

# **The ROADMAPPING Code: How to deal with "Real World" Issues in Action-based Dynamical Modelling the Milky Way**

W. Trick<sup>1,2</sup>, J. Bovy<sup>3,4</sup>, and H.-W. Rix<sup>1</sup>

trick@mpia.de

## **ABSTRACT**

Starting point for abstract: my old poster abstract. [TO DO] We aim to recover the Milky Way's gravitational potential using action-based dynamical modeling (cf. Bovy & Rix 2013, Binney & McMillan 2011, Binney 2012). This technique works by modeling the observed positions and velocities of disk stars with an equilibrium, three-integral quasi-isothermal distribution function. In preparation for the application to stellar phase-space data from Gaia, we create and analyze a large suite of mock data sets and we develop qualitative "rules of thumb" for which characteristics and limitations of data, model and code affect constraints on the potential most. We investigate sample size and measurement errors of the data set, size and shape of the observed volume, numerical accuracy of the code and action calculation, and deviations of the data from the assumed family of axisymmetric model potentials and distribution functions. This will answer the question: What kind of data gives the best and most reliable constraints on the Galaxy's potential?

*Subject headings:* Galaxy: disk — Galaxy: fundamental parameters — Galaxy: kinematics and dynamics — Galaxy: structure

## **Contents**

### **1 Introduction**

**3**

---

<sup>1</sup>Max-Planck-Institut für Astronomie, Königstuhl 17, D-69117 Heidelberg, Germany

<sup>2</sup>Correspondence should be addressed to trick@mpia.de.

<sup>3</sup>Institute for Advanced Study, Einstein Drive, Princeton, NJ 08540, USA

<sup>4</sup>Hubble fellow

<b>2</b>	<b>Dynamical Modelling</b>	<b>7</b>
2.1	Actions and Potential Models . . . . .	8
2.2	Distribution Function . . . . .	10
2.3	Selection Function . . . . .	12
2.4	Mock Data . . . . .	14
2.5	Likelihood . . . . .	17
2.6	Fitting Procedure . . . . .	22
<b>3</b>	<b>Results</b>	<b>25</b>
3.1	Model parameter estimates in the limit of large data sets . . . . .	29
3.2	The Role of the Survey Volume Geometry . . . . .	34
3.3	What if our assumptions on the (in-)completeness of the data set are incorrect? . . . . .	38
3.4	Effect of measurement errors on recovery of potential? . . . . .	41
3.5	The Impact of Deviations of the Data from the Idealized qDF . . . . .	42
3.6	What if our assumed potential model differs from the real potential? . . . . .	49
<b>4</b>	<b>Discussion and Summary</b>	<b>52</b>
<b>5</b>	<b>Conclusion</b>	<b>57</b>
<b>A</b>	<b>Appendix</b>	<b>58</b>
A.1	Influence of wrong assumptions about incompleteness of the data parallel to the Galactic plane . . . . .	59
<b>2</b>	<b>Questions that haven't been covered so far:</b>	<b>63</b>

## 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),[REF]). 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 [REF] [TO DO: References that explain that the modelling is scarce, or previous modelling approaches??] (previous modelling attempts were made e.g. by [TO DO]).

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 constraint 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 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 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 ([REF])? 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  $[\text{Fe}/\text{H}]$  and  $[\alpha/\text{FE}]$  (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 [REF]. This modelling also accounted for the complex, but known selection function of the kinematic tracers. For each MAP, the modelling resulted in a constraint of its DF, and an independent constraint on the gravitational potential, which members of all MAPs feel the same way.

Taken as an ensemble, the individual MAP models constrained the disk surface mass density over a wide range of radii ( $\sim 4 - 9$  kpc), and proved a powerful constraint on the disk mass scale length ( $\sim 2$  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 make, 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 explicitly 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 (§3.2). (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 (§??)?

One of the strongest assumptions is to restrict the dynamical modelling to a certain family of parametrized models. We investigate how well we can hope to recover the true potential, when our potential and DF models deviate from the true potential and DF. For the DF we specifically investigate two of our assumptions in §??: First, what would happen if the stars within *MAPs* do intrinsically not follow a single qDF as assumed by Ting et al. (2013); Bovy & Rix (2013). Second, and assuming *MAPs* do indeed follow the qDF, what would be the effect of pollution of *MAPs* through stars from neighbouring *MAPs* in the  $[\alpha/\text{Fe}]$ - $[\text{Fe}/\text{H}]$  plane due to too big abundance errors or bin sizes. And last but not least we test in §3.6 how well the modelling works, if our assumed potential family deviates 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 §?? following the procedure in §??, analysed according to the description of the machinery in §?? and the results are shown in §??.

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

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

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 can be build out of these Stäckel fudge actions, as described in Bovy (2015).

**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  $p_\Phi$ . The density distribution of these potentials is illustrated in Fig. 1.



Table 1. Gravitational potentials of the reference galaxies used throughout 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. In the subsequent analyses we aim to recover these potential parameters again. The parameters of "MW13-Pot" and "KKS-Pot" were found as best fits to the "MW14-Pot".

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

<sup>a</sup>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  $(\frac{a}{c})_p := \sqrt{\frac{\alpha_p}{\gamma_p}}$  describes the flatness of the corresponding Stäckel component.

## 2.2. Distribution Function

**Distribution Function.** Motivated by the findings of Bovy et al. 2012??? and Ting et al. (2013) about the simple phase-space structure of *MAPs*, 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  $\mathbf{J} = (J_R, J_z, L_z)$  and has the form

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

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

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

$$(5)$$

Here  $R_g \equiv R_g(L_z)$  and  $\Omega \equiv \Omega(L_z)$  are the (guiding-center) radius and the circular frequency of the circular orbit with angular momentum  $L_z$  in a given potential.  $\kappa \equiv \kappa(L_z)$  and  $\nu \equiv \nu(L_z)$  are the radial/epicycle ( $\kappa$ ) and vertical ( $\nu$ ) 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 value ( $L_0 = 10 \times R_\odot/8 \times v_{\text{circ}}(R_\odot)/220$ ).

For this qDF to be able to incorporate the findings by Bovy et al. 2012??? about the phase-space structure of *MAPs* summarized in §1, we set the functions  $n$ ,  $\sigma_R$  and  $\sigma_z$ , which

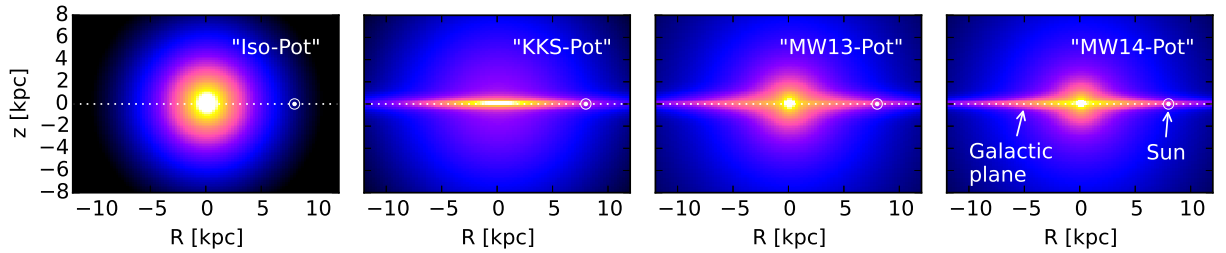


Fig. 1.— Density distribution of the four reference galaxy potentials in table 1, for illustration purposes. These potentials are used throughout this work for mock data creation and potential recovery. [TO DO: Halo sichtbarer machen, evtl. mit isodensity contours]

indirectly set the stellar number density and radial and vertical velocity dispersion profiles,

$$n(R_g | p_{\text{DF}}) \propto \exp\left(-\frac{R_g}{h_R}\right) \quad (6)$$

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

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

The qDF for each *MAP* has therefore a set of five free parameters  $p_{\text{DF}}$ : the density scale length of the tracers  $h_R$ , the radial and vertical velocity dispersion at the solar position  $R_\odot$ ,  $\sigma_{R,0}$  and  $\sigma_{z,0}$ , and the scale lengths  $h_{\sigma_R}$  and  $h_{\sigma_z}$ , that describe the radial decrease of the velocity dispersion. The *MAPs* 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  $\rho_{\text{DF}}(\mathbf{x} | p_\Phi, p_{\text{DF}})$  for a given qDF and potential . We do this by integrating the qDF at a given  $(R, z)$  over all three velocity components, using a  $N_{\text{velocity}}$ -th order Gauss-Legendre quadrature for each integral:

$$\rho_{\text{DF}}(R, |z| | p_\Phi, p_{\text{DF}}) = \int_{-\infty}^{\infty} \text{qDF}(\mathbf{J}[R, z, \mathbf{v} | p_\Phi] | p_{\text{DF}}) d^3\mathbf{v} \quad (9)$$

$$\approx \int_{-N_{\text{sigma}}\sigma_R(R|p_{\text{DF}})}^{N_{\text{sigma}}\sigma_R(R|p_{\text{DF}})} \int_{-N_{\text{sigma}}\sigma_z(R|p_{\text{DF}})}^{N_{\text{sigma}}\sigma_z(R|p_{\text{DF}})} \int_0^{1.5v_{\text{circ}}(R_\odot)} \text{qDF}(\mathbf{J}[R, z, \mathbf{v} | p_\Phi] | p_{\text{DF}}) dv_T dv_z dv_R, \quad (10)$$

where  $\sigma_R(R | p_{\text{DF}})$  and  $\sigma_z(R | p_{\text{DF}})$  are given by eq. (7) and (8) and the integration ranges are motivated by Fig. 2. For a given  $p_\Phi$  and  $p_{\text{DF}}$  we explicitly calculate the density on  $N_{\text{spatial}} \times N_{\text{spatial}}$  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  $N_{\text{spatial}}^2 \cdot N_{\text{velocity}}^3$ . Fig. ??? shows the importance of choosing  $N_{\text{spatial}}$ ,  $N_{\text{velocity}}$  and  $N_{\text{sigma}}$  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  $\mathbf{x} \equiv (R, \phi, z)$  and corresponding velocity components  $\mathbf{v} \equiv (v_R, v_\phi, v_z)$ . If the stellar phase-space data is given in observed coordinates, position  $\tilde{\mathbf{x}} \equiv (\alpha, \delta, m - M)$  in right ascension  $\alpha$ , declination  $\delta$  and distance modulus  $(m - M)$ , and velocity  $\tilde{\mathbf{v}} \equiv (\mu_\alpha, \mu_\delta, v_{\text{los}})$  as proper motions  $\boldsymbol{\mu} = (\mu_\alpha, \mu_\delta)$  [TO DO: cos somewhere??] and line-of-sight velocity  $v_{\text{los}}$ , the data  $(\tilde{\mathbf{x}}, \tilde{\mathbf{v}})$  has to be converted first into the Galactocentric rest-frame coordinates  $(\mathbf{x}, \mathbf{v})$  using the sun’s position and velocity. For simplicity we assume for the sun

$$\begin{aligned} (R_\odot, \phi_\odot, z_\odot) &= (8 \text{ kpc}, 0^\circ, 0 \text{ kpc}) \\ (v_{R,\odot}, v_{T,\odot}, v_{z,\odot}) &= (0, 230, 0) \text{ km s}^{-1}. \end{aligned}$$

**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

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

It’s value describes the probability to observe a star at  $\mathbf{x}$ .

For the observed volume we use simple geometrical shapes: Either a sphere of radius  $r_{\text{max}}$  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  $R \in [R_{\text{min}}, R_{\text{max}}], \phi \in [\phi_{\text{min}}, \phi_{\text{max}}], z \in [z_{\text{min}}, z_{\text{max}}]$  within the model galaxy.

The completeness is, in our framework, a function of position with  $0 \leq \text{completeness}(\mathbf{x}) \leq 1$  everywhere inside the observed volume. It could be understood as a position-dependent detection probability. Unless explicitly stated otherwise, we use everywhere

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

Table 2. Reference distribution function parameters for the qDF in eq. (2)-(8). These qDFs describe the phase-space distribution of stellar *MAPs* for which mock data is created and analysed throughout this work for testing purposes. The parameters of the "cooler" & "colder" ("hotter" & "warmer") *MAPs* were chosen such, that they have the same  $\sigma_R/\sigma_z$  ratio as the "hot" ("cool") *MAP*. The "colder" and "warmer" *MAPs* 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 <i>MAP</i>	qDF parameters $p_{\text{DF}}$				
	$h_R$ [kpc]	$\sigma_R$ [km s <sup>-1</sup> ]	$\sigma_z$ [km s <sup>-1</sup> ]	$h_{\sigma_R}$ [kpc]	$h_{\sigma_z}$ [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

## 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  $(\mathbf{J}_i, \boldsymbol{\theta}_i) \longrightarrow (\mathbf{x}_i, \mathbf{v}_i)$  is however difficult to perform and computationally much more expensive than the transformation  $(\mathbf{x}_i, \mathbf{v}_i) \longrightarrow (\mathbf{J}_i, \boldsymbol{\theta}_i)$ . 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  $\rho(R, |z| \mid p_\Phi, p_{\text{DF}})$  generated by the given qDF and according to §2.2 and Eq. 10. For the creation of the mock data we use  $N_{\text{spatial}} = 20$ ,  $N_{\text{velocity}} = 40$  and  $N_{\text{sigma}} = 5$ .

