in Appendix §A.1 if indeed most of the information is stored in the rotation curve. For this we use the same mock data sets as in fig. **??** and **??**, but this time were not including the tangential velocities in the modelling, rather 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 Fig. 18).

3.4. *Effect of measurement errors on recovery of potential?*

Collection of possible tests and plots—

- *Plot 1:* The plot I had on the poster, which shows the number of MC samples needed for given maximum error. However, we still haven't tested, if this plot depends on: * hotness of stars * number of stars
- *Plot 2:* Some plot that shows, that our approximation of ignoring distance errors works. Any ideas?
- *Test 1:* One selection function, one population, vary the size of the proper motion error (don't forget to adapt the number of MC samples needed) *Plot 3:* (width of pdf) vs. (maximum velocity error / temperature parameter)

3.5. *The Impact of Deviations of the Data from the Idealized qDF*

Our modelling approach assumes that each *MAP* follows a quasi-isothermal distribution function, qDF. In this Section we explore what happens if this idealization does not hold. This could be, because even in the limit of perfectly measured abundances, MAPs do not follow a qDF. Or, even if they did do that, because the finite abundance errors effectively mix different MAPs. We investigate both these issues by creating mock data sets (Fig. **??**) that are drawn from two distinct qDFs of different temperature, and analyze the composite mock data set by fitting a single qDF to it. These results are illustrated in Figs. 11 and 12. Following the observational evidence, MAPs with cooler qDFs also have longer tracer scale lengths. In the first set of test, we choose qDFs of widely different temperatures and vary their relative fraction (dubbed "examples 1a/b", Fig. 11) ; in the second set of tests ("examples 2a/b", Fig. 12), we always mix mock data points from two different qDFs in equal proportion, but vary by how much the qDF's temperatures differ.
The first set of tests mimicks a DF that has wider wings or a sharper core in velocity space than a qDF (Fig. **??**). The second test could be understood by mixing neighbouring *MAP*s due to too large bin sizes or abundance measurement errors.

It is worth considering separately the impact of the DF deviations on the recovery of the potential and of the qDF parameters.
We find from example 1 that the potential parameters can be better and more robustly recovered, if a mock-data *MAP* is polluted by a modest fraction ($\lesssim 30\%$) of stars drawn from a cooler qDF with a longer scale length, as opposed to the same pollution of stars drawn from a hotter qDF with a shorter scale length.

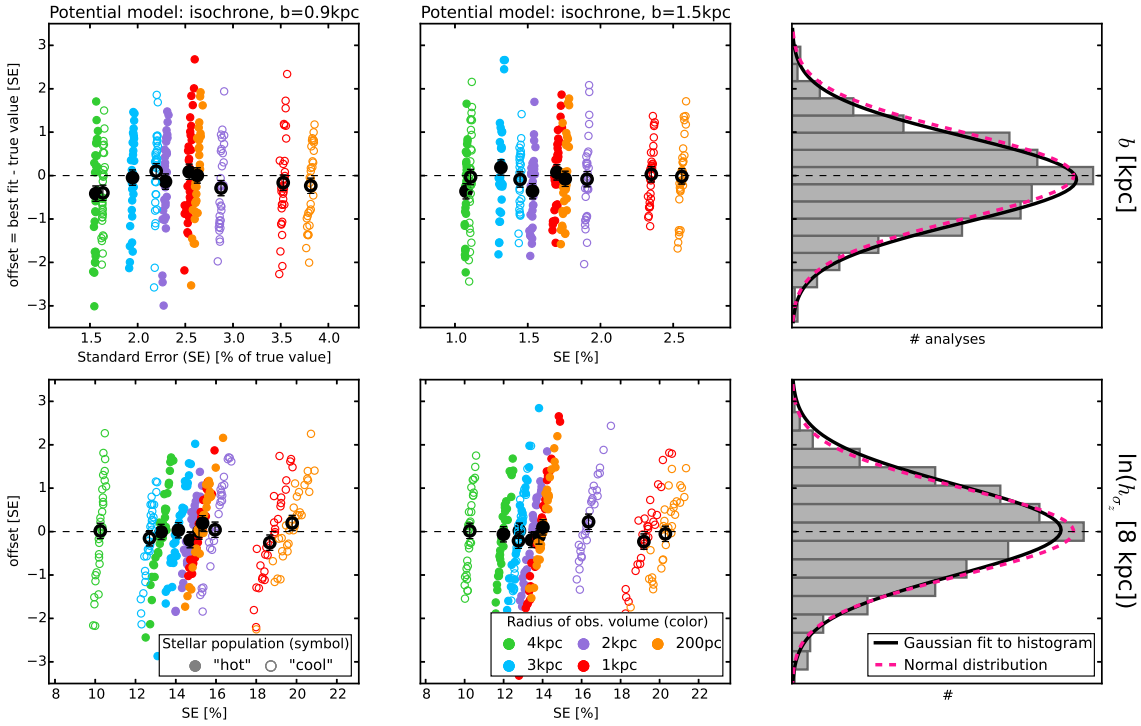

Figure 6. (Un-)bias of the parameter estimate: According to the central limit theorem the likelihood will follow a Gaussian distribution for a large number of stars. From this follows that also for a large number of data sets the corresponding best fit values for the model parameters have to follow a Gaussian distribution, centered on the true model parameters. That our method satisfies this and is therefore an unbiased estimator [TO DO: can I say that????] is demonstrated here. We create 640 mock data sets. They come from two different "Iso-Pot" potentials (first and second column), two different stellar populations ("hot" *MAP* (solid symbols) and "cool" *MAP* (open symbols)) and five spherical observation volumes of different sizes (color coded, see legend). All model parameters are summarized in Table 3 as test ③. We determine the best fit value and the standard error (SE) for each fit parameter by fitting a Gaussian to the marginalized likelihood. The offset is the difference between the best fit and the true value of each model parameter. In the first two columns the offset in units of the SE is plotted vs. the SE in % of the true model parameter. The first row shows the results for the isochrone scale length b and the second row the qDF parameter h_{σ_z}, which corresponds to the scale length of the vertical velocity distribution. [TO DO: rename isochrone potential in title to "Iso-Pot".]

Figure 6. (Continued.) The last column finally displays a histogram of the 640 offsets (in units of the corresponding SE). The black solid line is a Gaussian fit to a histogram. The dashed pink line is a normal distribution $\mathcal{N}(0,1)$. As they agree very well, 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???]

When considering the case of a 50/50 mix of contributions from different qDFs , there is a systematic, but only small, error in recovering the potential parameters, monotonically increasing with the qDF parameter difference (example 2); in particular for fractional differences in the qDF parameters of $\lesssim 20\%$ the systematics are insignificant even for samples sizes of 20,000, as used in the mock data.

The recovery of the effective qDF parameters, in light of non qDF mock data is quite intuitive: the effective qDF temperature lies between the two temperatures from which the mixed DF of the mock data was drawn; in all cases the scale length of the velocity dispersion fall-off, $h_{\sigma R}$ and $h_{\sigma,z}$, is shorter, because the stars drawn form the hotter qDF dominate at small radii, while stars form the cooler qDF (with its longer tracer scae length) dominate at large radii. The recovered tracer scale lengths, h_R vary smoothly between the input values of the two qDFs that enetered the mix of mock data, with again the impact of contamination by a hotter qDF (with its shorter scale length in this case) being more important.

Further notes:— [TO DO] Mention, that vcirc is more tightly constraint with cool MAP.

3.6. *What if our assumed potential model differs from the real potential?*

We inspect if we can give constraints on the true potential, if our beliefs about the overall parametric form of the MW's potential are slightly wrong. We ignore deviations from axisymmetry and focus on a test case where the mock data was drawn from one axisymmetric potential ("MW14-Pot") and is then analysed using another axisymmetric potential family ("KKS-Pot"), that does *not* incorporate the true potential (compare the second and fourth panel in Fig. 1). In the analysis we assume the circular velocity at the sun to be fixed and known and only fit the parametric potential form. The results are shown in Fig. 13.

The reference potential parameters of the "KKS-Pot" in Table 1 were found by adjusting the 2-component Kuzmin-Kutuzov Stäckel potential by Batsleer & Dejonghe (1994) such that it generates radial and vertical force profiles similar to the "MW14-Pot" from Bovy (2015) (dotted gray lines in Fig. 13). The analysis results from *RoadMapping* shown in Fig. 13, red for a "hot" mock data *MAP* and blue for a "cool" *MAP* ,

