Draft version March 31, 2016
Preprint typeset using LaTeX style emulateapj v. 12/16/11

ACTION-BASED DYNAMICAL MODELLING FOR THE MILKY WAY DISK

Wilma H. Trick[1,2], Jo Bovy[3], and Hans-Walter Rix[1]

Draft version March 31, 2016

ABSTRACT

We present *RoadMapping*, a full-likelihood dynamical modelling machinery that aims to recover the Milky Way's (MW) gravitational potential from large samples of stars in the Galactic disk. *RoadMapping* models the observed positions and velocities of stars with a parametrized, three-integral distribution function (DF) in a parametrized axisymmetric potential. We investigate how properties of data, model and machinery affect constraints on the potential and DF. By analyzing large sets of idealized mock data—using, e.g., the isochrone potential, axisymmetric DFs, contiguous survey selection functions (SF) and negligible measurement errors in many tests—we perform a differential diagnosis of different isolated modelling aspects. Overall, we find that the potential can be reliably recovered with *RoadMapping*, even if the model assumptions are slightly wrong. Specifically, our key results are: (i) If the MW's true gravitational potential is not included in the assumed family of parametrized model potentials, we can—at least in the axisymmetric case—still find a potential that robustly approximates the potential within the limitations of the model. (ii) Modest systematic differences between the true and best fit model DF are inconsequential. For example, when defining sub-populations by binning stars according to their chemical abundances, finite bin sizes and abundance errors do not affect the modelling as long as the DF parameters of neighbouring bins do not differ by more than 20%. Hotter populations are less affected by (iii) pollution or (iv) misjudgements of the proper motion uncertainty and (v) cooler populations recover the Galactic rotation curve more reliably. In general, we showed that *RoadMapping* gives constraints of high precision on both potential and DF parameters (vi) for large sample sizes, (vii) for survey volumes of large radial and vertical coverage, or (viii) as long as the proper motion uncertainties are well known and even as large as 5 mas yr^{-1}. Unbiased potential estimates are ensured, (ix) for small to moderate misjudgements of the spatial SF, (x) if distances are known only to within 10%, or (xi) if proper motion uncertainties are known within 10%. Challenges are the rapidly increasing computational costs for high precision likelihood evaluations required for large sample sizes. Overall, *RoadMapping* is well suited to making precise new measurements of the MW's potential with data from the 2017 Gaia release.

[TO DO: Relate error results with typical GAIA errors.] [TO DO: Abstract has to be shorter - 250 words maximum.] [TO DO: The referee writes: "I think section 3.5 and 3.6 are the most valuable in the paper as they really explore potential systematics. In my opinion, these are the key results." - Stress in Abstract.]

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

1. INTRODUCTION

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

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

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

[TO DO: Magorrian (2014) has provided a framework for constraining the potential without assuming a particular parametrized form for the DF. Whilst Magorrian's

[1] Max-Planck-Institut für Astronomie, Königstuhl 17, D-69117 Heidelberg, Germany
[2] Correspondence should be addressed to trick@mpia.de.
[3] Department of Astronomy and Astrophysics, University of Toronto, 50 St. George Street, Toronto, ON, M5S 3H4, Canada

method is computationally intensive, it should be referenced in the introduction as it relates to the later discussion of choosing a particular DF parametrization.]

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

Bovy & Rix (2013) (BR13 hereafter) put this in practice by implementing a rigorous modelling approach for so-called mono-abundance populations (MAPs), i.e, sub-sets of stars with similar [Fe/H] and [α/Fe] within the Galactic disk, which seem to follow simple DFs (Bovy et al. 2012b,c,d). Given an assumed (axisymmetric) model for the Galactic potential and action-based DF (Binney 2010; Binney & McMillan 2011; Ting et al. 2013) they calculated the likelihood of the observed $(\vec{x}, \vec{v})$ for each MAP, using SEGUE G-dwarf stars (Yanny et al. 2009). They also accounted for the complex, but known selection function of the kinematic tracers (Bovy et al. 2012d). For each MAP the modelling resulted in an independent estimate of the same gravitational potential. Taken as an ensemble, they constrained the disk surface density over a wide range of radii ($\sim 4-9$ kpc), and powerfully constrained the disk mass scale length and the stellar-disk-to-dark-matter ratio at the Solar radius.

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

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

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

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

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

For all of the above aspects we show some plausible and illustrative examples on the basis of investigating mock data. The mock data is generated from galaxy models outlined in Sections 2.1-**??** following the procedure in Appendix **??**and analysed according to the description of the *RoadMapping* machinery in Sections **??**-2.8. Section **??** compiles our results on the investigated modelling aspects. In particular, our key results about the systematics introduced by using wrong DF or potential models are presented in the Sections 3.5 and 3.6. Section 4 finally summarizes and discusses our findings.

2. DYNAMICAL MODELLING

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

2.1. *Coordinate system*

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

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

2.2. *Actions*

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

$$J_i \equiv \frac{1}{2\pi} \oint_{\mathrm{orbit}} p_i \, \mathrm{d}x_i, \tag{1}$$

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

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

evolve linearly in time and specify the position of the star along the orbit. (For a full introduction to angle-action variables see Binney & Tremaine 2008, §3.5.)

