

## ACTION-BASED DYNAMICAL MODELLING FOR THE MILKY WAY DISK

WILMA H. TRICK<sup>1,2</sup>, JO BOVY<sup>3</sup>, AND HANS-WALTER RIX<sup>1</sup>

*Draft version April 10, 2016*

### ABSTRACT

We present *RoadMapping*, a full-likelihood dynamical modelling machinery that aims to recover the Milky Way’s (MW) gravitational potential from large samples of stars in the Galactic disk. *RoadMapping* models the observed positions and velocities of stars with a parametrized, three-integral distribution function (DF) in a parametrized axisymmetric potential. We create and analyse a large suite of mock data sets and develop qualitative “rules of thumb” how data, model and machinery affect constraints on the potential and DF most. We investigate in isolated test cases how the breakdown of model assumptions and properties of the data affect constraints on the potential and DF by creating and analysing a large suite of idealized mock data sets. Overall, we find that the potential can be reliably recovered with *RoadMapping*, even if the model assumptions are slightly wrong. Our key results are: (i) If the MW’s true gravitational potential is not included in the assumed family of parametrized model potentials, we can—at least in the axisymmetric case—still find a robust approximation for the potential that robustly approximates the potential within the limitations of the model. (ii) Modest systematic differences between the true and best-fit model DF are inconsequential, e.g., when . When defining sub-populations with simple DFs by binning stars according to their chemical abundances, finite bin sizes and abundance errors, binning errors do not affect the modelling as long as the DF parameters of neighbouring bins do not differ by more than 20% < 20%. *RoadMapping* gives constraints of high precision (i) for large sample sizes, (ii) for survey volumes of large radial and vertical coverage, and (iii) as long as the proper motion uncertainties are well known and even as large as 5 mas yr<sup>-1</sup>. In addition, *RoadMapping* ensures unbiased potential estimates. Unbiased potential estimates are ensured, (iii)(i) for small to moderate misjudgements of the spatial selection function (i.e.,  $\lesssim 15\%$  at the survey volume’s edge), (iv)(ii) if distances are known only to within 10% (which is satisfied for stars with  $G < 15$  mag from Gaia), or (v)(iii) if proper motion uncertainties are known within 10%. Challenges are the rapidly increasing computational costs for high precision likelihood evaluations required for large sample sizes. Overall, *RoadMapping* is well suited to making precise new measurements of the MW’s potential with data from the 2017 Gaia release.

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

### 1. INTRODUCTION

Through dynamical modelling we can infer the Milky Way’s (MW) gravitational potential from stellar motions (Binney & Tremaine 2008; Binney 2011; Rix & Bovy 2013). Observational information on the 6D phase-space coordinates of stars is currently growing at a rapid pace, and will be taken to a whole new level in quantity and precision by the upcoming data from the Gaia mission (Perryman et al. 2001). Yet, rigorous and practical modelling tools that turn position-velocity data of individual stars into constraints both on the gravitational potential and on the distribution function (DF) of stellar orbits are scarce (Rix & Bovy 2013).

The Galactic gravitational potential is fundamental for understanding the MW’s dark matter and baryonic structure (McMillan 2012; Rix & Bovy 2013; Strigari 2013; Read 2014) and the stellar-population-dependent orbit DF is a basic constraint on the Galaxy’s formation history (Binney 2013; Sanders & Binney 2015).

There is a variety of practical approaches to dy-

namical modelling of discrete collisionless tracers, such as the stars in the MW, e.g., Jeans modelling (Kuijken & Gilmore 1989; Bovy & Tremaine 2012; Garbari et al. 2012; Zhang et al. 2013; Büdenbender et al. 2015), action-based DF modelling (with parametric DFs: Bovy & Rix 2013; Piffl et al. 2014; Sanders & Binney 2015; Das & Binney 2016; with non-parametric DFs: Magorrian 2014), torus modelling (McMillan & Binney 2008; McMillan & Binney 2012; McMillan & Binney 2013), or made-to-measure modelling (Syer & Tremaine 1996; de Lorenzi et al. 2007; or Hunt & Kawata 2014). Most of them—explicitly or implicitly—describe the stellar distribution through a DF. Not all of them avoid binning to exploit the full discrete information content of the data.

Recently, Binney (2012b) and Bovy & Rix (2013) proposed to constrain the MW’s gravitational potential by combining parametrized axisymmetric potential models with DFs that are simple analytic functions of the three orbital actions to model discrete data.

Bovy & Rix (2013) (BR13 hereafter) put this in practice by implementing a rigorous modelling approach for so-called mono-abundance populations (MAPs), i.e., sub-sets of stars with similar [Fe/H] and  $[\alpha/\text{Fe}]$  within the Galactic disk, which seem to follow simple DFs

<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> Department of Astronomy and Astrophysics, University of Toronto, 50 St. George Street, Toronto, ON, M5S 3H4, Canada

(Bovy et al. 2012a,b,c). Given an assumed (axisymmetric) model for the Galactic potential and action-based DF (Binney 2010; Binney & McMillan 2011; Ting et al. 2013) they calculated the likelihood of the observed  $(\vec{x}, \vec{v})$  for each MAP, using SEGUE G-dwarf stars (Yanny et al. 2009). They also accounted for the complex, but known selection function of the kinematic tracers (Bovy et al. 2012c). For each MAP the modelling resulted in an independent estimate of the same gravitational potential. Taken as an ensemble, they constrained the disk surface density over a wide range of radii ( $\sim 4 - 9$  kpc), and powerfully constrained the disk mass scale length and the stellar-disk-to-dark-matter ratio at the Solar radius.

BR13 made however a number of quite severe and idealizing assumptions about the potential, the DF and the knowledge of observational effects. These idealizations could plausibly translate into systematic errors on the inferred potential, well above the formal error bars of the upcoming surveys with their wealth and quality of data.

In this work we present *RoadMapping* (“Recovery of the Orbit Action Distribution of Mono-Abundance Populations and Potential INference for our Galaxy”)—an improved, refined, flexible, robust and well-tested version of the original dynamical modelling machinery by BR13. Our goal is to explore which of the assumptions BR13 made and which other aspects of data, model and machinery limit *RoadMapping*’s recovery of the true gravitational potential.

We investigate the following aspects of the *RoadMapping* machinery that become especially important for a large number of stars: (i) Numerical inaccuracies must not be an important source of systematics (Section ??2.6). (ii) As parameter estimates become much more precise, we need more flexibility in the potential and DF model and efficient strategies to find the best fit parameters. The improvements made in *RoadMapping* as compared to the machinery used in BR13 are presented in Section 2.8. (iii) *RoadMapping* should be an unbiased estimator (Section 3.1).

We also explore how different aspects of the observational experiment design impact the parameter recovery: (i) We consider the importance of the survey volume geometry, size, shape and position within the MW to constrain the potential (Section 3.2). (ii) We ask what happens if our knowledge of the sample selection function is imperfect, and potentially biased (Section 3.3). (iii) We investigate how to best account for individual, and possibly misjudged, measurement uncertainties (Section 3.4). (iv) We determine which of several stellar sub-populations is best for constraining the potential (Section 3.7).

One of the strongest assumptions is to restrict the dynamical modelling to a certain family of parametrized functions for the gravitational potential and the DF. We investigate how well we can hope to recover the true potential, when our models do not encompass the true DF (Section 3.5) and potential (Section 3.6).

For all of the above aspects we show some plausible and illustrative examples on the basis of investigating mock data. The mock data is generated from galaxy models presented outlined in Sections 2.1–??2.4 following the procedure in Section Appendices ??A-B, and analysed ac-

cording to the description of the *RoadMapping* machinery in Sections ??2.5–2.8. The results on the investigated modelling aspects are presented in Section 3 and summarized and discussed in Section 4. Section 3 compiles our results on the investigated modelling aspects. In particular, our key results about the systematics introduced by using wrong DF or potential models are presented in the Sections 3.5 and 3.6. Section 4 finally summarizes and discusses our findings.

## 2. DYNAMICAL MODELLING

In this section we summarize the basic elements of *RoadMapping*, the dynamical modelling machinery presented in this work, which in many respects follows BR13 and makes extensive use of the `galpy` Python package for galactic dynamics<sup>4</sup> (Bovy 2015).

### 2.1. Coordinate system

Our modelling takes place in the Galactocentric rest-frame with cylindrical coordinates  $\mathbf{x} \equiv (R, \phi, z)$  and corresponding velocity components  $\mathbf{v} \equiv (v_R, v_\phi, v_z)$ . If the stellar phase-space data is given in observed heliocentric coordinates, position  $\tilde{\mathbf{x}} \equiv (\text{RA}, \text{Dec}, m - M)$  in right ascension RA, declination Dec and distance modulus  $(m - M)$ , and velocity  $\tilde{\mathbf{v}} \equiv (\mu_{\text{RA}} \cdot \cos(\text{Dec}), \mu_{\text{Dec}}, v_{\text{los}})$  as proper motions and line-of-sight velocity, the data  $(\tilde{\mathbf{x}}, \tilde{\mathbf{v}})$  has to be converted into the Galactocentric rest-frame coordinates  $(\mathbf{x}, \mathbf{v})$  using the Sun’s position and velocity. We assume for the Sun

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

### 2.2. Actions

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

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

which is evaluated along the orbit with position  $\mathbf{x}(t)$  and momentum  $\mathbf{p}(t)$  in a given potential  $\Phi$ . Actions have several convenient properties which make them excellent orbit labels and ideal as arguments for orbit DFs: They are integrals of motion; they have an intuitive physical meaning as they quantify the amount of oscillation of the orbit in each coordinate direction; together with a set of angle coordinates  $\boldsymbol{\theta}$  actions form a set of canonical conjugate phase-space coordinates, i.e., the Jacobian determinant  $|\partial(\mathbf{J}, \boldsymbol{\theta})/\partial(\mathbf{x}, \mathbf{v})| = 1$ . The angles  $\boldsymbol{\theta}(t) \propto t$  evolve linearly in time and specify the position of the star along the orbit. (For a full introduction to angle-action variables see Binney & Tremaine 2008, §3.5.)

Orbits in an axisymmetric gravitational potential  $\Phi$  are best described and fully specified by the three actions  $\mathbf{J} \equiv (J_R, J_z, J_\phi \equiv L_z)$  (Binney & Tremaine 2008, §3.5).

Action calculation from a star’s phase-space coordinates,  $(\mathbf{x}, \mathbf{v}) \xrightarrow{\Phi} \mathbf{J}$ , is typically very computationally expensive. Only for or some special, separable

<sup>4</sup> `galpy` is an open-source code that is being developed on <http://github.com/jobovy/galpy>. The latest documentation can be found at <http://galpy.readthedocs.org/en/latest/>.

potentials Equation (1) simplifies significantly. The spherical isochrone potential 1959AnAp...22..126H and axisymmetric The triaxial Stäckel potentials (de Zeeuw 1985) are the most general (Galactic) potentials, that allow exact action calculations using a single quadrature (2008gady.book....B, §3.5.2 and §3.5.3). Some flattened axisymmetric Stäckel potentials are quite similar to our Galaxy's potential (Binney & Tremaine 2008, §3.5.3; Batsleer & Dejonghe 1994; Famaey & Dejonghe 2003). The spherical isochrone potential (Henon 1959; Binney & Tremaine 2008, §3.5.3) is the most general special case for which the action calculation is analytic without any integration. In all other potentials actions have to be numerically estimated; see Sanders & Binney (2016) for a recent review of action estimation methods. According to Sanders & Binney (2016) the best compromise of speed and accuracy for the Galactic disk is the We use the Stäckel fudge by Binney (2012a) for axisymmetric potentials. In addition we use and action interpolation grids (Binney 2012a; Bovy 2015) to speed up the calculation. The latter is one of the improvements employed by RoadMapping, which was not used in BR13.

### 2.3. Potential models

For the gravitational potential in our modelling we assume a family of parametrized models. We use: The MW-like potential from BR13 (MW13-Pot) with bulge, disks and halo A MW-like potential with disk, halo and bulge (DHB-Pot); the spherical isochrone potential (Iso-Pot); and the 2-component Kuzmin-Kutuzov Stäckel potential (Batsleer & Dejonghe 1994; KKS-Pot), which also displays a disk and halo structure. Table 1 summarizes all reference potentials used in this work together with their free parameters  $p_\Phi$ . The true circular velocity at the Sun for all potential models was chosen to be  $v_{\text{circ}}(R_\odot) = 230 \text{ km s}^{-1}$ . The density distribution of these potentials is illustrated in Figure 1. The Iso-Pot allows both accurate and particularly fast action calculations; we use it therefore for tests requiring a large number of analyses. The KKS-Pot and DHB-Pot were chosen for their more realistic galaxy shape and because their simple analytic form for  $\Phi(R, z)$  makes the computation of forces and densities quick and easy. The former allows also for exact action calculations, while the latter has physically more intuitive potential parameters. The parameter values of KKS-Pot and DHB-Pot in Table 1 were found by fitting them to the MW potential from Bovy (2015), MW14-Pot.

### 2.4. Stellar distribution functions

A stellar distribution function  $\text{DF}(\mathbf{x}, \mathbf{v})$  can be considered as the probability of a star to be found at  $(\mathbf{x}, \mathbf{v})$ . Using instead orbit DFs in terms of  $(\mathbf{J}, \boldsymbol{\theta})$  has the advantage that the distribution of stars in  $\boldsymbol{\theta}$  is uniform and the orbit DF reduces effectively to a function of the actions  $\mathbf{J}$  only. As  $|\partial(\mathbf{J}, \boldsymbol{\theta})/\partial(\mathbf{x}, \mathbf{v})| = 1$ , the function  $\text{DF}(\mathbf{J})$  can still be thought of as a probability in  $(\mathbf{x}, \mathbf{v})$ .

The action-based quasi-isothermal distribution function (qDF) by Binney (2010) and Binney & McMillan (2011) is a simple DF which we will employ as a specific example throughout this work to describe individual stellar sub-populations. This is motivated by the findings of Bovy et al. (2012a,b,c) and Ting et al. (2013) abouton

the simple phase-space structure of stellar MAPs and BR13's successful application. The qDF has the form

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

with some free parameters  $p_{\text{DF}}$  and

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

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

(Binney & McMillan 2011). Here  $R_g$ ,  $\Omega$ ,  $\kappa$  and  $\nu$  are functions of  $L_z$  and denote respectively the (guiding-center) guiding-center radius, circular, radial/epicycle and vertical frequency of the near-circular orbit with angular momentum  $L_z$  in a given potential. The term  $[1 + \tanh(L_z/L_0)]$  suppresses counter-rotation for orbits in the disk with  $L \gg L_0 L_z < L_0$  (with  $L_0 \sim 10 \text{ km s}^{-1} \text{ kpc}$ ).

Following BR13, we choose the functional forms

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

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

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

which indirectly set the stellar number density and radial and vertical velocity dispersion profiles. The qDF 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. RoadMapping allows to fit any number of DF parameters simultaneously, while BR13 kept  $\{\sigma_{R,0}, h_{\sigma,R}\}$  fixed. Throughout this work we make use of a few example stellar populations whose qDF parameters are given in in Table 2: Most tests use the hot and cool qDFs, which correspond to kinematically hot and cool populations, respectively. The warmer (cooler and colder) qDFs in Table 2 were chosen to have the same anisotropy  $\sigma_{R,0}/\sigma_{z,0}$  as the cool (hot) qDF, with  $X$  being a free parameter describing the temperature difference. Hotter populations have shorter tracer scale lengths (Bovy et al. 2012c) and the velocity dispersion scale lengths were fixed according to Bovy et al. (2012b).

One indispensable step in our dynamical modelling technique (Section ??2.5-2.6), as well as in creating mock data (Section ??Appendix A), is to calculate the (axisymmetric) spatial tracer density  $\rho_{\text{DF}}(\mathbf{x} | p_\Phi, p_{\text{DF}})$  for a

**Table 1**

Axisymmetric gravitational potential models used throughout this work. The potential parameters are fixed for the mock data creation at the values given in this table, which we subsequently aim to recover with *RoadMapping*. The parameters of **MW13-Pot**, **DHB-Pot** and **KKS-Pot** were chosen by eye to resemble the **MW14-Pot** (see Figure 1). We use  $v_{\text{circ}}(R_\odot) = 230 \text{ km s}^{-1}$  as the circular velocity at the Sun for all potentials in this work.

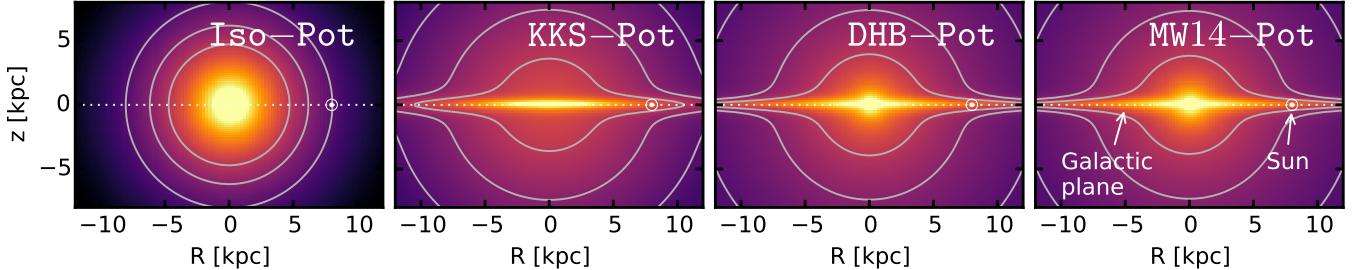
name	potential model	parameters $p_\Phi$		action calculation
<b>Iso-Pot</b>	isochrone potential <sup>(a)</sup> (Hénon 1959)	$b$	0.9 kpc	<b>analytic and exact</b> (Binney & Tremaine 2008, §3.5.2)
<b>KKS-Pot</b>	2-component Kuzmin-Kutuzov- Stäckel potential <sup>(b)</sup> (disk + halo) (Batsleer & Dejonghe 1994)	$\Delta$ $(\frac{a}{c})_{\text{Disk}}$ $(\frac{a}{c})_{\text{Halo}}$ $k$	0.3 20 1.07 0.28	<b>exact</b> using interpolation on action grid (Binney 2012a; Bovy 2015)
<b>DHB-Pot</b>	Disk+Halo+Bulge potential <sup>(c)</sup> : Miyamoto-Nagai disk, NFW halo, Hernquist bulge (same as <b>MW14-Pot</b> , except of bulge)	$a_{\text{disk}}$ $b_{\text{disk}}$ $f_{\text{halo}}$ $a_{\text{halo}}$ $f_{\text{bulge}}$ $a_{\text{bulge}}$	3 kpc 0.28 kpc (fixed) 0.35/0.95 16 kpc (fixed) 0.05/1.0 (fixed) 0.6 kpc (fixed)	<b>approximate</b> using <i>Stäckel fudge</i> (Binney 2012a) and interpolation on action grid
<b>MW14-Pot</b>	MW-like potential <sup>(d)</sup> : with Miyamoto-Nagaistar disk, NFW halo, cut-off power-law bulge (Bovy 2015)			<b>approximate</b> (same as <b>MW13-Pot</b> , <b>DHB-Pot</b> )

(a) The free parameter of the spherical **Iso-Pot** is the isochrone scale length  $b$ .