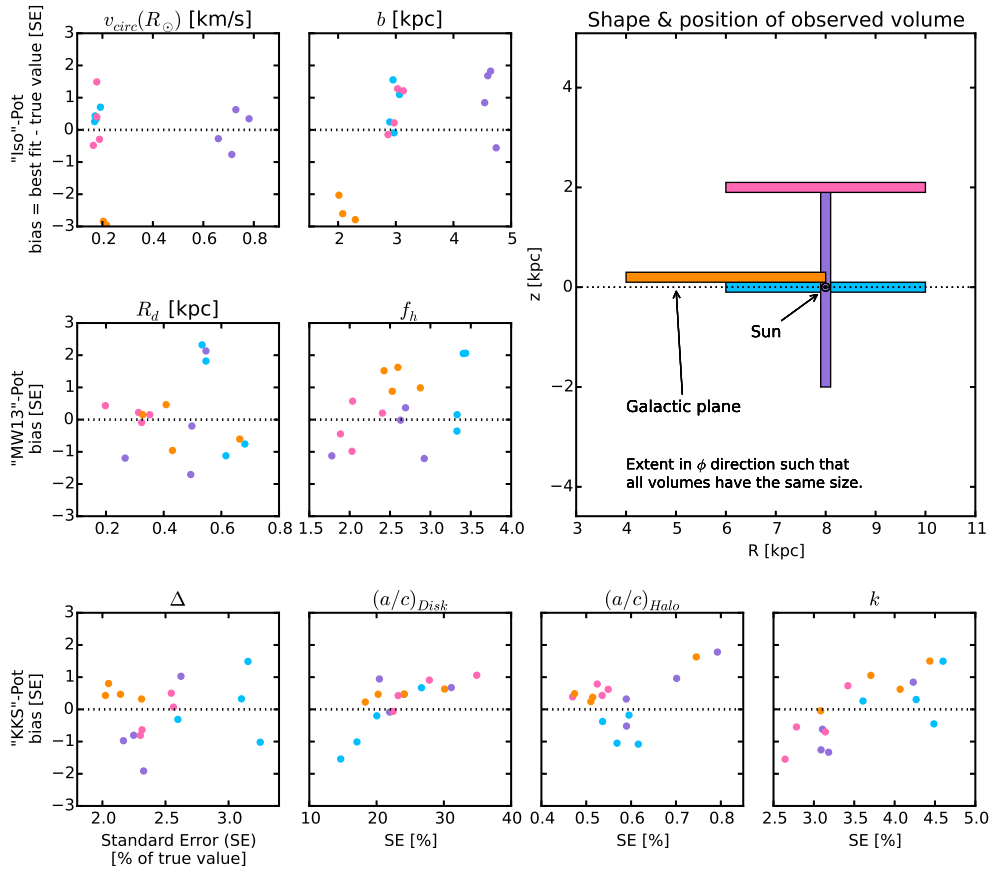

Figure 7. Bias vs. standard error in recovering the potential parameters for mock data stars drawn four different test observation volumes within the Galaxy (illustrated in the upper right panel) and three different potentials ("Iso-Pot", "MW13-Pot" and "KKS-Pot" from Table 1, top to bottom row). Standard error and offset were determined as in fig. 6. Per volume and potential we analyse four different mock data realisations; all model parameters are shown as Test ④ in Table 3. 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 the other volumes. [TO DO: MW-Pot and KKS-Pot analyses suffer from too low accuracy in action calculation (with StaeckelGrid). Used the StaeckelGrid for BOTH mock data and analysis, but the mock data distribution would actually not look exactly like the desired qDF distribution, i.e. this plot basically is created with a messed up DF. Don't know how higher accuracy would change the plot. The orange Iso-Pot analysis suffers from too small integration range in vT. More coding required, before redoing this.]

give an comparable good or even better agreement with the true potential than the (by-eye) fit directly to the potential: especially the force contours, to which the orbits are sensitive, and the rotation curve are very tightly constrained and reproduce the true potential even outside of the observed volume of the mock tracers. This demonstrates that *RoadMapping* provides an optimal best fit potential within the capabilities of the parametric potential model.

The density contours are less tightly constrained than the forces, but we still capture the essentials: The "hot" *MAP* from Table 2 constrains the halo; especially at smaller radii it is equally good or better than the "cool" *MAP* . The "cool" *MAP* gives tighter constraints on the halo in the outer region and recovers the disk better than the "hot" *MAP* . This is in concordance with expectations as the "cool" *MAP* has a longer tracer scale length and is more confined to the disk than the "hot" *MAP* and therefore also probes the Galaxy in these regions better. Overall the best fit disk is less dens in the midplane than the true disk.

Fig. 14 compares the true qDF parameters with the best fit parameters. While tracer scale length and radial velocity dispersion profile are very well recovered, we misjudge the radial profile of the vertical velocity dispersion: $\sigma_{0,z}$ and $h_{\sigma,z}$ are both underestimated, which leads to a steeper profile and a lower dispersion around the sun. This is a direct result of the surface density underestimation in the midplane, the corresponding lower vertical forces around the sun (see also Fig. 13) and therefore lower vertical actions. Fig. **??** demonstrates that even though the misjudgment of the potential lead to biases in the qDF parameters, the model is still a very good fit to the data.

[TO DO: Plot showing the true and best fit distribution]

[TO DO: Plot showing the radial profile of the vertical dispersion. ???]

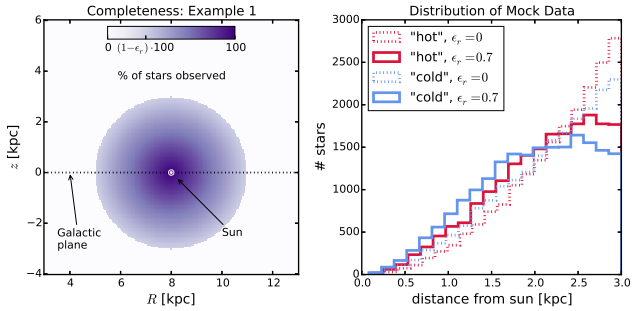

Figure 8. Selection function and mock data distribution for investigating radial incompleteness of the data. All model parameters are summarized as test ⑤, Example 1, in Table **??**. The survey volume is a sphere around the sun and the percentage of observed stars is decreasing linearly with radius 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 *MAP*s ("hot" in red and "cool" in blue, see table 2) and with two different ϵ_r, 0 and 0.7, are shown in the right panel for illustration purposes.

4. DISCUSSION AND SUMMARY

Recently implementations of action DF - based modelling of 6D data in the Galactic disk have been put forth, in part to lay the ground-work fo Gaia. [Binney,Sanders,Piffl,Bovy,Rix,McMillan??]. We present *RoadMapping* , an improved implementation of the dynamical modelling machinery by Bovy & Rix (2013), to recover the potential and orbit distribution function of stellar *MAP*s within the Galactic disk. In this work we investigated the capabilities, strengths and weaknesses of *RoadMapping* by testing its robustness against the breakdown of some of its assumptions - for well defined, isolated test cases using mock data. Overall the method works very well and reliable, also if there are small deviations of the model assumptions from the real world galaxy.

4.1. *Improved Computational Speed for Application to Larger Data Sets*

RoadMapping applies a full likelihood analysis and is statistically well-behaved. It allows for a straightforward implementation of different potential model families and a flexible number of free fit parameters in potential and qDF. It also accounts for selection effects by using full 3D selection functions (given some symmetries). *RoadMapping* is an asymptotically normal, un-biased estimator and the precision of parameter recovery increases by $1/\sqrt{N}$ with the number of stars.

Large data sets in the age of Gaia require more, and more accurate, likelihood evaluations for more flexible models. To be able to deal with these increased computational demands and explore larger parameter spaces, we sped up the code by combining a nested grid approach with MCMC and by faster action calculation using the Stäckel (Binney 2012) interpolation grid by Bovy (2015). Especially accurately determining the likelihood normalisation will be of crucial importance for large data sets. The nested-grid approach automatizes the search for the optimal normalisation integration ranges ("fiducial qDF") and start position for the MCMC walkers, which helps the MCMC to converge fast and to reduce biases due to insufficient accuracy. However, application of *RoadMapping* to millions of stars simultaneously with acceptable accuracy will still be a task for supercomputers and calls for even more improvements and speed-up in the fitting machinery.

4.2. *Modelling Sensitivity to Properties and Unaccounted Imperfections of the Data Set*

Choice of observation volume.— We found that the *position* of the survey volume maters little, in the sense that there are no regions in the Galaxy that contain intrinsically stars on manifestly more diagnostic orbits than others. Closer to the disk and at smaller Galactocentric radii it is only the increased number of stars that will lead to tighter constraints. Concerning the *shape* of the survey volume, a large radial *and* vertical coverage is best. In the axisymmetric case phi coverage doesn't matter. Making a volume cut for stars, that lie around $R_\odot$ but at larger *phi*, could therefore improve the results, if their measurements are very uncertain.
*MAP*s of different scale length and temperature probe different regions of the Galaxy (Bovy & Rix 2013). But there is no easy rule of thumb for which survey volume and stellar population which potential and DF parameter is constrained best.

Selection function misjudgment.— Surprisingly *RoadMapping* seems to be very robust against misjudgments in the selection function of the data. The reason for this robustness could be, that missing stars in the data set do not affect the connection between a star's velocity and position, which is given by the potential. A lot of information about the potential profile is stored in the rotation curve - but even when not including measurements of tangential velocities in the analysis, small misjudgments of the incompleteness do not affect the potential recovery.
That we reproduce the qDF equally well, could be due to the symmetry of our assumed incompleteness profiles around the sun. We investigated a decrease in knowledge of the data completeness in distance from the sun and Galactic plane. Our test with the radial incompleteness profile could be understood as a decreasing detection rate due to the lower apparent brightness of stars at larger distances. The test with the planar incompleteness profile could mimic a misjudgment of the dust obscuration within the Galactic plane. Both effects would show the same symmetries as tested in this work. This result is encouraging for future studies, but nevertheless surprising as it was previous believed that knowing the (spatial) selection function very precisely is of large importance for dynamical modelling (Rix & Bovy 2013).

Measurement errors.— Properly convolving the likelihood with measurement errors is computational very expensive. By ignoring positional errors and only including distance errors as part of the velocity error, we can drastically reduce the computational costs. [TO DO - No definite result after this disclaimer:] For stars within 3 kpc from the sun this approximation works well for distance errors of $\sim 10\%$ or smaller. The number of MC samples needed for the error convolution using MC integration scales by $N_{\rm error} \propto (\sigma_{v,\rm max})^2$ with the maximum velocity error at the edge of the sample. If we misjudge the size