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

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

**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  $\text{qDF}(R_i, z_i, \mathbf{v} \mid p_{\text{Phi}}, p_{\text{DF}})$  using a rejection method. To reduce the number of rejected velocities, we use a Gaussian in velocity space as an envelope function, from which we first randomly sample velocities and then apply the rejection method to shape the Gaussian velocity distribution towards the velocity distribution predicted by the qDF. We now have a mock data set according to the required:

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

**Example:** Fig. 2 shows examples of mock data sets in configuration space  $(\mathbf{x}, \mathbf{v})$  and action space, that demonstrate that the qDF indeeds produces ..... [TO DO]

**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  $(\alpha, \delta)$  and distance modulus  $(m - M)$  as proxy for the distance from the sun, the proper motion in both  $\alpha$  and  $\delta$  direction  $(\mu_\alpha, \mu_\delta)$  and the line-of-sight velocity  $v_{\text{los}}$ . 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<sup>-1</sup>. We assume Gaussian measurement errors in the observables  $\tilde{\mathbf{x}} = (\alpha, \delta, (m - M))$ ,  $\tilde{\mathbf{v}} = (\mu_\alpha, \mu_\delta, v_{\text{los}})$ .

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.

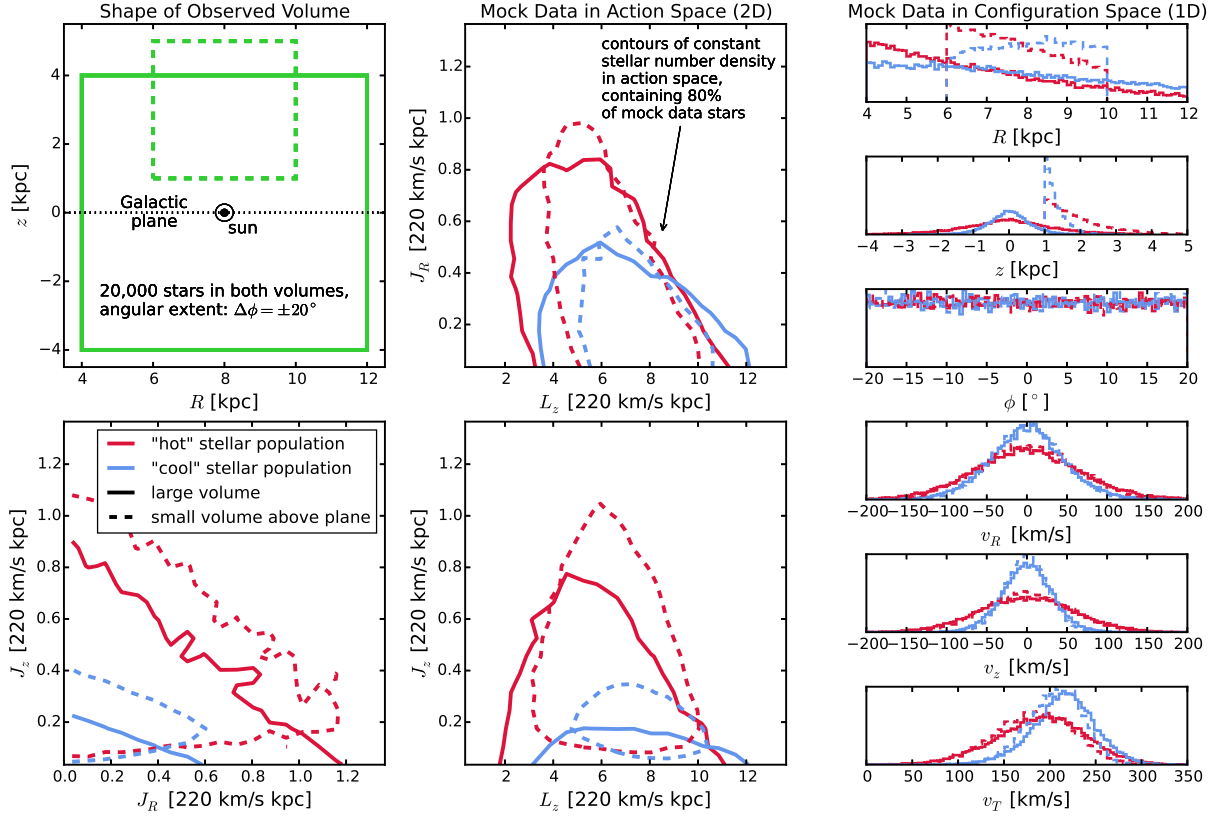


Fig. 2.— Distribution of mock data in action space (2D iso-density contours enclosing 80% of the stars, in the two central and the lower left panel) and configuration space (1D histograms in right panels), depending on shape and position of observation volume and temperature of the stellar population. 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 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). The stars of the "cool" population have in general lower radial and vertical actions, i.e. are on more circular orbits. The different ranges of  $L_z$ 's in the two volumes reflect  $L_z \sim Rv_{\text{circ}}$  and the different radial extent of both volumes. The volume above the plane contains no stars with  $J_z = 0$  and more with  $J_z$ : The higher a volume is located above the plane, the larger  $J_z$  has to be for the star's orbit to cross this volume. Circular orbits with  $J_R = 0$  and  $J_z = 0$  can obviously only be observed in the Galactic mid-plane. The smaller an orbit's  $L_z$ , the smaller also its mean orbital radius. For this orbit to be able to reach into a volume located at larger Galactocentric radius, it needs to be more eccentric and therefore have a larger  $J_z$ . This anti-correlation between  $L_z$  and  $J_R$  can be seen in the top central panel. Orbits with both large  $J_R$  and large  $J_z$  would be very energetic and are therefore less likely to be observed.



## 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  $\text{sf}(\mathbf{x})$ ,

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

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

$$M = \{p_{\text{DF}}, p_{\Phi}\},$$

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

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

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

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

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

where  $N$  is the number of stars in the data set  $D$ . In the last step we used eq. (9). The factor  $\prod_i \text{sf}(\mathbf{x}_i)$  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  $v_{\text{circ}}(R = 1) = 1$ . The circular velocity at the

sun’s radius,  $v_{\text{circ}}(R_{\odot} = 8.\text{kpc}) \sim 230\text{km s}^{-1}$ , 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 velocity units of  $v_o := v_{\text{circ}}(R_{\odot})$ . The prefactor  $1/(r_o v_o)^3$  in eq. (11) makes sure that the likelihood has the correct units to satisfy:

$$\int P(\mathbf{x}, \mathbf{v} \mid M) d^3x d^3v \propto 1$$

Including this prefactor is crucial when  $v_{\text{circ}}(R_{\odot})$  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,

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

In the case of an axisymmetric galaxy model and  $\text{sf}(\mathbf{x}) = 1$  everywhere inside the observed volume (i.e. a complete sample as assumed in most tests in this work), the normalisation is essentially a two-dimensional integral in  $R$  and  $z$  of the interpolated tracer density  $\rho_{\text{DF}}$  (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  $N_{\text{spatial}}^2 \times N_{\text{velocity}}^3$  action calculations. The accuracy has to be chosen high enough, such that a resulting numerical error

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

does not dominate the likelihood, i.e.

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

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,  $N_{\text{spatial}} = 16$ ,  $N_{\text{velocity}} = 24$  and  $N_{\text{sigma}} = 5$ , does satisfy this requirement.

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

$$N[\mathbf{y}_i, \sigma_{\mathbf{y},i}](\mathbf{y}') = N[\mathbf{y}', \sigma_{\mathbf{y},i}](\mathbf{y}_i) \equiv \prod_k \frac{1}{\sqrt{2\pi\sigma_{\mathbf{y},k}^2}} \exp\left(-\frac{(y_{i,k} - y'_{i,k})^2}{2\sigma_{\mathbf{y},k}^2}\right),$$

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

$$\tilde{P}(\mathbf{y}_i \mid p_\Phi, p_{\text{DF}}, \sigma_{\mathbf{y},i}) \equiv \frac{\text{sf}(\mathbf{y}_i) \cdot \int d^6 y' \text{DF}(\mathbf{y}') \cdot N[\mathbf{y}_i, \sigma_{\mathbf{y},i}](\mathbf{y}')}{\int d^6 y \text{DF}(\mathbf{y}) \cdot \int d^6 y' \text{sf}(\mathbf{y}') \cdot N[\mathbf{y}, \sigma_{\mathbf{y},i}](\mathbf{y}')}.$$

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

$$\tilde{P}(\mathbf{y}_i \mid p_\Phi, p_{\text{DF}}, \sigma_{\mathbf{y},i}) \approx \frac{\text{sf}(\mathbf{x}_i)}{\int d^6 y \text{DF}(\mathbf{y}) \cdot \text{sf}(\mathbf{x})} \cdot \frac{1}{N_{\text{error}}} \sum_n^{N_{\text{error}}} \text{DF}(\mathbf{x}_i, \mathbf{v}[\mathbf{y}'_{i,n}]) \quad (15)$$

$$\text{with } \mathbf{y}'_{i,n} \sim N[\mathbf{y}_i, \sigma_{\mathbf{y},i}](\mathbf{y}') \quad (16)$$

In doing so, we ignore errors in the star's position  $\mathbf{x}_i$  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  $N_{\text{error}}$  samples drawn from the full error Gaussian in observable space,  $\mathbf{y}'_{i,n}$ .

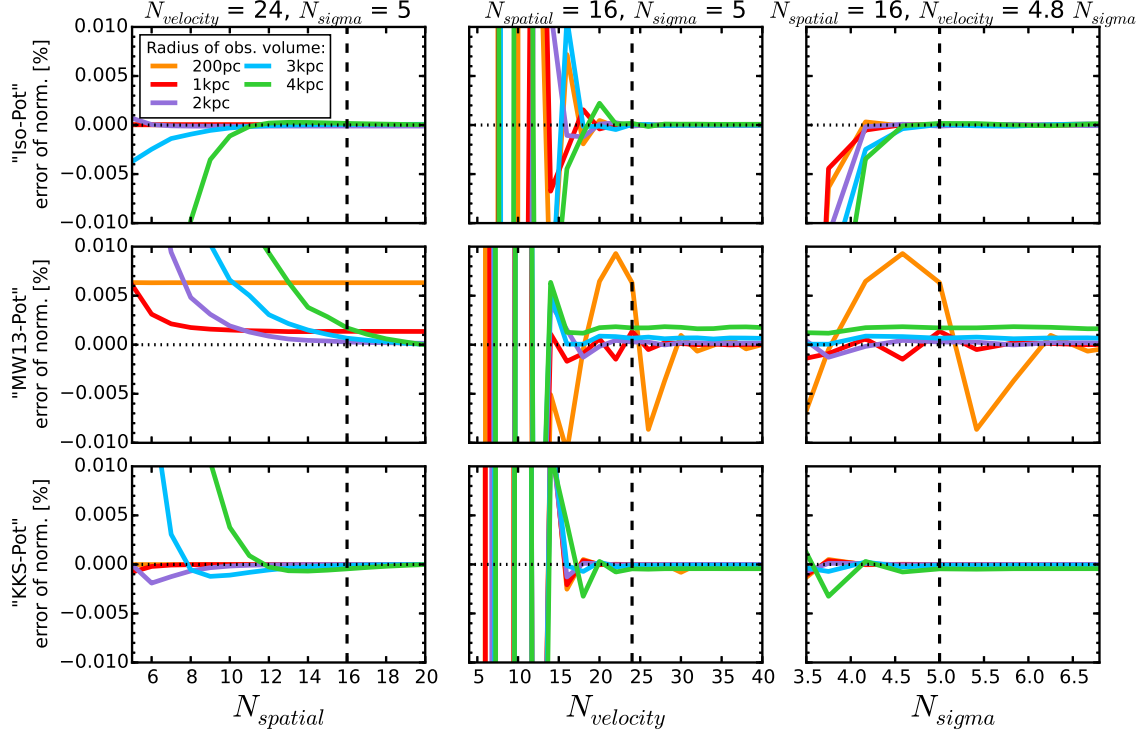


Fig. 3.— Relative error of the likelihood normalization in Eq. (13) depending on the accuracy of the density calculation in Eq. (10) (and surrounding text). The different colors represent calculations for different radii of the spherical observation volume around the sun, as indicated in the legend.  $N_{\text{spatial}}$  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  $N_{\text{velocity}}$  is performed over an integration range of  $\pm N_{\text{spatial}}$  times the dispersion in  $v_R$  and  $v_z$  and  $[0, 1.5v_{\text{circ}}(R_{\odot})]$  [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 ?? 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.)

Fig. 3.— (Continued.) We calculate the true normalization with high accuracy as  $M_{\text{tot,true}} \approx M_{\text{tot}}(N_{\text{spatial}} = 20, N_{\text{velocity}} = 56, N_{\text{sigma}} = 7)$ . 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  $N_{\text{spatial}}$ .???