(b) The coordinate system of each of the two Stäckel-potential components of the **KKS-Pot** is  $R^2/(\tau_{i,p} + \alpha_p) + z^2/(\tau_{i,p} + \gamma_p) = 1$  with  $p \in \{\text{Disk}, \text{Halo}\}$  and  $\tau_{i,p} \in \{\lambda_p, \nu_p\}$ . Both components have the same focal distance  $\Delta \equiv \sqrt{\gamma_p - \alpha_p}$ , to ensure that the superposition itself is a Stäckel potential. The axis ratio of the coordinate surfaces  $(a/c)_p := \sqrt{\alpha_p/\gamma_p}$  describes the flatness of each component.  $k$  is the relative contribution of the disk mass to the total mass.

(c) The free parameters of the **MW13-Pot** are stellar disk scale length  $R_d$  and height  $z_d$ , the relative halo contribution to  $v_{\text{circ}}^2(R_\odot)$ ,  $f_h$ , and the slope of the rotation curve,  $[d \ln(v_{\text{circ}}(R_\odot))]/[d \ln(R)]$ . The parameters of the **DHB-Pot** are the Miyamoto-Nagai disk scale length  $a_{\text{disk}}$  and height  $b_{\text{disk}}$ , the NFW halo scale length  $a_{\text{halo}}$  and its relative contribution to  $v_{\text{circ}}^2(R_\odot)$  (with respect to the total disk+halo contribution),  $f_{\text{halo}}$ , and the Hernquist bulge scale length  $a_{\text{bulge}}$  and its contribution to the total  $v_{\text{circ}}^2(R_\odot)$ ,  $f_{\text{bulge}}$ .

(d) The **MWPotential2014** by Bovy (2015) (see their Table 1) has  $v_{\text{circ}}(R_\odot) = 220 \text{ km s}^{-1}$ . We use however  $v_{\text{circ}}(R_\odot) = 230 \text{ km s}^{-1}$ .



**Figure 1.** [This is a new version of the original plot - using the DHB-Pot instead of the MW13-Pot. (TO DO: Remove this note for final version.)] Density distribution of the four reference galaxy potentials in Table 1. These potentials are used throughout this work to create and model mock data with *RoadMapping*.

given DF and potential. Analogously to BR13,

$$\begin{aligned} & \rho_{\text{DF}}(R, |z| \mid p_\Phi, p_{\text{DF}}) \\ &= \int_{-\infty}^{\infty} \text{DF}(\mathbf{J}[R, z, \mathbf{v} \mid p_\Phi] \mid p_{\text{DF}}) d^3v \\ &\approx \int_{-n_\sigma \sigma_R(R \mid p_{\text{DF}})}^{n_\sigma \sigma_R(R \mid p_{\text{DF}})} \int_{-n_\sigma \sigma_z(R \mid p_{\text{DF}})}^{n_\sigma \sigma_z(R \mid p_{\text{DF}})} \int_0^{1.5v_{\text{circ}}(R_\odot)} \\ & \quad \text{DF}(\mathbf{J}[R, z, \mathbf{v} \mid p_\Phi] \mid p_{\text{DF}}) dv_T dv_z dv_R, \end{aligned} \quad (8)$$

where  $\sigma_R(R \mid p_{\text{DF}})$  and  $\sigma_z(R \mid p_{\text{DF}})$  are given by Equations (6) and (7).<sup>5</sup> Each integral is evaluated using a

<sup>5</sup> The integration ranges over the velocities are motivated by

$N_v$ -th order Gauss-Legendre quadrature. For a given  $p_\Phi$  and  $p_{\text{DF}}$  we explicitly calculate the density on  $N_x \times N_x$  regular grid points in the  $(R, z)$  plane and interpolate  $\log \rho_{\text{DF}}$  in between using bivariate spline interpolation. The grid is chosen to cover the extent of the observations (for  $|z| \geq 0$ , because the model is symmetric in  $z$  by construction). The total number of actions to be calculated to set up the density interpolation grid is  $N_x^2 \times N_v^3$ , which is one of the factors limiting the computation speed. To complement the work by BR13, we will specifically work

Figure A2 and  $n_\sigma$  should be chosen as  $n_\sigma \sim 5$  (see Figure 2). The integration range  $[0, 1.5v_{\text{circ}}(R_\odot)]$  over  $v_T$  is in general sufficient, only for observation volumes with larger mean stellar  $v_T$  this upper limit needs to be increased.

**Table 2**

Reference parameters for the qDF in Equations (2)-(7), used to create 6D phase-space mock data sets for stellar populations of different kinematic temperature. The parameters of the cooler & colder (warmer) qDFs were chosen to have the same anisotropy  $\sigma_{R,0}/\sigma_{z,0}$  as the hot (cool) qDF, with  $X$  being a free parameter describing the temperature difference. Hotter populations have shorter tracer scale lengths (Bovy et al. 2012c) and the velocity dispersion scale lengths were fixed according to Bovy et al. (2012b).

name	qDF parameters $p_{\text{DF}}$				
	$h_R$ [kpc]	$\sigma_{R,0}$ [km s $^{-1}$ ]	$\sigma_{z,0}$ [km s $^{-1}$ ]	$h_{\sigma,R}$ [kpc]	$h_{\sigma,z}$ [kpc]
hot	2	55	66	8	7
cool	3.5	42	32	8	7
cooler	3	27.5	33	8	7
colder	$2 + X\%$	$55 - X\%$	$66 - X\%$	8	7
warmer	$3.5 - X\%$	$42 + X\%$	$32 + X\%$	8	7

out in Section 2.6 and Figure 2 how large  $N_x$ ,  $N_v$  and  $n_\sigma$  have to be chosen to get the density with a sufficiently high numerical accuracy.

### 2.5. Data likelihood

As data  $D$  we consider here the positions and velocities of a sub-population of stars within a given survey selection function  $\text{SF}(\mathbf{x})$ ,

$$D \equiv \{\mathbf{x}_i, \mathbf{v}_i \mid (\text{star } i \text{ in given sub-population}) \wedge (\text{SF}(\mathbf{x}_i) > 0)\}.$$

(For simplicity we assume in most tests of this study contiguous SFs of spherical shape around the Sun, which are functions of  $\mathbf{x}$  only and which we motivate in Appendix B. The maximum radius of this spherical observed volume is denoted by  $r_{\max}$ .)

We fit a model potential and DF (here: the qDF) which are specified by a number of fixed and free model parameters,

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

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

The likelihood of the data given the model is, following BR13,

$$\begin{aligned} & \mathcal{L}(D \mid p_M) \\ & \equiv \prod_i^{N_*} p(\mathbf{x}_i, \mathbf{v}_i \mid p_M) \\ & = \prod_i^{N_*} \frac{\text{DF}(\mathbf{J}_i \mid p_M) \cdot \text{SF}(\mathbf{x}_i)}{\int \text{DF}(\mathbf{J} \mid p_M) \cdot \text{SF}(\mathbf{x}) d^3x d^3v} \\ & \propto \prod_i^{N_*} \frac{\text{DF}(\mathbf{J}_i \mid p_M)}{\int \rho_{\text{DF}}(R, |z| \mid p_M) \cdot \text{SF}(\mathbf{x}) d^3x}, \end{aligned} \quad (9)$$

where  $N_*$  is the number of stars in  $D$ , and in the last step we used Equation (8).<sup>6</sup>  $\prod_i \text{SF}(\mathbf{x}_i)$  is independent of  $p_M$ , so we treat it as unimportant proportionality factor. We find the best fitting  $p_M$  by maximizing the posterior

<sup>6</sup> Because  $|\partial(\mathbf{J}, \boldsymbol{\theta})/\partial(\mathbf{x}, \mathbf{v})| = 1$ , we can treat  $\text{DF}(\mathbf{J} \mid p_M)$  equivalently as a function of  $(\mathbf{x}, \mathbf{v})$ . The integration over phase-space in the normalisation term can be performed either over  $(\mathbf{J}, \boldsymbol{\theta})$  or  $(\mathbf{x}, \mathbf{v})$ .

probability distribution  $pdf(p_M \mid D)$ , which is, according to Bayes' theorem, proportional to the likelihood  $\mathcal{L}(D \mid p_M)$  times a prior  $p(p_M)$ .

$$pdf(p_M \mid D) \propto \mathcal{L}(D \mid p_M) \cdot p(p_M),$$

where  $p(p_M)$  is some prior probability distribution on the model parameters. We assume flat priors in both  $p_\Phi$  and

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

(see Section 2.4) throughout this work. Then  $pdf$  and likelihood are proportional to each other and differ only in units.

In the limit of uninformative priors a maximum-likelihood estimation procedure (e.g., via the expectationmaximization (EM) algorithm and parameter uncertainty estimates from the Fisher information matrix) would lead to the same result as the Bayesian inferential procedure described in this work. We expect however that in due course increasingly informative priors will become available (like, e.g., rotation curve measurements from maser sources by Reid et al. (2009)) and Bayesian inference is therefore the preferred framework.

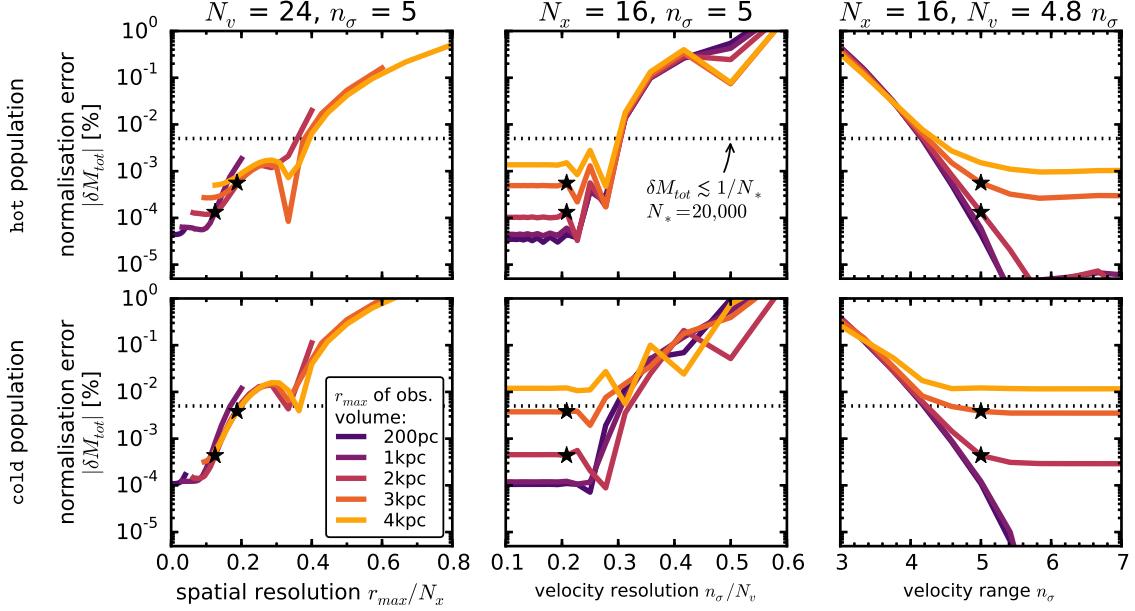
### 2.6. Likelihood normalisation

The normalisation in Equation (9) is a measure for the total number of tracers inside the survey volume,

$$M_{\text{tot}} \equiv \int \rho_{\text{DF}}(R, |z| \mid p_M) \cdot \text{SF}(\mathbf{x}) d^3x. \quad (11)$$

In the case of an axisymmetric Galaxy model and  $\text{SF}(\mathbf{x}) = 1$  within the observation volume (as in most tests in this work), the normalisation is essentially a two-dimensional integral in the  $(R, z)$  plane over  $\rho_{\text{DF}}$  with finite integration limits. We evaluate the integrals using Gauss-Legendre quadratures of order 40. The integral over the azimuthal direction can be solved analytically.

It turns out that a sufficiently accurate evaluation of the likelihood is computationally expensive, even for only one set of model parameters. This expense is dominated by the number of action calculations required, which in turn depends on  $N_*$  and the numerical accuracy of the tracer density interpolation grid with  $N_x^2 \times N_v^3$  grid points in Equation (8) needed for the likelihood normalisation in Equation (11). The accuracy of the normalisation has to be chosen high enough, such that the resulting numerical



**Figure 2.** [This is a new version of the original plot - using the DHB-Pot instead of the KKS-Pot, considering a more accurate calculation of the  $M_{\text{tot},\text{true}}$ , and we also considered a cool population to test the influence of the kinematic temperature. (TO DO: Remove this note for final version.)] Relative error of the likelihood normalisation,  $\delta M_{\text{tot}}$ , in Equation (12) depending on the accuracy of the grid-based density calculation in Equation (8) (and surrounding text) in five spherical observation volumes with different radius  $r_{\text{max}}$  and the kinematic temperature of the considered population in the DHB-Pot. (Test 1 in Table 3 summarizes the model parameters.) The tracer density in Equation (8) is calculated on  $N_x \times N_x$  spatial grid points in  $R \in [R_\odot \pm r_{\text{max}}]$  and  $|z| \in [0, r_{\text{max}}]$ . The integration over the velocities is performed with Gauss-Legendre quadratures of order  $N_v$  within an integration range of  $\pm n_\sigma$  times the dispersion  $\sigma_R(R)$  and  $\sigma_z(R)$  (and  $[0, 1.5v_{\text{circ}}]$  in  $v_T$ ). (We vary  $N_x$ ,  $N_v$  and  $n_\sigma$  separately and keep the other two fixed at the values indicated above each panel.) We calculate the “true” normalisation  $M_{\text{tot}}$  in Equation (12) with high accuracy as  $M_{\text{tot}} \equiv M_{\text{tot},\text{approx}}(N_x = 20, N_v = 56, n_\sigma = 7) M_{\text{tot}} \equiv M_{\text{tot},\text{approx}}(N_x = 32, N_v = 68, n_\sigma = 7)$ . The black dots indicate the accuracy used in our analyses: It is better than 0.001% (dotted line). The black stars indicate the accuracy used in analyses with the DHB-Pot, Tests 5 and 7: It is better than 0.005% (dotted line), which is required for  $N_* = 20,000$  stars. We find that the accuracy depends on the spatial resolution of the grid and requires more accurate integrations over the velocity for larger volumes within which the density varies more strongly. We find that the spatial resolution of the grid is important and depends on the kinematic temperature of the population, as cooler populations have a steeper density gradient in  $z$ -direction, which has to be sampled sufficiently.

error

$$\delta M_{\text{tot}} \equiv \frac{M_{\text{tot},\text{approx}}(N_x, N_v, n_\sigma) - M_{\text{tot}}}{M_{\text{tot}}} \quad (12)$$

does not dominate the numerically calculated log-likelihood, i.e.,

$$\begin{aligned} & \log \mathcal{L}_{\text{approx}}(D \mid p_M) \\ &= \sum_i^{N_*} \log \text{DF}(\mathbf{J}_i \mid p_M) - N_* \log(M_{\text{tot}}) \\ &\quad - N_* \log(1 + \delta M_{\text{tot}}), \end{aligned} \quad (13)$$

with

$$\log(1 + \delta M_{\text{tot}}) \leq \frac{1}{N_*}, \quad (14)$$

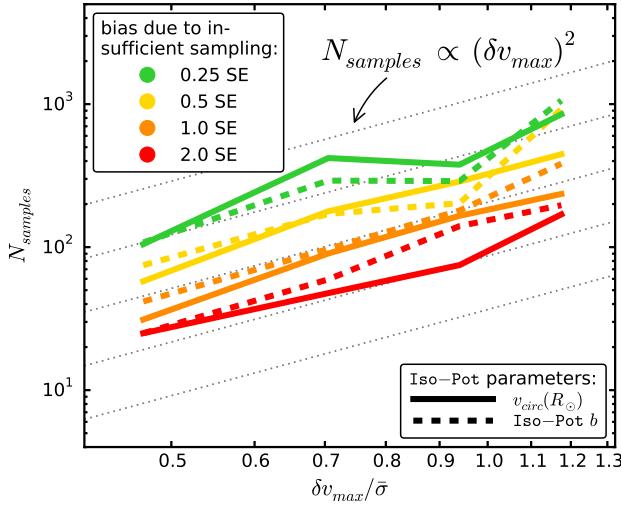
and therefore  $\delta M_{\text{tot}} \lesssim 1/N_*$ . Otherwise numerical inaccuracies could lead to systematic biases in the potential and DF recovery. For data sets as large as  $N_* = 20,000$  stars, which in the age of Gaia could very well be the case, one needs a numerical accuracy of 0.005% in the normalisation. We made sure that this is satisfied for all analyses in this work. Figure 2 demonstrates how that the numerical accuracy we use in the analysis for analyses with the DHB-Pot depends on the spatial and velocity resolution of the grid and that the accuracy we use,  $N_x = 16$ ,  $N_v = 24$  and  $n_\sigma = 5$ , does satisfy this require-

ment., is sufficient.<sup>7</sup> It has to be noted however, that the optimal values for  $N_x$ ,  $N_v$  and  $n_\sigma$  depend not only on  $N_*$ , but also on the kinematic temperature of the population (and to a certain degree even on the choice of potential<sup>8</sup>) and it has to be checked on a case-by-case basis what the optimal accuracy is. This is slightly higher than in BR13, where  $N_*$  was only a few  $\sim 100$ .

McMillan & Binney (2013), who use a similar modelling approach and likelihood normalisation, argued that the required accuracy for the normalisation scales as  $\log_{10}(1 + \delta M_{\text{tot}}) \leq 1/N_* \Rightarrow \delta M_{\text{tot}} \lesssim 2.3/N_*$ , which is satisfied for our tests as well. They evaluate the integrals in the normalisation via Monte-Carlo integration with  $\sim 10^9$  sample points in action space. Our approach uses a tracer density interpolation grid for which the resolution needs to be optimized by hand, but it has the advantage that it then only requires the calculation of  $N_x^2 \times N_v^3 \sim 4 \cdot 10^6 - 10^7$  actions per normalisation.

<sup>7</sup> The accuracy used in this work’s analyses is slightly higher than in BR13, where  $N_*$  was only a few  $\sim 100$ .

<sup>8</sup> In Figure 17 we will show a comparison for the qDF parameters of two very similar mock data distributions in two different potentials, the MW14-Pot and a best fit potential of the parametric form of the KKS-Pot. As some of the qDF parameters in both potentials are very different, and even more different from the actual physical scale lengths and velocity dispersions, an optimal  $n_\sigma$  has to be estimated first for a given potential model before running the RoadMapping analysis.