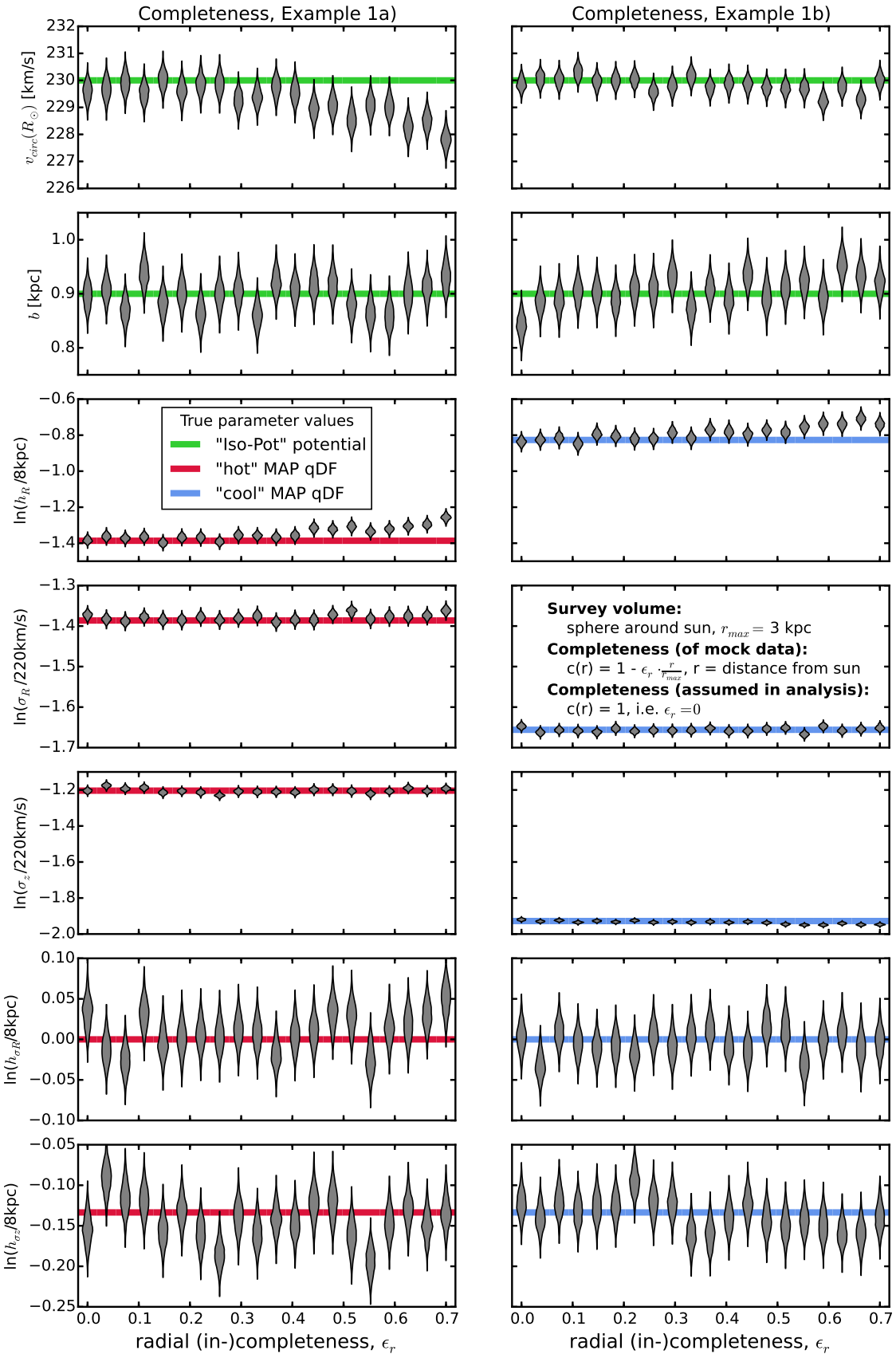

Figure 9. Influence of wrong assumptions about the radial incompleteness of the data on the parameter recovery with *RoadMapping* . Each mock data set was created having different incompleteness parameters ϵ_r (shown on the x-axis and illustrated in Fig. 8) and the model parameters are given as test ⑤, 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.

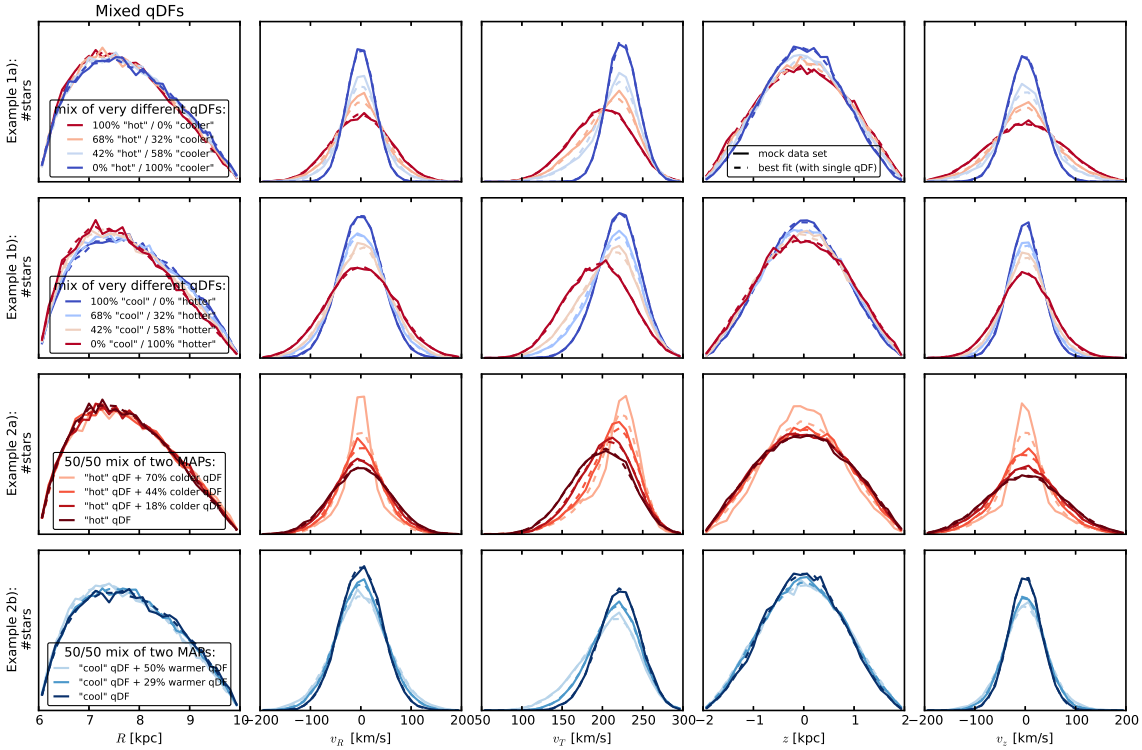

Figure 10. Distribution of mock data, created by mixing stars drawn from two different qDFs (solid lines), and the distribution predicted by the best fit of a single qDF and potential to the data (dashed lines). The model parameters to create the data are given in Table **??** as test (7), and the qDF parameters referenced in the figure's legend in Table 2. *Example 1:* Distribution of mock data drawn from a superposition of two very different (but fixed) qDFs at varying mixing rates. *Example 2:* Mock data distribution of two *MAP*s that were mixed at a fixed rate of 50%/50%, but the difference of the qDF parameters of one *MAP* was varied with respect to the qDF parameters of the other *MAP* by X% (see Table 2). The data sets are color coded in the same way as the corresponding analyses in Fig. 11 and 12. This figure demonstrates how mixing two qDFs can be used as a test case for changing the shape of the DF to not follow a pure qDF anymore, e.g. by adding wings or slightly changing the radial density profile. A second set of mock data was drawn from a single qDF and the best fit parameters found in Fig. 11 and 12 and overplotted as dashed lines. Especially for the most extreme deviations between mock data and best fit distribution it becomes obvious that a single qDF is a bad assumption for the stars' "true" DF. [TO DO: make larger distance between the top and bottom panels.] [TO DO: Write "mixture of 2 qDFs" in legend]

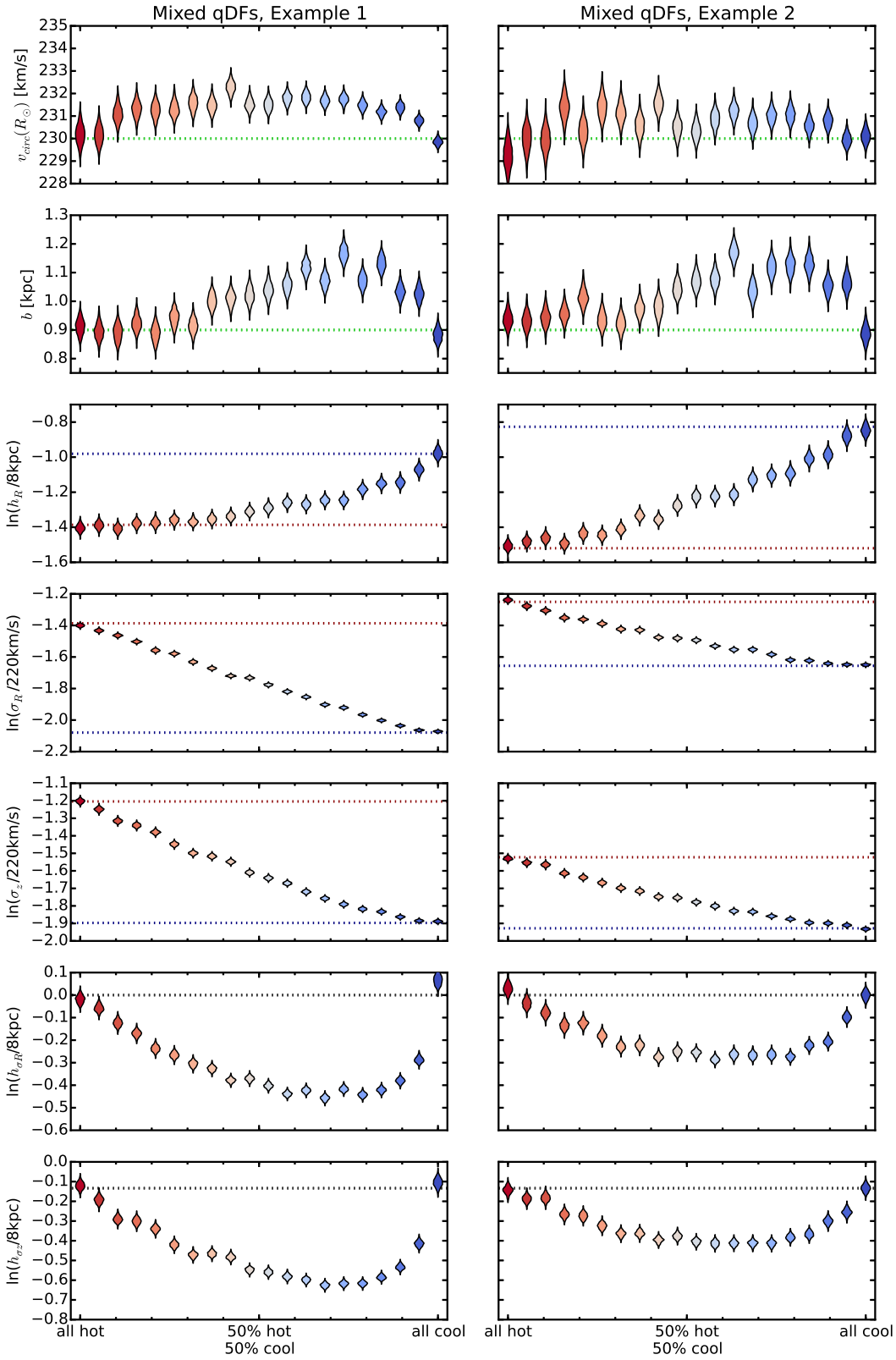

Figure 11. [TO DO: Update caption.] The dependence of the parameter recovery on degree of pollution and 'hotness' of the stellar population. To model the pollution of a hot stellar population by stars coming from a cool population and vice versa, we mix varying amounts of stars from two very different populations, as indicated on the x-axis. The composite mock data set is then fit with one single qDF. The violines represent the marginalized likelihoods found from the MCMC analysis. The mock data sets are shown in fig. **??**, in the same colors as the violins here. All mock data sets come from the same potential ("Iso-Pot") and selection function (sphere with $r_{\rm max} = 2$ kpc). The true potential parameters are indicated by green dotted lines. Example 1 (Example 2) in the left (right) panels mixes the "hot" ("cool") *MAP* with the "cooler" ("hotter") *MAP* in table 2. True parameters of the hotter (colder) of the two populations are shown as red (blue) dotted lines. We find, that a hot population is much less affected by pollution with stars from a cooler population than vice versa. [TO DO: This was done using the current qDF to set the fitting range. Nvelocity=24 and Nsigma=5 is high enough (though not perfect). Maybe redo with fiducial qDF to be consistent with MixDiff test. ???] [TO DO: Rename example 1 & 2 to example 1a/1b and example 3 & 4 to example 2a/2b] [TO DO: Legend]