## 2.6. Fitting Procedure

We search the  $(p_\Phi, p_{\text{DF}})$  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  $(p_\Phi, p_{\text{DF}})$  parameter space can be high-dimensional. To effectively minimizing the number of likelihood evaluations before finding its peak, we use a nested-grid approach:

- *Initialization.* For  $N$  free model parameters  $M = (p_\Phi, p_{\text{DF}})$ , we set up a sufficiently large initial grid with  $3^N$  regular grid points.<sup>1</sup>
- *Evaluation.* We evaluate the likelihood at each grid-point. Because of the many computationally expensive  $\mathbf{x}, \mathbf{v} \xrightarrow{p_\Phi} \mathbf{J}$  transformations that have to be performed for each new set of  $p_\Phi$  parameters, an outer loop iterates over the  $p_\Phi$  parameters and pre-calculates the actions, while an inner loop evaluates the likelihood Eq. (11) for all qDF parameters  $p_{\text{DF}}$  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 of the model parameter in  $M$  we marginalize the likelihood by summing over the grid. If the resulting 3 points all lie within  $4\sigma$  of a Gaussian, we fit a Gaussian to the 3 points and determine a new  $4\sigma$  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  $N_j \times N_{\text{error}}$  [TO DO: explain  $N_{\text{error}}$  ???] stellar data actions  $\mathbf{J}_i$ . To calculate

---

<sup>1</sup>To get a better feeling where in parameter space the true  $p_{\text{DF}}$  parameters lie, we fit eq. (???) directly to the data. This gives a very good initial 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  $h_R^{\text{out}}$  and the qDF tracer scale length  $h_R^{\text{in}} = h_R$ .

the normalisation in Eq. (11),  $N_{\text{spatial}}^2 \times N_{\text{velocity}}^3$  actions  $\mathbf{J}_n$  are needed. Formally the spatial coordinates at which the  $\mathbf{J}_n$  are calculated depend on the  $p_{\text{DF}}$  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  $p_{\text{DF}}$ 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  $p_\Phi$  makes the normalisation vary smoothly with different  $p_{\text{DF}}$ . 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  $p_{\text{DF}}$  parameters. We take care of this by setting, in each iteration step of the nested-grid search, the fiducial qDF simply to the  $p_{\text{DF}}$  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  $N_j \times N_{\text{error}} + N_{\text{spatial}}^2 \times N_{\text{velocity}}^3$  of actions to calculate, i.e. the number of stars, error samples and numerical accuracy of the normalisation calculations, c) the number of different potentials to investigate (i.e. the number of free potential parameters and number of grid points in each dimension) and d) the number of qDFs to investigate. The latter is also non-negligible, because for such a large number of actions the number of qDF-function evaluations also take some time.

**Fitting Step 2: MCMC.** After the nested-grid search is converged, the grid is centered at the peak of the likelihood and its extent contains the  $4\sigma$  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 ? 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 exactly 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. **[hat shouldl also be at the end of the introduction..** [say: except for the case of “errors” we assume that thne phase-space errors are negligible..]

Table 3. [TO DO: Caption] Parameters that are not left free in the analysis, are always fixed to their true value. Unless otherwise stated we calculate the likelihood by the nested-grid and MCMC approach outlined in ??? and use  $N_{\text{spatial}} = 16$ ,  $N_{\text{velocity}} = 24$ ,  $N_{\text{sigma}} = 5$  as numerical accuracy for the likelihood normalisation in Eq. ???. [TO DO: Change encircled numbers to proper order. Make sure the plots reference the right one.]

Test	Model for Mock Data		Model in Analysis	Figures
① Influence of survey volume on mock data distribution, also in action space	<i>Potential:</i> <i>MAP :</i> <i>Survey volume:</i> <i># stars per data set:</i> <i># data sets:</i>	"KKS-Pot" 2 MAPs "hot" or "cold" qDF a) $R \in [4, 12]$ kpc, $z \in [-4, 4]$ kpc, $\phi \in [-20^\circ, 20^\circ]$ . b) $R \in [6, 10]$ kpc, $z \in [1, 5]$ kpc, $\phi \in [-20^\circ, 20^\circ]$ . 20,000 4 ( $= 2 \times 2$ models)	-	Mock data: Fig. 2
⑨ Numerical accuracy in calculation of the likelihood normalisation	<i>Potential:</i> <i>MAP :</i> <i>Survey volume:</i> <i>Numerical accuracy:</i>	"Iso-Pot", "MW13-Pot" & "KKS-Pot" "hot" qDF sphere around sun, $r_{\text{max}} = 0.2, 1, 2, 3$ or 4 kpc $N_{\text{spatial}} \in [5, 20]$ , $N_{\text{velocity}} \in [6, 40]$ , $N_{\text{sigma}} \in [3.5, 7]$	-	Convergence of normalisation: Fig. 3
② Width of the likelihood scales with number of stars by $\propto 1/\sqrt{N}$ .	<i>Potential:</i> <i>MAP :</i> <i>Survey volume:</i> <i># stars per data set:</i> <i># data sets:</i> <i>Analysis method:</i> <i>Numerical accuracy:</i>	"Iso-Pot" "hot" qDF sphere around sun, $r_{\text{max}} = 3$ kpc between 100 and 40,000 132 likelihood on grid $N_{\text{velocity}} = 20$ and $N_{\text{sigma}} = 4$ (for speed)	"Iso-Pot", free parameter: $b$ "hot" qDF, free parameters: $\ln\left(\frac{h_R}{8\text{kpc}}\right), \ln\left(\frac{\sigma_R}{230\text{km s}^{-1}}\right), \ln\left(\frac{h_{\sigma,R}}{8\text{kpc}}\right)$ (fixed & known)	Fig. 5
③ Parameter estimates are unbiased.	<i>Potential:</i> <i>MAP :</i> <i>Survey volume:</i> <i># stars per data set:</i> <i># data sets:</i> <i>Analysis method:</i> <i>Numerical accuracy:</i>	2 "Iso-Pot" with $b = 0.8$ kpc or $b = 1.5$ kpc 2 MAPs, "hot" or "cool" qDF 5 spheres around sun, $r_{\text{max}} = 0.2, 1, 2, 3$ or 4 kpc 20,000 640 ( $= 2 \times 2 \times 5$ models $\times 32$ realisations) likelihood on grid $N_{\text{velocity}} = 20$ and $N_{\text{sigma}} = 4$ (for speed)	"Iso-Pot", free parameter: $b$ "hot"/"cool" qDF, free parameters: $\ln\left(\frac{h_R}{8\text{kpc}}\right), \ln\left(\frac{\sigma_R}{230\text{km s}^{-1}}\right), \ln\left(\frac{h_{\sigma,R}}{8\text{kpc}}\right)$ (fixed & known)	Fig. 6
④ Influence of position & shape of survey volume on parameter recovery	[TO DO]			
⑤ Influence of	<i>Potential:</i> <i>MAP :</i>	"Iso-Pot" 2 MAPs, a) "hot" or b) "cool" qDF	"Iso-Pot", all parameters free qDF, all parameters free	Illustration & mock data: Fig. 9 & 15

Table 3—Continued

Test		Model for Mock Data	Model in Analysis	Figures
wrong assumptions about the data set (in-)completeness on parameter recovery	<i>Survey volume:</i> <i>Completeness:</i>	sphere around sun, $r_{\max} = 3$ kpc <i>Example 1:</i> radial incompleteness, $\text{completeness}(r) = 1 - \epsilon_r \frac{r}{r_{\max}}$ , twenty $\epsilon_r \in [0, 0.7]$ $r \equiv$ distance from sun, <i>Example 2:</i> planar incompleteness, $\text{completeness}(z) = 1 - \epsilon_z \frac{ z }{r_{\max}}$ , $\epsilon_r \in [0, 0.7]$ , $z \equiv$ distance from Gal. plane.	(fixed & known) data set complete, $\text{completeness}(r) = 1$ , $\epsilon_r = 0$  data set complete, $\text{completeness}(r) = 1$ , twenty $\epsilon_z = 0$	Analysis results: Fig. 10 & ??? Analysis results when not using $v_T$ data: Fig. 17
	<i># stars per data set:</i> <i># data sets:</i>	20,000 40 ( $= 2 \times 2 \times 20$ )		
⑥ Measurement errors	[TO DO]			
⑦ Deviations in the assumed DF from the star's true DF	<i>Potential:</i> <i>MAP :</i>	"Iso-Pot" mix of two qDFs <i>Example 1:</i> with fixed qDF parameters, but 20 different mixing rates: a) "hot" & "cooler" qDF or b) "cool" & "hotter" qDF <i>Example 2:</i> 20 fixed 50/50 mixtures, with varying qDF parameters (by $X\%$ ): a) "hot" & "colder" qDF or b) "cool" & "warmer" qDF	"Iso-Pot", all parameters free single qDF, all parameters free      (fixed & known)	mock data: Fig. 11 Analysis results: 12 & Fig. 13
	<i>Survey volume:</i> <i># stars per data set:</i> <i># data sets:</i>	sphere around sun, $r_{\max} = 2$ kpc 20,000 40 ( $= 2 \times 2 \times 20$ )		

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

The individual *MAP* in Bovy & Rix (2013) contained typically 200 [CHECK] objects, so that each *MAP* implied a quite broad *pdf* for the  $p_M$ . Here we explore what happens in the limit of very much larger samples for each *MAP*, say 20,000 objects. As outlined in §[TO DO CHECK] the immediate consequence of larger samples is given by the likelihood normalization requirement,  $\log(1 + \text{rel.error}) \leq 1/N_{\text{sample}}$ , (see Eq. 5 [TO DO CHECK]), 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,  $\sigma_p$  that scales as  $1/N_{\text{sample}}$ . Further, we must verify that any bias in the *pdf* expectation value is far less than  $\sigma_p$ , even for quite large samples.

Using sets of mock data ([ TO DO: describe by referencing to Section]) and our fiducial model for  $p_M$ , we verified that the *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 of the  $p_M$  from with the mock data were drawn. Figure5 then illustrates that the *pdf* width,  $\sigma_p$  indeed scales as  $1/N_{\text{sample}}$ . Fig.6 illustrates even more, that the *RoadMapping* satisfies the central limit theorem. The average parameter estimates from many mock samples with identical underlying  $p_M$  are very close to the input  $p_M$ , and the distribution of the actual parameter estimates are a Gaussian around it.

**[TO DO] Stuff to explain about fig. 4 and 5:** The central limit theorem predicts that the likelihood will approach a Gaussian distribution  $\mathcal{N}(\mu, \sigma/\sqrt{N})$  with  $N$  being the number of data points.

**[TO DO] Stuff to explain about fig. 6:** Mention also that bigger volumes give most of the time better constraints and that there is no clear answer, if a hot or cooler population gives better constraints. Depends on parameter considered, selection function etc.

**[TO DO] Missing test and plot:** Would be cool to have a plot, that shows that for the Stäckel potential we don't get biases, but that there are some for the analytic Miyamoto-Nagai + power-law halo & interpolated MW potential and therefore this bias is probably due to incorrect action calculation.

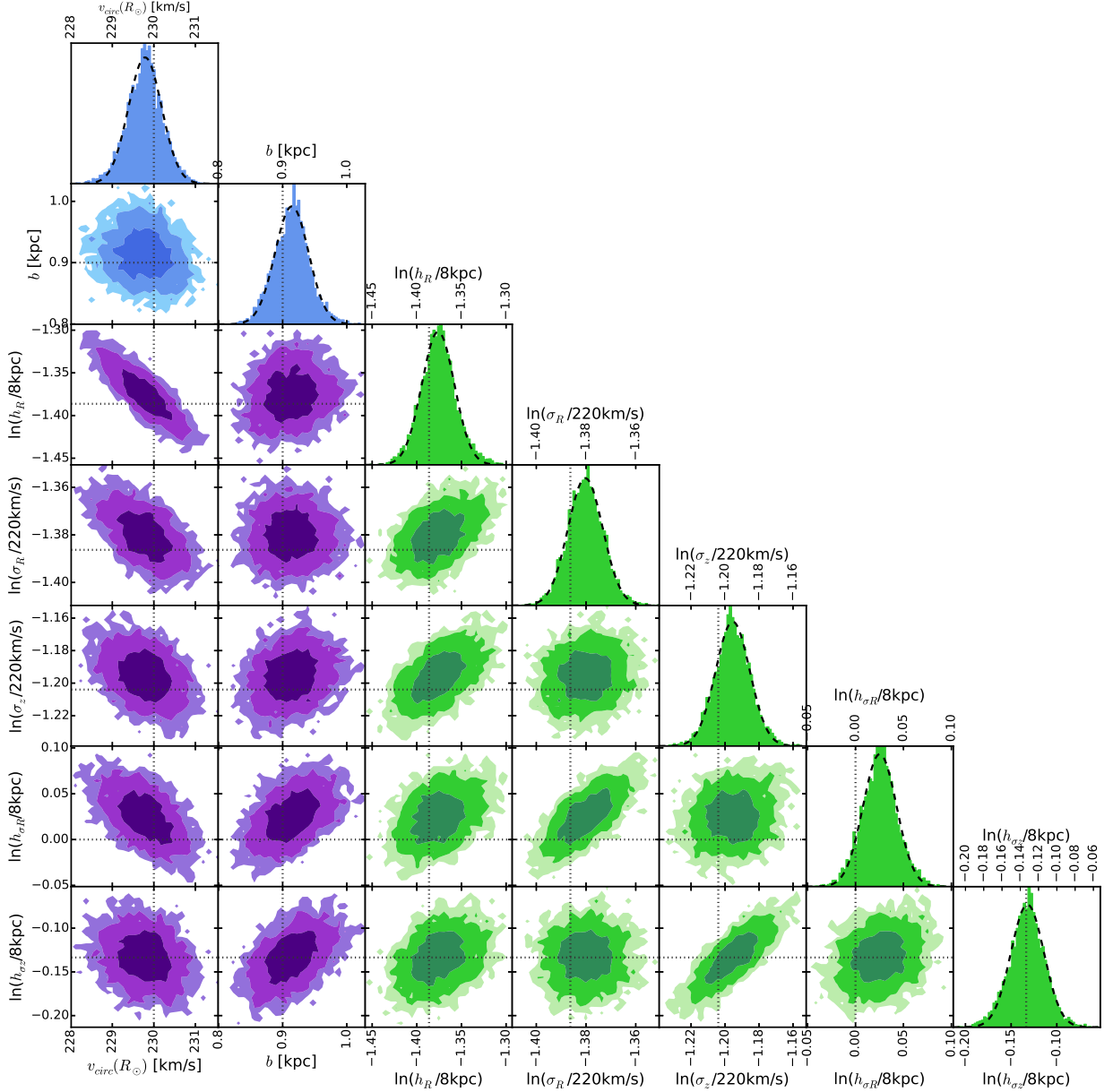


Fig. 4.— The likelihood in eq. (???) in the parameter space  $\{p_\Phi, \ln(p_{\text{DF}})\}$  for one example mock data set. This mock data set has 20,000 stars and was created in the potential "Iso-Pot" and from the "hot" qDF, and was observed within a spherical volume around the sun of radius  $r = 2$  kpc . 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. The likelihood here was sampled using MCMC (with flat priors in  $p_\Phi$  and  $\ln(p_{\text{DF}})$  to turn the likelihood into a full posterior distribution function). 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 multivariate Gaussian, as expected from the central limit theorem. [TO DO: Maybe re-do with higher accuracy??? This was done with  $N_{\text{sigma}} = 4$ .] [TO DO: Mention "Note: this was picked among 5 to have all 1sigma contours encompass the input values." ???] [TO DO: it's the cold population, not the hot one??? I'm not sure]

