

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

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

*Draft version September 22, 2015*

## ABSTRACT

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

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

## 1. INTRODUCTION

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

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

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

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

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

the stellar distribution through a distribution function.

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

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

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

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

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

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

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

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

One of the strongest assumptions is to restrict the dynamical modelling to a certain family of parametrized models. We investigate how well we can hope to recover the true potential, when our potential and DF models do not encompass the true potential and DF. First, we examine in Section 3.5 what would happen if the stars within MAPs do intrinsically not follow a single qDF as assumed by Ting et al. (2013) and Bovy & Rix (2013). Second, we test in Section 3.6 how well the modelling works, if our assumed potential family deviates from the true potential.

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

For all of the above aspects we show some plausible and illustrative examples on the basis of investigating mock data. The mock data is generated from galaxy models presented in Sections 2.1-2.4 following the procedure in Section 2.5, analysed according to the description of the *RoadMapping* machinery in Sections 2.6-2.7 and the results are presented in Section 3 and discussed in Section 4.

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

[TO DO: Comment from Hans-Walter: Make clear "new in this paper", "general background", "exactly as in BR13"]

## 2. DYNAMICAL MODELLING

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

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

### 2.1. Coordinate System

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

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

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

### 2.2. Actions and Potential Models

Orbits in axisymmetric potentials are best described and fully specified by the three actions  $\mathbf{J} \equiv (J_R, J_z, J_\phi = L_z)$  (Binney & Tremaine 2008, §[TO DO]). Their computation from a star's phase-space coordinates,  $(\mathbf{x}, \mathbf{v}) \rightarrow \mathbf{J}$ , is typically very expensive and depends on the choice of gravitational potential in which the star moves. The spherical isochrone potential (Henon 1959) and axisymmetric Stäckel potential [TO DO: REF] are the most general Galactic potentials, that allow exact action calculations (Binney & Tremaine 2008, §3.5.2 and [TO DO: REF]). In all other potentials actions have to be numerically estimated, e.g., by using the *Stäckel fudge* by Binney (2012) for axisymmetric potentials and action interpolation grids (Bovy 2015) to speed up the calculation. The latter is one of the improvements employed by *RoadMapping*, which was not used by Bovy & Rix (2013).

For the gravitational potential in our modelling we assume a family of parametrized potential models. We use: The Milky Way-like potential from Bovy & Rix (2013) (*MW13-Pot*) with bulge, disk and halo; 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 density distribution of these potentials is illustrated in Figure 1.

### 2.3. Stellar Distribution Functions

Throughout, we assume that each of our model stellar populations follows the single action-based quasi-isothermal distribution function (qDF) by Binney & McMillan (2011). This is motivated by the findings of Bovy et al. (2012b,c,d) and Ting et al. (2013) about the simple phase-space structure of stellar mono-abundance populations (MAP), and following Bovy & Rix (2013) and their successful application. The qDF by Binney & McMillan (2011) has the form

$$\text{qDF}(\mathbf{J} | p_{\text{DF}}) = f_{\sigma_R}(J_R, L_z | p_{\text{DF}}) \times f_{\sigma_z}(J_z, L_z | p_{\text{DF}}) \quad (1)$$

with

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

$$f_{\sigma_z}(J_z, L_z | p_{\text{DF}}) = \frac{\nu}{2\pi \sigma_z^2(R_g)} \exp\left(-\frac{\nu J_z}{\sigma_z^2(R_g)}\right). \quad (3)$$

Here  $R_g \equiv R_g(L_z)$  and  $\Omega \equiv \Omega(L_z)$  are the (guiding-center) radius and the circular frequency of the circular orbit with angular momentum  $L_z$  in a given potential.  $\kappa \equiv \kappa(L_z)$  and  $\nu \equiv \nu(L_z)$  are the radial/epicycle ( $\kappa$ ) and vertical ( $\nu$ ) orbital frequencies (Binney & Tremaine 2008, §3.2.3) [TO DO: ask someone, if I'm messing up different definitions of  $\kappa$ : frequencies with which the star would oscillate around the circular orbit in  $R$ - and  $z$ -direction when slightly perturbed]. The term  $[1 + \tanh(L_z/L_0)]$  suppresses counter-rotation for orbits in the disk with  $L \gg L_0$  (with  $L_0 = 10 \times R_\odot / 8 \times v_{\text{circ}}(R_\odot) / 220$  [TO DO: Jo said, galpy default is 10 km/s kpc. But I got the value actually from the code...]). Again following Bovy & Rix (2013), we choose the functional forms

$$n(R_g | p_{\text{DF}}) \propto \exp\left(-\frac{R_g}{h_R}\right) \quad (4)$$

$$\sigma_R(R_g | p_{\text{DF}}) = \sigma_{R,0} \times \exp\left(-\frac{R_g - R_\odot}{h_{\sigma,R}}\right) \quad (5)$$

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

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 Bovy & Rix (2013) 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 Table 2: Most tests use the *hot* and *cool* qDFs, which correspond to kinematically hot and cool populations, respectively.

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

**Table 1**

Gravitational potentials of the reference galaxies used throughout this work and the respective ways to calculate actions in these potentials. All four potentials are axisymmetric. The potential parameters are fixed for the mock data creation at the values given in this table. In the subsequent analyses we aim to recover these potential parameters again. The parameters of **MW13-Pot** and **KKS-Pot** were chosen to resemble the **MW14-Pot** (see Figure 1). We use  $v_{\text{circ}}(R_{\odot}) = 230 \text{ km s}^{-1}$  for all potentials in this work.

| name            | potential type  | potential parameters $p_{\Phi}$                      |         | action calculation  |
|-----------------|---|--|---------|---|
| <b>Iso-Pot</b>  | isochrone potential <sup>(a)</sup><br>(Henon 1959)  | $b$  | 0.9 kpc | <i>analytical and exact</i><br>(Binney & Tremaine 2008, §3.5.2)   |
| <b>KKS-Pot</b>  | 2-component   | $\Delta$   | 0.3     | <i>exact</i>  |
|                 | Kuzmin-Kutuzov-   | $\left(\frac{a}{c}\right)_{\text{Disk}}$             | 20      | using <i>Stäckel Fudge</i>  |
|                 | Stäckel potential <sup>(b)</sup>  | $\left(\frac{a}{c}\right)_{\text{Halo}}$             | 1.07    | (Binney 2012)   |
|                 | (disk + halo)<br>(Batsleer & Dejonghe 1994)   | $k$  | 0.28    | and interpolation<br>on action grid <sup>(e)</sup><br>(Bovy 2015) |
| <b>MW13-Pot</b> | MW-like potential <sup>(c)</sup> with<br>Hernquist bulge,<br>spherical power-law halo,<br>2 exponential disks<br>(stars + gas)<br>(Bovy & Rix 2013) | $R_d$  | 3 kpc   | <i>approximate</i><br>(same as KKS-Pot)                           |
|                 |   | $z_h$  | 0.4 kpc |   |
|                 |   | $f_h$  | 0.5     |   |
|                 |   | $\frac{d \ln(v_{\text{circ}}(R_{\odot}))}{d \ln(R)}$ | 0       |   |
| <b>MW14-Pot</b> | MW-like potential <sup>(d)</sup> with<br>cut-off power-law bulge,<br>Miyamoto-Nagai stellar disk,<br>NFW halo<br>(Bovy 2015)                        |  |         | <i>approximate</i><br>(same as KKS-Pot)                           |

<sup>(a)</sup> The isochrone potential **Iso-Pot** has one free parameter, the scale length  $b$ .

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

<sup>(c)</sup> The free parameters of the **MW13-Pot** are stellar disk scale length  $R_d$  and height  $z_d$ , as well as the relative halo contribution to  $v_{\text{circ}}^2(R_{\odot})$ ,  $f_h$ , and the slope of the rotation curve,  $\frac{d \ln(v_{\text{circ}}(R_{\odot}))}{d \ln(R)}$ .

<sup>(d)</sup> The **MWPotential2014** by Bovy (2015) (see their Table 1) has a circular velocity at the Sun of  $v_{\text{circ}}(R_{\odot}) = 220 \text{ km s}^{-1}$ . In this work we use however  $v_{\text{circ}}(R_{\odot}) = 230 \text{ km s}^{-1}$  for all potentials.

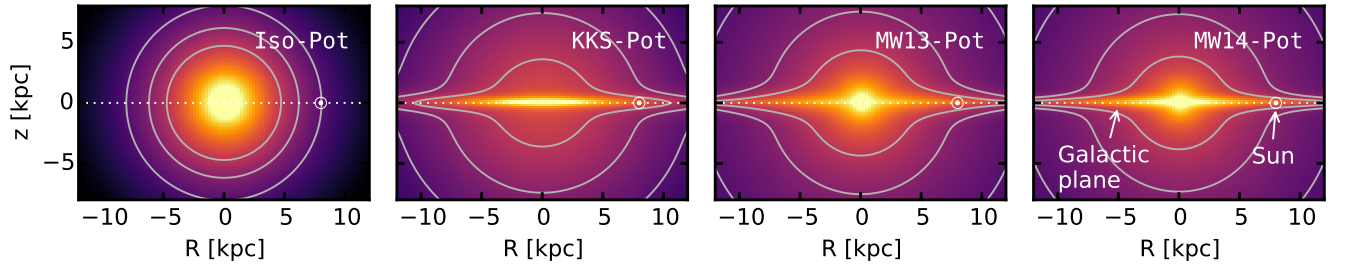
<sup>(e)</sup> We use a finely spaced action interpolation grid with  $R_{\text{max}} = 10$  [TO DO: What's that??? units???] and 50 grid points in  $E$  and  $\psi$  [TO DO: Find out what's that???], and 60 grid points in  $L_z$ .

**Table 2**

Reference distribution-function parameters for the qDF in Equations (1)-(6).

These DFs describe the phase-space distribution of stellar populations for which mock data is created and analysed throughout this work for testing purposes. The parameters of the **cooler & colder (hotter & warmer)** qDFs were chosen to have the same  $\sigma_{R,0}/\sigma_{z,0}$  ratio as the **hot (cool)** qDF. The **colder** and **warmer** qDF have a free parameter  $X$  that governs how much colder/warmer they are then the reference **hot** and **cool** qDFs. Hotter populations have shorter tracer scale lengths (Bovy et al. 2012d) and the velocity dispersion scale lengths were fixed according to Bovy et al. (2012c).

| name          | qDF parameters $p_{\text{DF}}$ |                                      |                                      |                      |                      |
|---------------|--------------------------------|--------------------------------------|--------------------------------------|----------------------|----------------------|
|               | $h_R$ [kpc]                    | $\sigma_{R,0}$ [km s <sup>-1</sup> ] | $\sigma_{z,0}$ [km s <sup>-1</sup> ] | $h_{\sigma,R}$ [kpc] | $h_{\sigma,z}$ [kpc] |
| <b>hot</b>    | 2                              | 55                                   | 66                                   | 8                    | 7                    |
| <b>cool</b>   | 3.5                            | 42                                   | 32                                   | 8                    | 7                    |
| <b>cooler</b> | 2 + 50%                        | 55-50%                               | 66-50%                               | 8                    | 7                    |
| <b>hotter</b> | 3.5-50%                        | 42+50%                               | 32+50%                               | 8                    | 7                    |
| <b>colder</b> | 2 + X%                         | 55-X%                                | 66-X%                                | 8                    | 7                    |
| <b>warmer</b> | 3.5-X%                         | 42+X%                                | 32+X%                                | 8                    | 7                    |



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



tential. Analogously to Bovy & Rix (2013),

$$\rho_{\text{DF}}(R, |z| \mid p_{\Phi}, p_{\text{DF}}) = \int_{-\infty}^{\infty} \text{qDF}(\mathbf{J}[R, z, \mathbf{v} \mid p_{\Phi}] \mid p_{\text{DF}}) d^3\mathbf{v} \quad (7)$$

$$\approx \int_{-n_{\sigma}\sigma_R(R \mid p_{\text{DF}})}^{n_{\sigma}\sigma_R(R \mid p_{\text{DF}})} \int_{-n_{\sigma}\sigma_z(R \mid p_{\text{DF}})}^{n_{\sigma}\sigma_z(R \mid p_{\text{DF}})} \int_0^{1.5v_{\text{circ}}(R_{\odot})} \text{qDF}(\mathbf{J}[R, z, \mathbf{v} \mid p_{\Phi}] \mid p_{\text{DF}}) dv_T dv_z dv_R, \quad (8)$$

where  $\sigma_R(R \mid p_{\text{DF}})$  and  $\sigma_z(R \mid p_{\text{DF}})$  are given by Equations 5 and 6.<sup>4</sup> Each integral is evaluated using a  $N_v$ -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 in between using bivariate spline interpolation. The grid is chosen to cover the extent of the observations (for  $|z| \leq 0$ , because the model is symmetric in  $z$  by construction). The total number of actions to be calculated to set up the density interpolation grid is  $N_x^2 \cdot N_v^3$ , which is one of the speed limiting factors. To complement the work by Bovy & Rix (2013), we will specifically work out in Section 2.6 and Figure 3 how large  $N_x$ ,  $N_v$  and  $n_{\sigma}$  have to be chosen to get the density with a sufficiently high numerical accuracy [TO DO: Ask Jo, if he really think that this is difficult to understand here (because we have not yet talked about the normalization)].

#### 2.4. Selection Functions

Any survey’s selection function (sf) can be understood as defining an effective sample sub-volume in the space of observables. We use simple spatial sfs, which describe the probability to observe a star at  $\mathbf{x}$ ,

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

The sf of the SEGUE survey [TO DO: Ask Jo for reference to SEGUE sf] used by Bovy & Rix (2013) consists of many pencil-beams. In anticipation of [TO DO: What is the word for surveys covering large volumes????] like Gaia, we use sfs that span large observed volumes of simple geometrical shapes: a sphere of radius  $r_{\text{max}}$  with the Sun at its center; or an angular segment of an cylindrical annulus (wedge), i.e. the volume with  $R \in [R_{\text{min}}, R_{\text{max}}]$ ,  $\phi \in [\phi_{\text{min}}, \phi_{\text{max}}]$ ,  $z \in [z_{\text{min}}, z_{\text{max}}]$  within the model Galaxy. The sharp outer 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. Here  $0 \leq \text{completeness}(\mathbf{x}) \leq 1$  everywhere inside the observed volume, so it can be understood as a position-dependent detection probability. Unless explicitly stated otherwise, we simplify to  $\text{completeness}(\mathbf{x}) = 1$ .