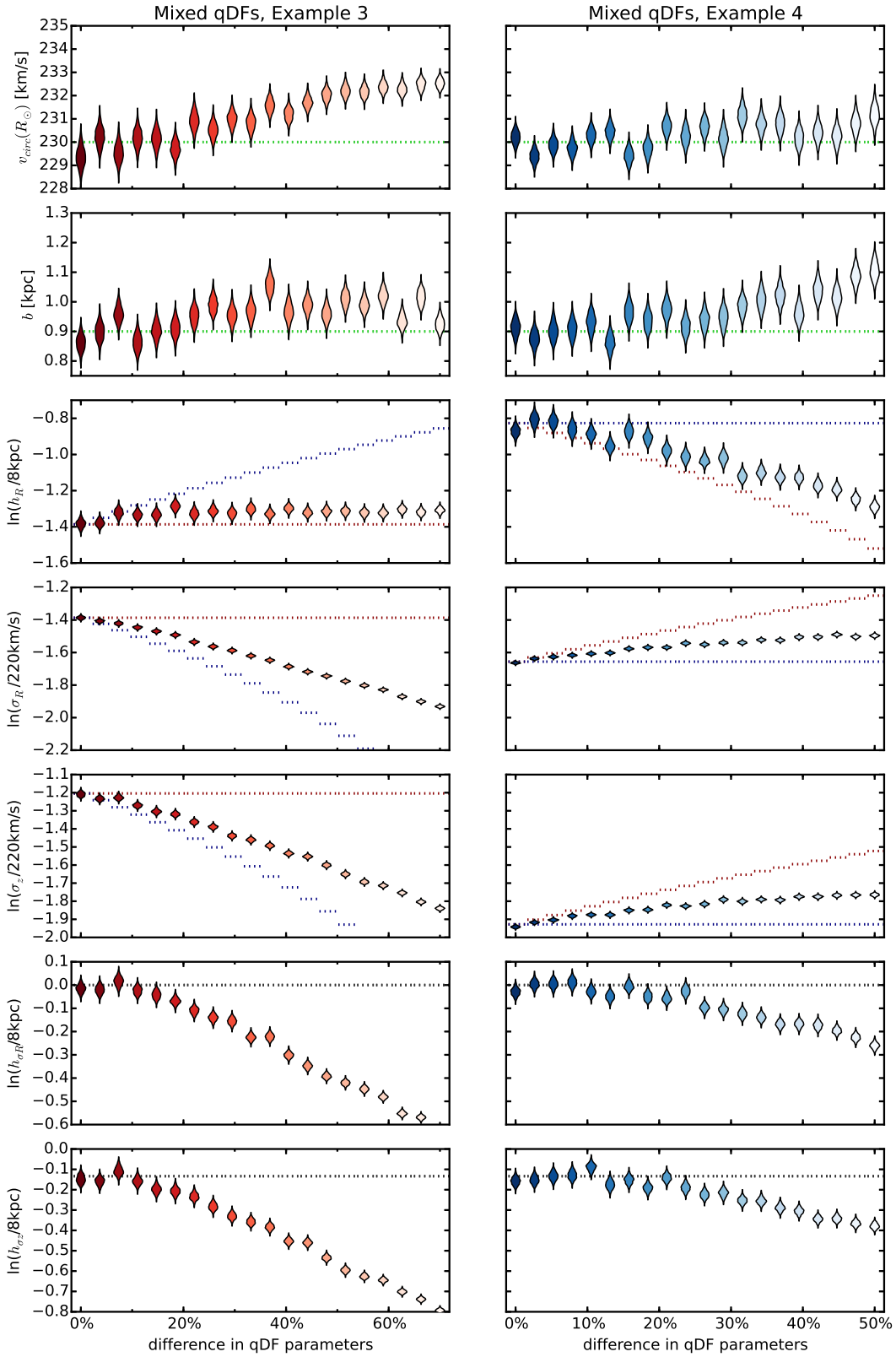

Figure 12. [TO DO: Update caption.] The dependence of the parameter recovery on the difference in qDF parameters of the 50%/50% mixture of two stellar populations and their 'hotness'. Each mock data set in Example 3 (Example 4) consists of 20,000 stars, half of them drawn from the "hot" ("cool") qDF in table 2, and the other half drawn from a "colder" ("warmer") population that has X% smaller (larger) σ_R and σ_z and X% larger (smaller) h_R. The difference X in these qDF parameters is indicated on the x-axis, and the true parameters of the two qDFs are indicated by the dotted red and blue lines. Each composite mock data set is fitted by a single qDF and the marginalized MCMC likelihoods for the best fit parameters are shown as violines in the third (fourth) column of panels. The mock data was created within the same potential ("Iso-Pot") and selection function (sphere with $r_{\rm max}$ = 2 kpc). The true potential parameters are indicated by green dotted lines. The data sets are shown in figure **??**, where the histograms have the same colors as the corresponding best fit violines here. By mixing *MAP*s with varying difference in their qDF parameters, we model the effect of bin size in the [Fe/H]-[α/Fe] plane when sorting stars into different *MAP*s : The smaller the bin size, the smaller the difference in qDF parameters of stars in the same bin. We find that the bin sizes should be chosen such that the difference in qDF parameters between neighbouring *MAP*s is less than 20%. [TO DO: Maybe different/same x-axis???] [TO DO: This was done using the current qDF to set the fitting range. Nvelocity=24 and Nsigma=5 is not high enough for the largest differences, i.e. grid search and MCMC converge to different values. Redo with fiducial qDF.] [TO DO: Add in plot a label, that it is a 50%/50% mix of a hot and a cold population.??] [TO DO: Rename example 1 & 2 to example 1a/1b and example 3 & 4 to example 2a/2b] [TO DO: Write in plot, that there is a 50/50 mix of cool and hot] [TO DO: Adapt colors to fit the residuals plot] [TO DO: Write free parameter X on x-axis.] [TO DO: Legend])

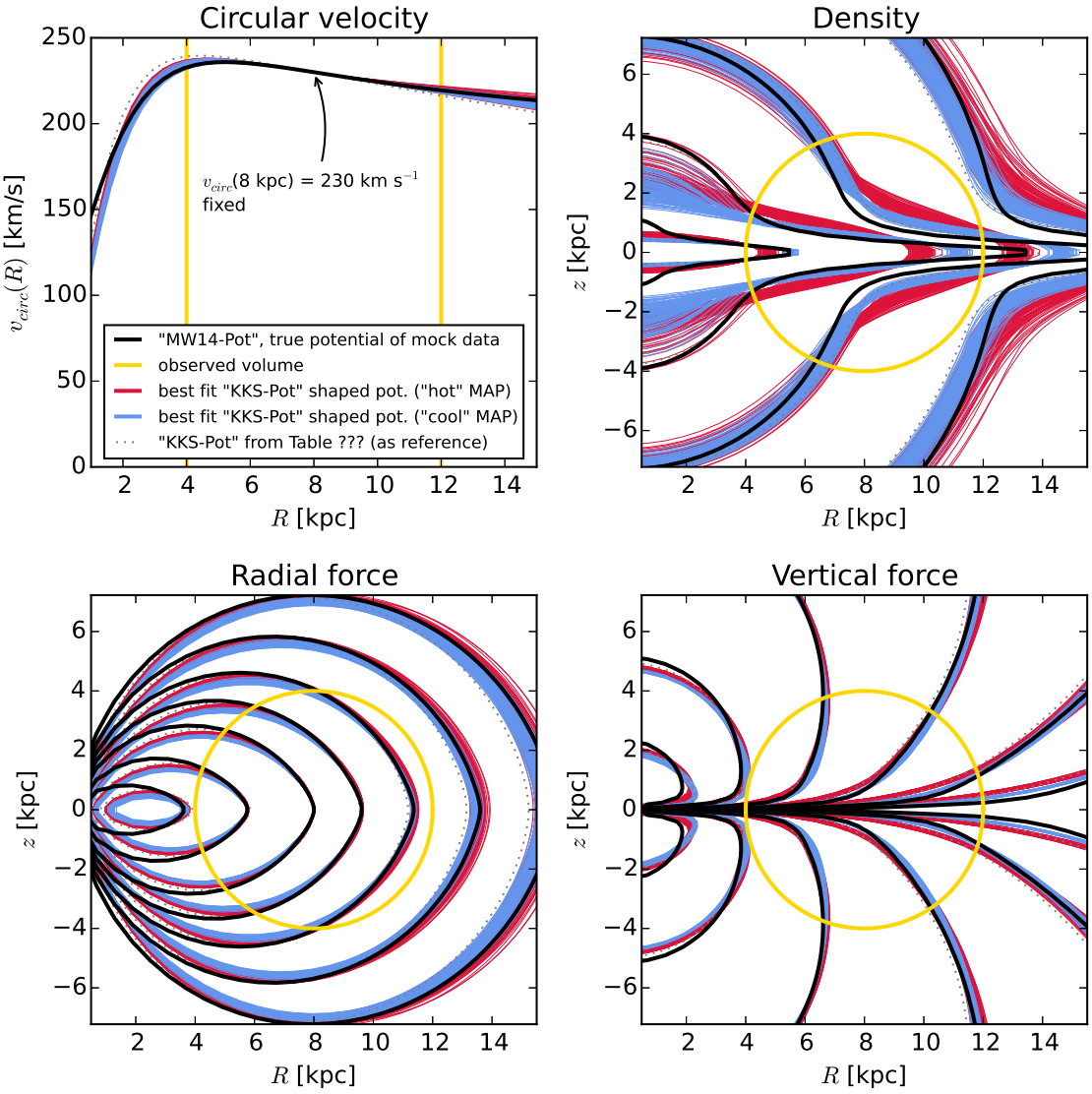

Figure 13. Recovery of the gravitational potential if the assumed potential model ("KKS-Pot" with fixed $v_{\rm circ}(R_\odot)$) and the true potential of the (mock) stars ("MW14-Pot" in Table 1) is slightly different. We show the circular velocity curve, as well as contours of equal density, radial and vertical force in the R-z-plane, and compare the true potential with 50 [TO DO: CHECK] sample potentials drawn from the posterior distribution function found with the MCMC for a "hot" (red) and a "cool" *MAP* (blue). All model parameters are given as Test ⑧ in Table 3. [TO DO: Do more analyses???]