**Figure 3.** Number of MC samples  $N_{\text{samples}}$  needed for the numerical convolution of the model probability with the measurement uncertainties in Equation (16), given the maximum velocity uncertainty  $\delta v_{\text{max}}$  within the stellar sample with respect to the sample's kinematic temperature  $\bar{\sigma}$ . Insufficient sampling introduces systematic biases in the parameter recovery; the size of the bias (in units of the standard error (SE) on the parameter estimate) is indicated in the legend. The relation found here,  $N_{\text{samples}} \propto \delta v_{\text{max}}^2$ , was distilled from analyses (with different  $N_{\text{samples}}$ ) of mock data sets with different proper motion uncertainties  $\delta\mu \in [2, 5] \text{ mas yr}^{-1}$  in the absence of position uncertainties (see Test 2 in Table 3). The proper motion uncertainty  $\delta\mu$  translates to heteroscedastic velocity uncertainties according to  $\delta v [\text{km s}^{-1}] \equiv 4.74047 \cdot r [\text{kpc}] \cdot \delta\mu [\text{mas yr}^{-1}]$ , with  $r$  being the distance of the star from the Sun. Stars with larger  $\delta v$  require more  $N_{\text{samples}}$  for the integral over its measurement uncertainties to converge; we therefore show how the  $N_{\text{samples}}$ —needed for the *pdf* of the *whole* data set to be converged—depends on the *largest* velocity error  $\delta v_{\text{max}} \equiv \delta v(r_{\text{max}})$  within the data set. We used As the reference for the converged convolution integral, we used  $N_{\text{samples}} = 800$  and  $1200$  for  $\delta\mu \leq 3 \text{ mas yr}^{-1}$  and  $\delta\mu > 3 \text{ mas yr}^{-1}$ , respectively, as the reference for the converged convolution integral (see also left panels in Figure 9). We plot  $\delta v_{\text{max}}$  in units of the sample temperature, which we quantify by  $\bar{\sigma} \equiv (\sigma_{R,0} + \sigma_{z,0})/2$  (see Table 2 for the hot qDF). This figure was generated from mock data sets with  $N_* = 10,000$ . We found that for  $N_* = 5,000$  the required  $N_{\text{samples}}$  remains similar for  $b$ , but gets smaller for  $v_{\text{circ}}(R_\odot)$ . Overall we expect that we need less accuracy and therefore smaller  $N_{\text{samples}}$  for smaller  $N_*$ .

### 2.7. Measurement errors

Measurement uncertainties of the data have to be incorporated in the likelihood. We assume Gaussian uncertainties in the observable space  $\mathbf{y} \equiv (\tilde{\mathbf{x}}, \tilde{\mathbf{v}}) = (\text{RA}, \text{Dec}, (m - M), \mu_{\text{RA}} \cdot \cos(\text{Dec}), \mu_{\text{Dec}}, v_{\text{los}})$ , i.e., the  $i$ -th star's observed  $\mathbf{y}_i$  is drawn from the normal distribution  $N[\mathbf{y}'_i, \delta\mathbf{y}_i] \equiv \prod_i^6 N[y_{i,k}', \delta y_{i,k}] = \prod_i^6 \exp\{-(y_k - y_{i,k}')^2 / (2\delta y_{i,k}^2)\} / \sqrt{2\pi\delta y_{i,k}^2}$ , with  $\mathbf{y}'$  being the star's true phase-space position,  $\delta\mathbf{y}_i$  its uncertainty, and  $y_k$  the  $k$ -th coordinate component of  $\mathbf{y}$ . Stars follow the  $\text{DF}(\mathbf{J}[\mathbf{y}' | p_\Phi] | p_{\text{DF}}) \equiv \text{DF}(\mathbf{y}') \equiv$  for short convolved with the measurement uncertainties  $N[0, \delta\mathbf{y}_i]$ . The selection function  $\text{SF}(\mathbf{y})$  acts on the space of (uncertainty affected) observables. Then the probability of one star becomes

$$\begin{aligned} \tilde{p}(\mathbf{y}_i | p_\Phi, p_{\text{DF}}, \delta\mathbf{y}_i) \\ = \frac{\text{SF}(\mathbf{y}_i) \cdot \int \text{DF}(\mathbf{y}') \cdot N[\mathbf{y}_i, \delta\mathbf{y}_i] d^6 y'}{\int (\text{DF}(\mathbf{y}') \cdot \int \text{SF}(\mathbf{y}) \cdot N[\mathbf{y}', \delta\mathbf{y}_i] d^6 y) d^6 y'}. \quad (15) \end{aligned}$$

In the case of uncertainties in distance or (RA, Dec) the evaluation of this is computational *very* expensive—especially if the stars have heteroscedastic  $\delta\mathbf{y}_i$  and the normalisation needs to be calculated for each star separately. In practice we compute the convolution using Monte Carlo (MC) integration with  $N_{\text{samples}}$  samples,

$$\begin{aligned} \tilde{p}_{\text{approx}}(\mathbf{y}_i | p_\Phi, p_{\text{DF}}, \delta\mathbf{y}_i) \\ \approx \frac{\text{SF}(\tilde{\mathbf{x}}_i)}{M_{\text{tot}}} \cdot \frac{1}{N_{\text{samples}}} \sum_n^{N_{\text{samples}}} \text{DF}(\tilde{\mathbf{x}}_i, \mathbf{v}[\mathbf{y}'_{i,n}]) \quad (16) \end{aligned}$$

with

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

The above approximation also assumes that the star's position  $\tilde{\mathbf{x}}_i$  is perfectly measured. As the SF is also velocity independent, this simplifies the normalisation drastically to Equation (11). Measurement uncertainties in RA and Dec are often negligible anyway. The uncertainties in the Galactocentric velocities  $\mathbf{v}_i = (v_{R,i}, v_{T,i}, v_{z,i})$  depend *besides not only* on  $\delta\mu$  and  $\delta v_{\text{los}}$  but also on the distance and its uncertainty, which we do *not* neglect when drawing MC samples  $\mathbf{y}'_{i,n}$  from the full uncertainty distribution  $N[\mathbf{y}_i, \delta\mathbf{y}_i]$ . An analogous but one-dimensional treatment of measurement uncertainties in only  $v_z$  was already applied by BR13. Similar approaches ignoring measurement uncertainties in the likelihood normalisation were for example also used by McMillan & Binney (2013) and Das & Binney (2016). In Section 3.4, Figure 9 (Test 6.2 in Table 3), we investigate the breakdown of this approximation for non-negligible distance uncertainties.

Figure 3 demonstrates that in the absence of position uncertainties the  $N_{\text{samples}}$  needed for the convolution integral to converge depends as

$$N_{\text{samples}} \propto (\delta v)^2$$

on the uncertainties in the (1D) velocities. Figure 3 is based on analyses of mock data sets with different proper motion uncertainties  $\delta\mu$  (see Test 2 in Table 3 for all parameters). The proper motion uncertainty  $\delta\mu$  translates to heteroscedastic velocity uncertainties according to

$$\delta v [\text{km s}^{-1}] \equiv 4.74047 \cdot r [\text{kpc}] \cdot \delta\mu [\text{mas yr}^{-1}],$$

with  $r$  being the distance star—Sun. Stars with larger  $\delta v$  require more  $N_{\text{samples}}$  for the integral over their measurement uncertainties to converge; Figure 3 therefore shows how the  $N_{\text{samples}}$ —needed for the *pdf* of the *whole* data set to be converged—depends on the *largest* velocity error  $\delta v_{\text{max}} \equiv \delta v(r_{\text{max}})$  within the data set. These mock data sets contained each  $N_* = 10,000$  stars. We found that for  $N_* = 5,000$  the required  $N_{\text{samples}}$  to reach a given accuracy gets smaller for  $v_{\text{circ}}(R_\odot)$ , but remains similar for  $b$ . The former is consistent with our expectation that we need higher accuracy and therefore more  $N_{\text{samples}}$  for larger data sets. The latter seems to be a special property of the Iso-Pot (see also the discussion in Section 3.3).

We found that the required  $N_{\text{samples}}$  to reach a given accuracy does not depend strongly on the number of stars in the sample. But in general we expect that we need higher accuracy and therefore more  $N_{\text{samples}}$  for larger

data sets. A similar but one-dimensional treatment of measurement uncertainties in only  $v_z$  was already applied by BR13.

### 2.8. Fitting procedure

To search the  $(p_\Phi, p_{\text{DF}})$  parameter space for the maximum of the  $\text{pdf}$  in Equation (9), we go beyond the single fixed grid search by BR13 and employ an efficient two-step procedure: Nested-grid search and Monte-Carlo Markov Chain (MCMC).

The first step employs a nested-grid search to find the approximate peak and width of the  $\text{pdf}$  in the high-dimensional  $p_M$  space with a low number of likelihood evaluations:

- *Initialization.* For  $N_p$  free model parameters  $p_M$  we start with a sufficiently large grid with  $3^{N_p}$  regular points.
- *Evaluation.* We evaluate the  $\text{pdf}$  at each grid-point similar to BR13 (their Figure 9): An outer loop iterates over the potential parameters  $p_\Phi$  and pre-calculates all  $N_* \times N_{\text{samples}} + N_x^2 \times N_v^3$  actions required for the likelihood calculation (see Equations (8), (9) and (16)). Then an inner loop evaluates Equation (9) (or (16)) for all DF parameters  $p_{\text{DF}}$  in the given potential.
- *Iteration.* For each of the model parameters  $p_M$  we marginalize the  $\text{pdf}$ . A Gaussian is fitted to the marginalized  $\text{pdf}$  and the peak  $\pm 4\sigma$  become the boundaries of the next grid with  $3^{N_p}$  grid points. The grid might be still too coarse or badly positioned to fit Gaussians. In that case we either zoom into the grid point with the highest probability or shift the current range to find new approximate grid boundaries. We proceed with iteratively evaluating the  $\text{pdf}$  on finer and finer grids, until we have found a reliable 4-sigma fit range in each of the  $p_M$  dimensions. The central grid point is then very close to the best fit  $p_M$ , and the grid range is of the order of the  $\text{pdf}$  width.
- *The fiducial qDF.* To save time by pre-calculating actions, they have to be independent of the choice of  $p_{\text{DF}}$ . However, the normalisation in Equation (11) requires actions on a  $N_x^2 \times N_v^3$  grid and the grid ranges in velocity space *do* depend on the current  $p_{\text{DF}}$  (see Equation (8)). To relax this, we follow BR13 and use a fixed set of qDF parameters (the *fiducial qDF*) to set the velocity grid boundaries in Equation (8) globally for a given  $p_\Phi$ . Choosing a fiducial qDF that is very different from the true DF can however lead to large biases in the  $p_M$  recovery. BR13 did not account for that. *RoadMapping* avoids this as follows: To get successively closer to the optimal fiducial qDF—with the (yet unknown) best fit  $p_{\text{DF}}$ —we use in each iteration step of the nested-grid search the central grid point of the current  $p_M$  grid as the fiducial qDF’s  $p_{\text{DF}}$ . As the nested-grid search approaches the best fit values, the fiducial qDF approaches its optimum as well.
- *Computational expense.* 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  $N_* \times N_{\text{samples}} + N_x^2 \times N_v^3$  actions required to be calculated per  $p_\Phi$ , c) the number of potential parameters and d) the number of DF parameters.

The second step samples the shape of the  $\text{pdf}$  using MCMC. Formally, calculating the  $\text{pdf}$  on a fine grid like BR13 (e.g., with  $K = 11$  grid points in each dimension) would provide the same information. However the number of expensive  $\text{pdf}$  evaluations scales as  $K^{N_p}$ . For a high-dimensional  $p_M$  ( $N_p > 4$ ), a MCMC approach might sample the  $\text{pdf}$  much faster: We use *emcee* by Foreman-Mackey et al. (2013) and release the walkers very close to the best fit  $p_M$  found by the nested-grid search, which assures fast convergence in much less than  $K^{N_p}$   $\text{pdf}$  evaluations. We also use the best fit  $p_M$  of the grid-search as fiducial qDF for the whole MCMC. In doing so, the normalisation varies smoothly with different  $p_M$  and is slightly less sensitive to the accuracy in Equation (8).

[TO DO: The referee writes: "Here a fixed sampling is used for the error samples. I think again you should reference McMillan & Binney (2013) as they discussed the numerical stability of this method." –j. Wilma: "Unfortunately I do not find this discussion of fixed error samples in the corresponding paper. As far as I understand, they used the same number of error samples for each star as well."]

## 3. RESULTS

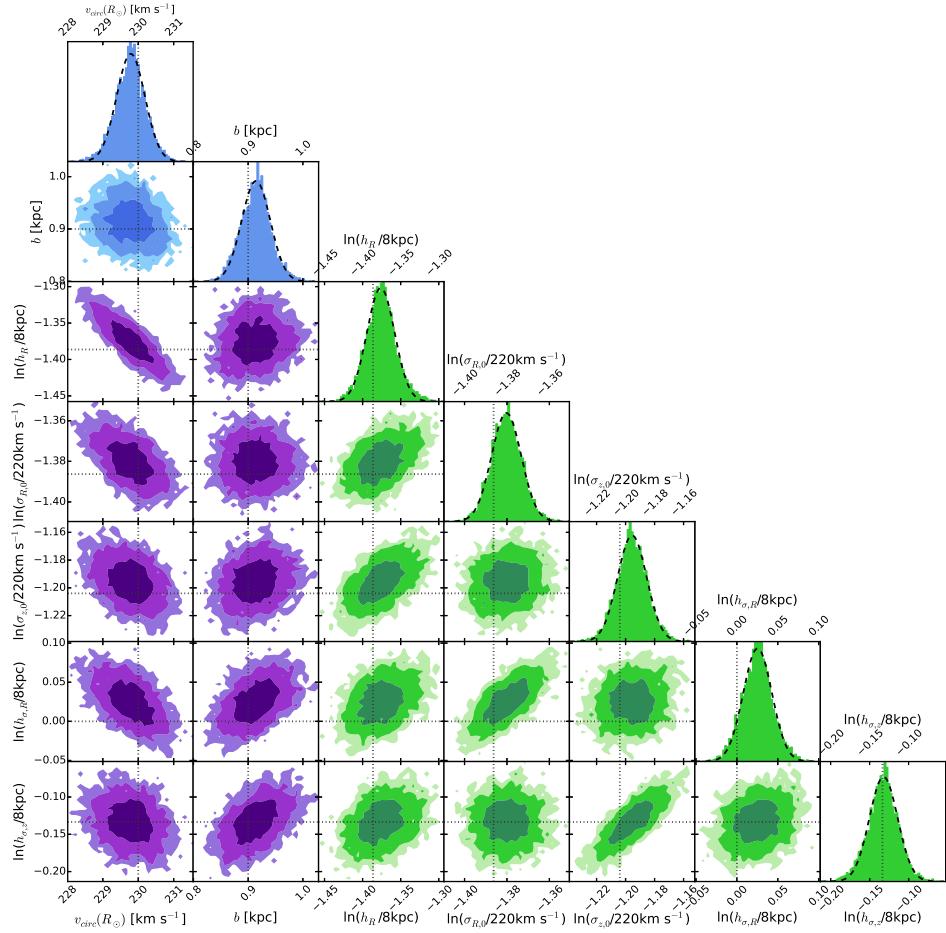
We are now in a position to examine the limitations of action-based modelling posed in the introduction using our *RoadMapping* machinery. We explore: (i) whether the parameter estimates are unbiased, (ii) the role of the survey volume, (iii) imperfect selection functions, (iv) measurement uncertainties, and what happens if the true (v) DF or (vi) potential are not included in the space of models.

We will rely on mock data as input to explore the limitations of the modelling. The mock data is generated directly from the fiducial potential and DF models introduced in Sections 2.3 and 2.4, following the procedure described in Appendix A. With the exception of the test suite on measurement uncertainties in Section 3.4, we assume that phase-space uncertainties are negligible. All tests are also summarized in Table 3.

We do not explore the breakdown of the assumption that the system is axisymmetric and in steady state nor the impact of resonances, which is not possible in the current setup using mock data drawn from axisymmetric galaxy models. (We plan however to investigate this in a future paper, where we will apply *RoadMapping* to N-body simulations of disk galaxies.)

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

The individual MAPs in BR13 contained typically between 100 and 800 objects, so that each MAP implied a quite broad  $\text{pdf}$  which implied broad  $\text{pdfs}$  for the model parameters  $p_M$ . Here we explore what happens in the limit of much larger samples, say  $N_* = 20,000$  objects. Several consequences arise in the limit of much larger samples, say  $N_* = 20,000$ : (i) As outlined in



**Figure 4.** The *pdf* in the parameter space  $p_M = \{p_\Phi, p_{\text{DF}}\}$  for one example mock data set (see Test 3.1 in Table 3). Blue indicates the *pdf* for the potential parameters  $p_\Phi$ , green the qDF parameters  $p_{\text{DF}}$ . The true parameters are marked by dotted lines. The dark, medium and bright contours in the 2D distributions represent  $1\sigma$ ,  $2\sigma$  and  $3\sigma$  confidence regions, respectively. The parameters are weakly to moderately covariant, but their level of covariance depends on the actual choice of the mock data's  $p_M$ . The *pdf* here was sampled using MCMC. The dashed lines in the 1D distributions are Gaussian fits to the histogram of MCMC samples. This demonstrates very well that for such a large number of stars, the *pdf* approaches the shape of a multi-variate Gaussian, that also projects into Gaussians when considering the marginalized *pdf* for all the individual  $p_M$ , as expected for a maximum likelihood estimator.

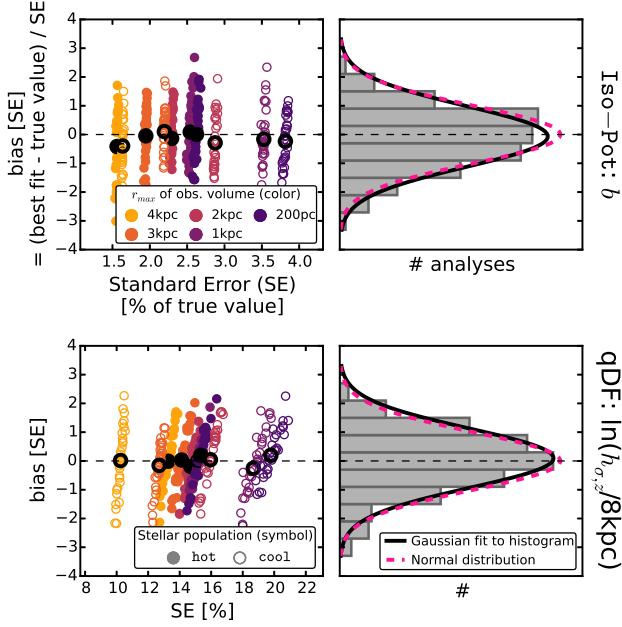
Section ??2.6 and investigated in Figure 2 (Test 1 in Table 3),, the immediate consequence of larger samples is given by higher numerical accuracy is needed due to the likelihood normalisation requirement  $\delta M_{\text{tot}} \lesssim 1/N_*$  (see Equation (14)), which is the modelling aspect that drives the computing time. (ii) The This issue aside, we would expect that in the limit of large data sets with vanishing measurement uncertainties the *pdfs* of the  $p_M$  become Gaussian, with a *pdf* width (i.e., the standard error (SE) on the parameter estimate) that scales as  $1/\sqrt{N_*}$ . The former is demonstrated in Figure 4 (Test 3.1 in Table 3) and we also verified that the latter is true. (iii) AnyFurther, we must verify that any bias in the *pdf* expectation value ishas to be considerably less than the SEerror, even for quite large samples. Using sets of mock data, created according to the procedure in Section ?? and a fiducial model for  $p_M$  (see Table 3, Tests ??, 3.2, and 3.1), we verified that *RoadMapping* satisfies all these conditions and expectations: Figure 4 illustrates the joint *pdfs* of all  $p_M$ . The *pdf* is a multi-variate Gaussian that projects into Gaussians when considering the marginalized *pdf* for all the individual  $p_M$ . Figure ?? then demonstrates that the *pdf* width indeed

scales as  $1/\sqrt{N_*}$ . Figure 5 (Test 3.2 in Table 3) illustrates even more that *RoadMapping* behaves like an unbiased maximum likelihood estimator: The average parameter estimates from many mock samples with identical underlying  $p_M$  data sets are very close to the input  $p_M$ , and the distribution of the actual parameter estimates are Gaussian around it.