#### 2.5. Mock Data

We will rely on mock data as input to explore the limitations of the modelling. We assume that the positions and velocities of our stellar mock sample are

<sup>4</sup> The integration ranges over the velocity are motivated by Figure 2. The integration range  $[0, 1.5v_{\text{circ}}(R_{\odot})]$  over  $v_T$  is in general sufficient, only for observation volumes with larger mean stellar  $v_T$  this upper limit needs to be increased.

indeed drawn from our assumed family of potentials and DFs (with given parameters  $p_{\Phi}$  and  $p_{\text{DF}}$ ). The DF is in terms of actions, while the transformation  $(\mathbf{x}_i, \mathbf{v}_i) \rightarrow \mathbf{J}_i$  is computationally much less expensive than the inversion. We therefore employ the following effective two-step method for creating mock data, which also accounts for a survey sf.

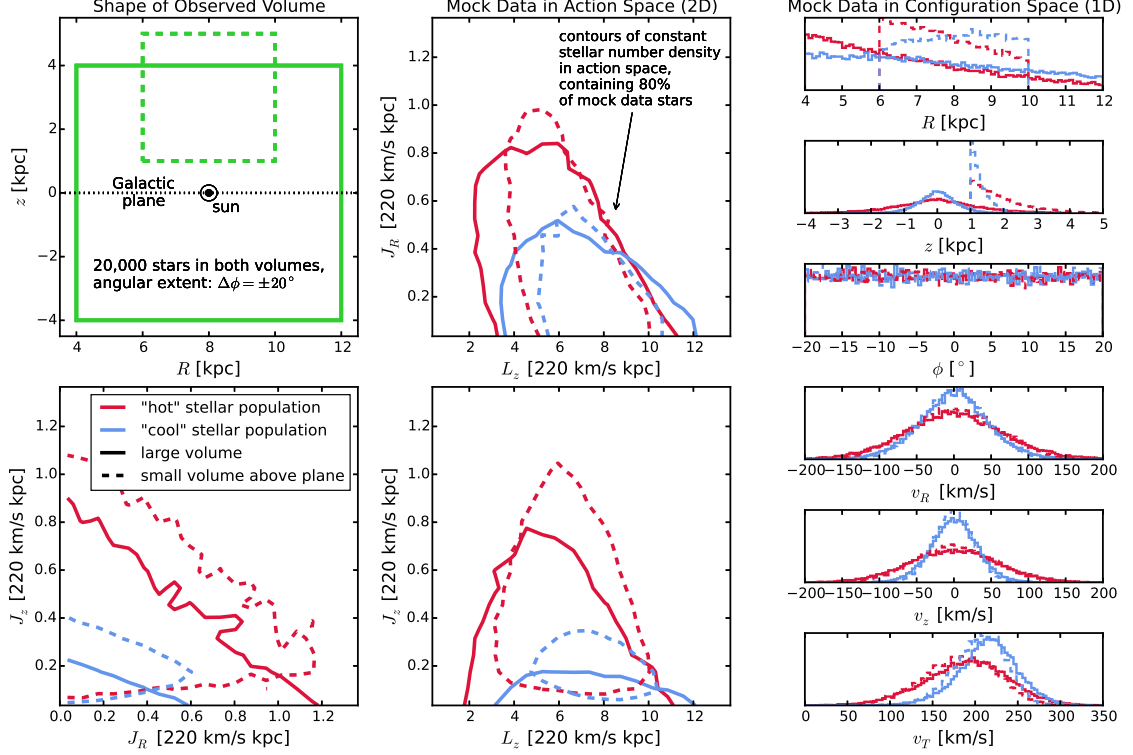
In the first step we draw stellar positions  $\mathbf{x}_i$ . We start by setting up the interpolation grid for the tracer density  $\rho(R, |z| \mid p_{\Phi}, p_{\text{DF}})$  generated according to Section 2.3.<sup>5</sup> Next, we sample random positions  $(R_i, z_i, \phi_i)$  uniformly within the observable volume. Using a Monte Carlo rejection method we then shape the sample to follow  $\rho(R, |z| \mid p_{\Phi}, p_{\text{DF}})$ . To apply a non-uniform sf( $\mathbf{x}$ ), 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  $(\mathbf{x}_i, \mathbf{v}_i) \rightarrow p(\mathbf{x}, \mathbf{v}) \propto \text{DF}(\mathbf{J}[\mathbf{x}, \mathbf{v} \mid p_{\Phi}] \mid p_{\text{DF}}) \times \text{sf}(\mathbf{x})$ .

Figure 2 shows examples of mock data sets in configuration space  $(\mathbf{x}, \mathbf{v})$  and action space. The mock data from the qDF lead to the expected distributions in configuration space. 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 2.

Measurement uncertainties can be added to the mock data by applying the following modifications to the above procedure. We assume Gaussian errors in the heliocentric phase-space coordinates  $\tilde{\mathbf{x}} = (\text{RA}, \text{DEC}, (m - M))$ ,  $\tilde{\mathbf{v}} = (\mu_{\text{RA}}, \mu_{\text{DEC}}, v_{\text{los}})$  (see Section 2.1), where we have taken  $(m - M)$  as a proxy for distance. In the case of distance uncertainties  $\delta(m - M)$ , 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.

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



**Figure 2.** Distribution of mock data in action space (2D iso-density contours, enclosing 80% of the stars, the two central and the lower left panel) and configuration space (1D histograms, right panels), depending on shape and position of the survey observation volume and temperature of the stellar population. The parameters of the mock data model is given as Test 1 in Table 3. In the upper left panel we demonstrate the shape of the two different wedge-like observation volumes within which we were creating each a hot (red) and cool (blue) mock data set: a large volume centred on the Galactic plane (solid lines) and a smaller one above the plane (dashed lines). The distribution in action space visualizes how orbits with different actions also reach into different regions within the Galaxy. The 1D histograms on the right illustrate that qDFs generate realistic stellar distributions in galactocentric coordinates  $(R, z, \phi, v_R, v_z, v_T)$ . [TO DO: fancybox Legend] [TO DO: Potential and/or population names in typewriter font] [TO DO: Jo suggests to make two or three separate figures out of this. I'm not yet convinced, as I think it is nice and tidy like this.]

### 2.6. Data Likelihood

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

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

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

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

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

$$p_{\text{DF}} := \{\ln h_R, \ln \sigma_{R,0}, \ln \sigma_{z,0}, \ln h_{\sigma,R}, \ln h_{\sigma,z}\}. \quad (9)$$

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

The likelihood of the data given the model is

$$\begin{aligned} \mathcal{L}(D \mid p_M) &= \prod_i^N p(\mathbf{x}_i, \mathbf{v}_i \mid p_M) \\ &= \prod_i^N \frac{\text{qDF}(\mathbf{J}_i \mid p_M) \cdot \text{sf}(\mathbf{x}_i)}{\int d^3x d^3v \text{qDF}(\mathbf{J} \mid p_M) \cdot \text{sf}(\mathbf{x})} \\ &\propto \prod_i^N \frac{\text{qDF}(\mathbf{J}_i \mid p_M)}{\int d^3x \rho_{\text{DF}}(R, |z| \mid p_M) \cdot \text{sf}(\mathbf{x})}, \end{aligned} \quad (10)$$

where  $N$  is the number of stars in the data set  $D$ , and in the last step we used Equation 8. The factor  $\prod_i \text{sf}(\mathbf{x}_i)$  is independent of the model parameters so we treat it as unimportant proportionality factor in the likelihood calculation. We find the best set of model parameters by maximizing the posterior probability distribution  $\text{pdf}(p_M \mid D)$ , which is according to Bayes' theorem proportional the likelihood  $\mathcal{L}(D \mid p_M)$  times the prior. We assume flat priors in both  $p_{\Phi}$  and  $p_{\text{DF}}$  (see Equation 9) through out this work, then  $\text{pdf}$  and likelihood can and will be used interchangeably for the remainder of the work.

The normalisation in Equation 10 is a measure for the

total number of tracers inside the survey volume,

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

In the case of an axisymmetric galaxy model and  $\text{sf}(\mathbf{x}) = 1$  everywhere inside the observed volume (i.e. a complete sample as assumed in most tests in this work), the normalisation is essentially a two-dimensional integral in  $R$  and  $z$  of the interpolated tracer density  $\rho_{\text{DF}}$  in Equation 8 over the differential survey volume, i.e.  $\frac{\partial M_{\text{tot}}}{\partial \phi}(R, z) = \int dR \, dz \, \rho_{\text{DF}} \times \frac{\partial V}{\partial \phi}$  [TO DO: missing factor of  $R$ ???]. We perform this integral as a Gauss Legendre quadrature of order 40 in each  $R$  and  $z$  direction. The angular integral, i.e.  $M_{\text{tot}} = \int R d\phi \frac{\partial M_{\text{tot}}}{\partial \phi}$ , can be solved analytically.

It turns out that the sufficiently accurate evaluation of the likelihood is computationally expensive, even for only one set of model parameters. This expense is dominated by the number of action calculations required, which in turn depends on the number of stars in the sample and the numerical accuracy of the integrals in Equation 8 needed for the normalisation, which requires  $N_x^2 \times N_v^3$  action calculations. The accuracy has to be chosen high enough, such that a resulting numerical error

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

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

$$\begin{aligned} & \log \mathcal{L}(p_M | D) \\ &= \sum_i^N \log \text{qDF}(\mathbf{J}_i | p_M) - 3N \log(r_o v_o) \\ & \quad - N \log(M_{\text{tot}}) - N \log(1 + \delta_{M_{\text{tot}}}), \end{aligned} \quad (13)$$

with

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

In other words, this error is only small enough if it does not affect the comparison of two adjacent models whose log-likelihoods differ, to be clearly distinguishable, by 1. Otherwise numerical inaccuracies could lead to systematic biases in the potential and DF fitting. For data sets as large as  $N = 20,000$  stars, which in the age of Gaia could very well be the case [TO DO: Really???], one needs a numerical accuracy of 0.005% in the normalisation. Figure 3 demonstrates that the numerical accuracy we use in the analysis,  $N_x = 16$ ,  $N_v = 24$  and  $N_{\text{sigma}} = 5$ , does satisfy this requirement.

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

If the data is affected by measurement errors, they have to be incorporated in the likelihood. We assume Gaussian errors in the observable space  $\mathbf{y} \equiv (\hat{\mathbf{x}}, \hat{\mathbf{v}}) = (\text{RA}, \text{DEC}, (m - M), \mu_{\text{RA}}, \mu_{\text{DEC}}, v_{\text{los}})$ , i.e. the  $i$ -th star's observed  $\mathbf{y}_i \sim N[\mathbf{y}'_i, \delta \mathbf{y}_i](\mathbf{y}) = N[\mathbf{y}, \delta \mathbf{y}_i](\mathbf{y}'_i)$  [TO DO: Talk to HW about best notation.], with  $\mathbf{y}'_i$  being the true position and velocity of the star. Stars follow

the (quasi-isothermal) distribution function ( $\text{DF}(\mathbf{y}') \equiv \text{qDF}(\mathbf{J}[\mathbf{y}' | p_\Phi] | p_{\text{DF}})$  for short), convolved with the error distribution  $N[0, \delta \mathbf{y}](\mathbf{y}')$  [TO DO: CHECK AGAIN]. The selection function  $\text{sf}(\mathbf{y})$  acts on the space of (error affected) observables. Then the probability of one star becomes

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

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

$$\begin{aligned} & \tilde{p}(\mathbf{y}_i | p_\Phi, p_{\text{DF}}, \delta \mathbf{y}_i) \\ & \approx \frac{\text{sf}(\mathbf{x}_i)}{\int d^6 \mathbf{y}' \, \text{DF}(\mathbf{y}') \cdot \text{sf}(\mathbf{x}')} \cdot \frac{1}{N_{\text{error}}} \sum_n^{N_{\text{error}}} \text{DF}(\mathbf{x}_i, \mathbf{v}[\mathbf{y}'_{i,n}]) \end{aligned} \quad (14)$$

with

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

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

Figure 4 demonstrates that in the absence of distance errors the  $N_{\text{error}}$  needed for the convolution integral to converge depends as

$$N_{\text{error}} \propto \delta v^2$$

on the uncertainties in the (1D) velocities.

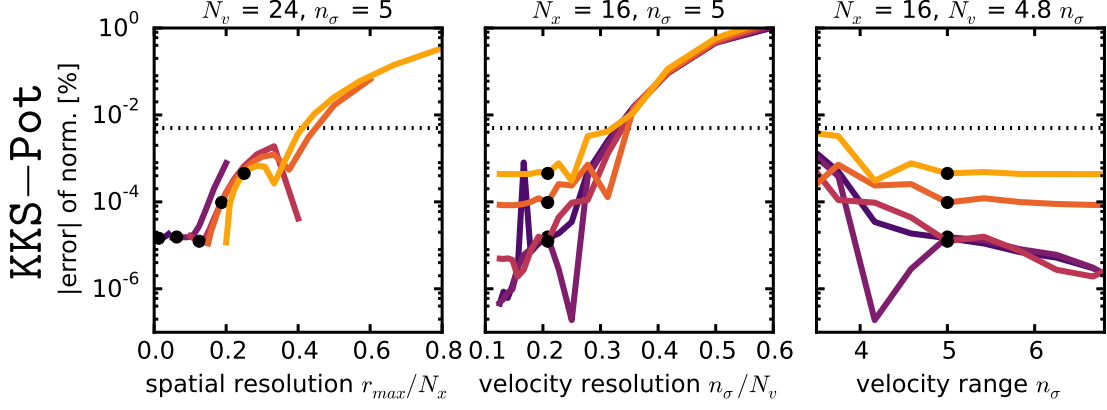
## 2.7. Fitting Procedure

To search the  $(p_\Phi, p_{\text{DF}})$  parameter space for the maximum of the likelihood in Equation 10, we go beyond the fixed grid search by Bovy & Rix (2013) and employ an effective two-step procedure: The first step finds the approximate peak and width of the likelihood using a nested-grid search, while the second step samples the shape of the likelihood using a Monte-Carlo Markov Chain (MCMC) approach.