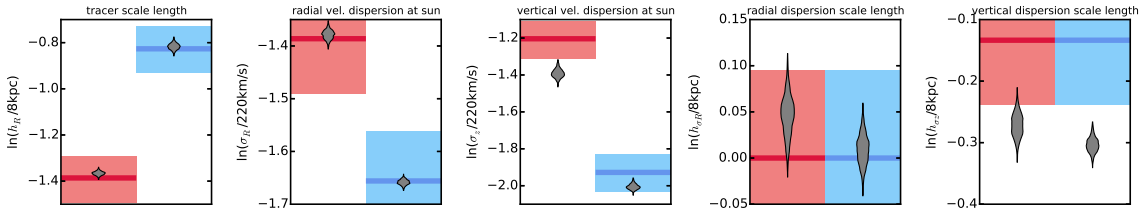

Figure 14. Recovery of the qDF parameters for the case where the true and assumed potential deviate from each other (Test ⑧ in Table 3). The thick red (blue) lines represent the true qDF parameters of the "hot" ("cool") qDF in Table 2 used to create the mock data, surrounded by a 10% error region. The grey violins are the marginalized likelihoods for the qDF parameters found simultaneously with the potential constraints shown in Fig. 13.

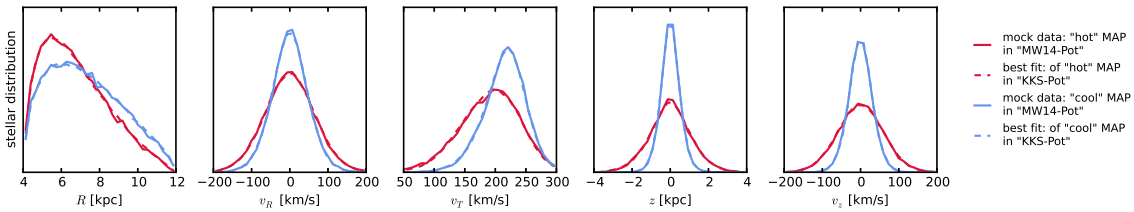

Figure 15. Comparison of the distribution of mock data in configuration space created in the "MW14-Pot" potential (solid lines) with a "hot" (red) and "cool" (blue) *MAP* (Test ⑧ in Table **??**), and the best fit distribution using a "KKS-Pot" potential (dashed lines). The best fit potentials are shown in Fig. 13 and the corresponding best fit qDF parameters in Fig. 14. The best fit

of the true measurement errors, we only can reproduce the true model parameters, as long as the velocity errors are smaller than the intrinsic velocity dispersion of the data set.

4.3. *Data Deviations from the Modelling Assumptions about the Distribution Function and the Potential*

Deviations from the qDF Assumption.— Our modelling is founded on the assumption, that we can identify *a priori* sub-components of the Galactic disk that follow a qDF (e.g. by considering *MAP*s). There are two reasons why any chosen sub-sample of star (here a *MAP*) may not be well described by any qDF. Either, because nature is more complex, or because even if perfect *MAP*s would be well described by qDFs finite abundance errors would mix *MAP*s . We have considered both cases.

In Example 1 in §??? we investigated how well we can recover the potential, if this assumption was not perfectly satisfied, i.e. the *MAP*s true DF does not perfectly follow a qDF. We considered two cases: a) a hot DF, that has less stars at small radii and more stars with low velocities than predicted by the qDF (reddish data sets in Fig. **??**), or b) a cool DF that has broader velocity dispersion wings and less stars at large radii than predicted by the qDF (bluish data sets). We find that case a) would give more reliable results for the potential parameter recovery. If we assumed that the distribution of stars from one *MAP* is caused by radial migration away from the initial location of star formation, it would more likely that the qDF overestimates the true number of stars at smaller radii than underestimating it at larger radii. [TO DO: Is this actually a sensible argument???]
Following this, focusing the analysis especially on hotter *MAP*s could be an advisable way to go in the future, if there is doubt that the stars truly follow the qDF.

Another critical point is the binning of stars into *MAP*s depending on their metallicity and α abundances. Example 2 in §??? could be understood as a model scenario for decreasing bin sizes in the metallicity-α plane when sorting stars in different *MAP*s , assuming that there is a smooth variation of qDF within the metallicity-α plane and each *MAP* indeed follows a qDF. We find that, in the case of 20,000 stars in each bin, differences of 20% in the qDF parameters of two neighbouring bins can still give quite good constraints on the potential parameters.
We compare this with the relative difference in the qDF parameters in the bins in Fig. 6 of Bovy & Rix (2013), which have sizes of $[Fe/H] = 0.1$ dex and $\Delta[\alpha/Fe] = 0.05$ dex. It seems that these bin sizes are large enough to make sure that $\sigma_{R,0}$ and $\sigma_{z,0}$ of neighbouring *MAP*s do not differ more than 20%. Fig. 11 and 12 suggest that especially the tracer scale length h_R needs to be recovered to get the potential right. For this parameter however the bin sizes in fig. 6 of Bovy & Rix (2013) might not yet be small enough to ensure no more than 20% of difference in neighbouring h_R. This is especially the case in the low-α ($[\alpha/Fe] \lesssim 0.2$), intermediate-metallicity ($[Fe/H] \sim -0.5$) *MAP*s . The above is valid for 20,000 stars per *MAP* . In case there are less than 20,000 stars in each bin the constraints are less tight and due to Poisson noise one could also allow larger differences in neighbouring *MAP*s while still getting reliable results. [TO DO: Discuss Binney & Sanders doubts about binning.]

Gravitational Potential beyond the Parameterized Functions Considered.— In the long run *RoadMapping* should incorporate a family of gravitational potential models that can reproduce the essential features of the MW's true mass distribution. While our fundamental assumption of the Galaxy's axisymmetry is at odds with the obvious existence of non-axisymmetries in the MW, we will not dive into investigating this implications in the scope of this paper. Instead we test how a misjudgment of the parametric potential form affects the recovery by fitting a potential of Stäckel form (Batsleer & Dejonghe 1994) to mock data from a different potential family with halo, bulge and exponential disk. The recovery is quite successful and we get the best fit within the limits of the model. However, even a strongly flattened Stäckel potential component has difficulties to recover the very flattened mass distribution of an exponential disk. This will lead to underestimation of the vertical velocity dispersion at the sun. aA the qDF parameter $\sigma_{z,0}$ corresponds to the physical vertical velocity dispersion at the sun, a comparison with direct measurements could be a valuable cross-checking reference. [TO DO: This might not be true. For isochrone and Staeckel potential I get this behaviour, but not for the MW14-Pot. Might be, because it's not separable.] In case of as many as 20,000 stars we should therefore already be able to distinguish between different potential models.
The advantage of using a Stäckel potential with *RoadMapping* is firstly the exact and fast action calculation via the numerical evaluation of a single integral, and secondly that the potential has a simple analytic form, which greatly speeds up calculations of forces and frequencies (as compared to potentials in which only the density has an easy description like the exponential disk). A superposition of several simple Kuzmin-Kutuzov Stäckel components can successfully produce MW-like rotation curves (see Batsleer & Dejonghe (1994), **?** and Fig. ???) and one could think of adding even more components for more flexibility, e.g. a small roundish component for the bulge. The potential model used by Bovy & Rix (2013) had only two free parameters (disk scale lentgh and halo contribution to $v_{\rm circ}(R_\odot)$. To circumvent the obvious disadvantage of this being at all not flexible enough, they fitted the potential separately for each *MAP* and recovered the mass distribution for each *MAP* only at that radius for which it was best constrained - assuming that *MAP*s of different scale length would probe different regions of the Galaxy best. Based on our results in Fig. 13 this seems to be indeed a sensible approach [TO DO: Check that this is indeed the case - it is not clear to me from the plot. ???].
We suggest that combining the flexibility and computational advantages of a superposition of several Stäckel potential components with probing the potential in different regions with different *MAP*s as done by Bovy & Rix (2013), could be a promising approach to get the best possible constraints on the MW's potential.

4.4. *Different Modelling Approaches using Action-based Distribution Functions*

We have focussed for the time being on *MAP*s for a number of reasons: First, they seem to permit simple DFs (Bovy et al. 2012b,c,d), i.e. approximately qDFs

(Ting et al. 2013). Second, all stars, e.g. those from different *MAP*s , must orbit in the same potential. Therefore each *MAP* will and can yield quite different DF parameters; but each *MAP* will also provide a (statistically) independent estimate of the potential parameters. At the same time – if all is well – those potential parameters, derived from different *MAP*s , should be mutually consistent. In some sense, this approach focusses on constraining the potential, treating the DF parameters as nuisance parameters.

The main drawback is that we have many astophysical reasons that the DF properties (for reasons of galaxy evolution and chemical evolution) are astrophysically linked between different *MAP*s . Ultimately, the goal is to do a fully consistent chemodynamical model that simultaneously fits the potential and DF($\boldsymbol{J}$, [X/H]) simultaneously (where [X/Fe] denotes the full abundance space) with a full likelihood analysis.
This has not yet been attempted.