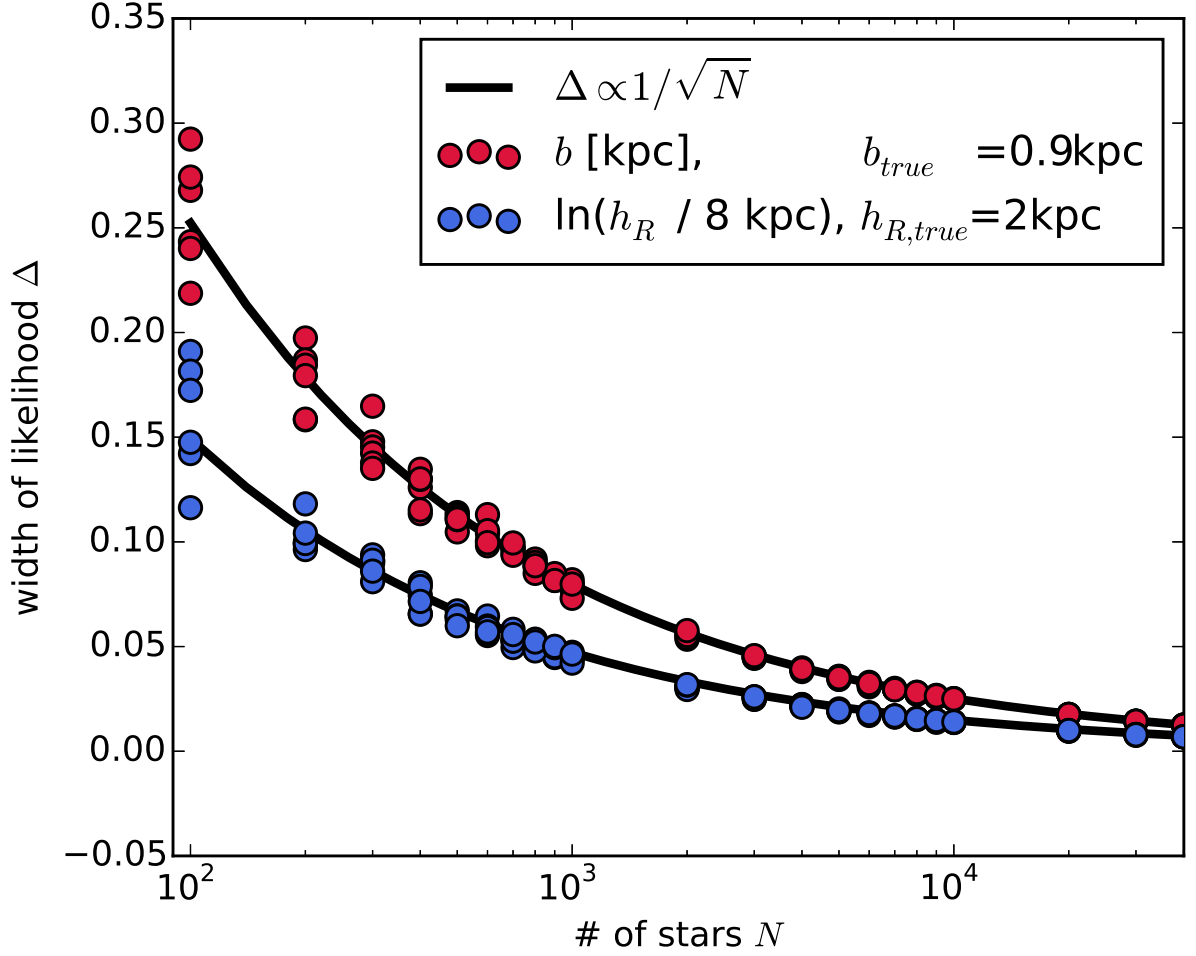


Fig. 5.— The width of the likelihood for two fit parameters found from analyses of 132 mock data sets vs. the number of stars in each data set. The mock data was created in the "Iso-Pot" potential and all model parameters are given as Test ② in Table ???. The likelihood in Eq. ??? was evaluated on a 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  $\Delta(N) \propto 1/\sqrt{N}$  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  $1/\sqrt{N}$  as predicted by the central limit theorem. [TO DO: Maybe re-do with higher accuracy??? This was done with  $N_{sigma} = 4$ .] [TO DO: rename width of likelihood into Standard Error (SE). Also x-axis:  $N$  (# of stars in data set)???

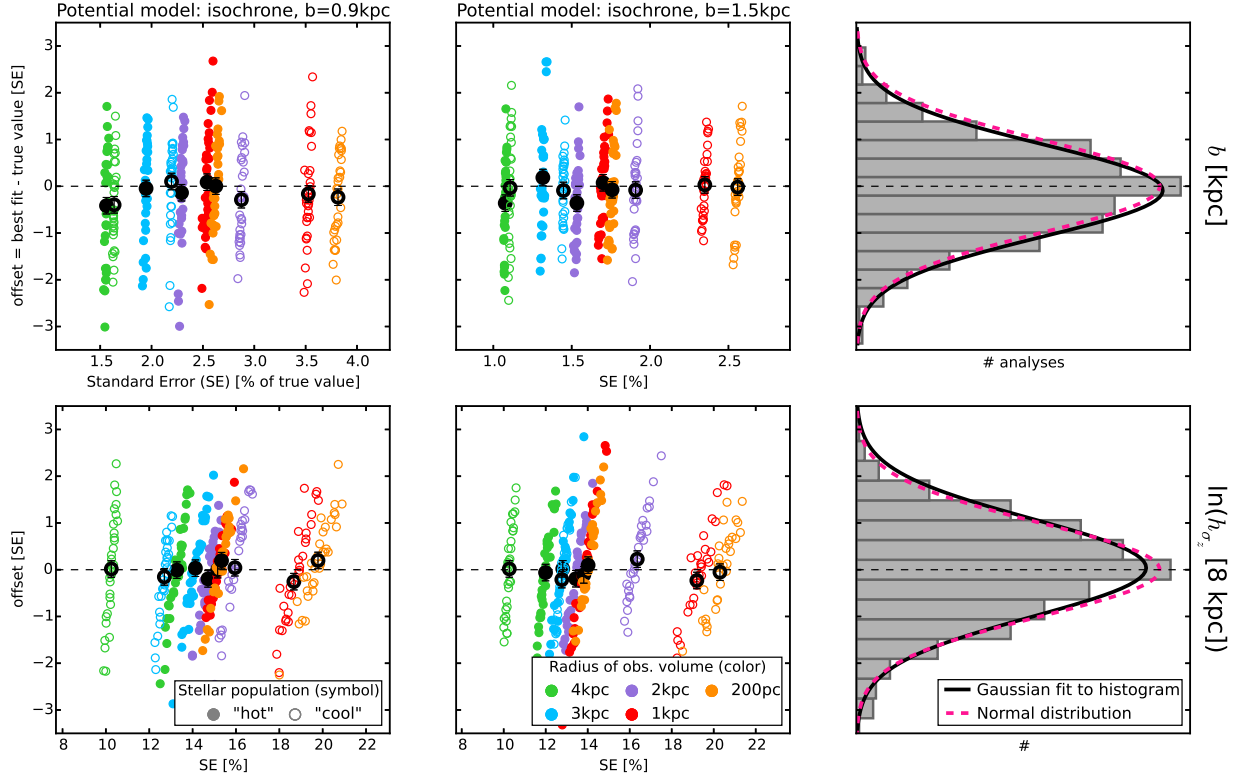


Fig. 6.— (Un-)bias of the parameter estimate: According to the central limit theorem the likelihood will follow a Gaussian distribution for a large number of stars. From this follows that also for a large number of data sets the corresponding best fit values for the model parameters have to follow a 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 ?? 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  $h_{\sigma_z}$ , which corresponds to the scale length of the vertical velocity distribution. [TO DO: rename isochrone potential in title to ”Iso-Pot”.]



Fig. 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???] [TO DO: units of b in title?????????]

### 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 a number mock data sets, drawn from a one single  $p_M$ , but drawn from six different volume wedges (see §[TO DO CHECK]), as illustrated in the left panels 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 ( $4.5 \text{ kpc}^3$ , achieved by adjusting the angular width of the edges). The right panels of Fig.7 the illustrate the ability of *RoadMapping* to constrain model parameters (in this case two  $p_\Phi$  parameters). The two top right panels of Fig.7 illustrate that the radial extent and the maximal height above the mid-plane matter. In the case shown, the standard error of the estimated parameters is twice as large for the volume with small  $\Delta R$  and  $\Delta|z|$ ; unsurprisingly, in the axisymmetric context the larger  $\Delta\phi$  extent of that volume does not help to constrain the parameters. The panels in the bottom row explore whether the radial or vertical extent plays a dominant role: it appears that substantive radial and vertical extent are comparably important to constrain the parameters.

This Figure also implies that for these cases volume offsets in the radial or vertical direction have at most modest impact. While we believe the argument for significant radial and vertical extent is generic, we have not done a full exploration of all combinations of  $p_M$  and volumina. Figure 6 amplifies the same point: it illustrates that at given sample size, drawing the data – more sparsely – from a larger volume provides better  $p_M$  constraints.

**Stuff that needs to be further examined in fig. 7:**

TO DO There are biases. Do they get smaller with higher accuracy? Do they disappear for KKS potential?

TO DO Maybe skip first row of plots?

TO DO 'Larger is better' is also demonstrated in fig. 6

TO DO We could compare these results with similar results for KKS pot. If the latter has no biases, we can state that to avoid biases when using an non-Staeckel potential, one should use a volume with comparable  $R$  *and*  $z$  coverage, because for this the biases seem to be smallest.