*Fitting Step 1: Nested-grid search.*— The  $(p_\Phi, p_{\text{DF}})$  parameter space can be high-dimensional. To effectively minimizing the number of likelihood evaluations before finding its peak, we use a nested-grid approach:

- *Initialization.* For  $N$  free model parameters  $M = (p_\Phi, p_{\text{DF}})$ , we set up a sufficiently large initial grid with  $3^N$  regular grid points.
- *Evaluation.* We evaluate the likelihood at each grid-point similar to Bovy & Rix (2013) (their Fig-





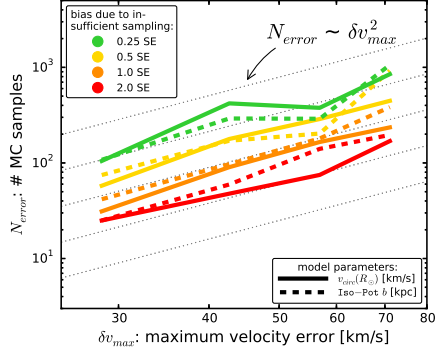
**Figure 3.** Relative error  $\delta M_{\text{tot}}$  of the likelihood normalization  $M_{\text{tot}}$  in Equation 12 depending on the accuracy of the grid-based density calculation in Equation 8 (and surrounding text). We show how  $\delta M_{\text{tot}}$  varies with the spatial resolution (first column), velocity resolution (second column) and velocity integration range (third column) for two different potentials (KKS-Pot in the first row and MW13-Pot in the second row) and five different spherical observation volumes with radius  $r_{\text{max}}$  (color coded according to the legend). (Test 2 in Table 3 summarizes all model parameters.)  $N_x$  is the number of spatial grid points in  $R \in R_{\odot} \text{ kpc} \pm r_{\text{max}}$  and  $|z| \in [0, r_{\text{max}}]$  on which the density is evaluated according to Equation 8. The spatial resolution in  $z$  is therefore  $r_{\text{max}}/N_x$  and  $2r_{\text{max}}/N_x$  in  $R$ . This choice is reasonable because the density is symmetric in  $z$  and varies less in  $R$  than in  $z$ , because the tracer scale length of the disk is much larger than its scale height. At each  $(R, z)$  of the grid a Gauss-Legendre integration of order  $N_v$  is performed over an integration range of  $\pm n_{\sigma}$  times the velocity dispersion in  $v_R$  and  $v_z$  and  $[0, 1.5v_{\text{circ}}(R_{\odot})]$  in  $v_T$ .  $n_{\sigma}/N_v$  is therefore a proxy for the velocity resolution of the grid. (We vary  $N_x$ ,  $N_v$  and  $n_{\sigma}$  separately and keep the other two fixed at the values indicated above the columns.) To arrive at the approximation  $M_{\text{tot,approx}}$  for  $M_{\text{tot}}$  in Equation 11, we perform a 40th-order Gauss-Legendre integration in each  $R$  and  $z$  direction of the interpolated density over the observed volume. We calculate the “true” normalization with high accuracy as  $M_{\text{tot}} \approx M_{\text{tot,approx}}(N_x = 20, N_v = 56, N_{\sigma} = 7)$ . The black dots indicate the accuracy used in our analyses: It is better than 0.002%. Only for the smallest volume in the MW13-Pot (yellow line) the error is only  $\sim 0.005\%$ . This could be due to the fact, that, while we have analytical formulas to calculate the actions for the Staechel potential KKS-Pot exactly, we have to resort to an approximate action calculation for the MW-like potential MW13-Pot (see Section 2.2). [TO DO: Write  $|\delta M_{\text{tot}}|$  on y-axis] [TO DO: Remove MW13-Pot completely from this plot, caption and test table] [TO DO: Caption too long] [TO DO: Rewrite caption, text and table, I changed the plot]

ure 9): Because of the many computationally expensive  $\mathbf{x}, \mathbf{v} \xrightarrow{p_{\Phi}} \mathbf{J}$  transformations that have to be performed for each new set of  $p_{\Phi}$  parameters, an outer loop iterates over the  $p_{\Phi}$  parameters and pre-calculates the actions, while an inner loop evaluates the likelihood Equation 10 for all qDF parameters  $p_{\text{DF}}$  with the actions in the given potential.

- *Iteration.* To find from the very sparse  $3^N$  likelihood grid a new grid, that is more centered on the likelihood and has a width of order of the width of the likelihood, we proceed as follows: For each of the model parameter in  $M$  we marginalize the likelihood by summing over the grid. If the resulting 3 points all lie within  $4\sigma$  of a Gaussian, we fit a Gaussian to the 3 points and determine a new  $4\sigma$  fitting range. Otherwise the boundaries of the grid point with the highest likelihood becomes the new fitting range. We proceed with iteratively evaluating the likelihood on finer and finer grids, until we have found a 4-sigma fit range in each of the model parameter dimensions.
- *The fiducial qDF.* For the above strategy to work properly, the action pre-calculations have to be independent of the choice of qDF parameters. This is clearly the case for the  $N_j \times N_{\text{error}}$  stellar data actions  $\mathbf{J}_i$ . To calculate the normalisation in Equation 10,  $N_x^2 \times N_v^3$  actions  $\mathbf{J}_n$  are needed. Formally the spatial coordinates at which the  $\mathbf{J}_n$  are calculated depend on the  $p_{\text{DF}}$  parameters via the integration ranges in Equation 8. To relax this dependence we instead use the same velocity integration limits in the likelihood calculations for all

$p_{\text{DF}}$ s in a given potential. This set of parameters, that sets the velocity integration range globally,  $(\sigma_{R,0}, \sigma_{z,0}, h_{\sigma,R}, h_{\sigma,z})$  in Equation 5 and 6, is referred to as the *fiducial* qDF. Using the same integration range in the density calculation for all qDFs at a given  $p_{\Phi}$  makes the normalisation vary smoothly with different  $p_{\text{DF}}$ . Choosing a fiducial qDF that is very different from the true qDF can however lead to large biases. The optimal values for the fiducial qDF are the (yet unknown) best fit  $p_{\text{DF}}$  parameters. We take care of this by setting, in each iteration step of the nested-grid search, the fiducial qDF simply to the  $p_{\text{DF}}$  parameters of the central grid point. As the nested-grid search approaches the best fit values, the fiducial qDF approaches automatically the optimal values as well. This is another advantage of the nested-grid search, because the result will not be biased by a poor choice of the fiducial qDF.

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

(a)  $N_* = 10,000$ (b)  $N_* = 5,000$ 

**Figure 4.** Number of Monte Carlo (MC) samples  $N_{\text{error}}$  needed for the numerical convolution of the model probability with the measurement uncertainties in Equation 14, given the maximum velocity error  $\delta v_{\text{max}}$  within the stellar sample. Unsufficient sampling introduces systematic biases in the parameter recovery as indicated in the legend. The relation found here,  $N_{\text{error}} \propto \delta v_{\text{max}}^2$ , was distilled from a set of analyses of mock data sets with different proper motion uncertainties  $\delta\mu \in [2, 5]$  mas yr $^{-1}$  in the absence of distance errors (see Test 6.1 in Table 3). The proper motion error  $\delta\mu$  translates to heteroscedastic [TO DO: make sure that this word is written correctly everywhere.] velocity errors 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  $N_{\text{error}}$  for the integral over its measurement uncertainties to converge. We therefore show how the  $N_{\text{error}}$  needed for the potential *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  $N_{\text{error}} = 800$  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 11). [TO DO: some of the 25 MC sample analyses have to be re-done.] [TO DO: Replace right plot with new plot with  $N_* = 5,000$ ] [TO DO: Use  $N_*$  everywhere where applicable] [TO DO: Introduce  $N_*$  somewhere.] [TO DO: Iso-Pot in legend] [TO DO: Comment from Jo: I think it is important to test, if the MC vs error plot depends on number of stars. Maybe test it with less stars (5000), to test this quickly. Naively, I would expect a large dependence on  $N_{\text{data}}$ .]

*Fitting Step 2: MCMC.*— After the nested-grid search is converged, the grid is centered at the peak of the likelihood and its extent contains the  $4\sigma$  confidence interval. To actually sample the full shape of the likelihood, we could do a grid search with much finer grid spacing (e.g.  $K = 11$  in each dimension). The number of grid points scales as a power of the free parameters  $N$ . For a large number of free parameters ( $N > 4$ ) a Monte Carlo Markov Chain (MCMC) approach might sample the *pdf* (which is here equivalent to the likelihood, see §2.6) much faster. We use *emcee* by Foreman-Mackey et al. (2013) and release the walkers very close to the *pdf* peak found by the nested-grid search, which will assure fast conver-

gence in much less than  $K^N$  likelihood evaluations. For a sufficiently high numerical accuracy in calculating the integrals in Equation 8 the current qDF at each walker position can be used as the fiducial qDF. To get reasonable results also for slightly lower accuracy, a single fiducial qDF can be used for all likelihood evaluations within the MCMC as well. As fiducial qDF we use the qDF parameters of the likelihood peak, found by the nested-grid search.

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

### 3. RESULTS

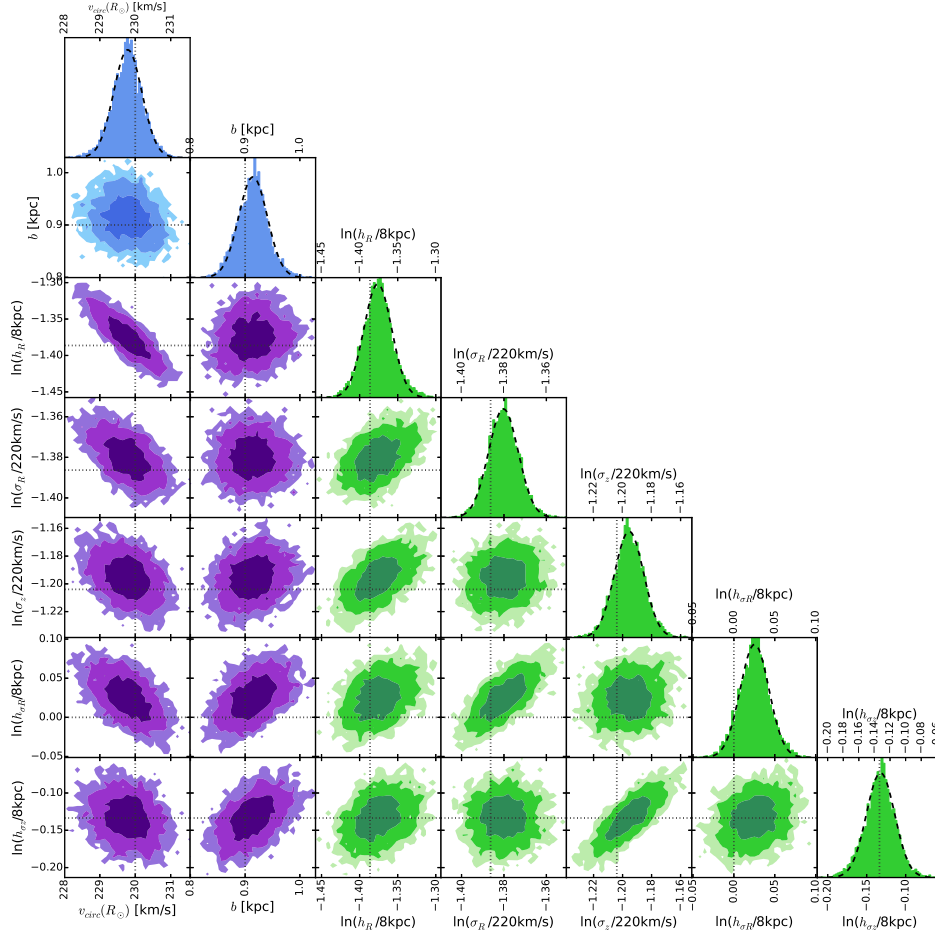
We are now in a position to explore the limitations of action based modelling posed in the introduction: (i) unbiased estimates; (ii) survey volume; (iii) imperfect selection function; (iv) measurement errors; (v) actual DF or (vi) Potential not spanned by the space of models. We do not explore the breakdown of the assumption that the system is axisymmetric and in steady state. With the exception of the test suite on measurement errors in §3.4, we assume that the phase-space errors are negligible. All tests are also summarized in Table 3.

[TO DO: Hans-Walter said that there are diagnostic plots in this papers that can be eliminated and their essence summarized in 1-2 sentences in the text. Fine. But which plots does he think can be eliminated? My plots contain either results or are only there to make the paper more readable for others.]

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

The individual MAPs in Bovy & Rix (2013) contained typically between 100 and 800 objects, so that each MAP implied a quite broad *pdf* for the model parameters  $p_M = \{p_\Phi, p_{\text{DF}}\}$ . Here we explore what happens in the limit of much larger samples for each MAP, say 20,000 objects. As outlined in §2.6 the immediate consequence of larger samples is given by the likelihood normalization requirement,  $\log(1 + \delta M_{\text{tot}}) \leq 1/N_{\text{sample}}$ , (see Equation 13)), which is the modelling aspect that drives the computing time. This issues aside, we would, however, expect that in the limit of large data sets with vanishing measurement errors the *pdf*s of the  $p_M$  become Gaussian, with a *pdf* width (i.e. standard error SE of the Gaussian) that scales as  $1/\sqrt{N_{\text{sample}}}$ . Further, we must verify that any bias in the *pdf* expectation value is considerably less than the error (SE), even for quite large samples.

Using sets of mock data, created according to §2.5 and the fiducial model for  $p_M$  (see Table 3, Tests 3.2, 3.3, and 3.1), we verified that *RoadMapping* satisfies all these conditions and expectations. Figure 5 illustrates the joint *pdf*s of all  $p_M$ . This figure illustrates that the *pdf* is a multivariate Gaussian that projects into Gaussians when considering the marginalized *pdf* for all the individual  $p_M$ . Note that some of the parameters are quite covariant, but the level of their actual covariance depends on the choice of the  $p_M$  from which the mock data were drawn. Figure 6 then illustrates that the *pdf* width, SE, indeed scales as  $1/\sqrt{N_{\text{sample}}}$ . Figure 7 illustrates even more that *RoadMapping* satisfies the central limit theorem. The average parameter estimates from many mock samples with identical underlying  $p_M$  are very close to the input  $p_M$ , and the distribution of the actual parameter estimates are a Gaussian around it.