Since the first application of *RoadMapping* by Bovy & Rix (2013) there have been two similar efforts to constrain the Galactic potential and/or orbit distribution using action-based distribution functions:

Piffl et al. (2014) fitted both potential and a $f(\boldsymbol{J})$ to giant stars from the RAVE survey (Steinmetz et al. 2006) and the vertical stellar number density profiles in the disk by Jurić et al. (2008). They did not include any chemical abundances in the modelling. Instead, they used a superposition of action-based DFs to describe the overall stellar distribution at once: a superposition of qDFs [TO DO: CHECK] for cohorts in the thin disk, a single qDF [TO DO: CHECK] for the thick disk stars and an additional DF for the halo stars. Taking proper care of the selection function requires a full likelihood analysis and the calculation of the likelihood normalisation is computational expensive. Piffl et al. (2014) choose to circumvent this problem by directly fitting a) histograms of the three velocity components in eight spatial bins to the velocity distribution predicted by the DF and b) the vertical density profile predicted by the DF to the profiles by Jurić et al. (2008). The vertical force profile of their best fit mass model nicely agrees with the results from Bovy & Rix (2013) for $R > 6.6$ kpc. The disadvantage of their approach is, that by binning the stars spatially, a lot of stellar information is not used.

Sanders & Binney (2015) have focussed on understanding the abundance-dependence of the DF, relying on a fiducial potential. They developed extended distribution functions, i.e. functions of both actions and metallicity for a superposition of thin and thisk disk, each consisting of several cohorts described by qDFs, a DF for the halo, a functional form of the metallicity of the interstellar medium at the time of birth, and a simple prescription for radial migration. They applied a full likelihood analysis accounting for selection effects and found a best fit for the eDF in a fixed fiducial potential by Dehnen & Binney (1998) to the stellar phase-space and metallicity [TO DO: CHECK] data of the Geneva-Kopenhagen Survey (GS) (Nordström et al. 2004; Holmberg et al. 2009) and the stellar density curves by Gilmore & Reid (1983). Their best fit predicted the velocity distribution of SEGUE G dwarfs quite well, but had biases in the metallicity distribution, which they accounted to being a problem with the SEGUE metallcities.

4.5. *On the assumption of axisymmetry*

The key assumption in our modelling, as well as in the approaches by Piffl et al. (2014) and Sanders & Binney (2015) described above, is the overall axisymmetry of Galaxy's potential and DF. This has the convenient advantage, that actions are conserved in axisymmetric potentials and can be calculated straightforward via the "Stäckel Fudge" by **?** and/or a single integration (in the case of separable potentials). Of course the Galactic disk is in reality not axisymmetric and actions are not conserved: Spiral arms in the disk and the Galactic bar lead to angular momentum exchange and therefore radial migration of stars, i.e. the orbit on which the stars were born on are modified (Minchev et al. 2011; Kawata et al. 2014). Apart from these obvious non-axisymmetries in the Galaxy, the disk itself is not smooth, as there is lot of sub-structure, streams and moving groups within the disk. A famous example is the Arcturus moving group, for which Navarro et al. (2004) found that the stars might have their origin in a disrupted satellite.
Spiral arms, the stirring of the Galactic bar and the disk sub-structure will undoubtedly affect our modelling with *RoadMapping* . How strong this effect will be and if the modelling can still work, needs to be investigated in detail in future work, e.g. by trying to recover the potential in N-body simulations. But as actions are conserved under adiabatic changes of the potential Binney & Tremaine (2008), and the vertical action under radial migration [TO DO: REF], there is some hope, that the modelling could still work.

The ultimate goal would be to theoretically describe those non-axisymmetries and sub-structures in terms of DFs in action/angle-abundance space. We see the current axisymmetric modelling approaches as intermediate step to this goal: Dehnen (1998) described the disk's overall stellar distribution as a smooth background distribution superimposed with sub-structure; the axisymmetric DFs used and to be found by *RoadMapping* and similar approaches could be treated as this smooth background. Firstly, this smooth background DF could help to actually find and identify the disk substructure in action space (see e.g. Sellwood (2010); Klement et al. (2008) for similar approaches to find disk substructure, which then helps to find DF descriptions. Secondly, and because simple superposition of action-based DFs is possible (see e.g. [TO DO: REF (see Payel's hint)]), the substructure DFs could then be directly added to the background DF and incorporated in the modelling. In other words, modelling the Galaxy together with its non-axisymmetries and substructure could be approached as applying perturbations to an axisymmetric equilibrium model. And *RoadMapping* will help finding this equilibrium model.

Such a Galaxy model will be also important in the meantime: Many studies of Galaxy structure and evolution use orbits as tracers and therefore require a reliable fiducial potentials to turn stellar positions and velocities into orbits. And as long as we are as far away from realistic Galaxy models as we are now, the axisymmetric case well need to be our reference.

5. ACKNOWLEDGMENTS

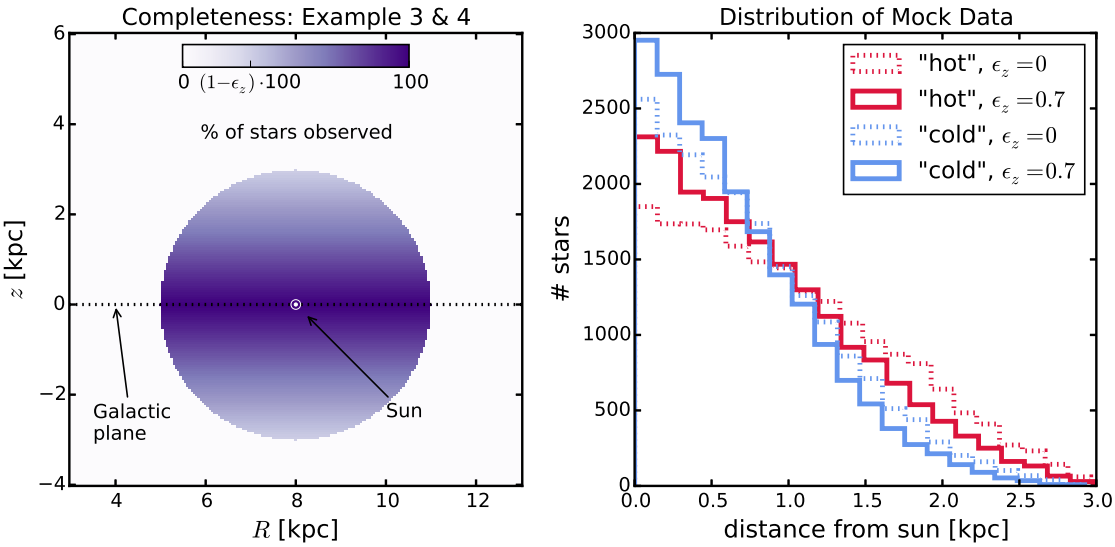

Figure 16. Selection function and mock data distribution for investigating vertical incompleteness of the data. All model parameters are summarized as test ⑤, Example 2, in Table **??**. The survey volume is a sphere around the sun and the percentage of observed stars is decreasing linearly with distance 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 *MAP*s ("hot" in red and "cool" in blue, see table 2) and with two different ϵ_z, 0 and 0.7, are shown in the right panel for illustration purposes. [TO DO: Re-do, if new analyses are in violin plot.] [TO DO: write distance from plane on x-axis]

[TO DO]

The authors thank Glenn van de Ven for the idea of using Kuzmin-Kutuzov Stäckel potentials in this case study.

APPENDIX

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 ⑤, Example 2, in Table **??** and Fig. **??**). We get a similar robust behaviour for small deviations, and only slightly less robustness for larger deviations. That an explantion 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 demonstratred in Fig. 18.

[TO DO: Explain how marginalization over a velocity coordinate is done.]

2. QUESTIONS THAT HAVEN'T BEEN COVERED SO FAR:

- What happens, when the errors are not uniform?
- What if errors in distance matter for selection?

Stuff that needs to be further examined in fig. 7 about the survey volume:—

[TO DO] Do we still get the same results, if we use an acceptable high numerical accuracy for the action grid?

[TO DO] Do the biases for the orange volume disappear, when we increase the integration range in vT?

[TO DO] Add the previous two volumes again, that had small extent in both R and z, or large extent in both.

[TO DO] Maybe add volume at smaller radius with large vertical extent?

[TO DO] Do we explicitely want to test, if it matters, if the radial coverage is larger or smaller the disk scale length, and the vertical coverage is larger or smaller than the disk scale height?

Stuff that needs to be further examined about the robustness against data incompleteness:—

[TO DO] Maybe instead of decreasing completeness with height above the plane, a completeness that INcreases with height above the plan, to model e.g. obscuration due to dust.

[TO DO] Make similar test as isoSphFlexIncompR, but with KKS potential, to test, if this robustness is a special case for the isochrone potential.