### 3.2. The role of the survey volume geometry

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

The first suite draws mock data for two different potentials (**Iso-Pot** and **MW13-PotDHB-Pot**) and four volume wedges (see Section ??Appendix B) with different extent and at different positions within the Galaxy, illustrated in the upper panel of Figure 6. Otherwise the data sets are generated from the same  $p_M$  (see Test 4 in Table 3). To isolate the role of the survey volume geometry, the mock data sets all have the same number of stars ( $N_* = 20,000$ )in all cases, and are drawn from identical total survey volumes ( $4.5 \text{ kpc}^3$ , achieved by adjusting the angular width of the wedges). We investigate these rather unrealistic survey volumes to test (i) if there



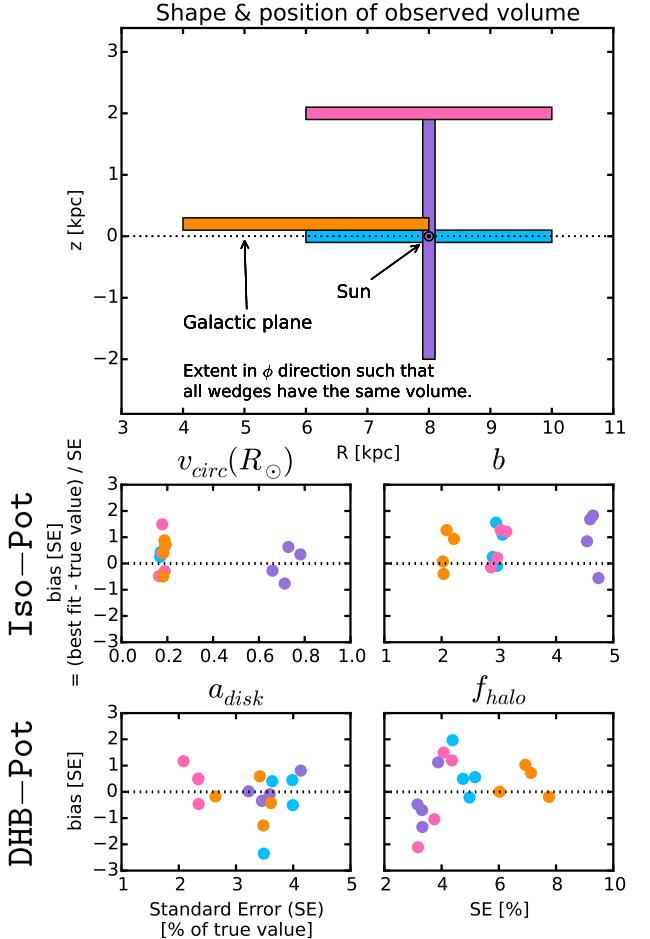
**Figure 5.** Lack of bias in the parameter estimates. Maximum likelihood estimators converge to the true parameter values for large numbers of data points and have a Gaussian spread—if the model assumptions are fulfilled. To test that these conditions are satisfied for *RoadMapping*, we create 320 mock data sets, which come from two different stellar populations and five spherical observation volumes (see legends). (All model parameters are summarized in Table 3 as Test 3.2.) Bias and relative standard error (SE) are derived from the marginalized *pdf* for two model parameters (isochrone scale length  $b$  in the first row and qDF parameter  $h_{\sigma,z}$  in the second row). The second column displays a histogram of the 320 bias offsets. As it closely follows a normal distribution, our modelling method is therefore well-behaved and unbiased. The black dots denote the *pdf* expectation value for the 32 analyses belonging to the same  $p_M$ .

are regions in the Galaxy where stars are on intrinsically more informative orbits (i.e., which give tighter potential constraints) and (ii) if spatial cuts applied to the survey volume (e.g., to avoid regions of large dust extinction or measurement uncertainties) could therefore strongly affect the modelling precision. To see the effect most dramatic we choose some extreme, but illustrative examples. The results are shown in Figure 6.

The second suite of mock data sets was already introduced in Section 3.1 (see also Test 3.2 in Table 3), where mock data sets were drawn from five spherical volumes around the Sun with different maximum radius, for two different stellar populations. The results of this second suite are shown in Figure 5 and exemplify the effect of the size of the survey volume.

Figure 5 demonstrates that, given a choice of  $p_{\text{DF}}$ , a larger volume always results in tighter constraints. There is no obvious trend that a hotter or cooler population will always give better results; it depends on the survey volume and the model parameter in question.

In Figure 6 the wedges all have the same volume and all give results of similar precision. Minor differences (e.g., the Iso-Pot potential being less constrained in the wedge with large vertical but small radial extent) are a special property of the considered potential and parameters, and not a global property of the corresponding survey volume. There are some minor differences (e.g.,  $v_{\text{circ}}(R_{\odot})$  and radial scale lengths,  $b$  or  $a_{\text{disk}}$ , being slightly better



**Figure 6.** [This is a new version of the original plot - using the DHB-Pot instead of the MW13-Pot. (TO DO: Remove this note for final version.)] Bias vs. standard error in recovering the potential parameters for mock data sets drawn from four different wedge-shaped test observation volumes within the Galaxy (illustrated in the upper panel; the corresponding analyses are colour-coded) and two different potentials (Iso-Pot and MW13-Pot<sub>DHB-Pot</sub> from Table 1; see also Test 4 in Table 3 for all model parameters used). Standard error and offset were determined from a Gaussian fit to the marginalized *pdf*. The angular extent of each wedge-shaped observation volume was adapted such that all have a volume of  $4.5 \text{ kpc}^3$ , even though their extent in  $(R, z)$  is different. (The recovery of the free potential parameter  $v_{\text{circ}}(R_{\odot})$  for the DHB-Pot in the different wedges is very similar for both potentials and therefore only shown for the Iso-Pot. Minor expected differences can be seen (e.g.,  $v_{\text{circ}}(R_{\odot})$  and tracer density scale lengths requiring larger radial extent), but overall there is no clear trend that an observation volume around the Sun, above the disk or at smaller Galactocentric radii should give remarkably better constraints on the potential than the other volumes.

recovered for large radial extent and the halo fraction at the Sun,  $f_{\text{halo}}$ , for volumes centered around  $R_{\odot}$ ). In the case of an axisymmetric model galaxy, the extent in  $\phi$  direction is not expected to matter. Overall radial extent and vertical extent seem to be equally important to constrain the potential. In addition, Figure 6 implies that volume offsets or spatial cuts of the survey volume in the radial or vertical direction have at most a modest impact—even in case of the very large sample size at hand.

While it appears that the argument for significant radial and vertical extent is generic, we have not done a

full exploration of all combinations of  $p_M$  and volumes.

That in reality different regions in the Galaxy have different stellar number densities and different measurement uncertainties, should therefore be the major factor to drive the precision of the potential recovery when choosing a survey volume.

### 3.3. Impact of misjudging the selection function of the data set

The survey SF (see Section also Appendix B) can be very complex and is therefore sometimes not perfectly known. Here we investigate how much this could affect the recovery of the potential. We do this by creating mock data in the DHB-Pot within a spherical survey volume with radius  $r_{\max}$  around the Sun (see Test 5 in Table 3) and a spatially varying completeness function

$$\text{completeness}(r) \equiv 1 - \epsilon_r \frac{r}{r_{\max}}, \quad (17)$$

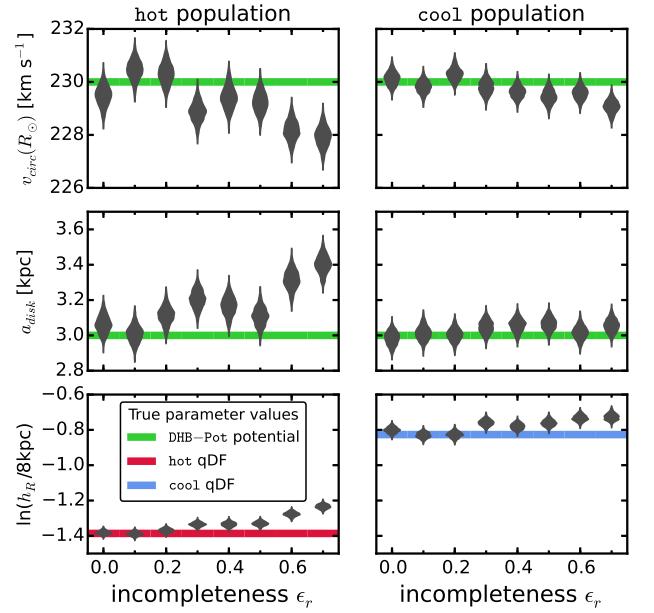
which drops linearly with distance  $r$  from the Sun. The completeness function can be understood as the probability of a star at distance  $r$  to be detected (see also Equation B1). In the RoadMapping analysis however on the other hand, we assume constant completeness ( $\epsilon_r = 0$ ). The incompleteness parameter  $\epsilon_r$  of the mock data quantifies therefore by how much we misjudge the SF. This mock test captures the relevant case of stars being less likely to be observed (than assumed) the further away they are (e.g., due to unknown dust obscuration).

Figure 7 demonstrates that the potential recovery with RoadMapping is very quite robust against somewhat wrong assumptions about the completeness of the data, i.e.,  $\epsilon_r \lesssim 0.15$  for the hot and  $\epsilon_r \lesssim 0.2$  for the cool population. The cool population is more robust, because it is less affected by the SF misjudgement at high  $|z|$  than the hot population. Our simple model SF affects stars at large and small radii in equal proportion. As long as the misjudgement is small, the tracer scale length parameter  $h_R$  can still be reliably recovered, and with it the potential. The robustness for the cool stellar population is even more striking than for the hot population. The reason for this robustness could be, that much information about the potential comes from the rotation curve measurements in the plane, which is not affected by the incompleteness of the sample. We test this by analysing the data sets from Figure ?? again, but this time not including tangential velocity measurements (which is done by marginalizing the likelihood in Equation (9) over  $v_T$ ). Figure B4 shows that in this case the potential is much less tightly constrained, even for 20,000 stars. For only minor deviations of true and assumed completeness ( $\epsilon_r \lesssim 0.15$ ) the true potential is however still included within the errors of our fitting result (see Figure B4).

We have also investigated several test suites using the Iso-Pot. The recovery of  $v_{\text{circ}}(R_\odot)$  and the qDF parameters at different  $\epsilon_r$  is qualitatively and quantitatively similar to Figure 7 for the DHB-Pot. The isochrone scale length  $b$  however is recovered independently of  $\epsilon_r$ —probably because rotation curve measurements in the plane alone, which are not affected by the SF cuts, give reliable constraints on  $b$ . When not including tangential velocity measurements in the analysis (which is done by marginalizing the likelihood in Equation (9) over  $v_T$ ),

the parameters are well recovered only for  $\epsilon_r \lesssim 0.15$  and  $\epsilon_r \lesssim 0.2$  for the hot and cool population respectively. As this is in concordance with our findings for the DHB-Pot, this result seems to be valid for different choices of potentials.

For spatial completeness functions varying with the distance from the plane  $|z|$  only, the Iso-Pot potential recovery is similarly robust as long as  $v_T$  measurements are included. We found similarly robust results also for a misjudgement of spatial completeness functions varying with the distance from the plane,  $|z|$ .



**Figure 7.** [This is a new version of the original plot (Figure B3) - using the DHB-Pot instead of the Iso-Pot. (TO DO: Remove this note for final version.)] Impact of misjudging the completeness of the data on the parameter recovery with RoadMapping. Each mock data set was created with a different incompleteness parameter  $\epsilon_r$  (shown on the x-axis, see Equation (17)). (The model parameters are given as Test 5 in Table 3.) The analysis however assumed that all data sets had constant completeness within the survey volume ( $\epsilon_r = 0$ ). The violins show the full shape of the projected pdfs for each model parameter, and the solid lines their true values. The RoadMapping method seems to be very robust against small to intermediate deviations between the true and the assumed data incompleteness. (The potential parameter  $f_{\text{halo}}$  and the other qDF parameters are recovered to a comparable accuracy and are therefore not shown here. The qDF parameters not shown here exhibit a similar robustness as  $h_R$ .)

### 3.4. Measurement uncertainties and their effect on the parameter recovery

Measurement uncertainties in proper motions and distance dominate over uncertainties in position on the sky (RA, Dec) and line-of-sight velocity, which can be more accurately determined.

The range of proper motion uncertainties we will investigate in this section,  $1 - 5 \text{ mas yr}^{-1}$ , is the approximate measurement accuracy that can be achieved by combining catalogues from ground-based surveys like the Sloan Digital Sky Survey (SDSS; Abazajian et al. 2003), the USNO-B catalog (Monet et al. 2003), 2MASS (Skrutskie et al. 2006) and Pan-STARRS1 photometric

catalog (PS1; Kaiser et al. 2010).<sup>9</sup> Space-based surveys can do even better: The Hipparcos (ESA 1997) and Tycho-2 (Høg et al. 2000) catalogues achieve  $\delta\mu \sim 2.5 \text{ mas yr}^{-1}$  (and even  $\delta\mu \lesssim 1 \text{ mas yr}^{-1}$  for all stars with  $V < 12$ ), which will be soon superseded by Gaia with only  $\delta\mu \sim 0.3 \text{ mas yr}^{-1}$  at its faint end at magnitude  $G \sim 20$  (de Bruijne et al. 2014).

We first investigate the impact of (perfectly known) proper motion uncertainties on the precision of the potential parameter recovery (see Test 6.1 in Table 3). Figure 8 demonstrates that for data sets with  $\delta\mu$  as high as  $5 \text{ mas yr}^{-1}$  the precision degrades by a factor of no more than  $\sim 2$  as compared to a data set without measurement uncertainties. The precision gets monotonically better for smaller  $\delta\mu$ , being larger only by a factor of  $\sim 1.15$  at  $\delta\mu = 1 \text{ mas yr}^{-1}$ . With relative standard errors on the recovered parameters of only a few percent at most for 10,000 stars, this means we still get quite precise constraints on the potential, as long as we know the proper motion uncertainties perfectly.

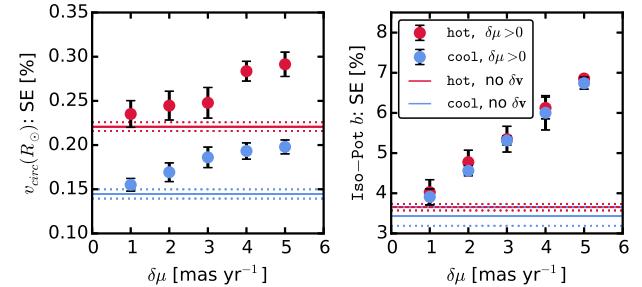
We also note that in this case the relative and absolute difference in recovered precision between the precise and the uncertainty-affected data sets does not seem to depend strongly on the kinematic temperature of the stellar population.

Secondly, we investigate the impact of additional measurement uncertainties in distance (modulus). In absence of distance uncertainties the uncertainty-convolved model probability given in Equation (16) is unbiased (see upper left panel in Figure 9). When including distance (modulus) uncertainties, Equation (16) is just an approximation for the true likelihood; the systematic bias thus introduced in the parameter recovery gets larger with the size of  $\delta(m - M)$ , as demonstrated in Figure 9, lower panels (see also Test 6.2 in Table 3). We find however that in case of  $\delta(m - M) \lesssim 0.2 \text{ mag}$  (if also  $\delta\mu \lesssim 2 \text{ mas yr}^{-1}$  and a maximum distance of  $r_{\max} = 3 \text{ kpc}$ , see Test 6.2 in Table 3) the potential parameters can still be recovered within  $2\sigma$ . This corresponds to a relative distance uncertainty of  $\sim 10\%$ . The overall precision of the potential recovery is also not degraded much by introducing distance uncertainties of less than 10%.

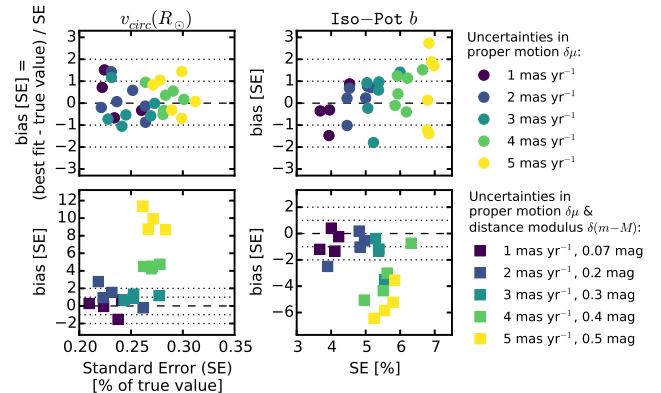
How does this compare with the distance uncertainties expected for Gaia? For a typical red clump giant star with  $M_I \sim 0 \text{ mag}$  and  $(V - I) \sim 1 \text{ mag}$  at a distance of  $r = 3 \text{ kpc}$  we estimate (using the magnitude transformation by Jordi et al. (2010) and the uncertainty parametrization by de Bruijne et al. (2014)) a parallax uncertainty of  $\delta\pi \sim 11 \mu\text{as}$ , which is consistent with a distance uncertainty of less than 5%, and a proper motion uncertainty of  $\delta\mu \sim 6 \mu\text{as yr}^{-1}$ , which is negligible. When not restricting the modelling to giant stars

<sup>9</sup> Combining observations from the SDSS Data Release 1 with the USNO-B catalog based on the Palomar Observatory Sky Survey's (POSS) photographic plates from the 1950s lead to proper motion measurements precise to  $\delta\mu \sim 3$  or  $5 \text{ mas yr}^{-1}$  depending on magnitude  $r < 18$  or  $r < 20$  respectively (Munn et al. 2004, 2008; Gould & Kollmeier 2004). The same accuracy can be achieved when using the four years of measurements by the PS1 only. By careful calibration of USNO-B and 2MASS with PS1, Sesar et al. (2015) even got proper motions as accurate as  $\delta\mu \sim 1.5 \text{ mas yr}^{-1}$  for  $r \lesssim 18$ . The Large Synoptic Survey Telescope (LSST, Ivezić et al. 2008) planned for 2021 might even achieve  $\delta\mu \lesssim 1 \text{ mas yr}^{-1}$  during its 10 years of scanning the sky (Ivezić et al. 2008).

only, a quick investigation by Rene Andrae of stars at  $r = 3 \text{ kpc} \pm 5 \text{ pc}$  from the Gaia Universe model snapshot catalogue (GUMS; Robin et al. 2012) revealed that a magnitude cut at  $G \sim 15 \text{ mag}$  in the overall Gaia data set should keep all distance uncertainties within 3 kpc below  $\sim 10\%$ , while also preserving Gaia's simple SF.