**Figure 5.** The  $pdf$  (proportional to the likelihood in Equation [TO DO]) in the parameter space  $p_M = \{p_\Phi, p_{DF}\}$  for one example mock data set created according to Test 3.1 in Table 3. Blue indicates the  $pdf$  for the potential parameters, green the qDF parameters. The true parameters are marked by dotted lines. The dark, medium and bright contours in the 2D distributions represent 1, 2 and 3 sigma confidence regions [TO DO: HW: "likelihood vs. pdf - This is where this matters: is this a confidence on the data or on the parameters?" Don't understand, what he means...], respectively, and show weak or moderate covariances. This analysis was picked among five similar analyses, to have all 1 sigma contours encompass the input values [TO DO: Jo didn't understand this sentence]. The  $pdf$  here was sampled using MCMC (with flat priors in  $p_\Phi$  and  $\ln(p_{DF})$ ) to turn the likelihood in Equation 10 into a full  $pdf$ . Because only 10,000 MCMC samples were used to create the histograms shown, the 2D distribution has noisy contours. The dashed lines in the 1D distributions are Gaussian fits to the histogram of MCMC samples. This demonstrates very well that for such a large number of stars, the  $pdf$  approaches the shape of a multi-variate Gaussian, as expected from the central limit theorem [TO DO: Jo wrote, that he is not sure if the central limit theorem is directly relevant here]. [TO DO: rename  $h_{\sigma R}$  to  $h_{\sigma, R}$ ,  $\sigma_R$  to  $\sigma_{R,0}$  and analogous for  $z$ ]

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

### 3.2. The Role of the Survey Volume Geometry

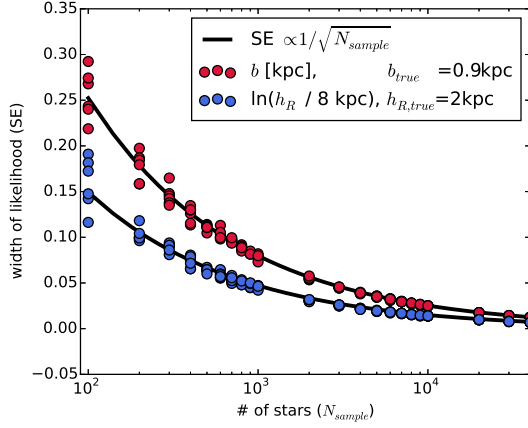
To explore the role of the survey volume (see Section 1) at given sample size, we devise two suites of mock data sets:

The first suite draws mock data from the same  $p_M$ , *two different potentials* (Iso-Pot and MW13-Pot, see Test 4 in Table 3), and volume wedges (see Section 2.4) at *different positions within the Galaxy*, illustrated in the right upper panel of Figure 8. To isolate the role of the survey volume geometry, the mock data sets are equally large (20,000) in all cases, and are drawn from identical total survey volumes ( $4.5 \text{ kpc}^3$ , achieved by adjusting the angular width of the wedges). The results are shown in Figure 8.

The second suite of mock data sets was already intro-

duced in Section 3.1 (see also Test 3.3), 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 7 and demonstrate the effect of the *size of the survey volume*.

Figures 7 and 8 illustrate the ability of *RoadMapping* to constrain model parameters, with the standard error of the  $pdf$  as measure of the precision on the  $x$ -axis. Figure 7 demonstrates that, given a choice of qDF, a larger volume always results in tighter constraints. There is no obvious trend that a hotter or cooler MAP will always give better results [TO DO: Comment from HW: The question of whether a hotter or a colder population gives tighter constraints is an important question, but it seems buried here in a section that is dedicated to another matter, namely the question of volume ... It's OK to leave it here, but somewhere we need to say clearly: whether the population is hot or cold does not make a big and generic difference...]; it depends on the survey volume and the



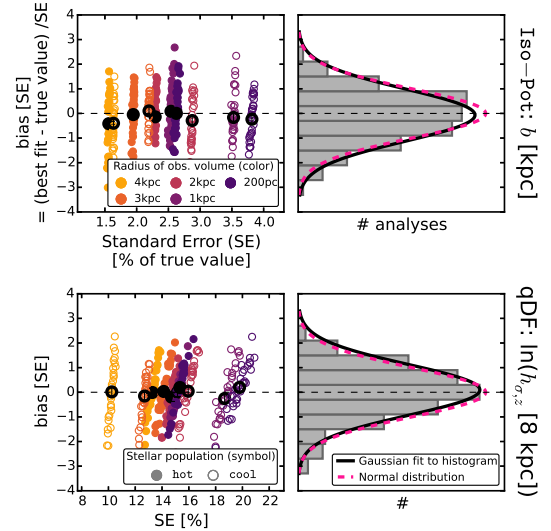
**Figure 6.** The width of the *pdf* for two fit parameters found from analyses of 132 mock data sets vs. the number of stars in each data set. The mock data was created in the Iso-Pot potential and all model parameters are given as Test 3.2 in Table 3. The *pdf* (using the likelihood in Equation 10 [TO DO: CHECK]) was evaluated and then a Gaussian was fitted to the marginalized *pdf* of each free fit parameter. The standard error (SE) of these best fit Gaussians is shown for the potential parameter  $b$  in kpc (red dots) and for the qDF parameter  $\ln(h_R/8\text{kpc})$  in dimensionless units (blue). The black lines are fits of the functional form  $\text{SE}(N_{\text{sample}}) \propto 1/\sqrt{N_{\text{sample}}}$  to the data points of both shown parameters. As can be seen, for large data samples the width of the *pdf* behaves as expected and scales with  $1/\sqrt{N_{\text{sample}}}$  as predicted by the central limit theorem. [TO DO: fancybox Legend] [TO DO: write pdf instead of likelihood on y-axis] [TO DO: axis labels similar to MC\_vs\_error plot.] [TO DO: Different colors (blue and green), and different markers (no black rings).]

model parameter in question. In Figure 8 the wedges all have the same volume and all give results of similar precision. Minor differences, e.g. with the Iso-Pot potential being less constrained in the wedge with large vertical, but small radial extent, are a special property of the considered potential and parameters, and not a global property of the corresponding survey volume. In the case of an axisymmetric model galaxy, the extent in  $\phi$  direction is not expected to matter. Overall radial extent and vertical extent seem therefore to be equally important to constrain the potential. In addition Figure 8 implies that for these cases volumes offsets in the radial or vertical direction have at most a modest impact - even in case of the very large sample size at hand.

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

### 3.3. Impact of Misjudging the Completeness of the Data Set

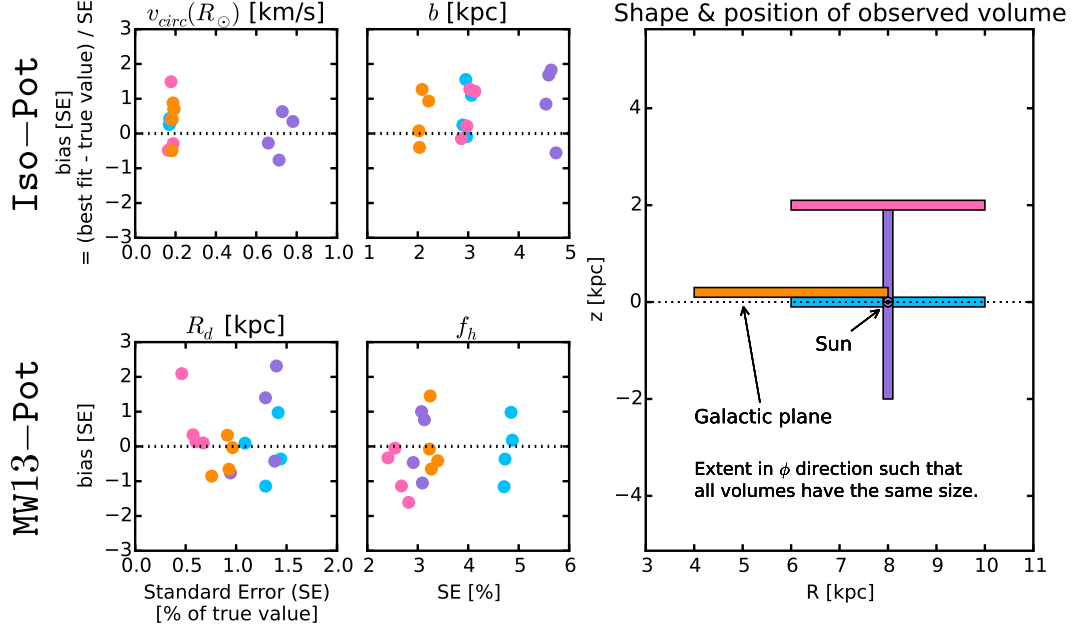
The completeness function (see Section 2.4) depends on the characteristics and mode of the survey. It can be very complex and is therefore sometimes not perfectly known. We investigate how much the recovery of the potential can be affected by imperfect knowledge of the selection function. We do this by creating mock data with varying incompleteness (within a maximal survey volume), while assuming constant completeness in the analysis. The mock data comes from a sphere around the sun with an incompleteness function that drops linearly with distance  $r$  from the sun (see Test 5, Example 1, in Table 3 and Figure 9). This captures the relevant case of stars



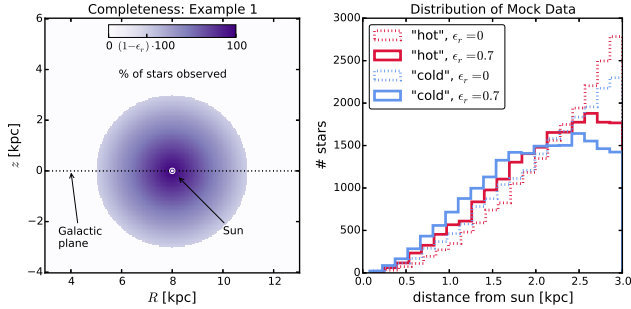
**Figure 7.** (Un-)bias of the parameter estimates: According to the central limit theorem the best fit values for a large number of data sets, each containing a large number of stars, will follow the Normal distribution. To test this, we create 320 mock data sets, which come from two different stellar populations and five spherical observation volumes (see legends). All model parameters are summarized in Table 3 as Test 3.3. Bias and relative standard error (SE) are derived from the marginalized *pdf* for one potential parameter (isochrone scale length  $b$  in first row) and one qDF parameter ( $h_{\sigma,z}$  in second row). The second column displays a histogram of the 320 offsets. As it closely follows a Normal distribution, our modelling method is therefore well-behaved and unbiased. For the 32 analyses belonging to one model we also determine the mean offset and SE, which are overplotted in black in the first two columns (with  $1/\sqrt{32}$  as error). [TO DO: Is the scatter of the black symbols too large??? Is the reason for this numerical inaccuracies???] [TO DO: Change test table accordingly, isochrone with  $b = 1.5$  is not used anymore] [TO DO: Caption is too long. Make shorter.] [TO DO:  $r_{\text{max}}$  instead of radius in legend] [TO DO: Leerzeichen fehlt in y-achsenbeschriftung]

being less likely to be observed (than assumed) the further away they are (e.g. due to unknown dust obscuration). We demonstrate that the potential recovery with *RoadMapping* is very robust against somewhat wrong assumptions about the radial completeness of the data (see Figure 10). Apparently, much information about the potential comes from the rotation curve measurements in the plane, which is not affected by the incompleteness of the sample. In Appendix .1 we also show that the robustness is somewhat less striking but still persists for small misjudgments of the incompleteness in vertical direction, parallel to the disk plane (Figures 20 and 21). This could model the effect of wrong corrections for interstellar extinction in the plane. We also investigate in Appendix .1 if indeed most of the information is stored in the rotation curve [TO DO: Comment by HW: I don't have an immediate solution for this, but again, it seems the interesting question of "how much of the information is in the rotation curve" is 'hidden' in the section on selection functions...]. For this we use the same mock data sets as analysed in Figures 10 and 21, but without including the tangential velocities in the modelling (by marginalizing the likelihood over  $v_T$ ). In this case the potential is much less tightly constrained, even for 20,000 stars. For only small deviations of true and assumed completeness ( $\lesssim 10\%$ ) we can however still incorporate the true potential in our fitting result (see Figure 22).





**Figure 8.** Bias vs. standard error in recovering the potential parameters for mock data stars drawn from four different test observation volumes within the Galaxy (illustrated in the upper right panel) and two different potentials (Iso-Pot and MW13-Pot from Table 1). Standard error and offset were determined as in Figure 7. Per volume and potential we analyse four different mock data realisations; all model parameters are given as Test 4 in Table 3. The colour-coding represents the different wedge-shaped observation volumes. The angular extent of each wedge-shaped observation volume was adapted such that all have the volume of  $4.5 \text{ kpc}^3$ , even though their extent in  $(R, z)$  is different. Overall there is no clear trend, that an observation volume around the sun, above the disk or at smaller Galactocentric radii should give remarkably better constraints on the potential than the other volumes. [TO DO: Write in Plot "... that all wedges have the same volume".]



**Figure 9.** Selection function and mock data distribution for investigating radial incompleteness of the data. All model parameters are summarized as Test 5, Example 1, in Table 3. The survey volume is a sphere around the sun and the percentage of observed stars is decreasing linearly with radius from the sun, as demonstrated in the left panel. How fast this detection/incompleteness rate drops is quantified by the factor  $\epsilon_r$ . Histograms for four data sets, drawn from two MAPs (hot in red and cool in blue, see Table 2) and with two different  $\epsilon_r$ , 0 and 0.7, are shown in the right panel for illustration purposes. [TO DO: Potential and/or population names in typewriter font]

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

### 3.4. Measurement Errors and their Effect on the Parameter Recovery

[TO DO: Comment from HW: This Section has three parts:

- convergence of the integral (ALREADY REMOVED)
- testing the approximation
- underestimating errors

It seems to me that the basic Section: What is the im-

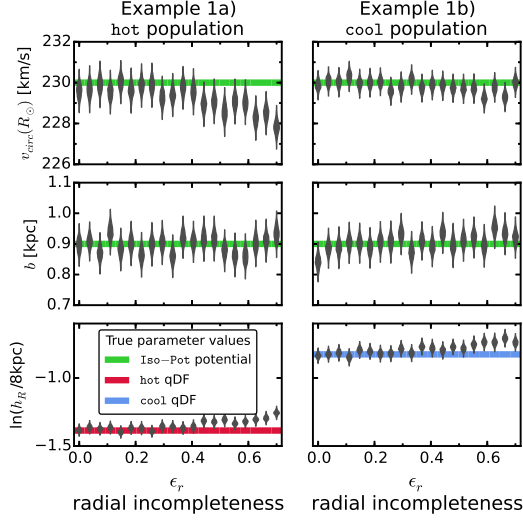
part of the errors? Is missing. That should be the center piece, and the other three aspects should be quick summary notes, only 1-2 sentences long.] [I'll try to address this with a plot mean(SE) vs. proper motion error - also for cold population (currently running on wolf).]

In absence of distance uncertainties the error convolved likelihood given in Equation 14 is unbiased. When including distance (modulus) errors, Equation 14 is just an approximation for the true likelihood. The systematic bias thus introduced in the parameter recovery gets larger with the size of the error. This is demonstrated in Figure 6.2. We find however that in case of  $\delta(m-M) \lesssim 0.3 \text{ mag}$  (if also  $\delta\mu \leq 2 \text{ mas yr}^{-1}$  and a maximum distance of  $r_{\text{max}} = 3 \text{ kpc}$ , see Test 6.2 in Table 3) the potential parameters can still be recovered within 2 sigma [TO DO: Make sure this is what I claim in abstract and discussion.]. This corresponds to a relative distance error of  $\sim 10\%$ .

[TO DO: Introduce a test and plot that demonstrates how the SE depends on proper motion error. Then write this little section.] Overall the standard errors on the recovered parameters are quite small (a few percent at most for 10,000 stars), which demonstrates that, if we perfectly knew the measurement errors, we still could get very precise constraints on the potential. The constraints also get tighter the smaller the proper motion error becomes. We found that for  $\delta\mu = 1 \text{ mas yr}^{-1}$  the precision of the recovered parameters reduce by  $\sim$  half compared to  $\delta\mu = 5 \text{ mas yr}^{-1}$ .