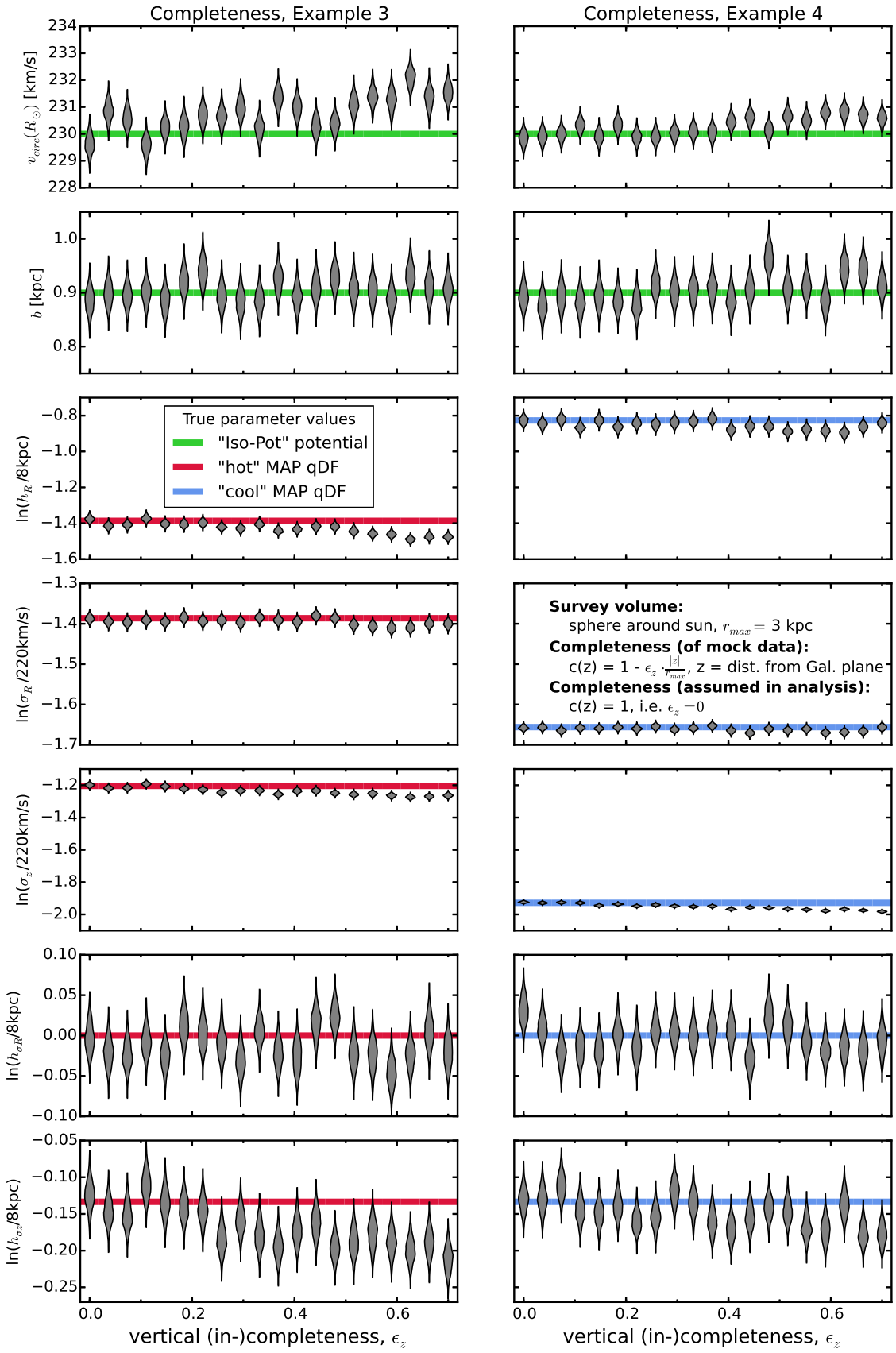

Figure 17. Influence of wrong assumptions about the incompleteness parallel to the Galactic plane of the data on the parameter reocovery with *RoadMapping* . Each mock data set was created having different incompleteness parameters ϵ_z (shown on the x-axis and illustrated in Fig. 16) and the model parameters are given as test ⑤, 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 Fig. 17. [TO DO: Rename Example 3 and 4 into 2a) and 2b)] [TO DO: This was done using the current qDF to set the fitting range. Nvelocity=24 and Nsigma=5 is high enough, though there is one analyse (hot, no3) for which it did not work. Maybe redo with fiducial qDF.]

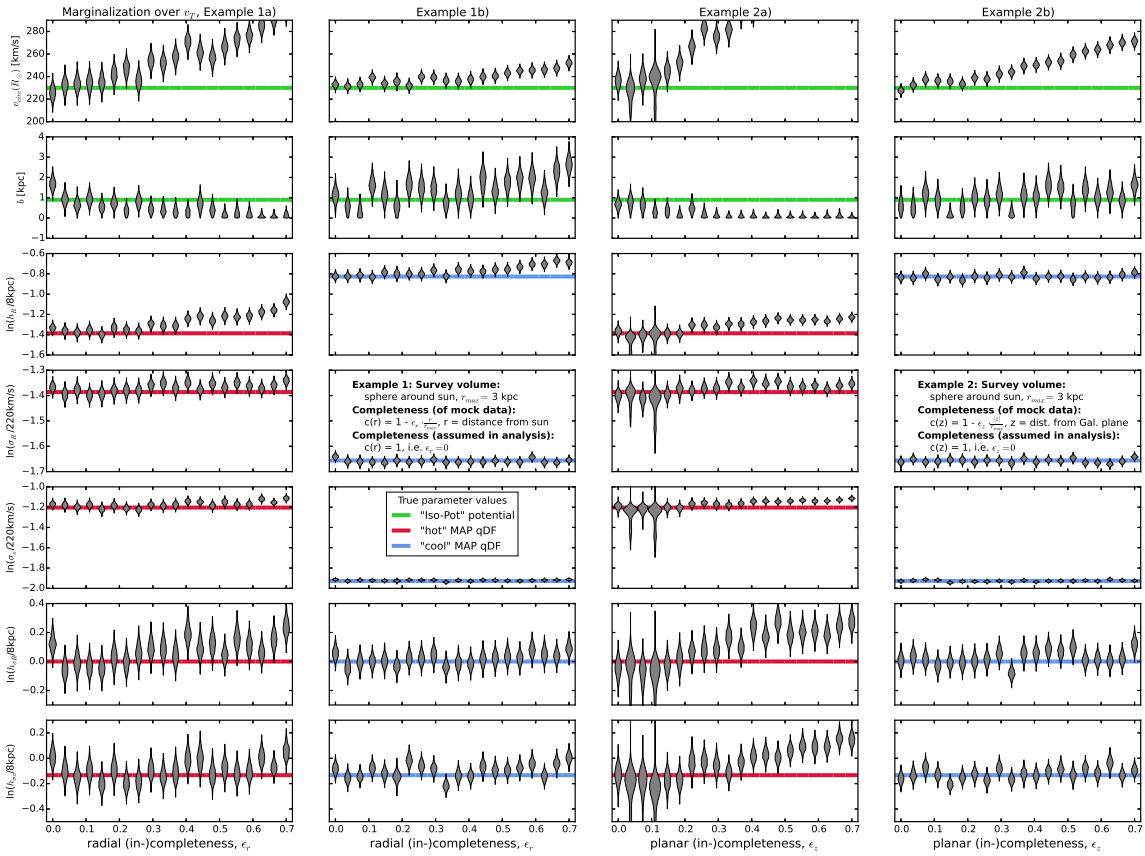

Figure 18. Influence of wrong assumptions about radial and vertical incompleteness on the parameter recovery, when *not* including information about the tangential velocities in the analysis. The mock data sets are the same as in Fig. 9 and 17, but this time we did not include the data coordinates v_T in the analysis and therefore marginalized the likelihood over v_T instead (see §???). 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 ($\epsilon_z \lesssim 10\%$). [TO DO: Redo all analyses for which MCMC did not converge to expected peak, and for which b ¡0 was not excluded. ???]) [TO DO: Rename Example 3 and 4 into 2a) and 2b), etc.]

General Stuff—

[TO DO:] Rename everywhere $N_{\rm sigma}$ to $n_{\rm interval}$ or something like this.

[TO DO:] Look up what McMillan & Binney 2013 have to say about the numerical accuracy of the normalisation. Sanders & Binney (2015) are quoting them on that matter.

[TO DO:] Consistent capitals in section titles.

[TO DO:] Make consistent: use of $\sigma_{R,0}$ and σ_R as profile or dispersion at sun.

[TO DO:] Make consistent h_{σ_R} –¿ $h_{\sigma,R}$

[TO DO:] Make consistent M –¿ p_M

[TO DO:] Make consistent MAP –¿ *MAP*

[TO DO:] Make consistent number of stars N –¿ $N_{\rm sample}$, introduce somewhere

[TO DO:] introduce *pdf* somewhere

[TO DO:] Rename all cite in citet and citep

[TO DO:] Make a backslash before the year in all references.

[TO DO:] Make sure, MW, DF, qDF, pdf are somewhere written and introduced explicitely

[TO DO:] Find out, if the bibitem references shold be the journal short cuts (e.g. to be able to be referenced on ADS)

[TO DO: Check if all references are actually used in paper. ???]

REFERENCES

Batsleer, P., & Dejonghe, H. 1994, A& A [TO DO], 287, 43

Binney, J. J. 2010, MNRAS, 401, 2318
Binney, J. J., & McMillan, P. 2011, MNRAS, 413, 1889
Binney, J. 2011, Pramana, 77, 39
Binney, J. J. 2012a, MNRAS, 426, 1324
Binney, J. J. 2012b, MNRAS, 426, 1328 (Princeton University Press)
Binney, J. & Tremaine, S. 2008, Galactic Dynamics: Second Edition
Bissantz, N., Debattista, V. P., & Gerhard, O. 2004, ApJ, 601, L155
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., 2012d, 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., & Binney, J. 1998, MNRAS, 294, 429
Dehnen, W. 1998, AJ, 115, 2384
De Lorenzi F., Debattista V.P., Gerhard O., Sambhus N. 2007, MNRAS, 376, 7
Foreman-Mackey, D., Hogg, D. W., Lang, D., & Goodman, J. 2013, PASP [TO DO], 125, 306
Gilmore, G. & Reid, N. 1983, MNRAS, 202, 1025
Holmberg, J., Nordström, B., & Andersen, J. 2009, A&A, 501, 941
Hunt, J. A. S., & Kawata, D. 2014, MNRAS, 443, 2112
Johnston, K. V., Zhao, H. S., Spergel, D. N., & Hernquist, L. 1999, ASPC [TO DO], 194, 15
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
Loebman, S. R., Ivezic, Z., Quinn, T. R. et al., 2012, ApJ, 758L, 23
McMillan, P. 2011, MNRAS, 414, 2446
Minchev, I., Famaey, B., Combes, F., et al. 2011, A&A, 527, A147
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
Piffl, T., Binney, J., & McMillan, P. J. et al., 2014, MNRAS, 455, 3133
Rix, H.-W., & Bovy, J. 2013, [TO DO] A& ARv, 21, 61
Roškar, R., Debattista, V. P., Stinson, G. S., Quinn, T. R., Kaufmann, T., & Wadsley, J. 2008a, ApJ, 675, L65
Roškar, R., Debattista, V. P., Quinn, T. R., Stinson, G. S., & Wadsley, J. 2008b, ApJ, 684, L79
Sackett, P. 1997, ApJ, 483, 103
Sanders J. L., Binney J. 2015, MNRAS, 449, 3479
Schönrich, R. & Binney, J. J. 2008, MNRAS, 396, 203
Sellwood, J. A. 2010, MNRAS, 409, 145
Sellwood, J. A., & Binney, J. J. 2002, MNRAS, 336, 785
Steinmetz, M. et al., 2006, AJ, 132, 1645
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., Newberg, H.-J., Johnson, J. A., et al. 2009, AJ, 137, 4377
Zhang, L., Rix, H.-W., van de Ven, G. et al., 2013, ApJ, 772, 108

[TO DO: Mit wie vielen J. wird Binney geschrieben?] [TO DO: Kommas nach letztem Namen oder nicht?] [TO DO: In welcher Reihenfolge soll ich sortieren?] [TO DO: Wie viele Autoren nennen, bevor et al.???]