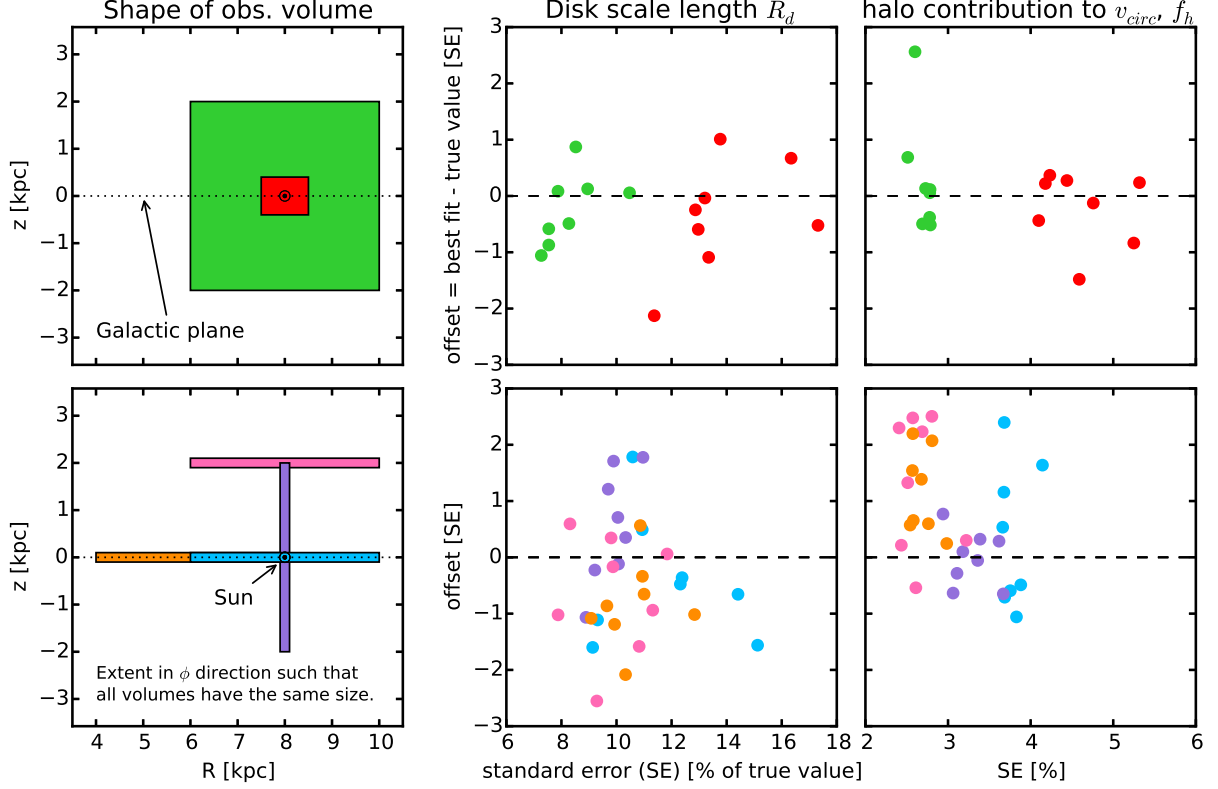


Fig. 7.— We demonstrate that for a given size of the observation volume the shape and position of the volume does not matter much as long as we have both large radial and/or vertical coverage. The left column shows the position of our test observation volumes within the Galaxy with respect to the Galactic plane and the sun. The angular extent of each wedge-shaped observation volume was adapted such that all have the volume of  $4.5 \text{ kpc}^3$ , even though their extent in  $(R, z)$  is different. Each data set contains 20,000 stars. We assume a Milky Way-like potential like in Bovy & Rix (2013), with  $p_\Phi = \{v_{\text{circ}}, R_d, z_h, f_h, \frac{d \ln v_c}{d \ln R}\} = \{230 \text{ km s}^{-1}, 2.5 \text{ kpc}, 400 \text{ pc}, 0.8, 0\}$  and a ‘hot’ stellar population with  $p_{\text{DF}} = \{h_R, \sigma_R, \sigma_z, h_{\sigma_R}, h_{\sigma_z}\} = \{2 \text{ kpc}, 55 \text{ km s}^{-1}, 66 \text{ km s}^{-1}, 8 \text{ kpc}, 7 \text{ kpc}\}$ . We evaluate the likelihood on a grid in the fit parameter  $\{R_d, f_h, \ln(h_R/8 \text{ kpc}), \ln(\sigma_R/230 \text{ km s}^{-1}), \ln(h_{\sigma_R}/8 \text{ kpc})\}$ . All other parameters are kept at their true values in the modelling. Standard error and offset were determined as in fig. 6. The accuracy of the analyses is  $N_{\text{velocity}} = 20$  and  $N_{\text{sigma}} = 4$ . In an axisymmetric potential the coverage in angular direction does not matter, as long as there are enough stars in the observation volume.

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

TO DO Do we explicitly 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?

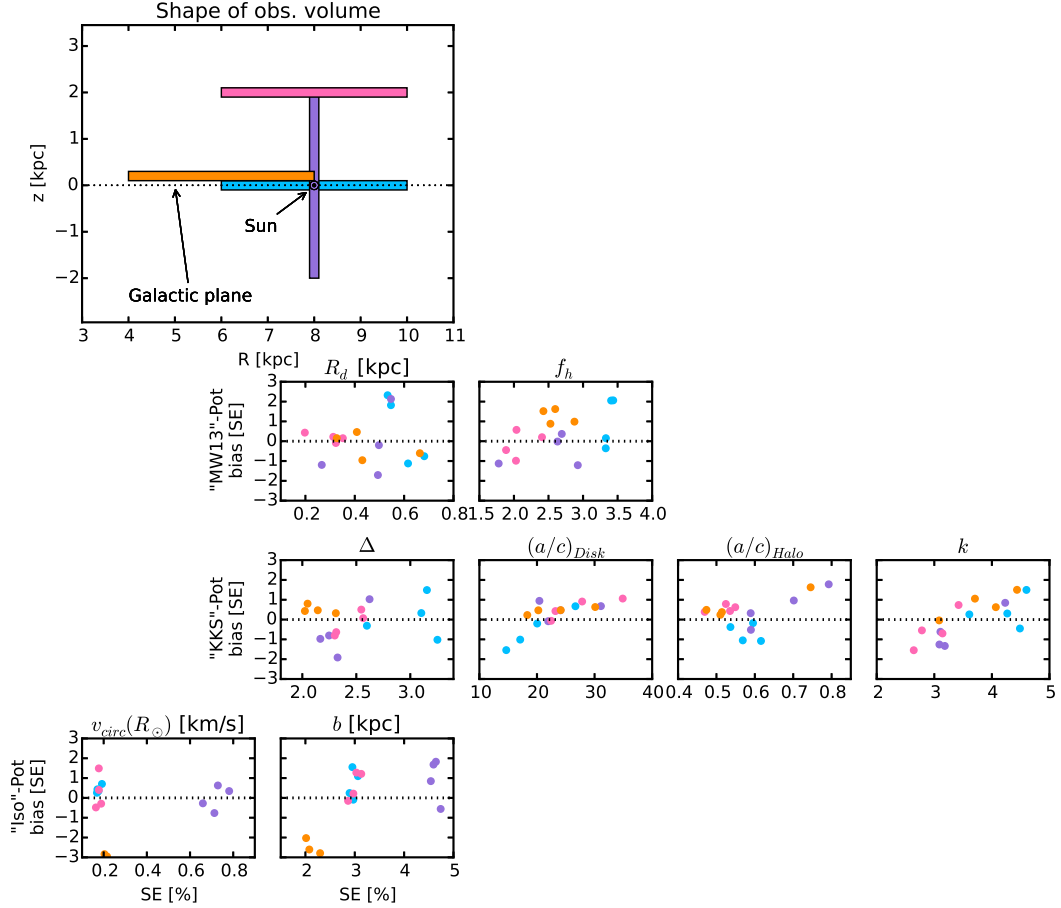


Fig. 8.— [TO DO: Caption] New plot (should replace Fig. 7) for 4 differently shaped volumes and 3 different kinds of potentials. Higher numerical accuracy in velocity integration. [TO DO: MW-Pot and KKS-Pot analyses suffer from too low accuracy in action calculation (with StaeckelGrid). Used the StaeckelGrid 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 tells us nothing. The orange Iso-Pot analysis suffers from too small integration range in  $vT$ . More coding required, before redoing this.]

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

The selection function of a survey is 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 §[TO DO CHECK]). 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 §?? we also show that the robustness is somewhat less striking but still given for small misjudgements of the incompleteness in vertical direction, parallel to the disk plane (fig. ?? and ??). This could model the effect of wrong corrections for dust obscurement in the plane. We also investigate in Appendix §?? 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. 17).

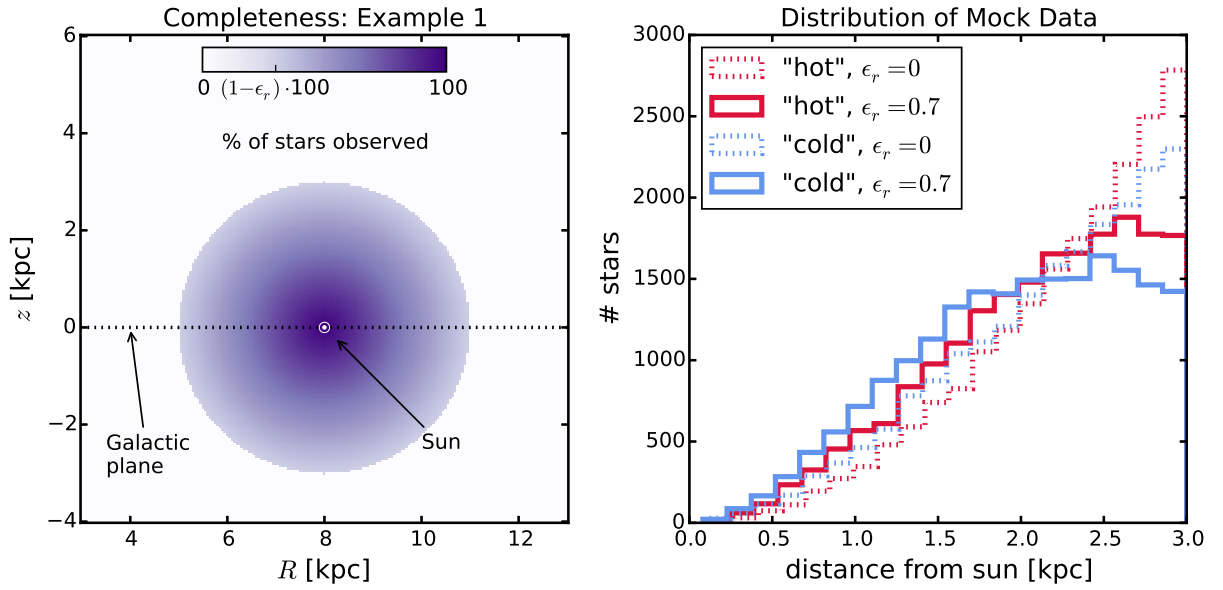


Fig. 9.— Selection function and mock data distribution for investigating radial incompleteness of the data. All model parameters are summarized as test ⑤, Example 1, in Table ???. The survey volume is a sphere around the sun and the percentage of observed stars is decreasing linearly with radius from the sun, as demonstrated in the left panel. How fast this detection/incompleteness rate drops is quantized by the factor  $\epsilon_r$ . Histograms for four data sets, drawn from two *MAPs* ("hot" in red and "cool" in blue, see table 2) and with two different  $\epsilon_r$ , 0 and 0.7, are shown in the right panel for illustration purposes.

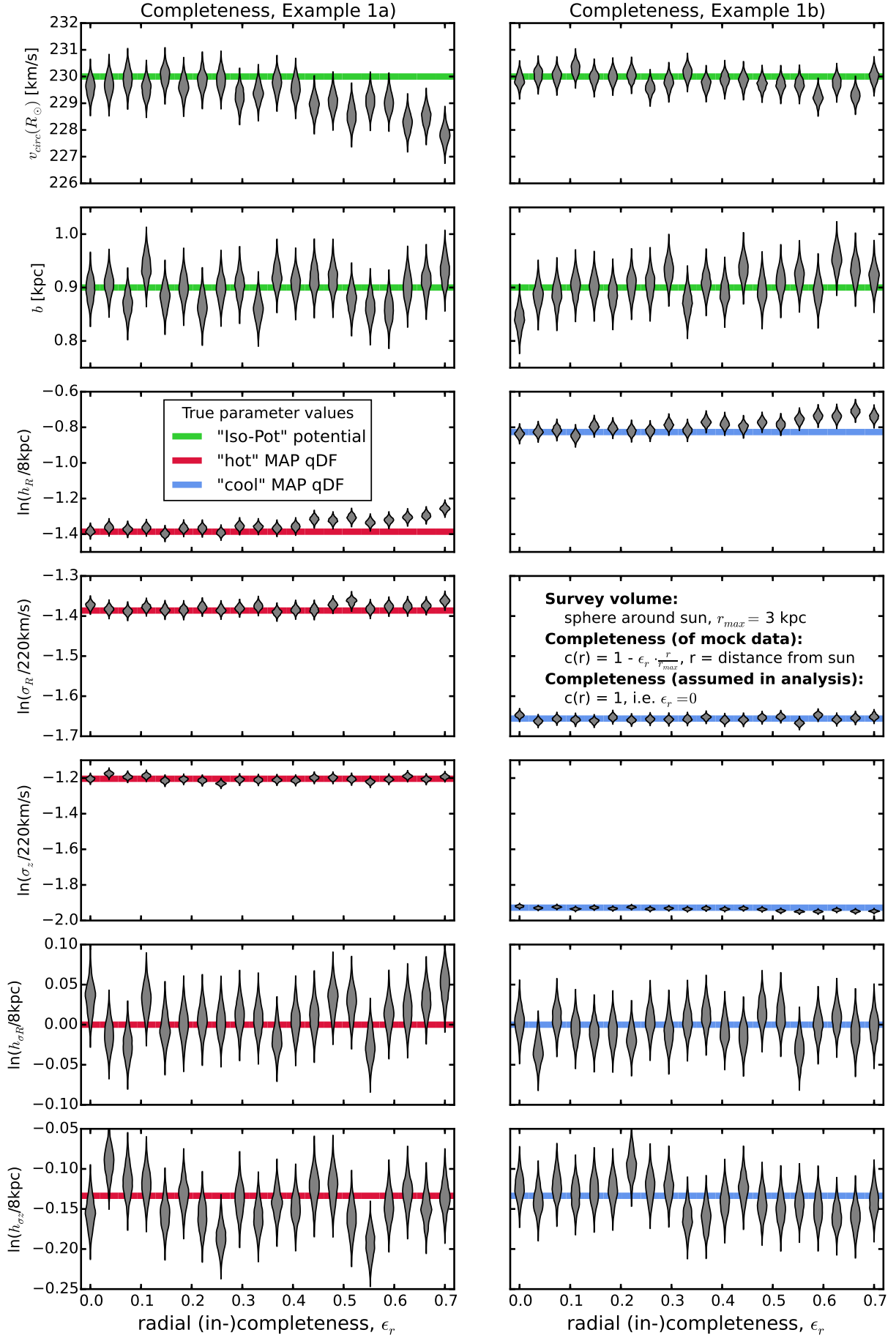


Fig. 10.— Influence of wrong assumptions about the radial incompleteness of the data on the



### 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. 12 and 13. 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. 12) ; in the second set of tests (“examples 2a/b”, Fig. 13), 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 MAPs 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.