We found that in case we perfectly knew the measurement errors (and the distance error is negligible), the convolution of the model probability with the measure-





**Figure 10.** Influence of wrong assumptions about the radial incompleteness of the data on the parameter recovery with *RoadMapping*. Each mock data set was created with different incompleteness parameters  $\epsilon_r$  (shown on the  $x$ -axis and illustrated in Figure 9) and the model parameters are given as Test 5, Example 1, in Table 3. The analysis however did not know about the incompleteness and assumed that all data sets had constant completeness within the survey volume ( $\epsilon_r = 0$ ). The marginalized likelihoods from the fits are shown as violins. The green lines mark the true potential parameters (Iso-Pot) and the red and blue lines the true qDF parameters (hot MAPin red and cool MAPin blue), which we tried to recover. The *RoadMapping* method seems to be very robust against small to intermediate deviations between the true and the assumed data incompleteness. [TO DO: Jo suggested to also remove the  $h_R$  panel, but I like, that one can see that it is the spatial tracer distribution that drives the little degradation of the recovery.]

ment uncertainties gives precise and accurate constraints on the model parameters - even if the error itself is quite large.

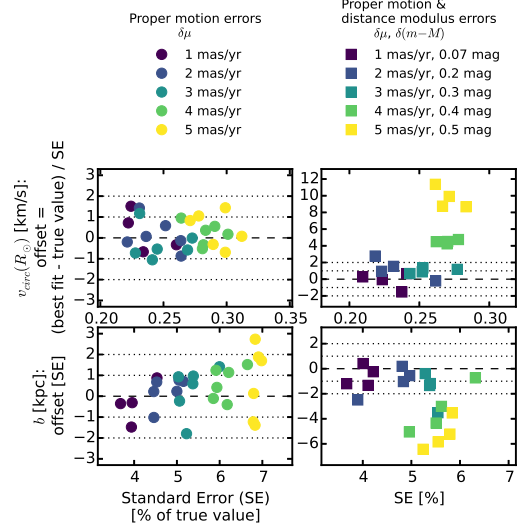
Figure 13 now investigates the effect of a systematic *underestimation* of the true proper motion uncertainties  $\delta\mu$  by 10% and 50%. We find that this causes a bias in the parameter recovery that grows seemingly linear with  $\delta\mu$ . For an underestimation of only 10% however, the bias is still  $\lesssim 2$  sigma for 10,000 stars [TO DO: Check] - even for  $\delta\mu \sim 3 \text{ mas yr}^{-1}$ .

The size of the bias also depends on the kinematic temperature of the stellar population and the model parameter considered (see Figure 13). 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.

[TO DO: Comment from Jo: Always use 'uncertainty' when describing how ou deal with the errors. 'Error' means the actual error (difference between observed and true).]

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

Our modelling approach assumes that each MAP follows a quasi-isothermal distribution function, qDF. In this Section we explore what happens if this idealization does not hold. We investigate this issue by creating mock data sets (Figure 14) that are drawn from two distinct

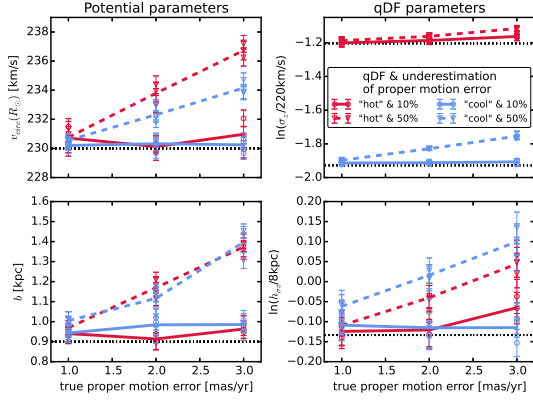


**Figure 11.** Potential parameter recovery using the approximation for the model probability convolved with measurement uncertainties in Equation 14. We show *pdf* offset and relative width (i.e., standard error SE) for the potential parameters derived from mock data sets, which were created according to Test 6.2 in Table 3). The data sets in the left panels have only uncertainties in line-of-sight velocity and proper motions, while the data sets in the right panels also have distance (modulus) uncertainties, as indicated in the legends in the first row. For data sets with proper motion error  $\delta(m - M) \leq 3 \text{ mas yr}^{-1}$  Equation 14 was evaluated with  $N_{\text{error}} = 800$ , for  $\delta(m - M) > 3 \text{ mas yr}^{-1}$  we used  $N_{\text{error}} = 1200$ . In absence of distance uncertainties Equation 14 gives unbiased results. For  $\delta(m - M) \geq 3 \text{ mas yr}^{-1}$  (which corresponds in this test to  $\delta v_{\text{max}} \lesssim 43 \text{ km s}^{-1}$ , see Equation ??) however biases of several sigma are introduced as Equation 14 is only an approximation for the true likelihood in this case.



**Figure 12.** [TO DO: This should be a figure that plots precision (SE) vs. proper motion error for a hot and a cool population (for no distance error). This is to demonstrate the effect of measurement errors in general. Currently running on cluster....]

qDFs of different temperature, and analyze the composite mock data set by fitting a single qDF to it. These results are illustrated in Figures 15 and 16. Following the observational evidence, MAPs with cooler qDFs also have longer tracer scale lengths. In the first set of test, we choose qDFs of widely different temperatures and vary their relative fraction (dubbed *Examples 1a/b* in Figure 15 and Test 7 in Table 3); in the second set of tests (*Examples 2a/b* in Figure 16 and Test 7 in Table 3), we always mix mock data points from two different qDFs in equal proportion, but vary by how much the qDF's



**Figure 13.** Effect of a systematic underestimation of proper motion errors in the recovery of the model parameters. The true model parameters used to create the mock data are summarized as Test 6.3 in Table 3, four of them are given on the  $y$ -axes and the true values are indicated as black dashed lines. The velocities of the mock data were perturbed according to Gaussian errors in the RA and DEC proper motions as indicated on the  $x$ -axis. The circles and triangles are the best fit parameters of several mock data sets assuming the proper motion uncertainty, with which the model probability was convolved, was underestimated in the analysis by 10% or 50%, respectively. The error bars correspond to 1 sigma confidence. The lines connect the mean of each two data realisations and are just to guide the eye. [TO DO: rename  $h_{\sigma_z}$  to  $h_{\sigma,z}$ ,  $\sigma_z$  to  $\sigma_{z,0}$ ] [TO DO: Potential and/or population names in typewriter font] [TO DO: Iso-Pot in Title] [TO DO: Delta mu on x-axis]

temperatures differ.

The first set of tests mimics a DF that has wider wings or a sharper core in velocity space than a qDF (Figure 14). The second test could be understood as mixing neighbouring MAPs due to large bin sizes or abundance measurement errors.

It is worth considering 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 better and more robustly recovered, if a mock-data MAP is polluted by a modest fraction ( $\lesssim 30\%$ ) of stars drawn from a much cooler qDF with a longer scale length, as opposed to the same pollution of stars drawn from a hotter qDF with a shorter scale length.

When considering the case of a 50/50 mix of contributions from different qDFs in Example 2, there is a systematic, but only small, error 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 samples sizes of 20,000, as used in the mock data.

Overall, mock data drawn from a cooler DF always seem to give tighter constraints on the circular velocity at the Sun [TO DO: Make sure that Sun is written everywhere with a capital S.], because the rotation curve can be constrained easier if more stars are on near-circular orbits. But we found the recovered  $v_{\text{circ}}(R_{\odot})$  not always to be unbiased at the implied precision.

The recovery of the effective qDF parameters, in light of non-qDF mock data is quite intuitive: the effective qDF temperature lies between the two temperatures from which the mixed DF of the mock data was drawn; in all cases the scale length of the velocity dispersion fall-off,

$h_{\sigma,R}$  and  $h_{\sigma,z}$ , is shorter, because the stars drawn from the hotter qDF dominate at small radii, while stars from the cooler qDF (with its longer tracer scale length) dominate at large radii. The recovered tracer scale lengths,  $h_R$  vary smoothly between the input values of the two qDFs that entered the mix of mock data, with again the impact of contamination by a hotter qDF (with its shorter scale length in this case) being more important. Overall, we find that the potential inference is quite robust to modest deviations of the data from the assumed DF.

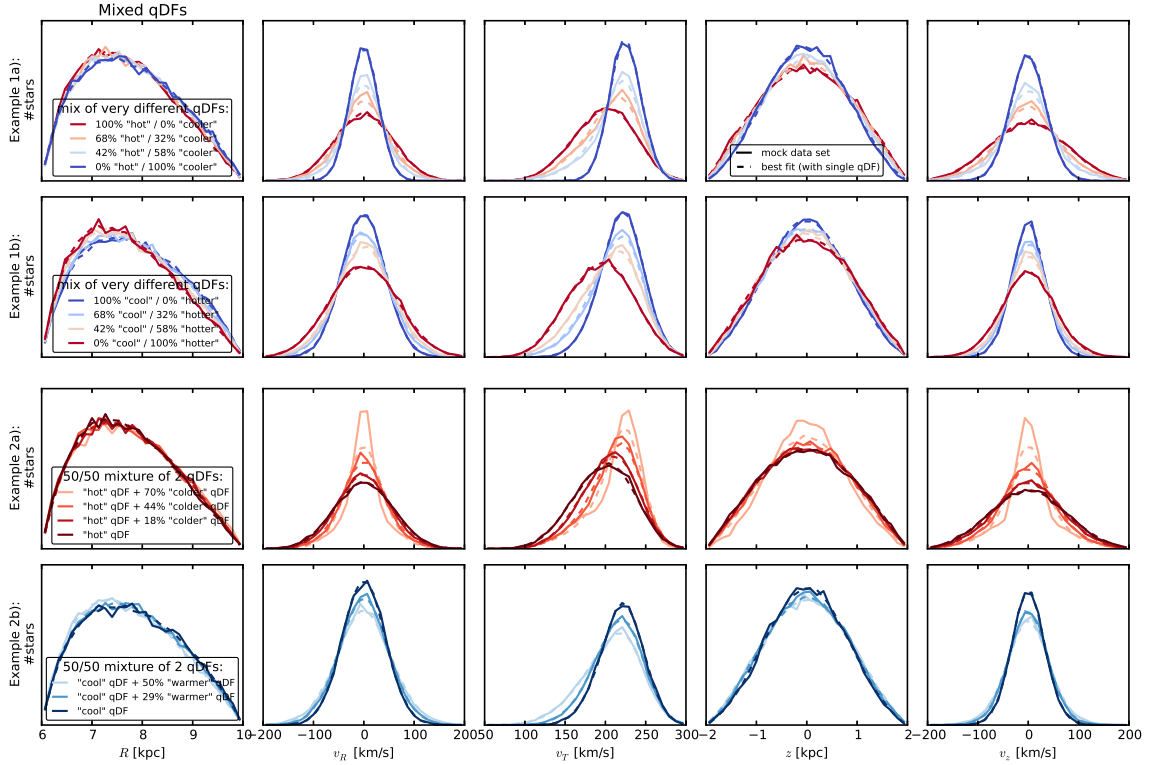
### 3.6. The Implications of a Gravitational Potential not from the Space of Model Potentials

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 only another axisymmetric family, here KKS-Pot (compare the second and fourth panel in Figure 1). In the analysis we assume the circular velocity at the Sun to be fixed and known [TO DO: Comment from Hans-Walter: Do we have reason to believe that this very restrictive assumption does not qualitatively impact our upshot (quantitative differences are OK).] and only fit the parametric potential form. The results are shown in Figure 18.

The reference potential parameters [TO DO: Comment from HW: What does "reference" mean here exactly? Is this an independent exercise, to ask which parameters are expected when fitting potential to potential? (I don't know, what he means...)] of the KKS-Pot in Table 1 were found by adjusting the 2-component Kuzmin-Kutuzov Stäckel potential by Batsleer & Dejonghe (1994) such that it generates radial and vertical force profiles similar to the MW14-Pot from Bovy (2015) (dotted gray lines in Figure 18). We then run *RoadMapping* using these "inconsistent" families of gravitational potentials, and find a good fit to the data in configuration space (see Figure 17). The results from *RoadMapping* analysis for the potential shown in Figure 18, red for a *hot* mock data MAP and blue for a *cool* MAP, give an comparable good or even better agreement with the true potential than the (by-eye) fit directly to the potential: especially the force contours, to which the orbits are sensitive, and the rotation curve are very tightly constrained and reproduce the true potential even outside of the observed volume of the mock tracers. This demonstrates that *RoadMapping* 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 contours are less tightly constrained than the forces, but we still capture the essentials: the *hot* MAP from Table 2 constrains the halo; especially at smaller radii it is equally good or better than the *cool* MAP. The *cool* MAP gives tighter constraints on the halo in the outer region and recovers the disk better than the *hot* MAP. This is in concordance with expectations as the *cool* MAP has a longer tracer scale length and is more confined to the disk than the *hot* MAP and therefore also probes the Galaxy in these regions better. Overall the best fit disk is less dense in the midplane than the true disk.

Figure 19 compares the true qDF parameters with the best fit parameters for this case. While tracer scale length and radial velocity dispersion profile are very well



**Figure 14.** Distribution of mock data, created by mixing stars drawn from two different qDFs (solid lines), and the distribution predicted by the best fit of a single qDF and potential to the data (dashed lines). The model parameters to create the mock data (solid lines) are given in Table 3 as Test 7, and the qDF parameters referenced in the figure’s legend are given in Table 2. The corresponding single qDF best-fit curves (dashed lines) were created by drawing mock data from the best fit parameters found in Figures 15 and 16. *Example 1:* Distribution of mock data drawn from a superposition of two very different (but fixed) qDFs at varying mixing rates. *Example 2:* Mock data distribution of two MAPs that were mixed at a fixed rate of 50%/50%, but the difference of the qDF parameters of one MAP was varied with respect to the qDF parameters of the other MAP by  $X\%$  (see Table 2). The data sets are color coded in the same way as the corresponding analyses in Figures 15 and 16. This figure demonstrates how mixing two qDFs can be used as a test case for changing the shape of the DF to not follow a pure qDF anymore, e.g. by adding wings or slightly changing the radial density profile. When comparing the mock data and best fit distribution, we see that especially for the most extreme deviations it becomes obvious that a single qDF is a bad assumption for the stars’ *true* DF. [TO DO: Potential and/or population names in typewriter font] [TO DO: include  $X$  somehow in figure to explain it better. Jo didn’t understand what I meant by it in this caption.] [TO DO: These are really many panels. Try to remove some.]

recovered, we misjudge the radial profile of the vertical velocity dispersion as  $\sigma_{0,z}$  and  $h_{\sigma,z}$  are both underestimated, which leads to a steeper profile and a lower dispersion around the Sun.

#### 4. DISCUSSION AND SUMMARY

[TO DO: Introduce DF somewhere - use DF wherever we don’t need qDF.]