Table 1

Gravitational potentials of the reference galaxies used troughout this work and the respective ways to calculate actions in these potentials. All four potentials are axisymmetric. The potential parameters are fixed for the mock data creation at the values given in this table. In the subsequent analyses we aim to recover these potential parameters again. The parameters of "MW13-Pot" and "KKS-Pot" were found as direct fits to the "MW14-Pot".

name	potential type	potential parameters p_Φ		action calculation	reference for potential type
"Iso-Pot"	isochrone potential	circular velocity at the sun isochrone scale length	$v_{\rm circ} = 230$ km s^{-1} $b = 0.9$ kpc	***analytical and exact*** J_r, J_ϑ, L_z; use $J_r \to J_R, J_\vartheta \to J_z$ in eq. (???)	Binney & Tremaine (2008)
"KKS-Pot"	2-component Kuzmin-Kutuzov- Stäckel potential	circular velocity at the sun focal distance of coordinate system[a] axis ratio of the coordinate surfaces[a] ...	$v_{\rm circ} = 230$ km s^{-1} $\Delta = 0.3$	***exact*** J_R, J_z, L_z using "Stäckel Fudge" (Binney 2012)	Batsleer & Dejonghe (1994)
	(disk + halo)	...of the disk component ...of the halo component	$\left(\frac{a}{c}\right)_{\rm Disk} = 20$ $\left(\frac{a}{c}\right)_{\rm Halo} = 1.07$	and interpolation on action grid	
	(analytic potential)	relative contribution of the disk mass to the total mass	$k = 0.28$	(Bovy 2015)	
"MW13-Pot"	MW-like potential with Hernquist bulge, 2 exponential disks	circular velocity at the sun stellar disk scale length stellar disk scale height	$v_{\rm circ} = 230$ km s^{-1} $R_d = 3$ kpc $z_h = 0.4$ kpc	***approximate*** J_R, J_z, L_z using "Stäckel Fudge" (Binney 2012)	Bovy & Rix (2013)
	(stars + gas), spherical power-law halo	relative halo contribution to $v^2_{\rm circ}(R_\odot)$ "flatness" of rotation curve	$f_h = 0.5$ $\frac{{\rm d}\ln(v_{\rm circ}(R_\odot))}{{\rm d}\ln(R)} = 0$	and interpolation on action grid	
	(interpolated potential)			(Bovy 2015)	
"MW14-Pot"	MW-like potential with cut-off power-law bulge, Miyamoto-Nagai stellar disk, NFW halo	-	-	***approximate*** J_R, J_z, L_z (see "MW13-Pot")	Bovy (2015)

[a] The coordinate system of each of the two Stäckel-potential components is $\frac{R^2}{\tau_{i,p}+\alpha_p} + \frac{z^2}{\tau_{i,p}+\gamma_p} = 1$ with $p \in \{\text{Disk}, /\text{Halo}\}$ and $\tau_{i,p} \in \{\lambda_p, \nu_p\}$. Both components have the same focal distance $\Delta = \sqrt{\gamma_p - \alpha_p}$, to make sure that the superposition of the two components itself is still a Stäckel potential. The axis ratio of the coordinate surfaces $\left(\frac{a}{c}\right)_p := \sqrt{\frac{\alpha_p}{\gamma_p}}$ describes the flattness of the corresponding Stäckel component.

Table 2

Reference distribution function parameters for the qDF in eq. (2)-(8). These qDFs describe the phase-space distribution of stellar *MAP*s for which mock data is created and analysed throughout this work for testing purposes. The parameters of the "cooler" & "colder" ("hotter" & "warmer") *MAP*s were chosen such, that the they have the same σ_R/σ_z ratio as the "hot" ("cool") *MAP* . The "colder" and "warmer" *MAP*s have a free parameter X that governs how much colder/warmer they are then the reference "hot" and "cool" qDFs. Hotter populations have shorter tracer scale lengths (Bovy et al. 2012d) and the velocity dispersion scale lengths were fixed according to Bovy et al. (2012c).

name of *MAP*	qDF parameters $p_{\rm DF}$				
	h_R [kpc]	σ_R [km s^{-1}]	σ_z [km s^{-1}]	h_{σ_R} [kpc]	h_{σ_z} [kpc]
"hot"	2	55	66	8	7
"cool"	3.5	42	32	8	7
"cooler"	2 +50%	55-50%	66-50%	8	7
"hotter"	3.5-50%	42+50%	32+50%	8	7
"colder"	2 +X%	55-X%	66-X%	8	7
"warmer"	3.5-X%	42+X%	32+X%	8	7

Table 3

Summary of test suites in this work: The first column indicates the test suite, the second column the potential, DF and selection function model etc. used for the mock data creation, the third model the corresponding model assumed in the analysis, and the last column lists the figures belonging to the test suite. Parameters that are not left free in the analyis, are always fixed to their true value. Unless otherwise stated we calculate the likelihood by the nested-grid and MCMC approach outlined in §2.6 and use $N_{\rm spatial} = 16$, $N_{\rm velocity} = 24$, $N_{\rm sigma} = 5$ as numerical accuracy for the likelihood normalisation in Eq. ???. [TO DO: Change encircled numbers to proper order. Make sure the plot references are the right ones.]

Test		Model for Mock Data	Model in Analysis	Figures
① Influence of survey volume on mock data distribution, also in action space	*Potential:*	"KKS-Pot"	-	Mock data: Fig. 2
	MAP :	2 *MAP*s "hot" or "cold" qDF		
	Survey volume:	a) $R \in [4, 12]$ kpc,$z \in [-4, 4]$ kpc, $\phi \in [-20°, 20°]$. b) $R \in [6, 10]$ kpc,$z \in [1, 5]$ kpc, $\phi \in [-20°, 20°]$.		
	# stars per data set:	20,000		
	# data sets:	4 (= 2 × 2 models)		
⑨ Numerical accuracy in calculation of the likelihood normalisation	*Potential:*	"Iso-Pot", "MW13-Pot" & "KKS-Pot"	-	Convergence of normalisation: Fig. 3
	MAP :	"hot" qDF		
	Survey volume:	sphere around sun, $r_{\rm max} = 0.2, 1, 2, 3$ or 4 kpc		
	Numerical accuracy:	$N_{\rm spatial} \in [5, 20]$, $N_{\rm velocity} \in [6, 40]$, $N_{\rm sigma} \in [3.5, 7]$		
⑩ *pdf* is a multivariate Gaussian for large data sets.	*Potential:*	"Iso-Pot"	"Iso-Pot", all parameters free	Fig. 4
	MAP :	"hot" qdf	qDF, all parameters free	
	Survey Volume:	sphere around sun, $r_{\rm max} = 2$ kpc	(fixed & known)	
	# stars per data set:	20,000		
	# data sets:	5 (only one is shown)		
	Numerical accuracy:		$N_{\rm velocity} = 20$ and $N_{\rm sigma} = 4$	
② Width of the likelihood scales with number of stars by $\propto 1/\sqrt{N}$.	*Potential:*	"Iso-Pot"	"Iso-Pot", free parameter: b	Fig. 5
	MAP :	"hot" qDF	"hot" qDF, free parameters: $\ln\left(\frac{h_R}{8\text{kpc}}\right), \ln\left(\frac{\sigma_R}{230\text{km s}^{-1}}\right), \ln\left(\frac{h_{\sigma,R}}{8\text{kpc}}\right)$	
	Survey volume:	sphere around sun, $r_{\rm max} = 3$ kpc	(fixed & known)	
	# stars per data set:	between 100 and 40,000		
	# data sets:	132		
	Analysis method:		likelihood on grid	
	Numerical accuracy:		$N_{\rm velocity} = 20$ and $N_{\rm sigma} = 4$ (for speed)	
③ Parameter estimates are unbiased.	*Potential:*	2 "Iso-Pot" with $b = 0.8$ kpc or $b = 1.5$ kpc	"Iso-Pot", free parameter: b	Fig. 6
	MAP :	2 *MAP*s , "hot" or "cool" qDF	"hot"/"cool" qDF, free parameters: $\ln\left(\frac{h_R}{8\text{kpc}}\right), \ln\left(\frac{\sigma_R}{230\text{km s}^{-1}}\right), \ln\left(\frac{h_{\sigma,R}}{8\text{kpc}}\right)$	
	Survey volume:	5 spheres around sun, $r_{\rm max} = 0.2, 1, 2, 3$ or 4 kpc	(fixed & known)	
	# stars per data set:	20,000		
	# data sets:	640 (= 2 × 2 × 5 models ×32 realisations)		
	Analysis method:		likelihood on grid	
	Numerical accuracy:		$N_{\rm velocity} = 20$ and $N_{\rm sigma} = 4$ (for speed)	
④ Influence of position & shape of survey volume on parameter recovery	*Potential:*	i) "Iso-Pot", ii) "MW13-Pot" or iii) "KKS-Pot"	i) "Iso-Pot", all parameters free ii) "MW13-Pot", R_d and f_h free iii) "KKS-Pot", all free except $v_{\rm circ}(R_\odot)$	Fig. 7
	MAP :	"hot" qDF	i) & iii) qDF, all parameters free ii) qDF, only h_R, $\sigma_{z,0}$ and h_{σ_R} free	
	Survey volume:	4 different wedges, see Fig. 7, upper right panel	(fixed & known)	
	# of stars per data set:	20,000		
	# data sets:	48 (= 4 × 3 models ×4 realisations)		
	Analysis method:		i) & ii) MCMC, iii) likelihood on grid	
	Action calculation:	ii) & iii) low accuracy "Stäckel Fudge" grid (Bovy 2015) for speed (# grid points: 25 in each E and ψ, 30 in L_z, $R_{\rm max} = 5$ [TO DO: What is psi and Rmax (units)?])	(same as mock data creation)	
⑤ Influence of wrong assumptions about the data set (in-)completeness	*Potential:*	"Iso-Pot"	"Iso-Pot", all parameters free	Illustration & mock data: Fig. 8 & 16 Analysis results: Fig. 9 & ??? Analysis results
	MAP :	2 *MAP*s , a) "hot" or b) "cool" qDF	qDF, all parameters free	
	Survey volume:	sphere around sun, $r_{\rm max} = 3$ kpc	(fixed & known)	
	Completeness:	*Example 1:* radial incompleteness, completeness$(r) = 1 - \epsilon_r \frac{r}{r_{\rm max}}$, twenty $\epsilon_r \in [0, 0.7]$	data set complete, completeness$(r) = 1$, $\epsilon_r = 0$	