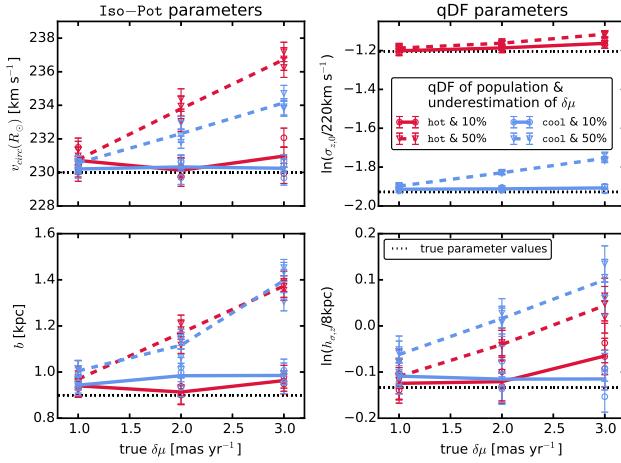


**Figure 8.** Effect of proper motion uncertainties  $\delta\mu$  on the precision of potential parameter recovery for two stellar populations of different kinematic temperature (see Test 6.1 in Table 3 for all model parameters). The relative standard error (SE) derived from the marginalized  $pdf$  for each model parameter was determined for precise data sets without measurement uncertainties (solid lines, with dotted lines indicating the error) and for data sets affected by different proper motion uncertainties  $\delta\mu$  and  $\delta v_{\text{los}} = 2 \text{ km s}^{-1}$  (data points with error bars), but no uncertainties in position. The errors come from taking the mean over several data sets.



**Figure 9.** Potential parameter recovery using the approximation for the model probability convolved with measurement uncertainties in Equation (16). We show  $pdf$  offset and relative width (i.e., standard error SE) for potential parameters recovered from mock data sets (which were created according to Test 6.2 in Table 3). The data sets in the upper panels are affected only by proper motion uncertainties  $\delta\mu$  (and  $\delta v_{\text{los}} = 2 \text{ mas yr}^{-1}$ ), while the data sets in the lower panels also have distance (modulus) uncertainties  $\delta(m - M)$ , as indicated in the legend. For data sets with  $\delta\mu \leq 3 \text{ mas yr}^{-1}$  Equation (16) was evaluated with  $N_{\text{samples}} = 800$ , for  $\delta\mu > 3 \text{ mas yr}^{-1}$  we used  $N_{\text{samples}} = 1200$ . In absence of distance uncertainties Equation (16) gives unbiased results. For  $\delta(m - M) > 0.2 \text{ mag}$  (i.e.,  $\delta r/r > 0.1$ ; for  $r \sim 3 \text{ kpc}$ ) however biases of several  $\sigma$  are introduced, as Equation (16) is only an approximation for the true likelihood in this case.

We therefore found that in case we perfectly know the measurement uncertainties (and the distance uncertainty is negligible or of the order of the uncertainties expected from Gaia within  $\sim 3 \text{kpc}$ ), the convolution of the model probability with the measurement uncertainties gives precise and accurate constraints on the model



**Figure 10.** Effect of a systematic underestimation of proper motion uncertainties  $\delta\mu$  on the recovery of the model parameters. (The true model parameters used to create the mock data are summarized as Test 6.3 in Table 3, four of them are indicated as black dotted lines in this figure.) The mock data was perturbed according to proper motion uncertainties  $\delta\mu = \delta\mu_{\text{Dec}} = \delta\mu_{\text{RA}}$  as indicated on the  $x$ -axis. In the *RoadMapping* analysis (see likelihood in Equation (16)) however, we underestimated the true  $\delta\mu$  by 10% (circles, solid lines) and 50% (triangles, dashed lines). The symbols denote the *pestbest* fit parameters with  $1\sigma$  error bars of several mock data sets. The lines connect the mean of corresponding data realisations to guide the eye.

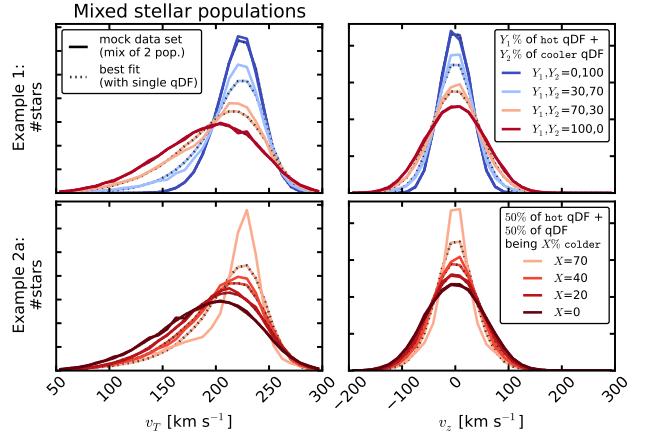
parameters—even if the measurement uncertainty itself is quite large.

Lastly, Figure 10 investigates the effect of a systematic *underestimation* of the true proper motion uncertainties  $\delta\mu$  by 10% and 50% (see also Test 6.3 in Table 3). We find that this causes a bias in the parameter recovery that grows seemingly linear with  $\delta\mu$ . For an underestimation of only 10% however, the bias is still  $\lesssim 2\sigma$  for 10,000 stars—even for  $\delta\mu \sim 3$  mas  $\text{yr}^{-1}$ .

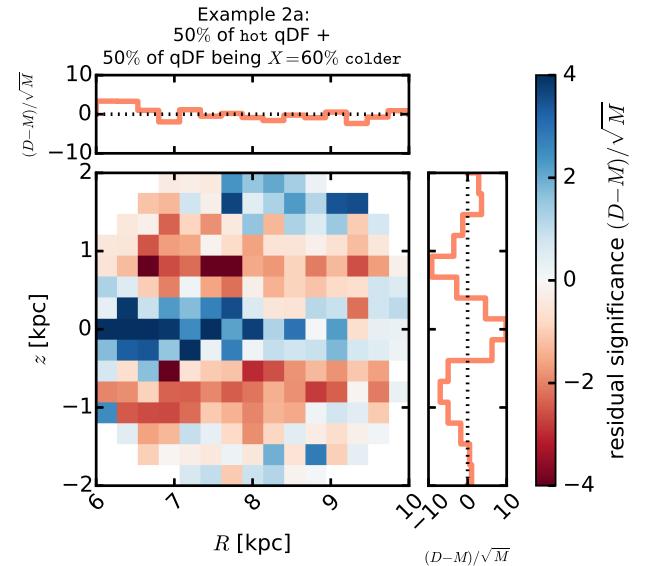
The size of the bias also depends on the kinematic temperature of the stellar population and the model parameter considered (see Figure 10). The qDF parameters are for example better recovered by hotter populations. This is, because the relative difference between the *true*  $\sigma_i(R)$  (with  $i \in \{R, z\}$ ) and *measured*  $\sigma_i(R)$  (which comes from the deconvolution with an underestimated velocity uncertainty) is smaller for hotter populations.

### 3.5. The impact of deviations of the data from the idealized distribution function

Our modelling approach assumes that each stellar population follows a simple DF; here we use the qDF. In this section we explore what happens if this idealization does not hold. We investigate this issue by creating mock data sets that are drawn from *two* distinct qDFs of different temperature<sup>10</sup> (see Table 2 and Test 7 in Table 3) in the **DHB-Pot**, and analyse the composite mock data set by fitting a *single* qDF to it. The velocity distribution of some mock data sets and their best fit qDFs are illustrated in Figure 11, while Figure 12 shows the tracer density residuals between data and best fit in the  $(R, z)$  plane. Figures 13 and 14 compare the input and best fit parameters and the comparison of input and best fit pa-



**Figure 11.** [This is a new version of the original plot - using the **DHB-Pot** instead of the **Iso-Pot**. (TO DO: Remove this note for final version.)] Distribution of mock data  $v_T$  and  $v_z$  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 (dotted lines). (The model parameters used to create the mock data are given in Table 3 as Test 7, *Example 1* & *2a*, with the qDF parameters referred to in the legend given in Table 2.) The corresponding single qDF *best fit*/*best-fit* curves were derived from the best fit parameters found in Figures 13 and 14. (The data sets are colour-coded in the same way as the corresponding analyses in Figures 13 and 14.) We use the mixtures of two qDFs to demonstrate how *RoadMapping* behaves for data sets following DFs with shapes slightly differing from a single qDF. For large deviations it might already become visible from directly comparing the mock data and best fit distribution, that a single qDF is a bad assumption for the stars' true DF.

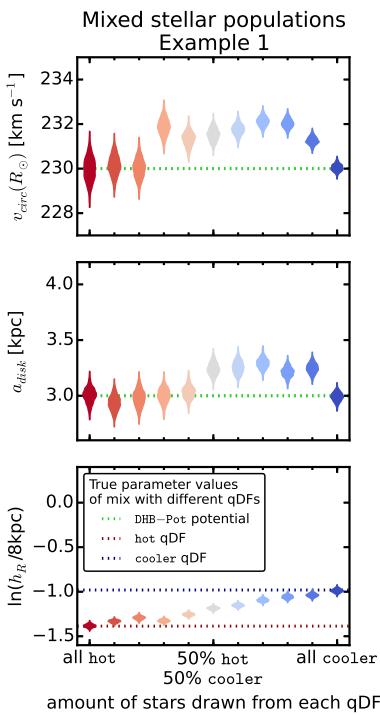


**Figure 12.** [This is a completely new plot. (TO DO: Remove this note for final version.)] Residual significance  $(D - M)/\sqrt{M}$  of one example mock data set  $D$  and its best fit single qDF model  $M$  in the  $(R, z)$  plane. The mock data set  $D$  was created by mixing a hot and a  $X\%$  colder population (see also Table 3, Test 7, *Example 2a*, with  $X = 60\%$ ). The best fit distribution  $M$  was derived analogously to the ones shown for the velocity components in Figure 11.) This is an extreme example where the best fit single qDF is not a good fit anymore (see Figure 14, *Example 2a*,  $X = 60\%$ ), but it illustrates how we constructed mock data distributions with radial and vertical density profiles differing from a single qDF by mixing two different qDF populations for the test suite 7.

<sup>10</sup> Following the observational evidence, our mock data populations with cooler qDFs also have longer tracer scale lengths.

rameters in Figures 13 and 14. In *Example 1* we choose qDFs of widely different temperature and vary their relative fraction of stars in the composite mock data set (Figure 13); in *Example 2* we always mix mock data stars from two different qDFs in equal proportion, but vary by how much the qDFs' temperatures differ (Figure 14).

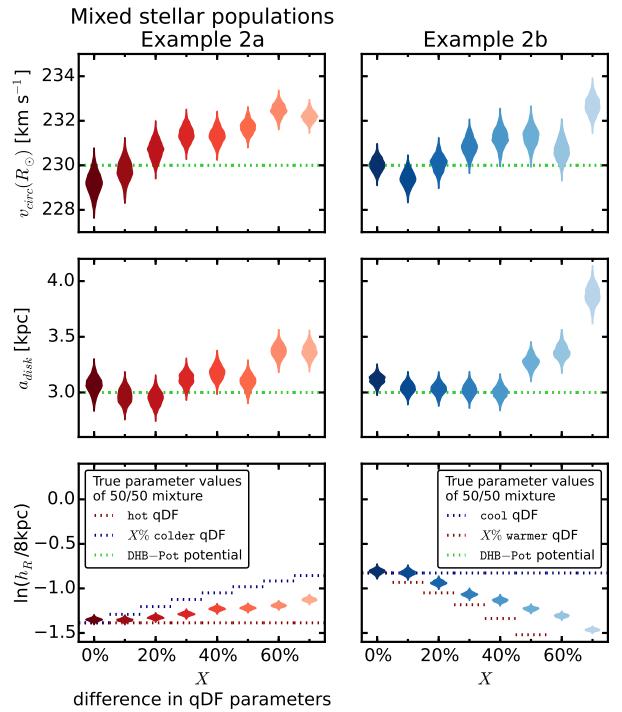
The first set of tests mimics a DF that has wider wings or a sharper core in velocity space than a qDF (see Figure 11) and slightly different radial and vertical tracer density profiles, similar to Figure 12. The second test could be understood as mixing neighbouring MAPs in the  $[\alpha/\text{Fe}]\text{-vs.-}[\text{Fe}/\text{H}]$  plane due to large bin sizes or abundance measurement errors (cf. BR13).



**Figure 13.** [This is a new version of the original plot (Figure B1)-using the DHB-Pot instead of the Iso-Pot. (TO DO: Remove this note for final version.)] The dependence of the parameter recovery on degree of pollution and temperature of the stellar population. We mix (i.e., “pollute”) varying amounts of stars from a hot stellar population with stars from a very different cooler population (see Table 2), as indicated on the  $x$ -axis. (All model parameters used to create the mock data are given as Test 7, *Example 1*, in Table 3.) The composite polluted mock data set follows a true DF that has a slightly different shape than the qDF. We then analyse it using *RoadMapping* and fit a single qDF only. The violins represent the marginalized *pdfs* for the best fit model parameters. Some mock data sets are shown in Figure 11, first row, in the same colours as the violins here. We find that a hot population is much less affected by pollution with stars from a cooler population than vice versa.

We consider the impact of the DF deviations on the recovery of the potential and of the qDF parameters separately.

We find from *Example 1* that the potential parameters can be more robustly recovered, if a mock data population is polluted by a modest fraction ( $\lesssim 30\%$ ) of stars drawn from a much cooler qDF, as opposed to the same pollution of stars from a hotter qDF. When considering

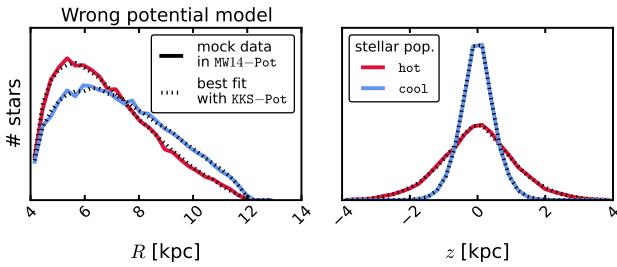


**Figure 14.** [This is a new version of the original plot (Figure B2) - using the DHB-Pot instead of the Iso-Pot. (TO DO: Remove this note for final version.)] The dependence of the parameter recovery on the difference in qDF parameters of a 50/50 mixture of two stellar populations and their temperature. The two qDFs from which the stars in each mock data set were drawn are indicated in the legend, with the qDF parameters  $\sigma_{R,0}, \sigma_{z,0}$  and  $h_R$  differing by  $X\%$  (see also Table 2 and Section 2.4), as indicated on the  $x$ -axis. (The model parameters used for the mock data creation are given as Test 7, *Example 2a & b*, in Table 3.) Each composite mock data set is fitted with a single qDF and the marginalized *pdfs* are shown as violins. Some mock data sets of *Example 2a* and their best fit distributions are shown in Figure 11, last row (colour-coded analogous to the violins here), and Figure 12 shows the corresponding residuals in the  $(R, z)$  plane. By mixing populations with varying difference in their qDF parameters, we model the effect of finite bin size or abundance errors when sorting stars into different MAPs in the  $[\alpha/\text{Fe}]\text{-vs.-}[\text{Fe}/\text{H}]$  plane and assuming they follow single qDFs (cf. BR13). We find that the bin sizes should be chosen such that the difference in qDF parameters between neighbouring MAPs is less than 20%.

the case of a 50/50 mix of contributions from different qDFs in *Example 2*, there is a systematic, but mostly small, bias in recovering the potential parameters, monotonically increasing with the qDF parameter difference. In particular for fractional differences in the qDF parameters of  $\lesssim 20\%$  the systematics are insignificant even for sample sizes of  $N_* = 20,000$ , as used in the mock data.

Overall, the circular velocity at the sunSun is very reliably recovered to within 2% in all these tests. But the best fit  $v_{\text{circ}}(R_\odot)$  is not always unbiased at the implied precision.

The recovery of the effective qDF parameters, in light of non-qDF mock data, is quite intuitive (in Figures 13 and 14 we therefore show only  $h_R$ ): 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 lengths of the velocity dispersion fall-off,  $h_{\sigma,R}$  and  $h_{\sigma,z}$ , are shorter than the true scale lengths, because the stars drawn from the hotter qDF dominate at small radii, while stars from the cooler qDF (with its



**Figure 15.** [This is a new version of the original plot. TO DO: Remove this note for final version. This plot uses the new analyses from Figure 16. Instead of the velocity distribution we show the spatial distribution of stars.] Comparison of the spatial distribution of mock data  $v_T$  and  $v_{\text{in}}$  in  $R$  and  $z$  created in the MW14-Pot potential and with two different stellar populations (see Test 8 in Table 3 for all mock data model parameters), and the best fit distribution recovered by fitting the family of KKS-Pot potentials to the data. The best fit potentials are shown in Figure 16 and the corresponding best fit qDF parameters in Figure 17. The data is very well recovered, even though the fitted potential family did not incorporate the true potential.

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. The latter is demonstrated in Figure 12: The radial tracer density profile of the mock data is steeper than a single qDF in the mid-plane and more shallow at higher  $|z|$ ; overall the best fit  $h_R$  lies therefore in between.

We note, that in the cases where the systematic bias in the potential parameter recovery becomes several  $\sigma$  large, a direct comparison of the true mock data set and best fit distribution (see Figure 11) can sometimes already reveal that the assumed DF is not a good model for the data.

We performed the same tests also using the spherical Iso-Pot instead of the galaxy-like DHB-Pot and for a much higher sampling of the input parameters. The results are qualitatively and quantitatively very similar and therefore independent of the exact choice of potential.

Overall, we find that the potential inference is quite robust to modest deviations of the data from the assumed DF.

### 3.6. The implications of a gravitational potential not from the space of model potentials

We now explore what happens when the mock data were drawn from one axisymmetric potential family, here MW14-Pot, and is then modelled considering potentials from another axisymmetric family, here KKS-Pot (see Table 1 and Figure 1). In the analysis we assume the circular velocity at the Sun to be fixed and known and only fit the parametric potential form.<sup>11</sup>

We analyse a mock data set from a hot and cool stellar population each (see Test 8 in Table 3) with high numerical accuracy. The distributions generated from the best fit parameters reproduce the data in configuration space very well (see Figure 15 for the spatial distribution and the circles in Figure 17 for the velocity distribution).

<sup>11</sup> We made sure that  $v_{\text{circ}}(R_\odot)$  can be very well recovered when included in the fit of a cool population. The model assumption that  $v_{\text{circ}}(R_\odot)$  is known does therefore not affect the discussion qualitatively.

The results for the potential The comparison between true and best fit potentials are shown in Figure 16. We find that the potential recovered by RoadMapping is in good agreement with the true potential. Especially the force contours, to which the orbits are sensitive, and the rotation curve are very tightly constrained and reproduce the true potential even outside of the observed volume of the mock tracers.