[TO DO: Compare these sections with the results. Points should be made detailed in the results section and short here in the discussion. Says Hans-Walter.]

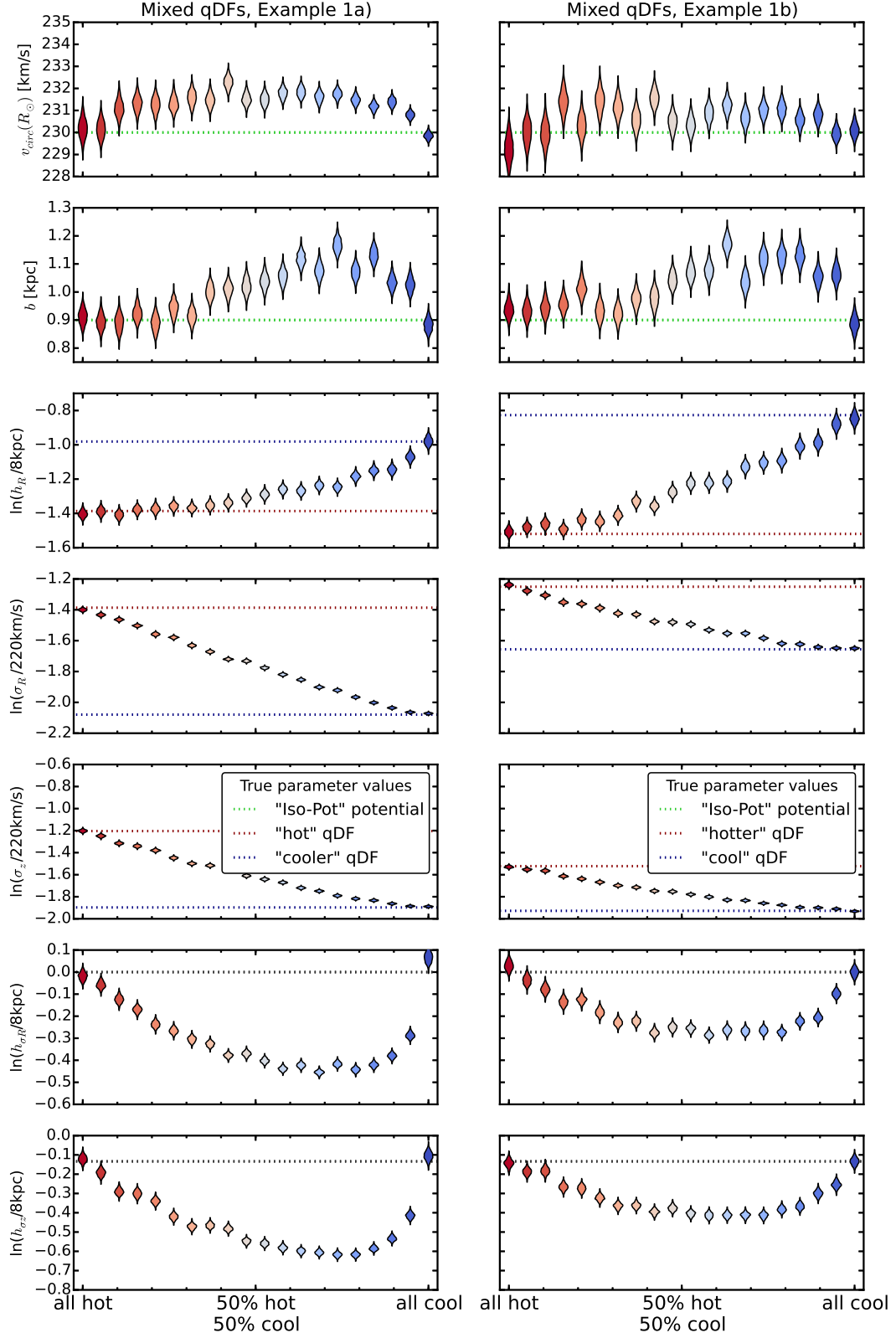
Recently implementations of action DF-based modelling of 6D data in the Galactic disk have been put forth, in part to lay the ground-work fo Gaia (Bovy & Rix 2013; McMillan & Binney 2013; Piffl et al. 2014; Sanders & Binney 2015).

We present *RoadMapping*, an improved implementation of the dynamical modelling machinery of Bovy & Rix (2013), to recover the potential and orbit distribution function of stellar MAPs 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 reliable, even when

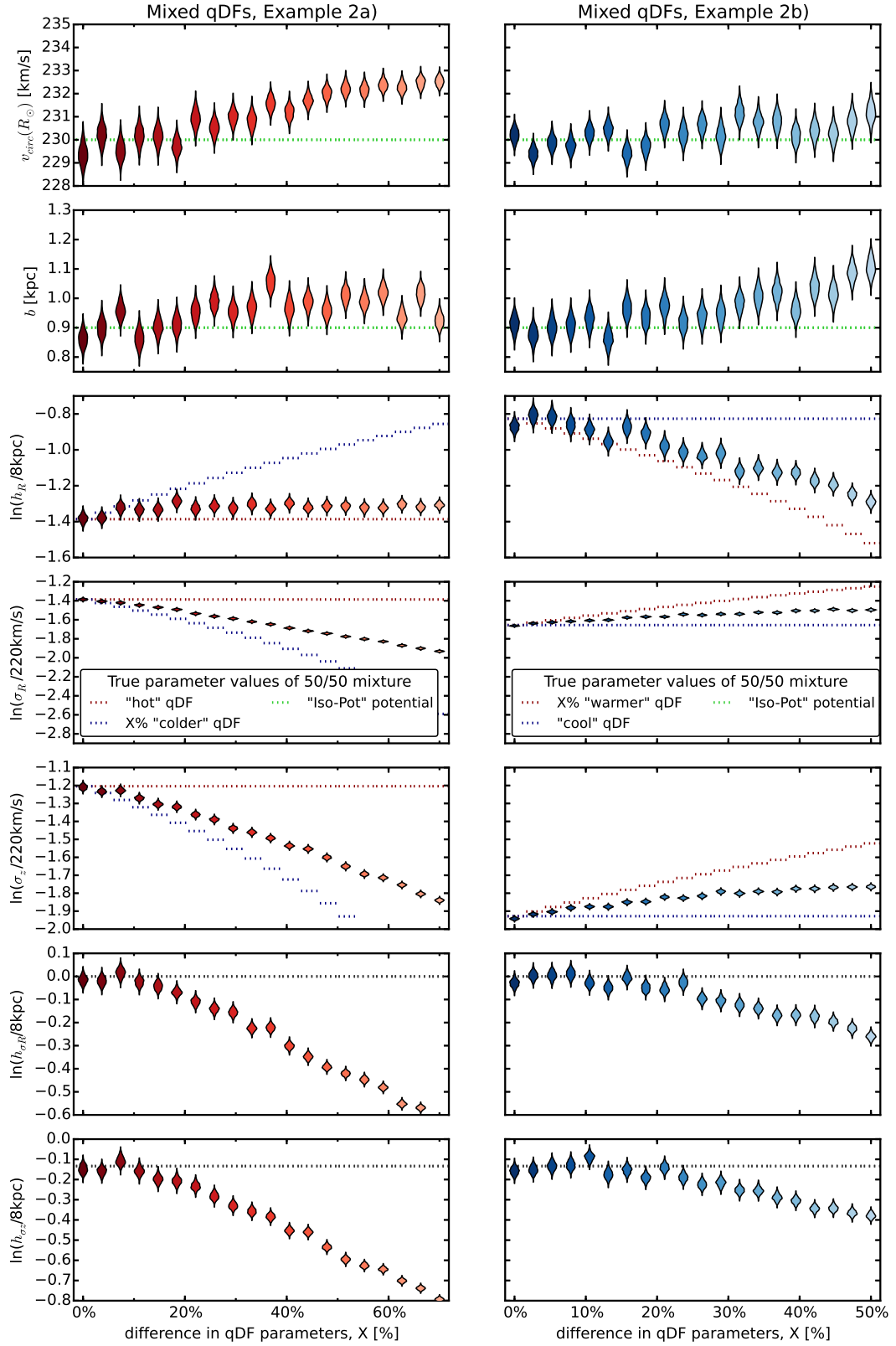
there are small deviations of the model assumptions from the real world Galaxy.

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

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



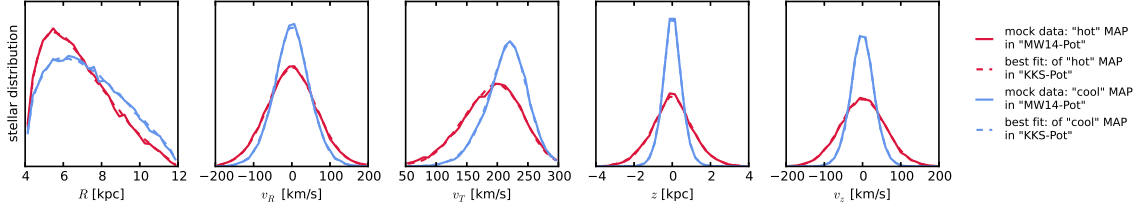
**Figure 15.** The dependence of the parameter recovery on degree of pollution and temperature of the stellar population. To model the pollution of a hot stellar population by stars coming from a cool population and vice versa, we mix varying amounts of stars from two very different populations, as indicated on the X-axis. The composite mock data set is then fit with one single qDF. The violins represent the marginalized likelihoods found from the MCMC analysis. *Example 1a* (*Example 1b*) in the left (right) panels mixes the *hot* (*cool*) MAP with the *cooler* (*hotter*) MAP in Table 2. All model parameters used to create the mock data are given in Test 7, *Example 1a*) & *b*) in Table 3. Some mock data sets are shown in Figure 14, first two rows, in the same colors 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. [TO DO: rename  $h_{\sigma_R}$  to  $h_{\sigma,R}$ ,  $\sigma_R$  to  $\sigma_{R,0}$  and analogous for  $z$ ] [TO DO: Potential and/or population names in typewriter font] [TO DO: Comment from Jo: I feel like just showing one of these examples might be clearer, because they essentially demonstrate the same thing.] [TO DO: Remove  $\sigma_R$  and  $h_{\sigma,R}$  panels. Then make two columns with only one example, potential and DF parameters separately]



**Figure 16.** (Caption on next page.)



**Figure 16.** The dependence of the parameter recovery on the difference in qDF parameters of a 50%/50% mixture of two stellar populations and their temperature. Half of the star in each mock data set in *Example 2a* (*Example 2b*) was drawn from the *hot* (*cool*) qDF in Table 2, and the other half drawn from a *colder* (*warmer*) population that has  $X\%$  smaller (larger)  $\sigma_{R,0}$  and  $\sigma_{z,0}$  and  $X\%$  larger (smaller)  $h_R$ . Each composite mock data set is then fitted by a single qDF and the marginalized MCMC likelihoods for the best fit parameters are shown as violins. The model parameters used for the mock data creation are given as Test 7, *Example 2a*) & *b*) in Table 3. Some mock data sets are shown in figure 14, last two rows, where the distributions have the same colors as the corresponding best fit violins here. By mixing MAPs with varying difference in their qDF parameters, we model the effect of bin size in the  $[\text{Fe}/\text{H}]-[\alpha/\text{Fe}]$  plane when sorting stars into different MAPs: The smaller the bin size, the smaller the difference in qDF parameters of stars in the same bin. We find that the bin sizes should be chosen such that the difference in qDF parameters between neighbouring MAPs is less than 20%. [TO DO: rename  $h_{\sigma R}$  to  $h_{\sigma,R}$ ,  $\sigma_R$  to  $\sigma_{R,0}$  and analogous for  $z$ ] [TO DO: Potential and/or population names in typewriter font]



**Figure 17.** Comparison of the distribution of mock data in configuration space created in the MW14-Pot potential (solid lines) with a *hot* (red) and *cool* (blue) MAP (Test 8 in Table 3), and the best fit distribution using a KKS-Pot potential (dashed lines). The best fit potentials are shown in Figure 18 and the corresponding best fit qDF parameters in Figure 19. The best fit [TO DO: Continue Caption, Jo suggests add something about the fit being good.] [TO DO: Potential and/or population names in typewriter font]

**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. We also found that the location of the survey volume matters little. At given sample size, the volume over which the data are sampled also matters little, if the modelling assumptions are fulfilled. Concerning the *shape* of the survey volume, a large radial and vertical coverage is best, because in the axisymmetric regime the azimuthal coverage does not matter.

Stellar populations of different scale length and temperature probe different regions of the Galaxy (Bovy & Rix 2013). But there is no easy rule of thumb for which survey volume and stellar population which potential and DF parameter is constrained best.

Surprisingly, (cf. Rix & Bovy 2013) *RoadMapping* seems to be very robust against misjudgments in the selection function of the data. We speculate that this is because missing stars in the data set do not affect the connection between a star's velocity and position, which is given by the potential. Much of the information about the potential profile is stored in the rotation curve, but we find that even when we do not include measurements of tangential velocities in the analysis, small misjudgments of the incompleteness do not affect the potential recovery.