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 from the hotter qDF dominate at small radii, while stars from the cooler qDF (with its longer tracer scale length) dominate at large radii. The recovered tracer scale lengths,  $h_R$  vary smoothly between the input values of the two qDFs that entered 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  $v_{\text{circ}}$  is more tightly constraint with cool MAP.

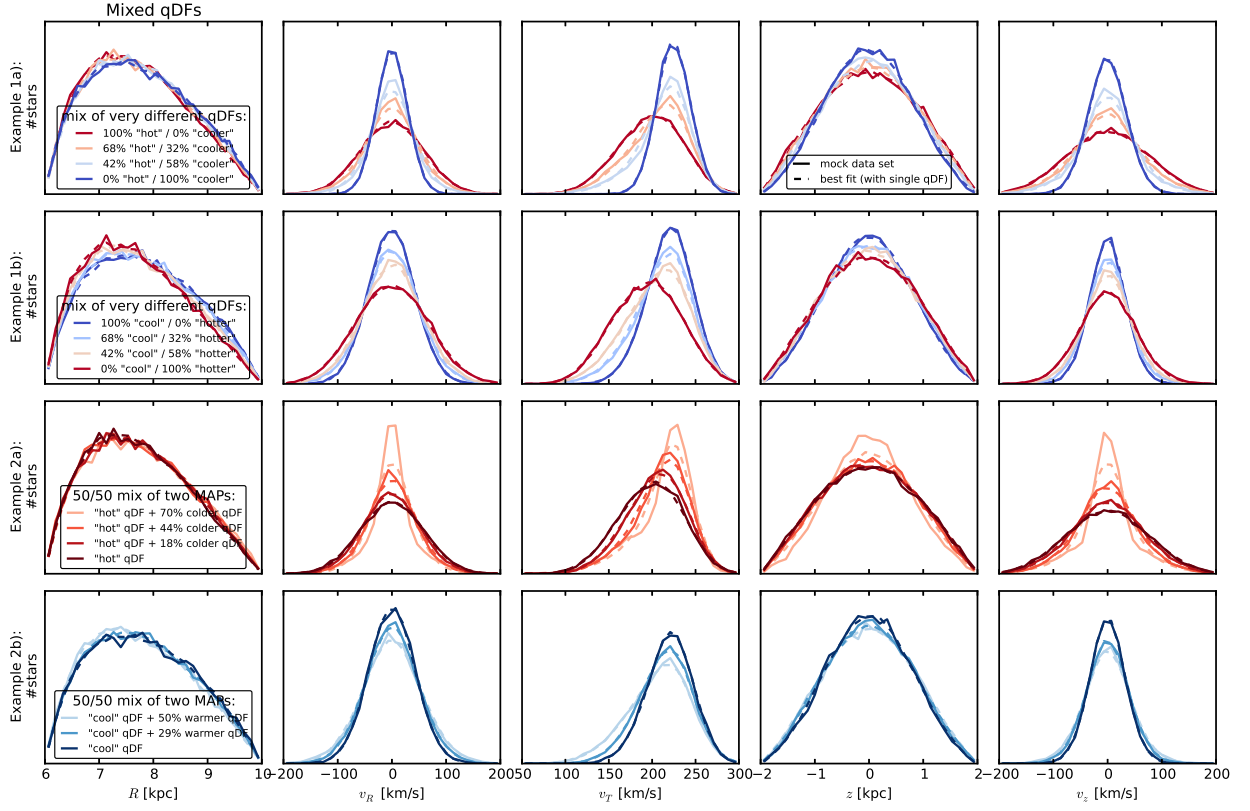


Fig. 11.— Distribution of mock data, created by mixing stars drawn from two different qDFs (solid lines), and the distribution predicted by the best fit of a single qDF and potential to the data (dashed lines). The model parameters to create the data are given in Table ?? as test ⑦, and the qDF parameters 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 MAPs 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. 12 and 13. 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. 12 and 13 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]

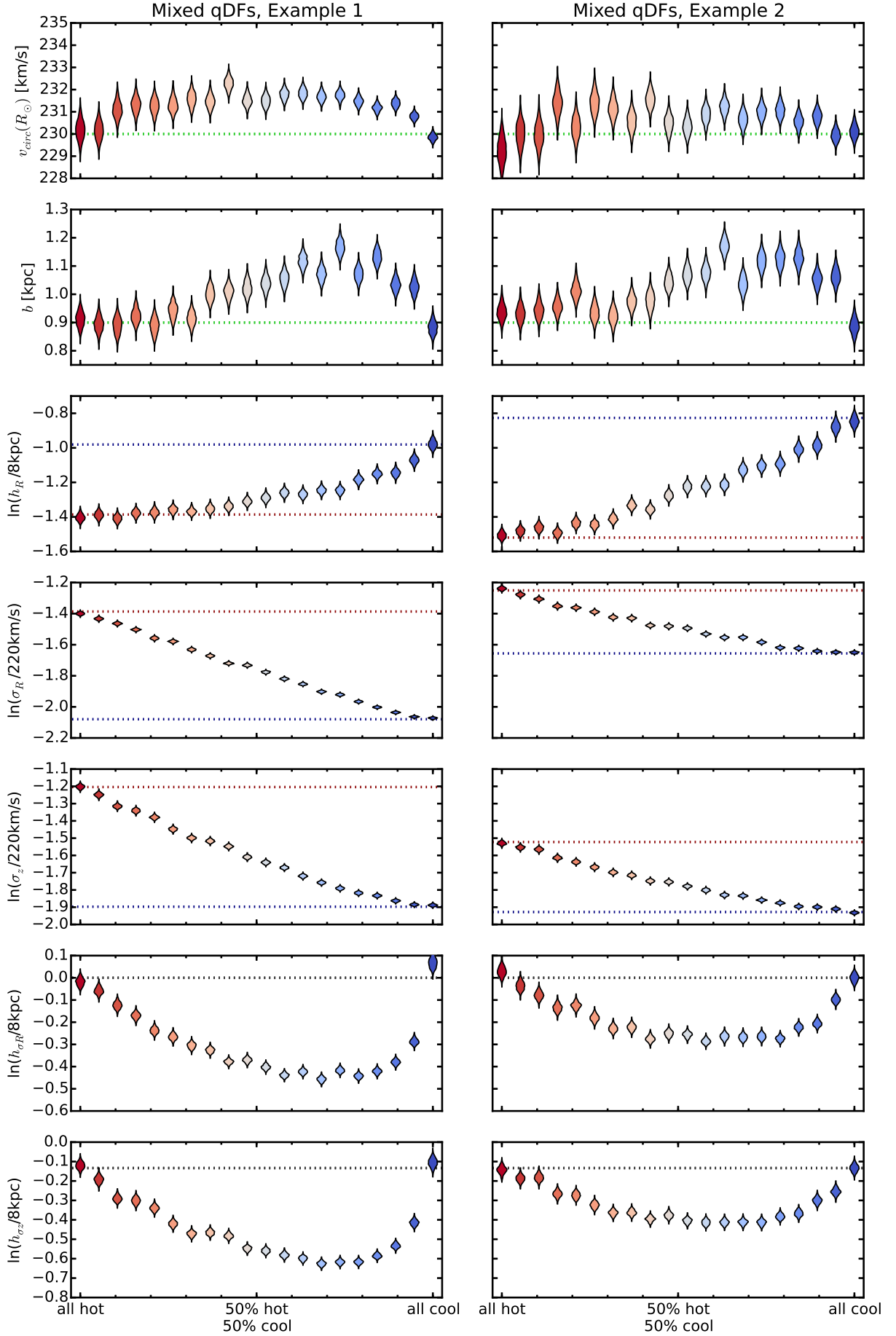


Fig. 12.— (Caption on next page.)

Fig. 12.— (Continued.) [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  $r_{\text{max}} = 2$  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]

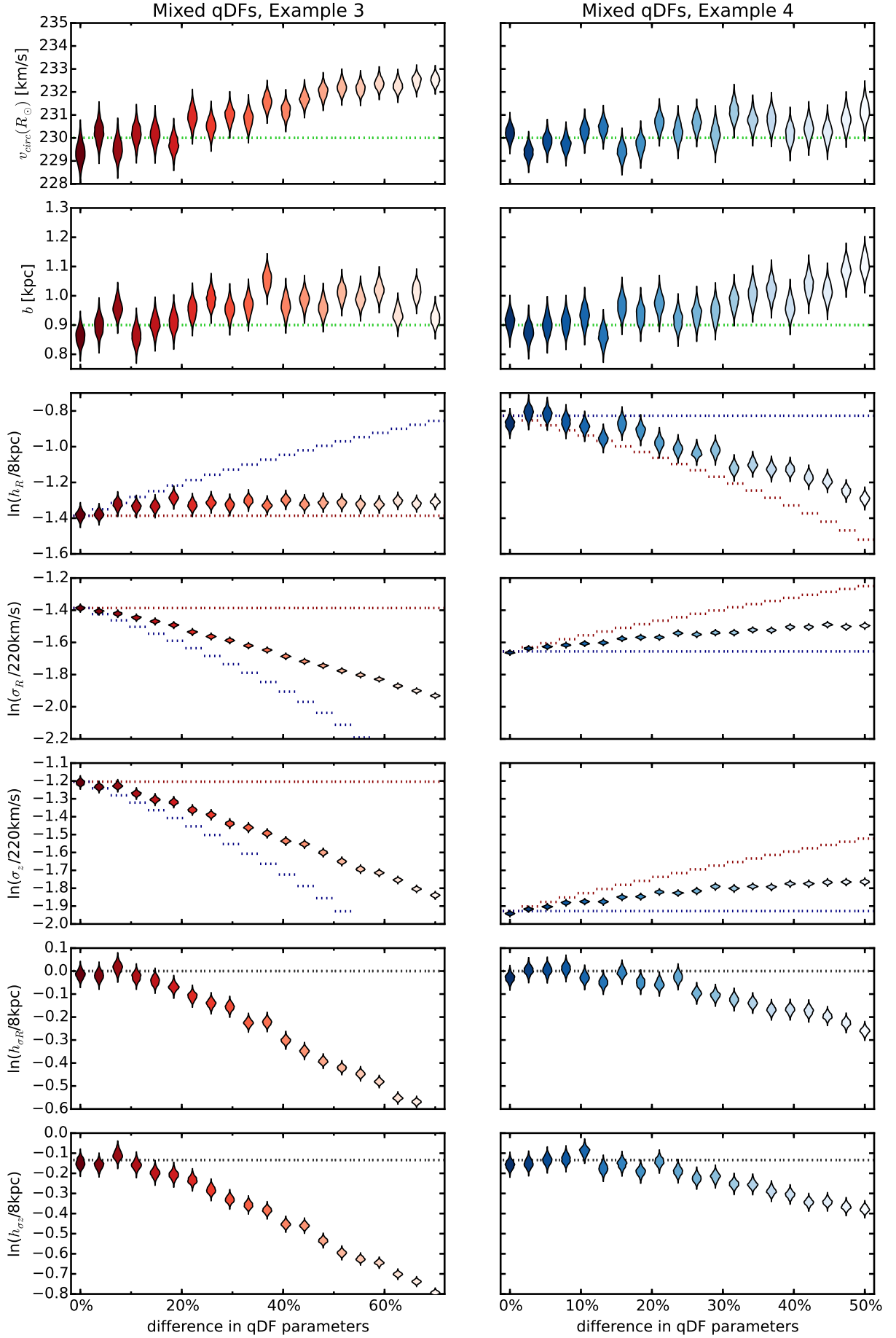


Fig. 13.— (Caption on next page.)

Fig. 13.— (Continued.) [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)  $\sigma_R$  and  $\sigma_z$  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  $r_{\max} = 2$  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 *MAPs* with varying difference in their qDF parameters, we model the effect of bin size in the  $[\text{Fe}/\text{H}]-[\alpha/\text{Fe}]$  plane when sorting stars into different *MAPs*: 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 *MAPs* 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])



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

In the long run we would like to incorporate a family of gravitational potential models in *RoadMapping* that is flexible enough to reproduce the essential features of the MW’s true mass distribution. Here we want to inspect if we can already give constraints on the true potential, even if our assumed potential is still too rigid - be it because of a low number of free potential parameters, or because our beliefs about the overall shape of the MW’s potential are slightly wrong. While our fundamental assumption of axisymmetry springs immediately to mind, being at odds with the obvious existence of a bar and spiral arms in the MW, we will not dive into investigating the implications in the scope of this paper. We rather focus on the 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. ???). The results are shown in Fig. 14.

The set of reference potential parameters of the “KKS-Pot” in Table ??? were found by adjusting the 2-component Kuzmin-Kutuzov Stäckel potential by Batsleer & Dejonghe (1994) such that it looks like the “MW14-Pot” from Bovy (2015): the radial and vertical force in  $R \in [4, 12]$  kpc,  $|z| \in [0, 4]$  kpc, and the rotation curve in  $R \in [0, 16]$  kpc (blue??? lines in Fig. 14). This could be understood as optimum, i.e. a fitting result from *RoadMapping* will most likely not be better than a fit directly to the potential. Even though the analysis results from *RoadMapping* shown in Fig. 14 (yellow??? lines) fit the overall density shape less than the optimum (blue???), we only used tracers within the survey volume (marked in red???). And within the survey volume we actually capture the radial and vertical gravitational force very well - and it is the forces to which the stars’ orbit are sensitive to. We also get the density structure of the disk inside the survey volume right, as well as the slope of the rotation curve at the sun. We used the “hot” MAP from Table ???, which has a short tracer scale length, i.e. probes the inner regions better than the outer regions. This could explain why the halo shape in the outer regions of the survey volume is less well recovered than the disk in the inner regions.