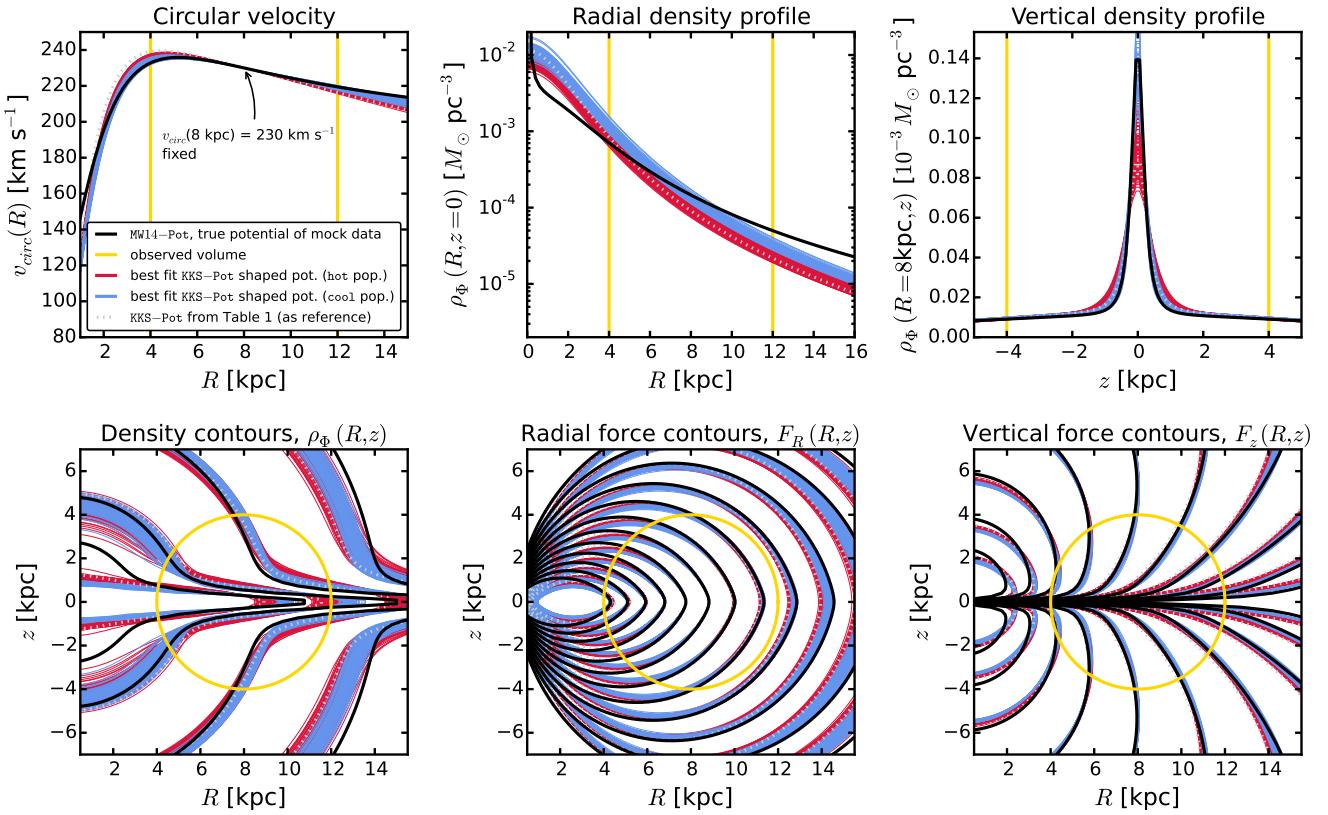
Overplotted in Figure 16 is also the KKS-Pot with the parameters from Table 1, which were fixed based on a (by-eye) fit directly to the force field (within  $r_{\max} = 4$  kpc from the Sun) and rotation curve of the MW14-Pot. The potential found with the RoadMapping analysis is an equally good or even slightly better fit. This demonstrates that RoadMapping fitting infers a potential that in its actual properties resembles the input potential for the mock data as closely as possible, given the differences in functional forms.

The density profiles and contours are less tightly constrained than the forces, but we still capture the essentials. The radial density profile of the KKS-Pot is quite different from the MW14-Pot's density profile at larger radii and outside of the survey volume, and the hot populations has more difficulties to recover the very flattened density distribution of the disk than the cool population. The disk-to-halo surface density fraction within  $z = 1.1$  kpc,  $f_{\Sigma,1.1 \text{ kpc}}(R) \equiv \Sigma_{\text{disk}}(R, |z| \leq 1.1 \text{ kpc}) / \Sigma_{\text{halo}}(R, |z| \leq 1.1 \text{ kpc})$ , at the Sun is  $\Sigma, 1.1 \text{ kpc}(R = 8 \text{ kpc}) = 3.08$  for the true MW14 – Pot, but 9% (11%) smaller for the best fit KKS-Pot and the cool (hot) population. Using a wrong potential model can therefore lead to biases in local dark matter measurements. However, at  $R = 5.5$  kpc, where most of the stars are located (see Figure 15), the recovered  $\Sigma, 1.1 \text{ kpc}(R = 5.5 \text{ kpc})$  is only 0.1% (0.4%) smaller than the true  $f_{\Sigma,1.1}(R = 5.5 \text{ kpc}) = 4.45$  for the analysis with the cool (hot) population.

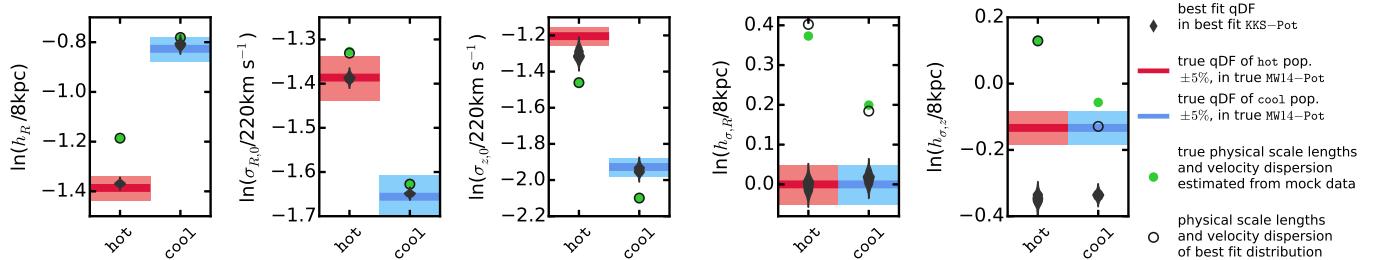
The obvious limitations of a wrongly parametrized potential model aside (e.g., in this case the difficulty to simultaneously have a roundish halo and a very flat disk in the KKS-Pot when it is required that both Stäckel components have the same focal distance (see Table 1)), it appears that we get the most reliable potential constraints where the tracer density is largest: inside the survey volume, at  $R \sim 5.5$  kpc and—especially for the cool population—in the midplane.

Overall the best fit disk is less dense in the mid-plane than the true disk. While it is in general possible to generate very flattened density distributions from Stäckel potentials, it might be difficult to simultaneously have a roundish halo and to require that both Stäckel components have the same focal distance (see Table 1).

Figure 17 compares the true qDF parameters with the best fit qDF parameters belonging to the best fit potentials from Figure 16, and we also overplot the actual physical scale lengths and velocity dispersion as estimated directly from the mock data. While we recover  $h_R$ ,  $\sigma_{R,0}$  and  $h_{\sigma,R}$  within the errors, we misjudge the parameters of the vertical velocity dispersion ( $\sigma_{0,z}$  and especially  $h_{\sigma,z}$ ), even though the actual mock data distribution is well reproduced. This discrepancy could be connected to the KKS-Pot not being able to reproduce the flatness of the disk. Also,  $\sigma_z$  and  $\sigma_R$  in Equations (6)-(7) are scaling profiles for the qDF (cf. BR13) and



**Figure 16.** [This is a new version of the original plot. TO DO: Remove this note for final version. Changes that have been made: (i) As I made some small adjustments in RoadMapping considering the numerical accuracy of the ActionAngleStaeckelGrid and the likelihood normalisation, I thought I re-run these two analyses to see if it changes something. The result is actually better than the original analysis. (ii) I added two more panels, showing cuts through the density profile along  $R$  and  $z$  to better show the quantitative difference between true and best fit potential (as requested by the referee).] Recovery of the gravitational potential if the assumed potential model family (KKS-Pot with fixed  $v_{\text{circ}}(R_\odot)$ ) and the true potential of the (mock data) stars (MW14-Pot in Table 1) have slightly different parametric forms. We show the circular velocity curve, density profiles along  $(R, z = 0)$  and  $(R = 8 \text{ kpc}, z)$ , as well as contours of equal density, radial and vertical force in the  $R$ - $z$ -plane, and compare the true potential (black lines) with 100 sample potentials (red and blue lines) drawn from the  $pdf$  found with MCMC for a hot (red) and a cool (blue) stellar population. (All mock data model parameters are given as Test 8 in Table 3.) Overall, the true potential is well recovered—especially in regions where most of the observed stars are.



**Figure 17.** [This is a new version of the original plot. TO DO: Remove this note for final version. Changes: (i) Displayed the new analyses from Figure 16. (ii) Added points for the physical velocity dispersion and scale lengths and a legend.] Recovery of the qDF parameters for the case where the true and assumed potential deviate from each other (see Test 8 in Table 3). The thick red (blue) lines represent the true qDF parameters of the hot (cool) qDF in Table 2 used to create the mock data, surrounded by a 10% (5%) error region. The grey violins are the marginalized  $pdf$ s for the qDF parameters found simultaneously with the potential constraints shown in Figure 16. We compare the qDF parameters with the actual physical scale lengths and velocity dispersion at the Sun estimated from the mock data and the best fit distribution by fitting exponential functions to the data. Firstly this shows that—apart from some small deviations in the velocity dispersion scale lengths—the velocity distribution of the mock data is very well reproduced by the best fit. Secondly, this demonstrates how the qDF parameters in different potentials do not necessarily agree with each other or with the actual physical velocity distribution.

how close they are to the actual velocity profile depends on the choice of potential; that is, the *physical* velocity dispersion is well recovered, even if the *qDF* velocity-dispersion parameters are not. Figure 17 stresses and demonstrates once more that the actual values of

action-based DF parameters do not easily translate to the physical characteristics of a stellar distribution and have always to be considered together with the potential in which they were derived. This is of importance in studies that use a fiducial potential to fit action-based to

DFs to stellar data, like, e.g., Sanders & Binney (2015) and Das & Binney (2016).

### 3.7. The influence of the stellar population's kinematic temperature

Overall, we found that it does not make a big difference if we use hot or cool stellar populations in our modelling.

How precise and reliable model parameters can be recovered does to a certain extent depend on the kinematic temperature of the data, as well as on the model parameter in question and on the observation volume. But there is no easy rule of thumb, what combination would give the best results (see Figure 5). There are two exceptions.

First, the circular velocity at the Sun,  $v_{\text{circ}}(R_\odot)$ , is always best recovered with cooler populations (see Figures 8, 10, 13, 14 and 16), because more stars are on near-circular orbits (see Figure A1). As cooler populations probe the rotation curve better, which in turn probes the gravitational potential, the potential recovery using cool stellar populations is less sensitive to misjudgements of (spatial) selection functions (see Figures 7 and B4). There is however the caveat, that cool populations are more susceptible to non-axisymmetric streaming motions in the disk.

Second, hotter populations seem to be less sensitive to misjudgements of proper motion measurement uncertainties (see Figure 10) and pollution with stars from a cooler population (see Figures 13 and 14), because of their higher intrinsic velocity dispersion (see Figure A2).

In addition we found indications in Figure 16, that different regions within the Galaxy are probed best by populations of different kinematic temperature: The hot stellar population, with more stars reaching to high  $|z|$  and a shorter tracer scale length, constrained force and density contours in the halo better—especially at smaller radii; the cool population, with more stars in the plane and longer tracer scale length, gave tighter force and density constraints in the outer regions of the halo and recovered the disk more reliably.

## 4. SUMMARY AND DISCUSSION

Recently, implementations of action DF-based modelling of 6D data in the Galactic disk have been put forth, in part to lay the ground-work for Gaia (BR13; McMillan & Binney 2013; Piffl et al. 2014; Sanders & Binney 2015).

We present *RoadMapping*, an improved implementation of the dynamical modelling machinery of BR13, to recover the MW's gravitational potential by fitting an orbit DF to stellar populations within the Galactic disk. In this work we investigated the capabilities, strengths and weaknesses of *RoadMapping* by testing its robustness against the breakdown of some of its assumptions—for well-defined, isolated test cases using mock data. Overall the method works very well and is robust, even when there are small deviations of the model assumptions from the “real world” Galaxy.

*RoadMapping* applies a full likelihood analysis and is statistically well-behaved. It goes beyond BR13 by allowing for a straightforward and flexible implementation of different model families for potential and DF. It also accounts for selection effects by using full 3D selection functions (given some symmetries).

**Computational speed:** Large data sets in the age of Gaia require increasingly accurate likelihood evaluations and flexible models. To be able to deal with these computational demands, we sped up the *RoadMapping* code by combining a nested-grid approach with MCMC and by faster action calculation using the Stäckel (Binney 2012a) interpolation grid by Bovy (2015). However, application of *RoadMapping* to millions of stars will still be a task for supercomputers and calls for even more improvements and speed-up in the fitting machinery.

**Properties of the data set:** We could show that *RoadMapping* can provide potential and DF parameter estimates that are very accurate (i.e., unbiased) and precise in the limit of large datasets, as long as the modelling assumptions are fulfilled.

In case the data set is affected by substantive measurement uncertainties, the potential can still be recovered to high precision, as long as these uncertainties are perfectly known and distance uncertainties are negligible. For large proper motion uncertainties, e.g.,  $\delta\mu \sim 5 \text{ mas yr}^{-1}$ , the formal errors on the parameters are only twice as large as in the case of no measurement uncertainties. However, properly accounting for measurement uncertainties is computationally expensive.

For the results to be accurate within  $2\sigma$  (for 10,000 stars), we need to know to within 10% both the true stellar distances (at  $r_{\text{max}} \leq 3 \text{ kpc}$  and  $\delta\mu \lesssim 2 \text{ mas yr}^{-1}$ ) and the true proper motion uncertainties (with  $\delta\mu \lesssim 3 \text{ mas yr}^{-1}$ ).

The former restriction is an artefact of the likelihood approximation, Equation (16), that *RoadMapping* uses to save computation time, and the reason why we will have to restrict the *RoadMapping* modelling to stars with small distance uncertainties.

Fortunately, the measurement uncertainties of the final Gaia data release with  $\delta\mu \lesssim 0.3 \text{ mas yr}^{-1}$  at  $G \lesssim 20 \text{ mag}$  and  $\delta r/r \lesssim 5\%$  at  $r \sim 3 \text{ kpc}$  and for  $G < 15 \text{ mag}$  (see Section 3.4 and de Bruijne et al. 2014) will be well below these limits and promise accurate potential constraints.

The main caveat of Tests 2 and 6.1- 6.3 (see Table 3) investigating the effect of measurement uncertainties is the use of the **Iso-Pot**, which we choose for its advantages concerning computational speed. However, Tests 5 and 7, which we run for both the **DHB-Pot** and the **Iso-Pot**, gave qualitatively and quantitatively very similar results for both potentials. This makes us confident that also our results considering the measurement uncertainties are independent of the actual choice of potential.

We also found that the location of the survey volume within the Galaxy matters little. At given sample size a larger survey volume with large coverage in both radial and vertical direction will give the tightest constraints on the model parameters.

Surprisingly (cf. Rix & Bovy 2013), the potential recovery with *RoadMapping* seems to be very robust against misjudgements of the spatial data SF. We speculate that this is because missing stars in the data set do not affect the measured rotation curve, which contains information about the potential.

We found indications that populations of different scale lengths and temperature probe different regions of the Galaxy best. This supports the approach by

BR13, who constrained for each MAP the surface mass density only at one single best radius to account for missing flexibility in their potential model. While cooler populations probe the Galaxy rotation curve better and hotter populations are less sensitive to pollution, overall stellar populations of different kinematic temperature seem to be equally well-suited for dynamical modelling.

**Deviations from the DF assumption:** *RoadMapping* assumes that stellar sub-populations can be described by simple DFs. We investigated how much the modelling would be affected if the assumed family of DFs would differ from the stars' true DF.

In *Example 1* in Section 3.5 we considered true stellar DFs being (i) hot with more stars with low velocities and less stars at small radii than assumed (reddish data sets in Figure 11 and 13), or (ii) cool with broader velocity dispersion wings and less stars at large radii than assumed (bluish data sets). We find that case (i) would give more reliable results for the potential parameter recovery.

Binning of stars into MAPs in  $[\alpha/\text{Fe}]$  and  $[\text{Fe}/\text{H}]$ , as done by BR13, could introduce systematic errors due to abundance uncertainties or too large bin sizes—always assuming MAPs follow simple DF families (e.g., the qDF). In *Example 2* in Section 3.5 we found that, in the case of 20,000 stars per bin, differences of  $\lesssim 20\%$  in the qDF parameters of two neighbouring bins can still give quite good constraints on the potential parameters.

The relative differences in the qDF parameters  $\sigma_{R,0}$  and  $\sigma_{z,0}$  of neighbouring MAPs in Figure 6 of BR13 (which have bin sizes of  $[\text{Fe}/\text{H}] = 0.1 \text{ dex}$  and  $\Delta[\alpha/\text{Fe}] = 0.05 \text{ dex}$ ) are indeed smaller than 20%. Figure 13 and 14 suggest that especially the tracer scale length  $h_R$  needs to be recovered to get the potential scale length right. For this parameter however the bin sizes in Figure 6 of BR13 might not yet be small enough to ensure no more than 20% of difference in neighbouring  $h_R$ .

The qDF is a specific example for a simple DF for stellar sub-populations which we used in this paper. But it is not essential for the *RoadMapping* approach. Future studies might apply slight alternatives or completely different DFs to data.

**Gravitational potential beyond the parametrized functions considered:** In addition to the DF, *RoadMapping* also assumes a parametric model for the gravitational potential. We test how using a potential of Stäckel form (KKS-Pot, Batsleer & Dejonghe 1994) affects the *RoadMapping* analysis of mock data from a different potential family with halo, bulge and exponential disk. The potential recovery is quite successful: We properly reproduce the mock data distribution in configuration space; and the best fit potential is—with the limits of the model—as close as it gets to the true potential, even outside of the observation volume of the stellar tracers.

For as many as 20,000 stars constraints become already so tight that it should presumably be possible to distinguish between different parametric MW potential models (e.g., the MW13-Pot used by BR13DHB-Pot and the KKS-Pot).

BR13 fitted a MW-like model potential and calculated actions using the Stäckel approximation Fudge

(Binney 2012a); in this case study we directly fitted a Stäckel potential to the data, with exact actions in the model potential. The latter is computationally much less expensive due to the simple analytic form of the potential. It would also allow flexibility by expressing the MW potential as a superposition of many more simple Kuzmin-Kutuzov Stäckel components (Famaey & Dejonghe (2003) used for example 3 components). The former approach by BR13 however allows to parametrize the potential with intuitive and physically motivated building blocks (exponential disks, power-law dark matter halo etc.). While both approaches are formally similar, it remains to decide which is better. Fitting parametrized potentials of Stäckel form to MW data (see, e.g., Batsleer & Dejonghe 1994; Famaey & Dejonghe 2003) has the advantage of allowing action calculations that are accurate and—due to the simple analytic expression for  $\Phi(R, z)$ —computationally very fast. It does however limit the space of potentials that can be investigated, as different potential components are all required to have the same focal distance. Using the Stäckel Fudge (Binney 2012a) together with parametrized potentials made up from physically motivated building blocks (exponential disks, power-law dark matter halo etc.), as was done by BR13, seems to be the most promising approach. Even though the Stäckel Fudge is a good compromise between speed and accuracy (Sanders & Binney 2015), the machinery is still too slow to fit potentials with many free parameters and  $\Phi(R, z)$  of non-trivial shape (e.g., an exponential disk); we expect computing time scales of several days to weeks on 50 cores. In this study we had to resort to potentials for which analytic expressions exist for  $\Phi(R, z)$  directly (like the Iso-Pot, DHB-Pot and KKS-Pot) and to using an action interpolation grid (Binney 2012a; Bovy 2015) with fixed focal length  $\Delta$  for all stars, which introduces further action inaccuracies. This sped up the analysis considerably: Each of the analyses in the Tests 5 and 7 (see Table 3) took “only” between  $\sim 25 - 30$  hours on 25 cores to complete. Having to restrict the *RoadMapping* analysis to suitable analytic potential parametrizations and a limited number of parameters due to constraints of computation speed remains a problem that needs to be solved ultimately.