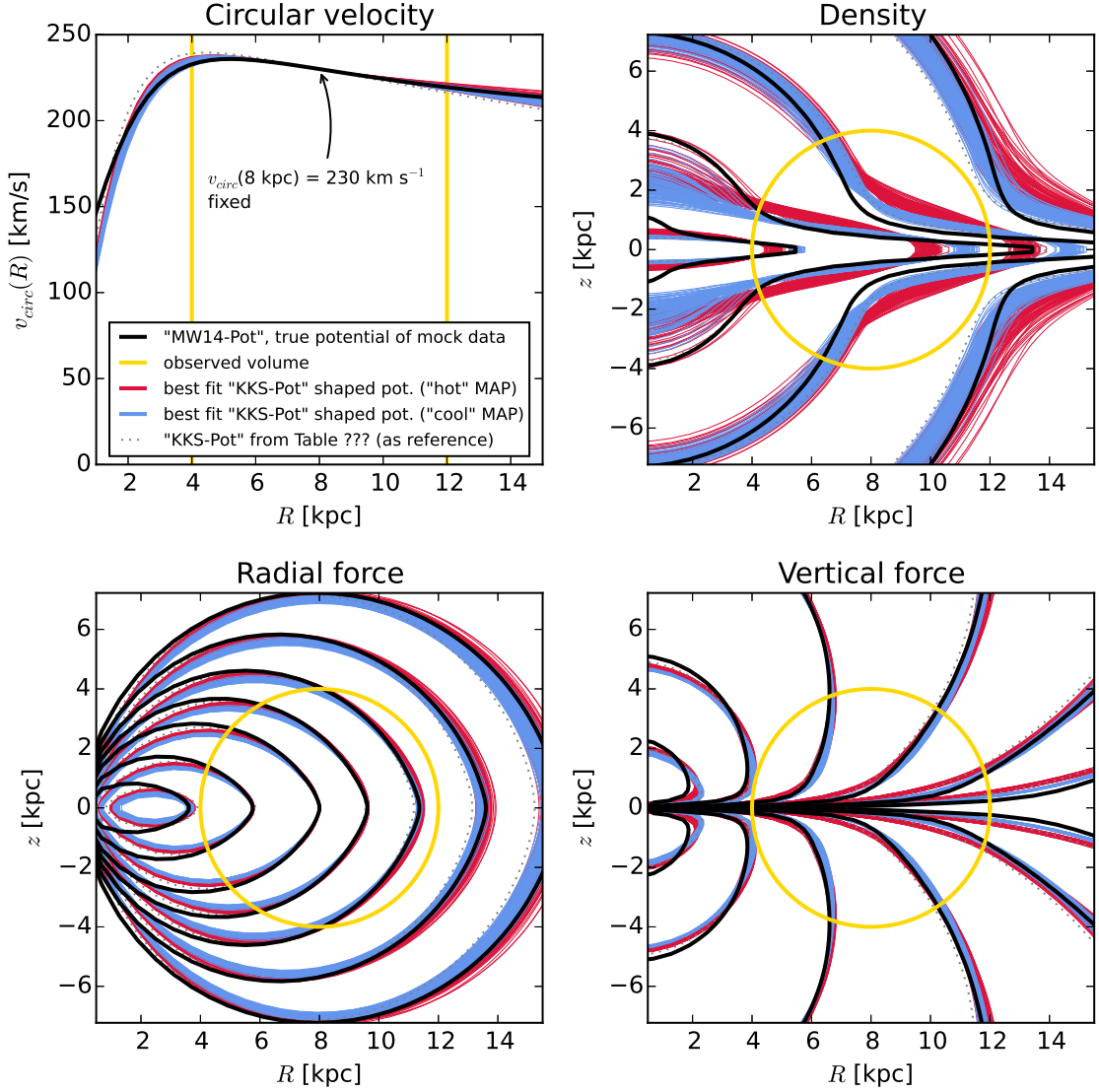
[TO DO: Comment from HW: Author: rix Subject: This paragraph should be 1 or 2 sentences, following the first paragraph on "Sample/Data Properties". This – at the moments – reads to be quite confusing. I don't quite get what the "upshot" is; there is technical detail on  $N_{\text{error}}$  [enough to say it's expensive]; and, as noted earlier; I don't understand why the error convolution for a nearby data point needs to know about  $\delta v_{\text{max}}$ ] Properly convolving the likelihood with measurement errors is computationally very expensive. By ignoring positional errors and only including distance errors as part of the velocity error, we can drastically reduce the computational costs. For stars within 3 kpc from the

sun this approximation works well for distance errors of  $\sim 10\%$  or smaller. The number of MC samples needed for the error convolution using MC integration scales by  $N_{\text{error}} \propto (\delta v_{\text{max}})^2$  with the maximum velocity error at the edge of the sample. If we did not know the true size of the proper motion measurement errors perfectly, we can only reproduce the true model parameters to within  $\lesssim 2$  sigma [TO DO: Check??] as long as we do not underestimate it by more than 10% and for proper motion errors  $\lesssim 2$  mas yr $^{-1}$ .

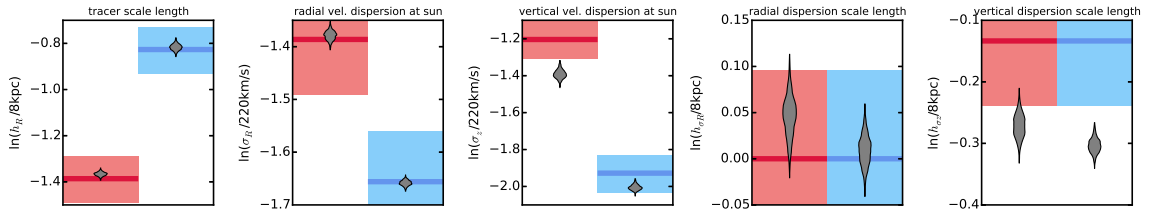
**Deviations from the qDF Assumption:** Our modelling is founded on the assumption, that we can identify *a priori* sub-components of the Galactic disk that follow a qDF (e.g., by considering MAPs). There are two reasons why any chosen sub-sample of stars (here a MAP) may not be well described by any qDF. Either, because nature is more complex, or because even if perfect MAPs would be well described by qDFs finite abundance errors would mix MAPs. We have considered both cases. [TO DO: Comment from HW: it feels to me that this is the 3rd time you said this. It's OK to say, but in 1 line at most.]

In Example 1 in §3.5 we investigated how well we can recover the potential, if this assumption was not perfectly satisfied, i.e., the MAP's true DF does not perfectly follow a qDF. We considered two cases: a) a hot DF, that has less stars at small radii and more stars with low velocities than predicted by the qDF (reddish data sets in Figure 14), or b) a cool DF that has broader velocity dispersion wings and less stars at large radii than predicted by the qDF (bluish data sets). We find that case a) would give more reliable results for the potential parameter recovery, but in both cases biases are small if the contamination is less than [TO DO: CHECK].

If we assumed that the distribution of stars from one MAP is caused by radial migration away from the initial location of star formation, it would more likely that the qDF overestimates the true number of stars at smaller radii than underestimating it at larger radii. [TO DO: Is this actually a sensible argument??? Jo is not convinced that this is right.]



**Figure 18.** Recovery of the gravitational potential if the assumed potential model (KKS-Pot with fixed  $v_{\text{circ}}(R_{\odot})$ ) and the true potential of the (mock) stars (MW14-Pot in Table 1) is slightly different. We show the circular velocity curve, as well as contours of equal density, radial and vertical force in the  $R$ - $z$ -plane, and compare the true potential with 50 [TO DO: CHECK] sample potentials drawn from the posterior distribution function found with the MCMC for a hot (red) and a cool MAP (blue). All model parameters are given as Test 8 in Table 3. [TO DO: Do more analyses??] [TO DO: fancybox Legend] [TO DO: Potential and/or population names in typewriter font] [TO DO: Reference correct Table in Plot - don't forget!] [TO DO: Redo whole analysis with  $v_{\text{circ}}$  not being fixed (HW is not sure if this really doesn't make a difference.)] [TO DO: Comment from Jo: Maybe better with twice as many contour levels? Now not that many within the yellow curve.]



**Figure 19.** Recovery of the qDF parameters for the case where the true and assumed potential deviate from each other (Test 8 in Table 3). The thick red (blue) lines represent the true qDF parameters of the hot (cool) qDF in Table 2 used to create the mock data, surrounded by a 10% error region. The grey violins are the marginalized likelihoods for the qDF parameters found simultaneously with the potential constraints shown in Figure 18. [TO DO: rename  $h_{\sigma R}$  to  $h_{\sigma, R}$ ,  $\sigma_R$  to  $\sigma_{R,0}$  and analogous for  $z$ ]

Following this, focusing the analysis especially on hotter MAPs could be an advisable way to go in the future, if there is doubt that the stars truly follow the qDF.

Another critical point is the binning of stars into MAPs depending on their metallicity and  $\alpha$  abundances. Example 2 in §3.5 could be understood as a model scenario for decreasing bin sizes in the metallicity- $\alpha$  plane when sorting stars in different MAPs, assuming that there is a smooth variation of qDF within the metallicity- $\alpha$  plane and each MAP indeed follows a qDF. We find that, in the case of 20,000 stars in each bin, differences of 20% in the qDF parameters of two neighbouring bins can still give quite good constraints on the potential parameters. We compare this with the relative difference in the qDF parameters in the bins in Figure 6 of Bovy & Rix (2013), which have sizes of  $[\text{Fe}/\text{H}] = 0.1$  dex and  $\Delta[\alpha/\text{Fe}] = 0.05$  dex. It seems that these bin sizes are large enough to make sure that  $\sigma_{R,0}$  and  $\sigma_{z,0}$  of neighbouring MAPs do not differ more than 20%. Figure 15 and 16 suggest that especially the tracer scale length  $h_R$  needs to be recovered to get the potential right. For this parameter however the bin sizes in Figure 6 of Bovy & Rix (2013) might not yet be small enough to ensure no more than 20% of difference in neighbouring  $h_R$ . This is especially the case in the low- $\alpha$  ( $[\alpha/\text{Fe}] \lesssim 0.2$ ), intermediate-metallicity ( $[\text{Fe}/\text{H}] \sim -0.5$ ) MAPs, which were however not used in the dynamical modelling by Bovy & Rix (2013). The above is valid for 20,000 stars per MAP. In case there are less than 20,000 stars in each bin the constraints are less tight and due to Poisson noise one could also allow larger differences in neighbouring MAPs while still getting reliable results.

[TO DO: Include the following comments by Jo somewhere: This is a general approach of fitting action-based disk DFs for getting the potential. The qDF is a specific example that we use in this paper. In future studies different forms of DFs might be fitted to data. That the results are quite robust to the form of the DF not entirely correct motivates this further.]

**Gravitational Potential beyond the Parameterized Functions Considered:** [TO DO: Comment from HW: In style and content this seems very similar to the RESULTS section. We should either drastically shorten the text in the results section (probably not), or here.] In the long run *RoadMapping* should incorporate a family of gravitational potential models that can reproduce the essential features of the MW's true mass distribution. While our fundamental assumption of the Galaxy's axisymmetry is at odds with the obvious existence of non-axisymmetries in the MW, we will not dive into investigating these implications in this paper. Instead we test how a misjudgment of the parametric potential form affects the recovery by fitting a potential of Stäckel form (Batsleer & Dejonghe 1994) to mock data from a different potential family with halo, bulge and exponential disk. The recovery is quite successful and we get the best fit within the limits of the model. However, even a strongly flattened Stäckel potential component has difficulties to recover the very flattened mass distribution of an exponential disk. This leads to misjudgment of the qDF parameters describing the vertical action distribution,  $\sigma_{z,0}$  and  $h_{\sigma,z}$ . As the qDF parameter  $\sigma_{z,0}$  corresponds to the physical vertical velocity dispersion at the sun, a comparison with

direct measurements could be a valuable cross-checking reference. [TO DO: This might not be true. For isochrone and Staeckel potential I get this behaviour, but not for the MW14-Pot!!! Might be, because it's not separable??? Check!!!] [TO DO: best to simply remove it...] In case of as many as 20,000 stars we should therefore already be able to distinguish between different potential models.

The advantage of using a Stäckel potential with *RoadMapping* is firstly the exact and fast action calculation via the numerical evaluation of a single integral, and secondly that the potential has a simple analytic form, which greatly speeds up calculations of forces and frequencies (as compared to potentials in which only the density has an easy description like the exponential disk). A superposition of several simple Kuzmin-Kutuzov Stäckel components can successfully produce MW-like rotation curves (see Batsleer & Dejonghe (1994), Famaey & Dejonghe (2003) and Figure 18) and one could think of adding even more components for more flexibility, for example a small roundish component for the bulge. [TO DO: Comment by Jo: In a sense the two approaches (a) using the Staeckel action approximation with a MW like potential and (b) using a Staeckel potential directly are doing the same thing (approximating the true potential as a Staeckel potential). The question is which is best.]

The potential model used by Bovy & Rix (2013) had only two free parameters (disk scale length and halo contribution to  $v_{\text{circ}}(R_{\odot})$ ). To circumvent the obvious disadvantage of this being at all not flexible enough, they fitted the potential separately for each MAP and recovered the mass distribution for each MAP only at that radius for which it was best constrained - assuming that MAPs of different scale length would probe different regions of the Galaxy best. Based on our results in Figure 18 this seems to be indeed a sensible approach [TO DO: Check that this is indeed the case - it is not clear to me from the plot. ???].

We suggest that combining the flexibility and computational advantages of a superposition of several Stäckel potential components with probing the potential in different regions with different MAPs as done by Bovy & Rix (2013), could be a promising approach to get the best possible constraints on the MW's potential.

**Different Modelling Approaches using Action-based Distribution Functions:** We have focussed for the time being on MAPs for a number of reasons: First, they seem to permit simple DFs (Bovy et al. 2012b,c,d), i.e., approximately qDFs (Ting et al. 2013). Second, all stars, e.g., those from different MAPs, must orbit in the same potential. Therefore each MAP will and can yield quite different DF parameters; but each MAP will also provide a (statistically) independent estimate of the potential parameters. At the same time—if all is well—those potential parameters, derived from different MAPs, should be mutually consistent. In some sense, this approach focusses on constraining the potential, treating the DF parameters as nuisance parameters. The main drawback is that we have many astrophysical reasons that the DF properties (for reasons of galaxy evolution and chemical evolution) are astrophysically linked between different MAPs. Ultimately, the goal

is to do a fully consistent chemodynamical model that simultaneously fits the potential and  $DF(\mathbf{J}, [X/H])$  simultaneously (where  $[X/Fe]$  denotes the full abundance space) with a full likelihood analysis. This has not yet been attempted here, because the behaviour is quite complex.

Since the first application of *RoadMapping* by Bovy & Rix (2013) there have been two similar efforts to constrain the Galactic potential and/or orbit distribution using action-based distribution functions: Piffl et al. (2014) fitted both potential and a  $f(\mathbf{J})$  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 [TO DO: CHECK] for the thick disk stars and an additional DF for the halo stars. Taking proper care of the selection function requires a full likelihood analysis and the calculation of the likelihood normalisation is computationally expensive. Piffl et al. (2014) choose to circumvent this problem by directly fitting a) histograms of the three velocity components in eight spatial bins to the velocity distribution predicted by the DF and b) the vertical density profile predicted by the DF to the profiles by Jurić et al. (2008). The vertical force profile of their best fit mass model nicely agrees with the results from Bovy & Rix (2013) for  $R > 6.6$  kpc. The disadvantage of their approach is, that by binning the stars spatially, a lot of information is not used. Sanders & Binney (2015) have focussed on understanding the abundance-dependence of the DF, relying on

a fiducial potential. They developed extended distribution functions, i.e., functions of both actions and metallicity for a superposition of thin and 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, and a simple prescription for radial migration. They applied a full likelihood analysis accounting for selection effects and found a best fit for the eDF in a fixed fiducial potential by Dehnen & Binney (1998) to the stellar phase-space and metallicity [TO DO: CHECK] data of the Geneva-Copenhagen Survey (Nordström et al. 2004; Holmberg et al. 2009) and the stellar density curves by Gilmore & Reid (1983). Their best fit predicted the velocity distribution of SEGUE G dwarfs quite well, but had biases in the metallicity distribution, which they accounted to being a problem with the SEGUE metallicities.

We know that real galaxies, including the Milky Way, are not axisymmetric. Using N-body models, we will explore in a subsequent paper what when data from a non-axisymmetric system get interpreted through axisymmetric modelling.