[TO DO:] Also do the same thing for a cold population + redo the hot population analysis with using fiducial qDF. Show violines for the qDF parameters.

We note that the precision of the potential recovery (as opposed to its accuracy) is very tight. This means that 20,000 stars seem already to be enough stars per MAP to be able to distinguish a “KKS-Pot”-like potential from a “MW14-Pot”-like potential, i.e. should encourage us to probe and compare different potential model families when actually fitting to real data sets of this size.

The potential model used by Bovy & Rix (2013) had only two free parameters (disk scale length and halo contribution to  $v_{\text{circ}}(R_{\odot})$ ). 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 *MAPs* of different scale length would probe different regions of the Galaxy best. Based on our results in Fig. 14 this seems to be indeed a sensible approach [TO DO: Check that this is indeed the case].

Our choice of fitting a superposition of two Stäckel potentials to the mock data was motivated by the work of Batsleer & Dejonghe (1994) and ?, who aimed to create MW-like Stäckel potentials from a superposition of several Kuzmin-Kutuzov Stäckel potentials. The big advantage of this approach is the exact and fast action calculation in such a potential, which allows to explore a bigger potential parameter space in the same computation time. Our results in Fig. 14 are also very encouraging that already two components alone can give relatively good constraints. Using more components could allow us also to model the bulge or to include more flexibility in modelling the disk structure.

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 *MAPs* as done by Bovy & Rix (2013), could be a promising approach to get the best possible constraints on the MW's potential.

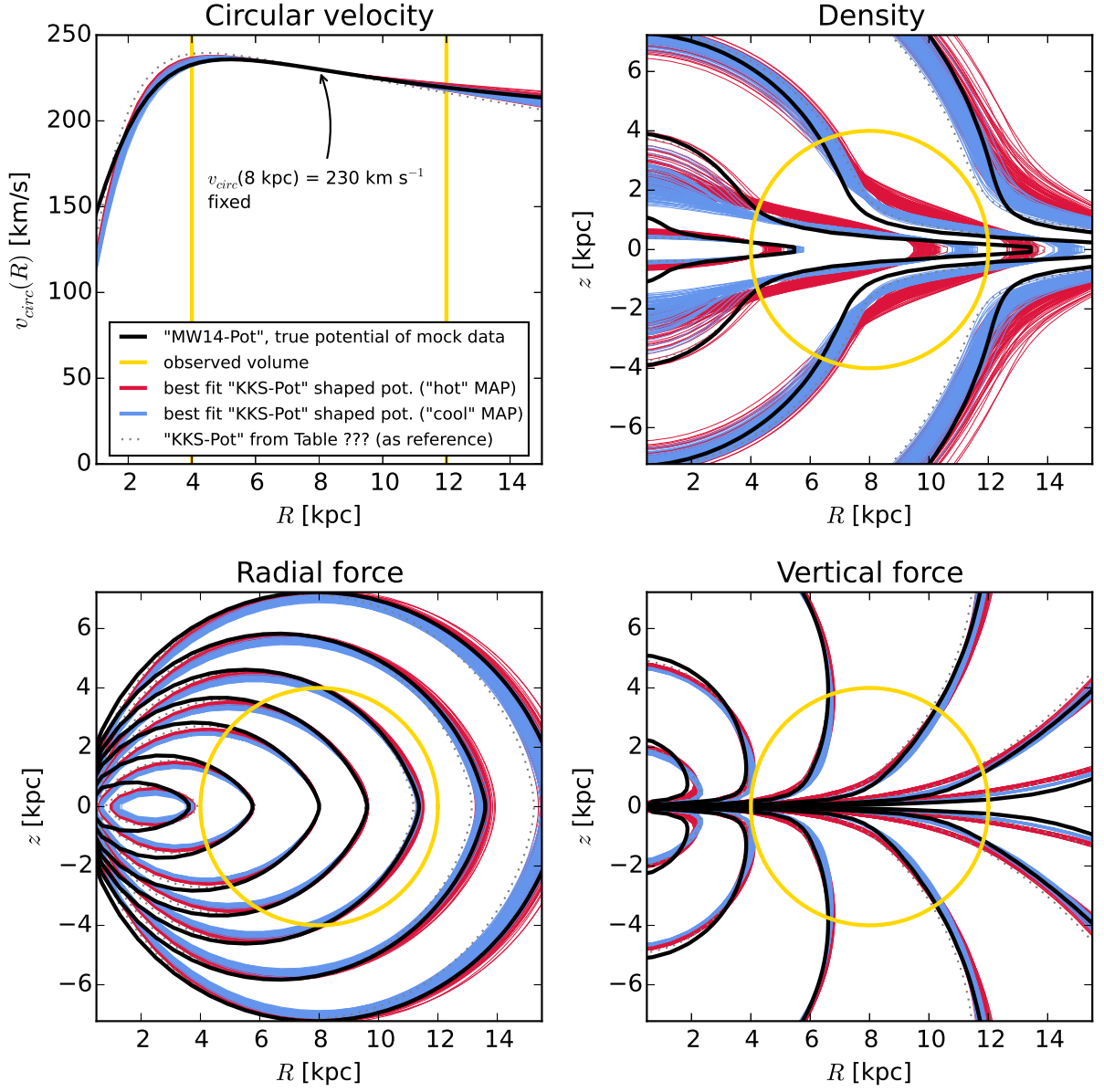


Fig. 14.— Caption [TO DO] [TO DO: Do more analyses??]

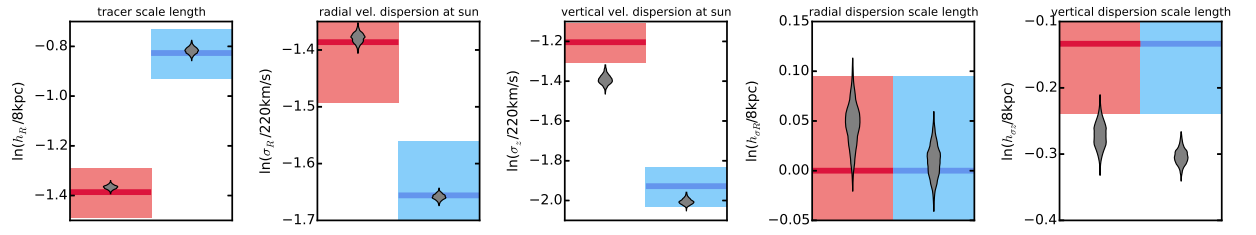


Fig. 15.— Caption [TO DO]

## 4. Discussion and Summary

[TO DO]

### • Objective of this Paper in a Nutshell

- 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 MAPs within the Galactic disk. – speed-up various aspects (Jo’s) Stckel fudge grid – sampling of parameters via MCMC – more free parameters – nested grid to initialize MCMC: no need to set fiducial DF by hand – potential source of bias; speed-up; – straightforward implementation of different potential families.. – straightforward implementation of more potential/DF parameters, full 3D selection function (some symmetries) – far higher accuracy – needed for larger samples; better convergence.....
- 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.

### • Summary of Results - Machinery

- Our fitting uses a full likelihood analysis, accounting for selection effects etc.
- The method is an asymptotically normal un-biased estimator and precision of the parameter recovery increases by  $1/\sqrt{N}$  with the number of stars.
- Getting un-biased results also for very large data sets, requires us to calculate the normalisation with very high precision → analysis of GAIA data will be a task for huge supercomputers and/or other ways to speed up the action calculation or the fitting routine have to be made.
- This was not yet a problem for the small sample sizes in BR13. Improvements in *RoadMapping* to make it more adept for large data sets and more free parameters: Nested-grid search with MCMC combination.

### • Summary of Results - Data

- *Survey volume position:* The position of the survey volume matters little i.e. there are no regions that are intrinsically manifestly more diagnostic than others. Only difference: closer to the disk & bulge there are more stars.

- *Survey volume & stellar population:* MAPs of different scale length and temperature probe different regions of the Galaxy (Bovy & Rix 2013). However, there is no easy rule of thumb for which selection function and stellar population which potential and DF parameter is constrained best.
- *Survey volume shape:* [No definite result yet] A large coverage in both  $R$  and  $z$  however 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.
- *Incompleteness:* Surprisingly the method seems to be very robust against small to medium misjudgments in the incompleteness of the data. (We investigated a radial and a planar decrease in completeness while still assuming a complete data set.) This could be, because missing stars in the data set do not affect the connection between a star's velocity at a given position. A lot of information about the potential shape is stored in the rotation curve - but even when not including  $vT$  in the analysis, small misjudgments are not a deal breaker. This result is especially encouraging for future studies, because selection functions usually depend in a non-trivial, and often not perfectly known way on the observables.
- *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. [No definite result yet] 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 to do the error convolution using MC integration scales by  $N_{\text{error}} \propto (\sigma_{v,\text{max}})^2$  with the maximum velocity error at the edge of the sample.

## • Summary of Results - Model

- *Distribution Function - MAPs not following qDF:*
  - \* Assumption to test: Our modelling crucially depends on the assumption, that the Galactic disk consists of many MAPs, each following a qDF.
  - \* Test case: In Example 1 in §??? we investigated how well we can recover the potential parameters, if this assumption was not perfectly satisfied, i.e. the MAPs true DF does not perfectly follow a qDF. We consider 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).

- \* Result: We find that the former would give more reliable results for the potential parameter recovery.
  - \* Additional thought: At the same time, if we assume that the distribution of stars from one *MAP* is caused by radial migration away from the initial location of star formation, it is 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???
  - \* Conclusion: From this follows that focusing the analysis especially on hotter *MAPs* could be an advisable way to go in the future, if there is doubt if the stars truly follow the qDF.
- *Distribution Function - MAPs polluted by neighbouring MAPs*:
- \* Assumption to test: Another critical point is the binning of stars into *MAPs* depending on their metallicity and  $\alpha$  abundances.
  - \* Test case: Example 2 in §??? could be understood as a model scenario for decreasing bin sizes in the metallicity- $\alpha$  plane when sorting stars in different *MAPs*, assuming that there is a smooth variation of qDF within the metallicity- $\alpha$  plane and each *MAP* indeed follows a qDF.
  - \* Result & Conclusion: 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.
  - \* Comparison with BR13: 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 *MAPs* do not differ more than 20%. As fig. 12 and 13 suggests 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  $h_R$ , especially in the low- $\alpha$  ( $[\alpha/Fe] \lesssim 0.2$ ), intermediate-metallicity ( $[Fe/H] \sim -0.5$ ) *MAPs* - provided of course, that each bin contains 20,000 stars. 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 *MAPs* while still getting reliable results.
  - \* Additional thought: Discussing Sanders & Binney’s doubts here?
- *Potential model misjudgment*:
- \* Assumption to test: What happens if the true potential does not exactly be-

long to the family of fit potentials?

- \* Test case: Fitting 2-component Staechel potential to bulge+halo+disk potential.
- \* Result: [Not a definite result yet.] Quite successful. Of course an exact reproduction is not possible, but the best fit potential is not much worse than a fit of the potential directly to the forces and rotation curve.
- \* Additional thought: Using a Staechel potential as potential model would have the advantage of fast and exact action calculations. Plus, the model can be made more flexible by adding additional Staechel components (see e.g. Famaey & Dejonghe 2003 ???), e.g. more disk components of another roundish component for the bulge. And the model potential family will always deviate from the true potential model, so using a flexible enough Staechel has no disadvantages. Only problem: Disentangling the potential in distinct halo and disk etc. components would be more difficult.

• **Other Action/DF-based modelling** Compare to our own approach

- Piffl et al. 2014: They fitted a superposition of DFs (superposition of qDFs for cohorts in thin disk, single qDF for thick disk, other DF for halo) to the full RAVE data set. No chemical information at all. No treatment of selection function. Fitting of velocity dispersion histograms in density bins. Advantage: Avoiding problem of accurate normalisation calculation. Disadvantage: A lot of stellar information is not used.
- Sanders & Binney 2015: developed extended distribution functions, i.e. functions of both actions and metallicity for thin/thick disk + halo. Including of selection function. BUT: potential not fitted.
- Ultimate goal: Fitting both a very flexible distribution function in both action/abundance space for the whole galaxy + flexible potential.
- Binney’s group’s focus: rather on developing eDFs, potential recovery more secondary.
- Our focus: getting very good constraints on potential with an optimum of simplicity in DF and flexibility in capturing the actual distribution.
- Our philosophy: We see our approach as intermediate/first step before ultimately using eDFs. We don’t think that there is an easy to find and formulate eDF that describes the distribution of stars in both action and abundance space (metallicity AND alpha), see fig. 6 in Bovy & Rix (2013). Bovy & Rix 2013 were fitting each



*MAP* separately with a qDF and potential and used the best fit potential model only to give a mass constraint at one radius. By doing so they/we

- a) separate out complexity / avoid explicitly dealing with the substructure of the DF in abundance space while still including abundance information (it's easier to see what goes right and wrong; less assumptions)
- b) make use of different *MAPs* constraining different regions of the potential best and are therefore less limited by having to get the potential model right.
- c) Quote HW: "It provides true cross-checking redundancy w.r.t. the potential estimates." [What does he mean by that?]