**Different modelling approaches using action-based DFs:** BR13 focussed on MAPs for a number of reasons: First, they seem to permit simple DFs (Bovy et al. 2012a,b,c), i.e., approximately qDFs (Ting et al. 2013). Second, all stars must orbit in the same potential. While each MAP can yield different DF parameters, it will also provide a (statistically) independent estimate of the potential. This allows for a valuable cross-checking reference. In some sense, the *RoadMapping* approach focusses on constraining the potential, treating the DF parameters as nuisance parameters. That we were able to show in this work that *RoadMapping* results are quite robust to the form of the DF not being entirely correct motivates this approach further.

Magorrian (2014) introduced a framework which avoids specific parametrizations of action-based DFs and marginalizes over all possible DFs to constrain the potential. While this might be the proper way to treat

the nuisance DF, it appears to be computationally very challenging.

For reasons of galaxy and chemical evolution, the DF properties are astrophysically linked between different MAPs (Sanders & Binney 2015). In its current implementation, *RoadMapping* treats all MAPs as independent and does not exploit such correlations. Ultimately, the goal is to do a consistent chemodynamical model that simultaneously fits the potential and DF( $\mathbf{J}, [\mathrm{X}/\mathrm{H}]$ ) (where  $[\mathrm{X}/\mathrm{Fe}][\mathrm{X}/\mathrm{H}]$  is  $[\mathrm{Fe}/\mathrm{H}]$  and other elements either referenced to H or Fe, i.e.,  $[\mathrm{X}/\mathrm{H}]$  denotes the whole abundance space) with a full likelihood analysis. This has not yet been attempted with *RoadMapping*, because the behaviour is quite complex.

Since the first application of *RoadMapping* by BR13 there have been two similar efforts to constrain the Galactic potential and/or orbit DF:

Piffl et al. (2014) fitted both potential and a  $f(\mathbf{J})$  to giant stars from the RAVE survey (Steinmetz et al. 2006) and the vertical stellar number density profiles in the disk by Jurić et al. (2008). They did not include any chemical abundances in the modelling. Instead, they used a superposition of action-based DFs to describe the overall stellar distribution at once: a superposition of qDFs for cohorts in the thin disk, a single qDF for the thick disk stars and an additional DF for the halo stars. Taking proper care of the selection function requires a full likelihood analysis, which is computationally expensive. Piffl et al. (2014) choose to circumvent this difficulty by directly fitting a) histograms of the three velocity components in eight spatial bins to the velocity distribution predicted by the DF and b) the vertical density profile predicted by the DF to the profiles by Jurić et al. (2008). The vertical force profile of their best fit mass model nicely agrees with the results from BR13 for  $R > 6.6$  kpc. The disadvantage of their approach is, that by binning the stars spatially, a lot of information is not used.

Sanders & Binney (2015) have focussed on understanding the abundance-dependence of the DF, relying on a fiducial potential. They developed extended distribution functions (eDF), i.e., functions of both actions and metallicity for a superposition of thin and thick disk, each consisting of several cohorts described by qDFs, a DF for the halo, a functional form of the metallicity of the interstellar medium at the time of birth of the stars, and a simple prescription for radial migration. They applied a full likelihood analysis accounting for selection effects and found a best fit for the eDF in the fixed fiducial potential by Dehnen & Binney (1998) to the stellar phase-space data of the Geneva-Copenhagen Survey (Nordström et al. 2004; Holmberg et al. 2009), metallicity determinations by Casagrande et al. (2011) and the stellar density curves by Gilmore & Reid (1983). Their best fit predicted the velocity distribution of SEGUE G-dwarfs (Ahn et al. 2014) quite well, but had biases in the metallicity distribution, which they accounted to being a problem with the SEGUE metallicities.

**Future work:** We know that real galaxies, including the MW, are not axisymmetric. Using N-body models, we will explore in a subsequent paper how the recovery of the gravitational potential with *RoadMapping* will be affected when data from a non-axisymmetric **systemdisk galaxy system with spiral arms** get interpreted through

axisymmetric models. In this context further investigations of questions that came up in Section 3.3 (“How much of the information on the potential is stored in the rotation curve?”) and earlier in Section 4 (“What is better: fitting a MW-like potential using approximate Stäckel actions, or fitting a Stäckel potential to the MW using exact actions?”) should also be conducted. There are several interesting scientific questions for which a *RoadMapping* investigation of galaxy simulations could be a pragmatic approach to address them: (i) What is the influence of spiral arms and resonances on the modelling outcome? (ii) Can we recover the potential well enough to calculate actions so accurate that clumps in orbit space can be identified? This is important to be able to compare clumps in action space to clustering of stars in abundance space. (iii) How do results from *RoadMapping*, i.e., the potential and DF, compare with results from Jeans models?

[TO DO: mention (somewhere???) that machinery is not yet ready for the application to actual data because of computational speed]

[TO DO: mention (somewhere???) that if we use suitable giant tracers (which are bright and have therefore small errors) within 4 kpc the proper motion errors are probably negligible (compare with tabe 1 in de bruijne) and the distance errors are only 5% which is small enough for our method. There are therefore sub-sets of Gaia stars with almost no errors to which we could RM apply. / APGEE: Giant stars, absolut helligkeit 0 mag, bei 3 kpc distance modulus of 12,  $-i$  m=12 mag  $i$ – bright limit of gaia De Bruijne Paper  $-i$  Post launch the proper motions are infinitely accurate bei 3 kpc  $-i$  nachrechnen]

## 5. ACKNOWLEDGEMENTS

We thank Glenn van de Ven for suggesting the use of Kuzmin-Kutuzov Stäckel potentials in this case study. We also thank the anonymous referee for her/his very helpful comments and suggestions, James J. Binney and Payel Das (University of Oxford) for valuable discussions, and Rene Andrae (MPIA) for an estimate of Gaia distance uncertainties. W.H.T. and H.-W.R. acknowledge funding from the European Research Council under the European Union’s Seventh Framework Programme (FP 7) ERC Grant Agreement n. [321035]. J.B. acknowledges the financial support from the Natural Sciences and Engineering Research Council of Canada.

## REFERENCES

- Abazajian, K., Adelman-McCarthy, J. K., Agüeros, M. A., et al. 2003, AJ, 126, 2081
- Ahn, C. P., Alexandroff, R., Allende Prieto, C., et al. 2014, ApJS, 211, 17
- Batsleer, P., & Dejonghe, H. 1994, A&A, 287, 43
- Binney, J. 2010, MNRAS, 401, 2318
- . 2011, Pramana, 77, 39
- . 2012a, MNRAS, 426, 1324
- . 2012b, MNRAS, 426, 1328
- . 2013, NewAR, 57, 29
- Binney, J., & McMillan, P. 2011, MNRAS, 413, 1889
- Binney, J., & Tremaine, S. 2008, Galactic Dynamics: Second Edition (Princeton University Press)
- Bovy, J. 2015, ApJS, 216, 29
- Bovy, J., & Rix, H.-W. 2013, ApJ, 779, 115
- Bovy, J., Rix, H.-W., Green, G. M., Schlafly, E. F., & Finkbeiner, D. P. 2016, ApJ, 818, 130
- Bovy, J., Rix, H.-W., & Hogg, D. W. 2012a, ApJ, 751, 131

- Bovy, J., Rix, H.-W., Hogg, D. W., et al. 2012b, *ApJ*, 755, 115  
 Bovy, J., Rix, H.-W., Liu, C., et al. 2012c, *ApJ*, 753, 148  
 Bovy, J., & Tremaine, S. 2012, *ApJ*, 756, 89  
 Büdenbender, A., van de Ven, G., & Watkins, L. L. 2015, *MNRAS*, 452, 956  
 Casagrande, L., Schönrich, R., Asplund, M., et al. 2011, *A&A*, 530, A138  
 Das, P., & Binney, J. 2016, ArXiv e-prints, arXiv:1603.09332  
 de Bruijne, J. H. J., Rygl, K. L. J., & Antoja, T. 2014, in EAS Publications Series, Vol. 67, EAS Publications Series, 23–29  
 de Lorenzi, F., Debattista, V. P., Gerhard, O., & Sambhus, N. 2007, *MNRAS*, 376, 71  
 de Zeeuw, T. 1985, *MNRAS*, 216, 273  
 Dehnen, W., & Binney, J. 1998, *MNRAS*, 294, 429  
 ESA, ed. 1997, ESA Special Publication, Vol. 1200, The HIPPARCOS and TYCHO catalogues. Astrometric and photometric star catalogues derived from the ESA HIPPARCOS Space Astrometry Mission  
 Famaey, B., & Dejonghe, H. 2003, *MNRAS*, 340, 752  
 Foreman-Mackey, D., Hogg, D. W., Lang, D., & Goodman, J. 2013, *PASP*, 125, 306  
 Garbari, S., Liu, C., Read, J. I., & Lake, G. 2012, *MNRAS*, 425, 1445  
 Gilmore, G., & Reid, N. 1983, *MNRAS*, 202, 1025  
 Gould, A., & Kollmeier, J. A. 2004, *ApJS*, 152, 103  
 Henon, M. 1959, *Annales d'Astrophysique*, 22, 126  
 Høg, E., Fabricius, C., Makarov, V. V., et al. 2000, *A&A*, 355, L27  
 Holmberg, J., Nordström, B., & Andersen, J. 2009, *A&A*, 501, 941  
 Hunt, J. A. S., & Kawata, D. 2014, *MNRAS*, 443, 2112  
 Ivezić, Ž., Monet, D. G., Bond, N., et al. 2008, in IAU Symposium, Vol. 248, A Giant Step: from Milli- to Micro-arcsecond Astrometry, ed. W. J. Jin, I. Platais, & M. A. C. Perryman, 537–543  
 Ivezic, Z., Tyson, J. A., Abel, B., et al. 2008, ArXiv e-prints, arXiv:0805.2366  
 Jordi, C., Gebran, M., Carrasco, J. M., et al. 2010, *A&A*, 523, A48  
 Jurić, M., Ivezić, Ž., Brooks, A., et al. 2008, *ApJ*, 673, 864  
 Kaiser, N., Burgett, W., Chambers, K., et al. 2010, in Proc. SPIE, Vol. 7733, Ground-based and Airborne Telescopes III, 77330E  
 Kuijken, K., & Gilmore, G. 1989, *MNRAS*, 239, 605  
 Magorrian, J. 2014, *MNRAS*, 437, 2230  
 McMillan, P. J. 2012, in European Physical Journal Web of Conferences, Vol. 19, European Physical Journal Web of Conferences, 10002  
 McMillan, P. J., & Binney, J. 2012, *MNRAS*, 419, 2251  
 McMillan, P. J., & Binney, J. J. 2008, *MNRAS*, 390, 429  
 —. 2013, *MNRAS*, 433, 1411  
 Monet, D. G., Levine, S. E., Canzian, B., et al. 2003, *AJ*, 125, 984  
 Munn, J. A., Monet, D. G., Levine, S. E., et al. 2004, *AJ*, 127, 3034  
 —. 2008, *AJ*, 136, 895  
 Nordström, B., Mayor, M., Andersen, J., et al. 2004, *A&A*, 418, 989  
 Perryman, M. A. C., de Boer, K. S., Gilmore, G., et al. 2001, *A&A*, 369, 339  
 Piffl, T., Binney, J., McMillan, P. J., et al. 2014, *MNRAS*, 445, 3133  
 Read, J. I. 2014, *Journal of Physics G Nuclear Physics*, 41, 063101  
 Reid, M. J., Menten, K. M., Zheng, X. W., et al. 2009, *ApJ*, 700, 137  
 Rix, H.-W., & Bovy, J. 2013, *A&A Rev.*, 21, 61  
 Robin, A. C., Luri, X., Reylé, C., et al. 2012, *A&A*, 543, A100  
 Sanders, J. L., & Binney, J. 2015, *MNRAS*, 449, 3479  
 —. 2016, *MNRAS*, 457, 2107  
 Sesar, B., Bovy, J., Bernard, E. J., et al. 2015, *ApJ*, 809, 59  
 Skrutskie, M. F., Cutri, R. M., Stiening, R., et al. 2006, *AJ*, 131, 1163  
 Steinmetz, M., Zwitter, T., Siebert, A., et al. 2006, *AJ*, 132, 1645  
 Strigari, L. E. 2013, *Phys. Rep.*, 531, 1  
 Syer, D., & Tremaine, S. 1996, *MNRAS*, 282, 223  
 Ting, Y.-S., Rix, H.-W., Bovy, J., & van de Ven, G. 2013, *MNRAS*, 434, 652  
 Yanny, B., Rockosi, C., Newberg, H. J., et al. 2009, *AJ*, 137, 4377  
 Zhang, L., Rix, H.-W., van de Ven, G., et al. 2013, *ApJ*, 772, 108

**Table 3**

Summary of test suites in this work: The first column indicates the test suite, the second column the potential, DF and SF model, etc., used for the mock data creation, the third column the corresponding model assumed in the *RoadMapping* analysis, and the last column lists the figures belonging to the test suite. Reference potentials and qDFs are introduced in Tables 1 and 2, respectively. Parameters that are not left free in the *analysis*, are always fixed to their true value. Unless stated otherwise, all mock data sets have SFs with completeness( $\mathbf{x}$ ) = 1 and no measurement uncertainties, and we use  $N_x = 16$ ,  $N_v = 24$ ,  $n_\sigma = 5$  as numerical accuracy for calculating the likelihood normalisation.

Test	Model for Mock Data			Model in Analysis
Test 1 : Numerical accuracy in calculating the likelihood normalisation	Potential: DF: Survey volume: Numerical accuracy:	KKS-Pot hot or cool qDF sphere around Sun, $r_{\max} = 0.2, 1, 2, 3$ or 4 kpc $N_x \in [5, 20]$ , $N_v \in [6, 40]$ , $n_\sigma \in [3.5, 7]$	$N_x \in [5, 32]$ , $N_v \in [4, 48]$ , $n_\sigma \in [3, 7]$	-
Test 2 : Numerical convergence of convolution with measurement uncertainties	Potential: DF: Survey Volume: Uncertainties: Numerical accuracy: $N_*$ :	Iso-Pot hot qDF sphere around Sun, $r_{\max} = 3$ kpc $\delta RA = \delta Dec = \delta(m - M) = 0$ $\delta v_{los} = 2 \text{ km s}^{-1}$ $\delta \mu = 2, 3, 4$ or $5 \text{ mas yr}^{-1}$ 10,000	$N_{\text{samples}} \in [25, 1200]$	Iso-Pot, all parameters free qDF, all parameters free (fixed & known) (fixed & known)
Test 3.1 : Shape of the model The pdf is a parameters' pdf/multivariate for large data sets Gaussian for large data sets.	Potential: DF: Survey Volume: $N_*$ :	Iso-Pot hot qDF sphere around Sun, $r_{\max} = 2$ kpc 20,000		Iso-Pot, all parameters free qDF, all parameters free (fixed & known)
Test 3.2 : Parameter estimates are unbiased; Influence of survey volume size	Potential: DF: Survey volume: $N_*$ :	Iso-Pot hot or cool qDF sphere around Sun, $r_{\max} = 0.2, 1, 2, 3$ or 4 kpc 20,000		Iso-Pot, free parameter: $b$ qDF, free parameters: $\ln h_R, \ln \sigma_{R,0}, \ln h_{\sigma,R}$ (fixed & known)
Test 4 : Influence of position & shape of survey volume on parameter recovery	Potential: DF: Survey volume: $N_*$ :	(i) Iso-Pot or (ii) MW13-Pot hot qDF 4 different wedges, see Figure 6, upper panel 20,000		(i) Iso-Pot, all parameters free (ii) MW13-Pot, $R_d$ and $f_h$ free $v_{circ}(R_\odot)$ , $a_{disk}$ , $f_{halo}$ (i) qDF, all parameters free (ii) qDF, only $h_R$ , $\sigma_{R,0}$ and $h_{\sigma,R}$ free (fixed & known)
Test 5 : Influence of wrong assumptions about the spatial SF on parameter recovery	Potential: DF: Survey volume: Completeness: $N_*$ :	Iso-Pot hot or cool qDF sphere around Sun, $r_{\max} = 3$ kpc Equation (17) with $\epsilon_r \in [0, 0.7]$ 20,000		Iso-Pot, all parameters free $v_{circ}(R_\odot)$ , $a_{disk}$ , $f_{halo}$ qDF, all parameters free (fixed & known) completeness( $\mathbf{x}$ ) = 1, i.e., $\epsilon_r = 0$
Test 6.1 : Effect of proper motion uncertainties on precision of potential recovery	Potential: DF: Survey volume: Uncertainties: $N_*$ :	Iso-Pot hot or cool qDF sphere around Sun, $r_{\max} = 3$ kpc (i) $\delta RA = \delta Dec = \delta(m - M) = 0$ , $\delta v_{los} = 2 \text{ km s}^{-1}$ , $\delta \mu = 1, 2, 3, 4$ or $5 \text{ mas yr}^{-1}$ (ii) no measurement uncertainties 10,000		Iso-Pot, all parameters free qDF, all parameters free (fixed & known) (fixed & known)

**Table 3** — *Continued*

Test	Model for Mock Data			Model in Analysis
Test 6.2 : Testing the convolution with measurement uncertainties in Equation (16) with & without distance uncertainties	<i>Potential:</i> <i>DF:</i> <i>Survey Volume:</i> <i>Uncertainties:</i>	Iso-Pot hot qDF sphere around Sun, $r_{\max} = 3$ kpc $\delta\text{RA} = \delta\text{Dec} = 0$ , $\delta v_{\text{los}} = 2 \text{ km s}^{-1}$ , $\delta\mu = 1,2,3,4$ or $5 \text{ mas yr}^{-1}$ , (i) $\delta(m - M) = 0$ or (ii) $\delta(m - M) \neq 0$ (see Figure 9) $N_*:$ 10,000		Iso-Pot, all parameters free qDF, all parameters free (fixed & known) (fixed & known)
Test 6.3 : Underestimation of proper motion uncertainties	<i>Potential:</i> <i>DF:</i> <i>Survey volume:</i> <i>Uncertainties:</i>	Iso-Pot hot or cool qDF sphere around Sun, $r_{\max} = 3$ kpc only proper motion uncertainties 1, 2 or 3 mas $\text{yr}^{-1}$ $N_*:$ 10,000		Iso-Pot, all parameters free qDF, all parameters free (fixed & known) proper motion uncertainties 10% or 50% underestimated
Test 7 : Deviations of the assumed DF from the stars' true DF	<i>Potential:</i> <i>DF:</i>	Iso-PotDHB-Pot mix of two qDFs... (i) <i>Example 1:</i> ... with different mixing rates and fixed qDF parameters (hot & cooler qDF from Table 2) (ii) <i>Example 2:</i> ... with 50/50 mixing rate and varying qDF parameters (by X%): a) hot & colder qDF or b) cool & warmer qDF (see Table 2)		Iso-Pot, all parameters free $v_{\text{circ}}(R_{\odot}), a_{\text{disk}}, f_{\text{halo}}$ single qDF, all parameters free
	<i>Survey volume:</i> $N_*:$	sphere around Sun, $r_{\max} = 2$ kpc 20,000		(fixed & known)
Test 8 : Deviations of the assumed potential model from the stars' true potential	<i>Potential:</i> <i>DF:</i> <i>Survey volume:</i> $N_*:$	MW14-Pot hot or cool qDF sphere around Sun, $r_{\max} = 4$ kpc 20,000		KKS-Pot, all parameters free, only $v_{\text{circ}}(R_{\odot}) = 230 \text{ km s}^{-1}$ fixed qDF, all parameters free (fixed & known)
	<i>Numerical accuracy:</i>			$N_x = 20, N_v = 28, n_{\sigma} = 5.5$

## APPENDIX

## A. MOCK DATA

We will rely on mock data as input to explore the limitations of the modelling. The mock data in this work is generated according to the following procedure:

We assume that the positions and velocities of our stellar mock sample are indeed drawn from our assumed family of potentials (Section 2.3) and DFs (Section 2.4), (with given parameters  $p_\Phi$  and  $p_{\text{DF}}$ ). The DF is in terms of actions, while the transformation  $(\mathbf{x}_i, \mathbf{v}_i) \xrightarrow{\Phi} \mathbf{J}_i$  is computationally much less expensive than its inversion. We therefore employ the following efficient two-step method for creating mock data, which also accounts for a survey SF( $\mathbf{x}$ ).