[TO DO: Comment from Jo: Maybe we also want a conclusion with a simple bullet-point list of the main conclusions discussed in detail in the Discussion section.]

## 5. ACKNOWLEDGMENTS

The authors thank Glenn van de Ven for suggesting the use of Kuzmin-Kutuzov Stäckel potentials in this case study. [TO DO: What else do we to have acknowledge??]

## APPENDIX

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

In §3.3 we found a striking robustness of the *RoadMapping* modelling approach against wrong assumptions about the radial incompleteness of the data set. To further test this result, we investigate a different completeness function that drops with distance from the Galactic plane (see Test 5, Example 2, in Table 3 and Figure 20). We get a similar robust behaviour for small deviations, and only slightly less robustness for larger deviations. That an explanation for this robustness could be, that much of the information about the potential comes from the rotation curve, which is not affected by incompleteness, is demonstrated in Figure 22.

*Marginalization over  $v_T$ .* — The likelihood in Equation 10 is marginalized over the coordinate  $v_T$  as follows

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

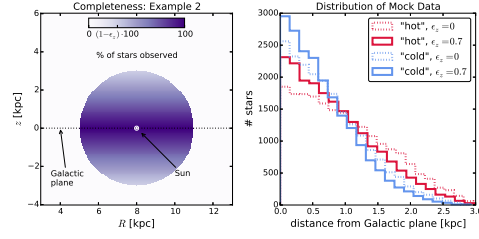
where  $P(\mathbf{x}, \mathbf{v} | p_M)$  is the same as in Equation 10 and the numerical integral over  $v_T$  is performed as a 24th order Gauss-Legendre quadrature. The additional factor of  $v_0$  is needed to get the units of  $P_{(v_T \text{ marg.})}(\mathbf{x}_i, v_{R,i}, v_{z,i} | p_M)$  right.

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

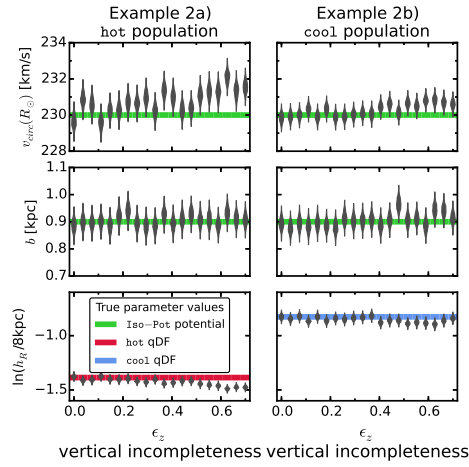
## REFERENCES

- |   |   |
|---|---|
| Batsleer, P., & Dejonghe, H. 1994, A&A, 287, 43   | Binney, J. 2012, MNRAS, 426, 1328                                 |
| Binney, J. 2010, MNRAS, 401, 2318                 | Binney, J. 2013, NAR [TO DO: emulateapj doesn't know NAR], 57, 29 |
| Binney, J., & McMillan, P. 2011, MNRAS, 413, 1889 |   |
| Binney, J. 2011, Pramana, 77, 39                  |   |
| Binney, J. 2012, MNRAS, 426, 1324                 |   |





**Figure 20.** Selection function and mock data distribution for investigating vertical incompleteness of the data. All model parameters are summarized as Test 5, Example 2, in Table 3. The survey volume is a sphere around the sun and the percentage of observed stars is decreasing linearly with distance from the Galactic plane, as demonstrated in the left panel. How fast this detection/incompleteness rate drops is quantized by the factor  $\epsilon_z$ . Histograms for four data sets, drawn from two MAPs (hot in red and cool in blue, see Table 2) and with two different  $\epsilon_z$ , 0 and 0.7, are shown in the right panel for illustration purposes. [TO DO: Potential and/or population names in typewriter font]



**Figure 21.** Influence of wrong assumptions about the incompleteness parallel to the Galactic plane of the data on the parameter recovery with *RoadMapping*. Each mock data set was created having different incompleteness parameters  $\epsilon_z$  (shown on the  $x$ -axis and illustrated in Figure 20) and the model parameters are given as Test 5, Example 2, in Table 3. The analysis however didn't know about the incompleteness and assumed that all data sets had constant completeness within the survey volume ( $\epsilon_z = 0$ ). The marginalized likelihoods from the fits are shown as violins. The green lines mark the true potential parameters (Iso-Pot) and the red and blue lines the true qDF parameters (hot MAP in red and cool MAP in blue), which we tried to recover. The *RoadMapping* method seems to be robust against small to intermediate deviations between the true and the assumed vertical data incompleteness, as well as the radial incompleteness in Figure 21.

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

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

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

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

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

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

Bovy, J. 2015, *ApJS*, 216, 29

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

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

de Lorenzi, F., Debattista, V. P., Gerhard, O., & Sambhus, N. 2007, *MNRAS*, 376, 71

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

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

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

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

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

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

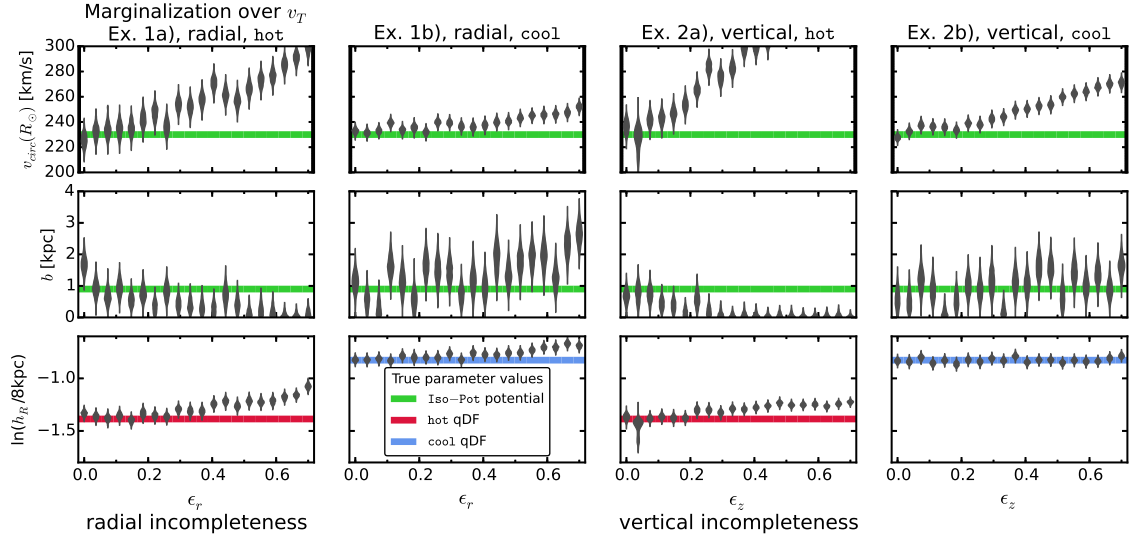
Kawata, D., Hunt, J. A. S., Grand, R. J. J., Pasetto, S., & Cropper, M. 2014, *MNRAS*, 443, 2757

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

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

McMillan, P. J. 2011, *MNRAS*, 414, 2446





**Figure 22.** Influence of wrong assumptions about radial and vertical incompleteness on the parameter recovery, when *not* including information about the tangential velocities in the analysis. The mock data sets are the same as in Figure 10 and 21, but this time we did not include the data coordinates  $v_T$  in the analysis and therefore marginalized the likelihood over  $v_T$  instead (see §.1). This demonstrates that 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 ( $\epsilon_z \lesssim 10\%$ ).

McMillan, P. J. 2012, European Physical Journal Web of Conferences, 19, 10002  
 McMillan, P. J., & Binney, J. J. 2008, MNRAS, 390, 429  
 McMillan, P. J., & Binney, J. 2012, MNRAS, 419, 2251  
 McMillan, P. J., & Binney, J. J. 2013, MNRAS, 433, 1411  
 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

[TO DO: In which order should I give the references????]

[TO DO: replace the references which I typed myself with the ones from ADS.]

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

Rix, H.-W., & Bovy, J. 2013, A&A Rev., 21, 61  
 Sanders, J. L., & Binney, J. 2015, MNRAS, 449, 3479  
 Sellwood, J. A. 2010, MNRAS, 409, 145  
 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 selection function model etc. used for the mock data creation, the third model the corresponding model assumed in the analysis, and the last column lists the figures belonging to the test suite. Parameters that are not left free in the analysis, are always fixed to their true value. Unless otherwise stated we calculate the likelihood by the nested-grid and MCMC approach outlined in §2.7 and use  $N_x = 16$ ,  $N_v = 24$ ,  $n_\sigma = 5$  as numerical accuracy for the likelihood normalisation in Equations (10) and (8).

| Test   | Model for Mock Data  |   | Model in Analysis   | Figures  |
|--|--|---|---|--|
| Test 1 :<br>Influence of<br>survey volume on<br>mock data distribution,<br>also in action space      | <i>Potential:</i><br><i>DF:</i><br><i>Survey volume:</i><br><br><i># stars per data set:</i><br><i># data sets:</i>  | KKS-Pot<br>hot or cool qDF<br>a) $R \in [4, 12]$ kpc, $z \in [-4, 4]$ kpc, $\phi \in [-20^\circ, 20^\circ]$ .<br>b) $R \in [6, 10]$ kpc, $z \in [1, 5]$ kpc, $\phi \in [-20^\circ, 20^\circ]$ .<br>20,000<br>4 ( $= 2 \times 2$ models) | -   | Mock data:<br>Figure 2   |
| Test 2 :<br>Numerical accuracy<br>in calculation<br>of the likelihood<br>normalisation               | <i>Potential:</i><br><i>DF:</i><br><i>Survey volume:</i><br><i>Numerical accuracy:</i>   | Iso-Pot, MW13-Pot & KKS-Pot<br>hot qDF<br>sphere around sun, $r_{\max} = 0.2, 1, 2, 3$ or 4 kpc<br>$N_x \in [5, 20]$ , $N_v \in [6, 40]$ , $n_\sigma \in [3.5, 7]$  | -   | Convergence<br>of normalisation:<br>Figure 3   |
| Test 3.1 :<br><i>pdf</i> is a<br>multivariate<br>Gaussian<br>for large data sets.                    | <i>Potential:</i><br><i>DF:</i><br><i>Survey Volume:</i><br><i># stars per data set:</i><br><i># data sets:</i><br><i>Numerical accuracy:</i>                                    | Iso-Pot<br>hot qDF<br>sphere around sun, $r_{\max} = 2$ kpc<br>20,000<br>5 (only one is shown)  | Iso-Pot, all parameters free<br>qDF, all parameters free<br>(fixed & known)<br><br>$N_v = 20$ and $n_\sigma = 4$  | Figure 5   |
| Test 3.2 :<br>Width of the<br>likelihood scales<br>with number of stars<br>by $\propto 1/\sqrt{N}$ . | <i>Potential:</i><br><i>DF:</i><br><br><i>Survey volume:</i><br><i># stars per data set:</i><br><i># data sets:</i><br><i>Analysis method:</i><br><i>Numerical accuracy:</i>     | Iso-Pot<br>hot qDF<br><br>sphere around sun, $r_{\max} = 3$ kpc<br>between 100 and 40,000<br>132<br><br>likelihood on grid<br>$N_v = 20$ and $n_\sigma = 4$ (for speed)   | Iso-Pot, free parameter: $b$<br>hot qDF, free parameters:<br>$\ln\left(\frac{h_R}{8\text{kpc}}\right), \ln\left(\frac{\sigma_{R,0}}{230\text{km s}^{-1}}\right), \ln\left(\frac{h_{\sigma,R}}{8\text{kpc}}\right)$<br>(fixed & known)   | Figure 6   |
| Test 3.3 :<br>Parameter estimates<br>are unbiased.   | <i>Potential:</i><br><br><i>DF:</i><br><br><i>Survey volume:</i><br><i># stars per data set:</i><br><i># data sets:</i><br><i>Analysis method:</i><br><i>Numerical accuracy:</i> | 2 Iso-Pot with<br>$b = 0.8$ kpc or $b = 1.5$ kpc<br>hot or cool qDF<br><br>5 spheres around sun, $r_{\max} = 0.2, 1, 2, 3$ or 4 kpc<br>20,000<br>640 ( $= 2 \times 2 \times 5$ models $\times 32$ realisations)                         | Iso-Pot, free parameter: $b$<br><br>hot/cool qDF, free parameters:<br>$\ln\left(\frac{h_R}{8\text{kpc}}\right), \ln\left(\frac{\sigma_{R,0}}{230\text{km s}^{-1}}\right), \ln\left(\frac{h_{\sigma,R}}{8\text{kpc}}\right)$<br>(fixed & known)<br><br>likelihood on grid<br>$N_v = 20$ and $N_\sigma = 4$ (for speed) | Figure 7   |
| Test 4 :<br>Influence of<br>position & shape<br>of survey volume<br>on parameter recovery            | <i>Potential:</i><br><br><i>DF:</i><br><br><i>Survey volume:</i><br><i># of stars per data set:</i><br><i>Analysis method:</i>   | i) Iso-Pot or ii) MW13-Pot<br><br>hot qDF<br><br>4 different wedges, see Figure 8, upper right panel<br>20,000  | i) Iso-Pot, all parameters free<br>ii) MW13-Pot, $R_d$ and $f_h$ free<br>i) qDF, all parameters free<br>ii) qDF, only $h_R$ , $\sigma_{z,0}$ and $h_{\sigma,R}$ free<br>(fixed & known)<br><br>i) MCMC, ii) likelihood on grid  | Figure 8   |
| Test 5 :<br>Influence of<br>wrong assumptions<br>about the data set<br>(in-)completeness             | <i>Potential:</i><br><i>DF:</i><br><i>Survey volume:</i><br><i>Completeness:</i>   | Iso-Pot<br>a) hot or b) cool qDF<br>sphere around sun, $r_{\max} = 3$ kpc<br><i>Example 1:</i> radial incompleteness,<br>$\text{completeness}(r) = 1 - \epsilon_r \frac{r}{r_{\max}}$ , twenty $\epsilon_r \in [0, 0.7]$                | Iso-Pot, all parameters free<br>qDF, all parameters free<br>(fixed & known)<br>data set complete,<br>$\text{completeness}(r) = 1$ , $\epsilon_r = 0$  | Illustration & mock data:<br>Figures 9 & 20<br>Analysis results:<br>Figures 10 & 21<br>Analysis results: |

TRICK ET AL.

[TO DO: Remove # data sets, where it actually is not important.] [TO DO: Jo suggested to make many tables from this. But I actually like one big table at the end of the paper. Otherwise we had 6 additional tables interrupting the flow of text and figures all the time. And the parameters in the table are really just for reference.] [TO DO: Overall, this table could do with a little less information.]