- **On the assumption of axisymmetry**

- Key assumption of the modelling: axisymmetry of the MW. Needed for action calculation/conservation.
- Reality: real disk has spiral arms and ring-like structures, with a warp and a flare in the outer disk. Also the Milky Way's halo has substructure, streams and shell-like overdensities. In the case of non-axisymmetry actions are not conserved anymore, so this could indeed affect the modelling.
- Does modelling still work? Has to be investigated in detail in future work, by trying to recover the potential in N-body simulations. Because actions are at least conserved under adiabatic changes of the potential, and the vertical action under radial migration, there is some hope, that it could still work.
- Intermediate step to Ultimate goal: Ultimate goal: Finding substructures observationally and describing theoretically the structure and evolution of potential perturbations. Including non-axisymmetries also in modelling. Could be approached as applying perturbations to an equilibrium model - the axisymmetric model.
- Other applications: a) Axisymmetric model also helps finding and explaining substructures. b) As we are still far away from a completely realistic MW potential model, the axisymmetric case will be our reference to turn  $x, v$  into orbits, needed as tracers of Galaxy evolution.

## 5. Conclusion

[TO DO]

## A. Appendix

### A.1. 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 explanation for this robustness could be, that a lot of information about the potential comes from the rotation curve, which is not affected by incompleteness, is demonstrated in Fig. 17.

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

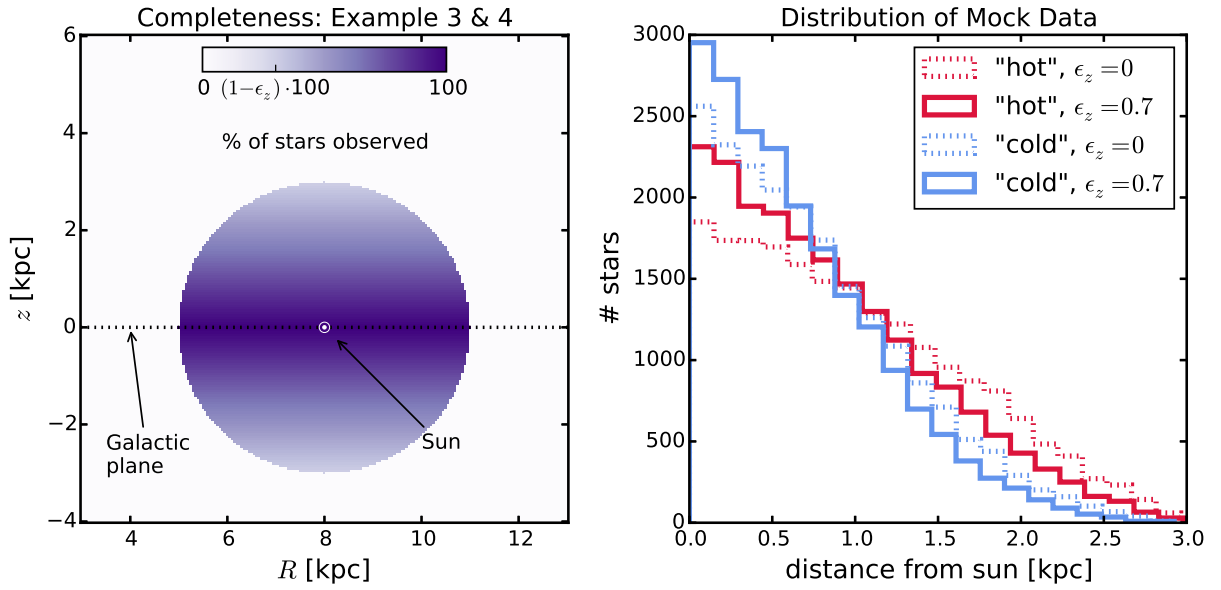


Fig. 16.— Selection function and mock data distribution for investigating vertical incompleteness of the data. All model parameters are summarized as test ⑤, Example 2, in Table ???. The survey volume is a sphere around the sun and the percentage of observed stars is decreasing linearly with distance from the Galactic plane, as demonstrated in the left panel. How fast this detection/incompleteness rate drops is quantized by the factor  $\epsilon_z$ . Histograms for four data sets, drawn from two MAPs (“hot” in red and “cool” in blue, see table 2) and with two different  $\epsilon_z$ , 0 and 0.7, are shown in the right panel for illustration purposes. [TODO: Re-do, if new analyses are in violin plot.]

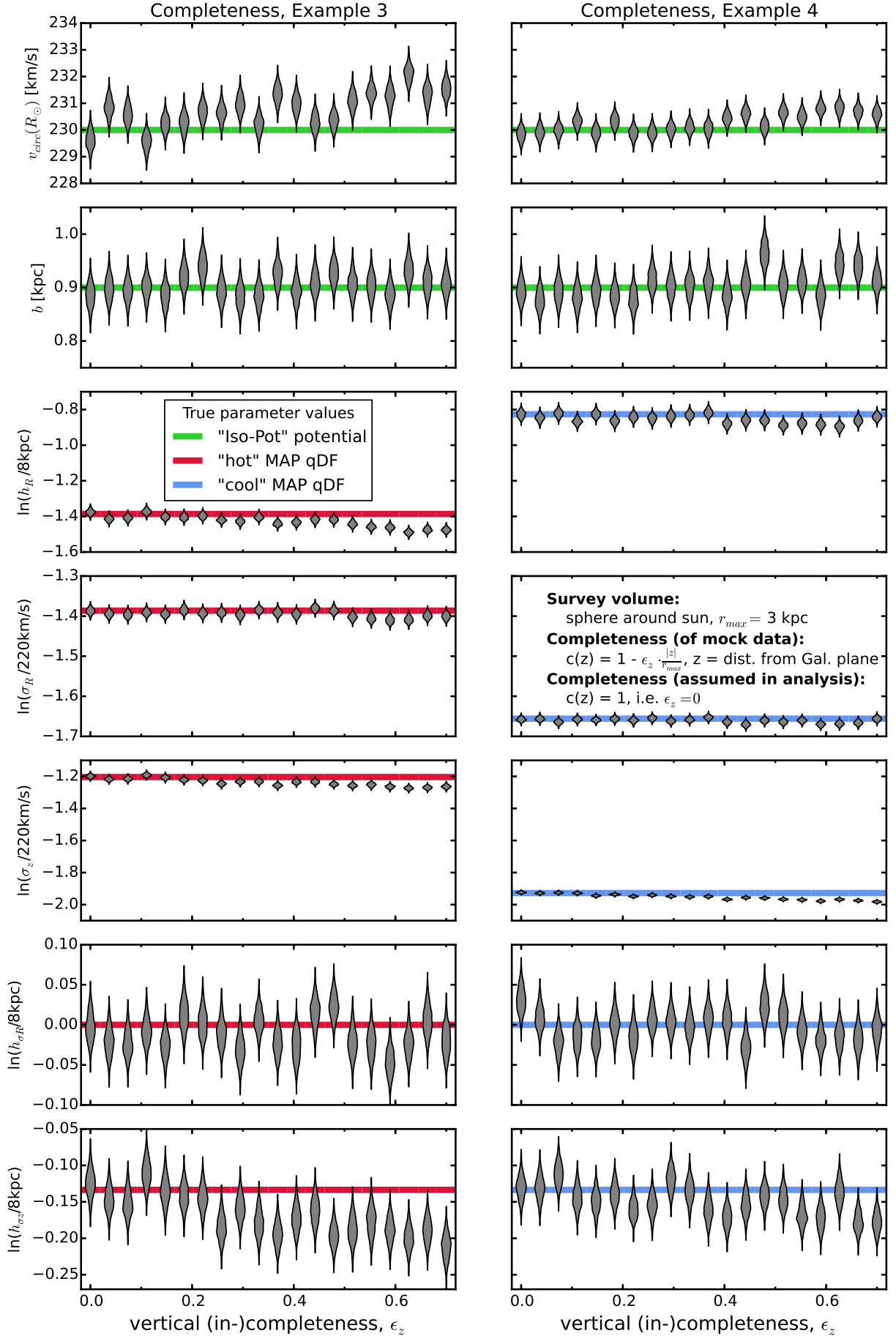


Fig. 17.— Influence of wrong assumptions about the incompleteness parallel to the Galactic

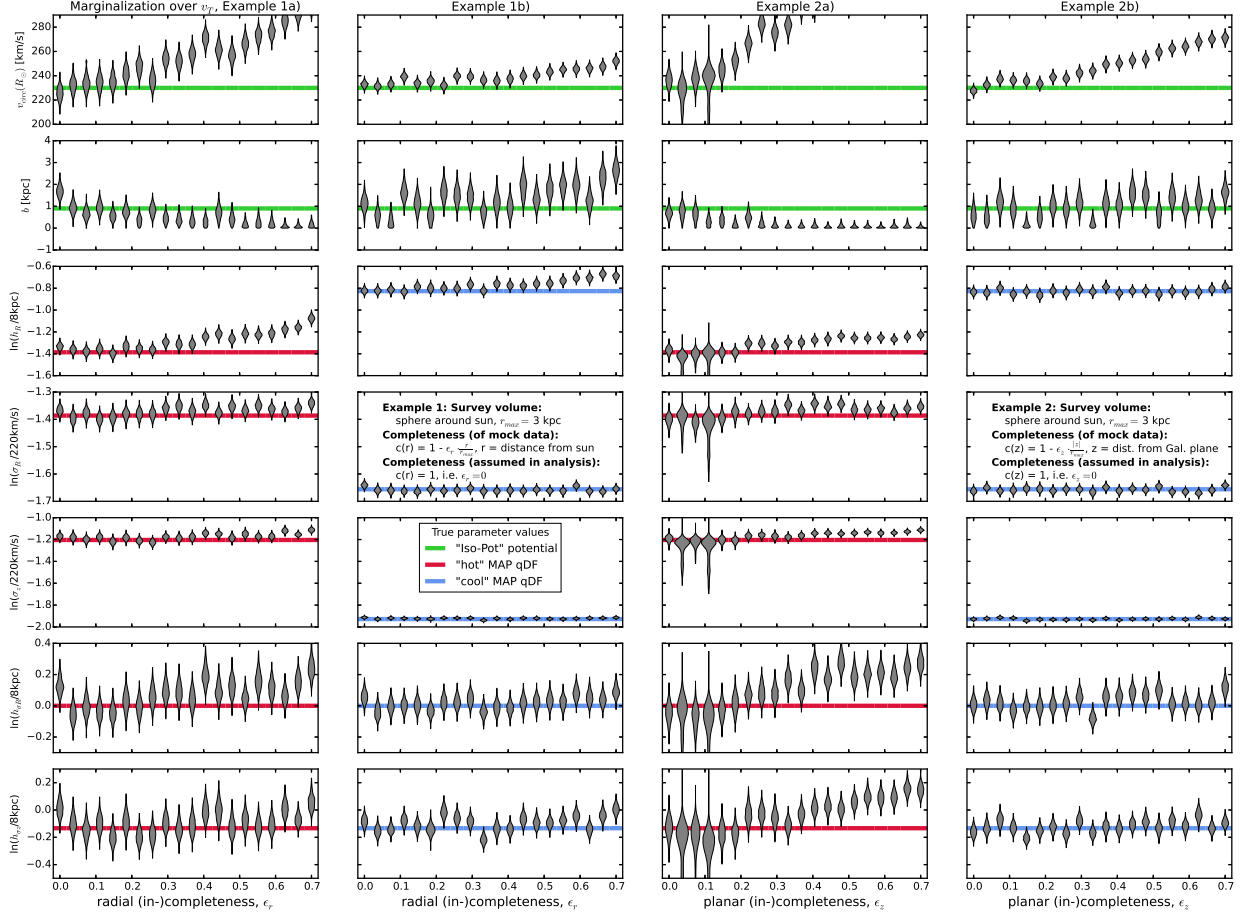


Fig. 18.— Influence of wrong assumptions about radial and vertical incompleteness on the parameter recovery, when *not* including information about the tangential velocities in the analysis. The mock data sets are the same as in Fig. 10 and 16, but this time we did not include the data coordinates  $v_T$  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  $b \neq 0$  was not excluded. ???] [TO DO: Rename Example 3 and 4 into 2a) and 2b), etc.]

## 2. Questions that haven't been covered so far:

- What limits the overall code speed?
- What happens, when the errors are not uniform?
- What if errors in distance matter for selection?
- Deviations from axisymmetry: Take numerical simulations.

**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.

## General Stuff

[TO DO: ] Rename everywhere  $N_{\text{sigma}}$  to  $n_{\text{interval}}$  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  $\sigma_R$  as profile or dispersion at sun.

[TO DO: ] Make consistent  $h_{\sigma_R}$   $\rightarrow$   $h_{\sigma,R}$



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

## REFERENCES

[TO DO]

Binney, J. J. 2010, MNRAS, 401, 2318

Binney, J. J., & McMillan, P. 2011, MNRAS, 413, 1889

Binney, J. J. 2012a, MNRAS, 426, 1324

Binney, J. J. 2012b, MNRAS, 426, 1328 (Princeton University Press)

Binney, J., & Tremaine, S. 2008, [TO DO: Galactic Dynamics???

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

[TO DO] Bovy (2015) Galpy paper

Dehnen, W., & Binney, J. 1998, MNRAS, 294, 429

Gilmore, G. & Reid, N. 1983, MNRAS, 202, 1025

McMillan, P. 2011, MNRAS, 414, 2446

Ness, M., Hogg, D. W., Rix, H.-W. et al., 2015 [TO DO????]

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

Sackett, P. 1997, ApJ, 483, 103

[TO DO] Sanders & Binney (2015) Extended distribution functions for our Galaxy

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

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

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.???