In the first step we draw stellar positions  $\mathbf{x}_i$ . We start by setting up the interpolation grid for the tracer density  $\rho(R, |z| \mid p_\Phi, p_{\text{DF}})$  generated according to Section 2.4.<sup>12</sup> Next, we sample random positions  $(R_i, z_i, \phi_i)$  uniformly within the observable volume. Using a Monte Carlo rejection method we then shape the samples distribution to follow  $\rho(R, |z| \mid p_\Phi, p_{\text{DF}})$ . To apply a non-uniform completeness function, we use the rejection method a second time. The resulting set of positions  $\mathbf{x}_i$  follows the distribution  $p(\mathbf{x}) \propto \rho_{\text{DF}}(R, |z| \mid p_\Phi, p_{\text{DF}}) \times \text{SF}(\mathbf{x})$ .

In the second step we draw velocities  $\mathbf{v}_i$ . For each of the positions  $(R_i, z_i)$  we first sample velocities from a Gaussian envelope function in velocity space which is then shaped towards  $\text{DF}(\mathbf{J}[R_i, z_i, \mathbf{v} \mid p_\Phi] \mid p_{\text{DF}})$  using a rejection method. We now have a mock data set satisfying  $(\mathbf{x}_i, \mathbf{v}_i) \longrightarrow p(\mathbf{x}, \mathbf{v}) \propto \text{DF}(\mathbf{J}[\mathbf{x}, \mathbf{v} \mid p_\Phi] \mid p_{\text{DF}}) \times \text{SF}(\mathbf{x})$ .

[TO DO: The referee wrote: "The discussion of selection on very erroneous  $\mathbf{x}$  coordinates is interesting but surely this isn't the way the data will actually be handled?" - What does he mean? What's wrong? - we are not 100% sure that we understand the referee's main point but given 10% distance errors at 3kpc (surveys edge) the volume from which data can get in practice it's likely to be a magnitude cut but for intrinsic standard candles like red clump stars a magnitude cut translates to a distance cut we realise it's idealized and there won't be a sharp distance cut scatter in and out of survey volume will play a role and we explore an ideal case]

Measurement uncertainties can be added to the mock data by applying the following modifications to the above procedure. We assume Gaussian uncertainties in the heliocentric phase-space coordinates  $\tilde{\mathbf{x}} = (\text{RA}, \text{Dec}, (m-M)), \tilde{\mathbf{v}} = (\mu_{\text{RA}} \cdot \cos \text{Dec}, \mu_{\text{Dec}}, v_{\text{los}})$  (see Section 2.1). In the case of distance and position uncertainties stars virtually scatter in and out of the observed volume. To account for this, we draw the *true*  $\mathbf{x}_i$  from a volume that is larger than the actual observation volume, perturb the  $\mathbf{x}_i$  according to the position uncertainties and then reject all stars that lie now outside of the observed volume. This mirrors the Poisson scatter around the detection threshold for stars whose distances are determined from the apparent brightness and the distance modulus. We then sample *true*  $\mathbf{v}_i$  (given the *true*  $\mathbf{x}_i$ ) as described above and perturb them according to the velocity uncertainties.

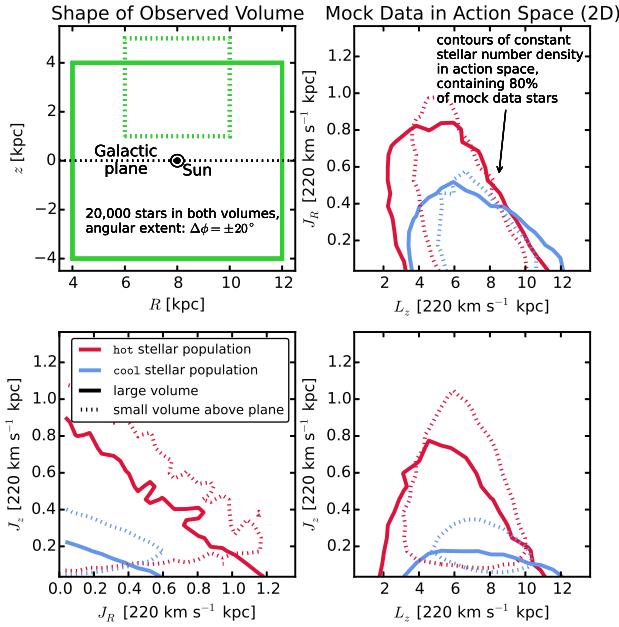
We show examples of mock data sets (without measurement uncertainties) in action space (Figure A1) and configuration space ( $\mathbf{x}, \mathbf{v}$ ) (Figure A2). The mock data generated from the qDF follow the expected distributions in configuration space. The distribution in action space illustrates the intuitive physical meaning of actions: The stars of the cool population have in general lower radial and vertical actions, as they are on more circular orbits. Circular orbits with  $J_R = 0$  and  $J_z = 0$  can only be observed in the Galactic mid-plane. The different ranges of angular momentum  $L_z$  in the two example observation volumes reflect  $L_z \sim R \times v_{\text{circ}}$  and the volumes' different radial extent. The volume at larger  $z$  contains stars with higher  $J_z$ . An orbit with  $L_z \ll$  or  $\gg L_z(R_\odot)$  can only reach into a volume at  $\sim R_\odot$ , if it is more eccentric and has therefore larger  $J_R$ . This together with the effect of asymmetric drift explains the asymmetric distribution of  $J_R$  vs.  $L_z$  in Figure A1.

## B. SELECTION FUNCTIONS

Any survey's selection function (SF) can be understood as defining an effective sample sub-volume in the space of observables, e.g., position on the sky (limited by the pointing of the survey), distance from the Sun (limited by brightness and detector sensitivity), colors and metallicity of the stars (limited by survey mode and targeting). The SF can therefore be thought of as having both spatial small scale structure (due to pencil beam pointing, dust obscuration, etc.) and some overall spatial characteristics (e.g., mean height above the plane and mean Galactocentric radius of the stars). The treatment of realistic and complex SFs was already demonstrated in BR13 (who used the pencil-beam SF of the SEGUE survey (Bovy et al. 2012c)) and Bovy et al. (2016) (who investigated the effect of dust extinction). In this work we aim to make a generic and basic exploration of search volume shapes and, as shown by Bovy et al. (2016), this should be possible without explicitly considering spatial SF substructure. Inspired by the contiguous nature of the Gaia SF, which is basically only limited by a magnitude cut, and the fact that this magnitude cut would—in the absence of small scale structure—translate to a sharp distance cut for standard candle tracer populations like red clump stars, we therefore use in our modelling a simple spatial SF of spherical shape with radius  $r_{\text{max}}$  around the Sun. In our modelling we use simple spatial SFs, describes the probability to observe a star at position  $\mathbf{x}$ ,

$$\text{SF}(\mathbf{x}) \equiv \begin{cases} \text{completeness}(\mathbf{x}) & \text{if } \mathbf{x} \text{ within obs. volume, if } |\mathbf{x} - \mathbf{x}_\odot| \leq r_{\text{max}}, \\ 0 & \text{if } \mathbf{x} \text{ outside. otherwise,} \end{cases} \quad (\text{B1})$$

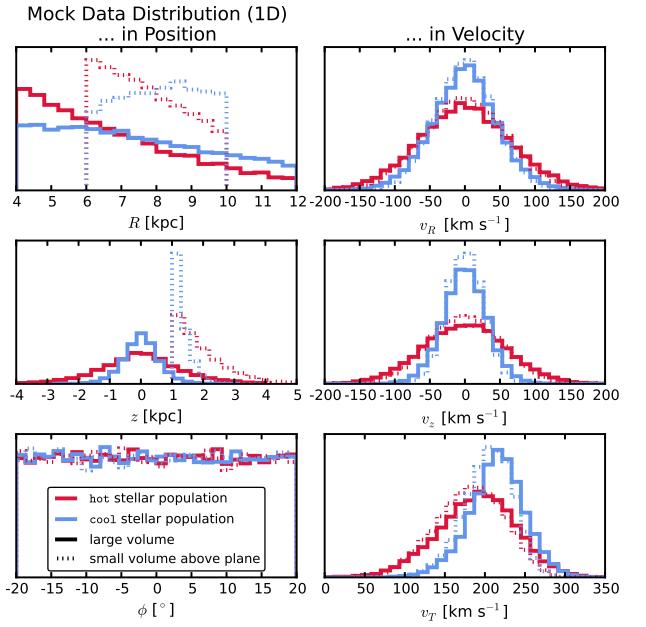
<sup>12</sup> For the creation of the mock data we use  $N_x = 20$ ,  $N_v = 40$  and  $n_\sigma = 5$  in Equation (8).



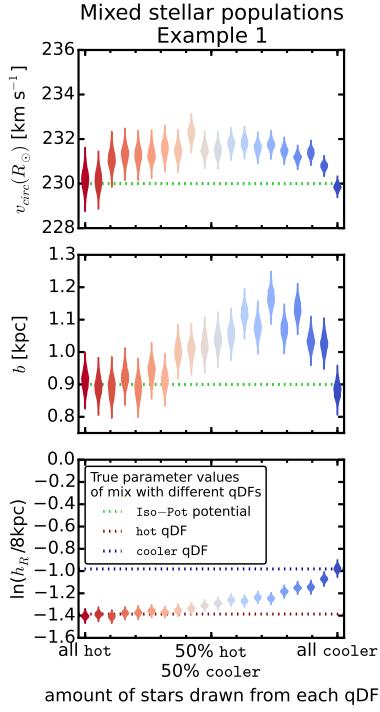
**Figure A1.** Distribution of mock data in action space (2D iso-density contours, enclosing 80% of the stars), depending on shape and position of a wedge-like survey observation volume (upper left panel) and temperature of the stellar population (indicated in the legend). The  $p_M$  of the mock data, created in the KKS-Pot potential, are given as Test 1 in Table 3. The four mock data sets are generated in the KKS-Pot from Table 1 from either the hot or cool DF in Table 2. The distribution in action space visualizes how orbits with different actions reach into different regions within the Galaxy. The corresponding mock data in configuration space is shown in Figure A2.

The SF of the SEGUE survey (Bovy et al. 2012c) used by BR13 consists of many pencil-beams. In anticipation of large contiguous volume surveys like Gaia, we use SFs that span large observed volumes of simple geometrical shapes: a sphere of radius  $r_{\max}$  with the Sun at its center; or an angular segment of an cylindrical annulus (wedge), i.e., the volume with  $R \in [R_{\min}, R_{\max}], \phi \in [\phi_{\min}, \phi_{\max}], z \in [z_{\min}, z_{\max}]$  within the model Galaxy. The sharp outer edge of the survey volume could be interpreted as a detection limit in apparent brightness in the case where all stars have the same luminosity. We set  $0 \leq \text{completeness}(\mathbf{x}) \leq 1$  everywhere inside the observed volume, so it can be understood as a position-dependent detection probability for a star at  $\mathbf{x}$ . Unless explicitly stated otherwise, we simplify to  $\text{completeness}(\mathbf{x}) = 1$ . Additionally, we use in Figures 6 (Test 4), A1 and A2 for illustrative purposes some rather unrealistic survey volumes which are angular segments of a cylindrical annulus (wedge), i.e., the volume with  $R \in [R_{\min}, R_{\max}], \phi \in [\phi_{\min}, \phi_{\max}], z \in [z_{\min}, z_{\max}]$  within the model Galaxy.

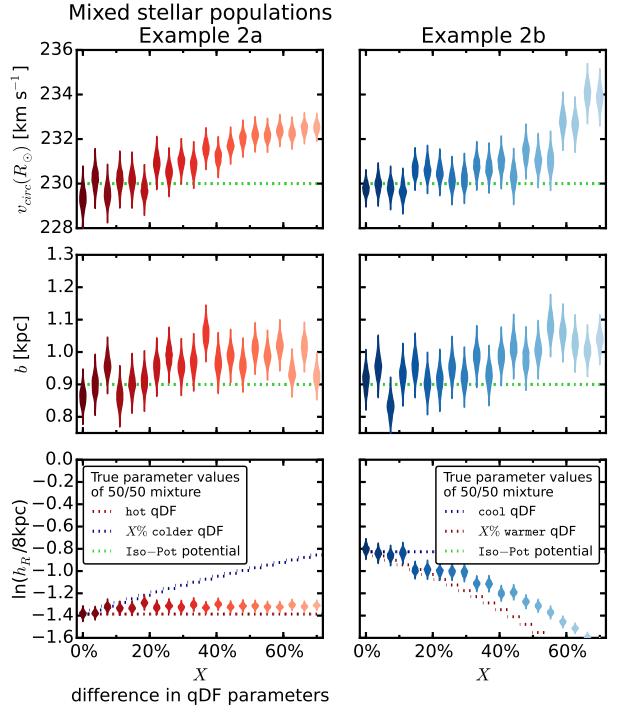
[TO DO: Make sure numbering in Table 3 is correct.] [TO DO: Remove latex comments] [TO DO: Make sure SF is still somewhere properly introduced.] [TO DO: Reference on Staeckel pancakes] [TO DO: Look for double spaces and replace them with one space. Make sure that this is the case in the PDF, also with all the corrections.] [TO DO: Make sure that nowhere in this work the MW13-Potential is still used.]



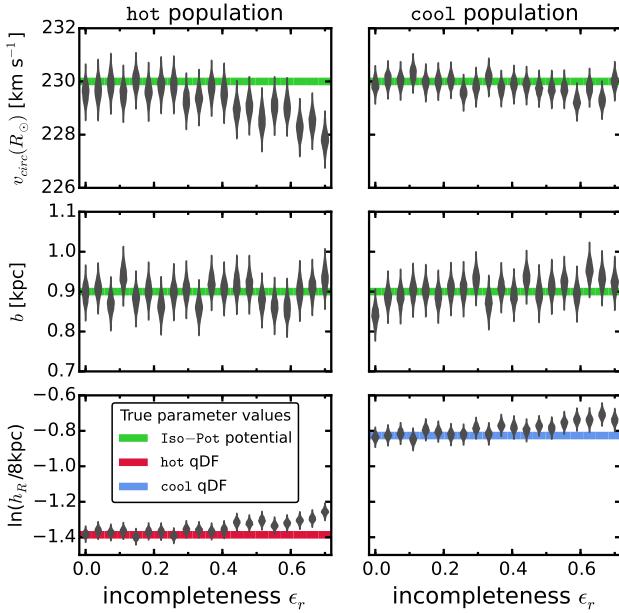
**Figure A2.** Distribution of the mock data from Figure A1 in configuration space. The corresponding observation volumes (as indicated in the legend) are shown in Figure A1, upper left panel. The 1D histograms illustrate that qDFs generate realistic stellar distributions in Galactocentric coordinates ( $R, z, \phi, v_R, v_z, vT$ ): More stars are found at smaller  $R$  and  $|z|$ , and are distributed uniformly in  $\phi$  according to our assumption of axisymmetry. The distribution in radial and vertical velocities,  $v_R$  and  $v_z$ , is approximately Gaussian with the (total projected) velocity dispersion being of the order of  $\sim \sigma_{R,0}$  and  $\sim \sigma_{z,0}$  (see Table 2). The distribution of tangential velocities  $v_T$  is skewed because of asymmetric drift.



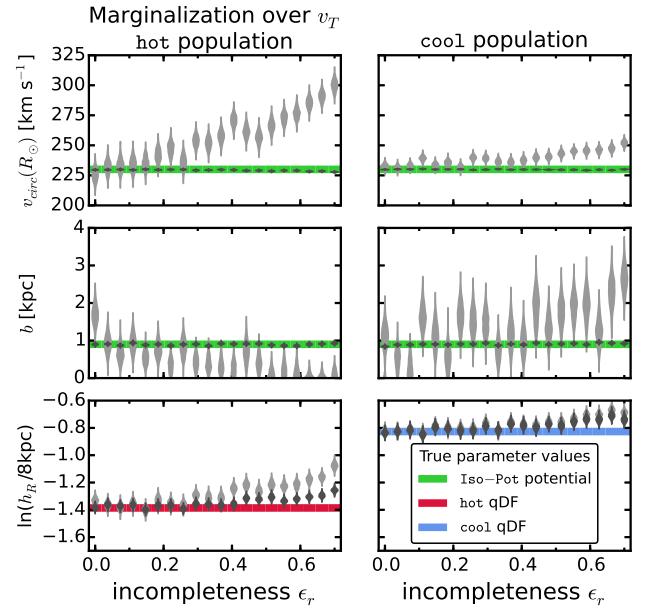
**Figure B1.** [TO DO: This is an old plot, with analyses using the isochrone potential. Delete it later. The new plot in shown in Figure 13.]



**Figure B2.** [TO DO: This is an old plot, with analyses using the isochrone potential. Delete it later. The new plot in shown in Figure 14.]



**Figure B3.** [TO DO: This is an old plot, with analyses using the isochrone potential. Delete it later. The new version of this plot is Figure 7.]



**Figure B4.** [TO DO: This is an old plot, with analyses using the isochrone potential. I do not need it in the new version of the paper. Delete it later.] Same as Figure 7, but without including information about the tangential velocities in the analysis. This was done by marginalizing the likelihood in Equation (9) over  $v_T$  (bright grey violins; the dark grey violins are the same as in Figure 7 for comparison). The parameter recovery is much worse than in Figure 7. This could indicate that much of the 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 ( $\epsilon_r \lesssim 0.15$ ).