Action calculation from a star's phase-space coordinates, $(\boldsymbol{x}, \boldsymbol{v}) \xrightarrow{\Phi} \boldsymbol{J}$, is typically very computationally expensive. Only for or some special, separable potentials Equation (1) simplifies significantly. The triaxial Stäckel potentials (de Zeeuw 1985) are the most general potentials, that allow exact action calculations using a single quadrature . Some flattened axisymmetric Stäckel potentials are quite similar to our Galaxy's potential (Binney & Tremaine 2008, §3.5.3; Batsleer & Dejonghe 1994; Famaey & Dejonghe 2003). The spherical isochrone potential (Henon 1959; Binney & Tremaine 2008, §3.5.3) is the most general special case for which the action calculation is analytic without any integration. In all other potentials actions have to be numerically estimated; see **?** for a recent review of action estimation methods. According to **?** the best compromise of speed and accuracy for the Galactic disk is the *Stäckel fudge* by Binney (2012a) for axisymmetric potentials. In addition we use action interpolation grids (Binney 2012a; Bovy 2015) to speed up the calculation. The latter is one of the improvements employed by *RoadMapping*, which was not used in BR13.

2.3. *Potential models*

For the gravitational potential in our modelling we assume a family of parametrized models. We use: The MW-like potential from BR13 (`MW13-Pot`) with bulge, disks and halo; 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. The true circular velocity at the Sun for all potential models was chosen to be $v_{\rm circ}(R_\odot) = 230$ km s^{-1}. Table 1 summarizes all reference potentials used in this work together with their free parameters p_Φ. The density distribution of these potentials is illustrated in Figure 1. Many tests are performed for speed reasons with the `Iso-Pot`, which allows the fastest action calculations, and the `KKS-Pot`, whose analytic form makes the computation of forces and densities quick and easy. Both potentials also have the advantage of allowing for accurate action estimation. [TO DO: Rewrite to account for DHB-Pot.] [TO DO: Make sure that nowhere in this work the MW13-Potential is still used.]

2.4. *Stellar distribution functions*

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

The action-based quasi-isothermal distribution function (qDF) by Binney (2010) and Binney & McMillan (2011) is a simple DF which we will employ as a specific example throughout this work to describe individual stellar sub-populations. This is motivated by the findings of Bovy et al. (2012b,c,d) and Ting et al. (2013) on the simple phase-space structure of stellar MAPs and BR13's successful application. The qDF has the form

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

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

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

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

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

Following BR13, we choose the functional forms

$$n(R_g \mid p_{\rm DF}) \propto \exp\left(-\frac{R_g}{h_R}\right) \tag{5}$$

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

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

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

One indispensable step in our dynamical modelling technique (Section **??**), as well as in creating mock data (Appendix **??**), is to calculate the (axisymmetric) spatial tracer density $\rho_{\rm DF}(\boldsymbol{x} \mid p_\Phi, p_{\rm DF})$ for a given DF and potential. Analogously to BR13,

$$\begin{aligned}&\rho_{\rm DF}(R, |z| \mid p_\Phi, p_{\rm DF})\\ &= \int_{-\infty}^{\infty} \mathrm{DF}(\boldsymbol{J}[R, z, \boldsymbol{v} \mid p_\Phi] \mid p_{\rm DF})\, \mathrm{d}^3 v\\ &\approx \int_{-n_\sigma\sigma_R(R|p_{\rm DF})}^{n_\sigma\sigma_R(R|p_{\rm DF})} \int_{-n_\sigma\sigma_z(R|p_{\rm DF})}^{n_\sigma\sigma_z(R|p_{\rm DF})} \int_0^{1.5 v_{\rm circ}(R_\odot)}\\ &\qquad \mathrm{DF}(\boldsymbol{J}[R, z, \boldsymbol{v} \mid p_\Phi] \mid p_{\rm DF})\, \mathrm{d}v_T\, \mathrm{d}v_z\, \mathrm{d}v_R,\end{aligned} \tag{8}$$

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

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

[(a)] The free parameter of the spherical `Iso-Pot` is the isochrone scale length b.
[(b)] The coordinate system of each of the two Stäckel-potential components of the `KKS-Pot` is $R^2/(\tau_{i,p}+\alpha_p) + z^2/(\tau_{i,p}+\gamma_p) = 1$ with $p \in \{\text{Disk}, \text{Halo}\}$ and $\tau_{i,p} \in \{\lambda_p, \nu_p\}$. Both components have the same focal distance $\Delta \equiv \sqrt{\gamma_p - \alpha_p}$, to ensure that the superposition itself is a Stäckel potential. The axis ratio of the coordinate surfaces $(a/c)_p := \sqrt{\alpha_p/\gamma_p}$ describes the flatness of each component. k is the relative contribution of the disk mass to the total mass.
[(c)] The parameters of the `DHB-Pot` are the Miyamoto-Nagai disk scale length $a_{\rm disk}$ and height $b_{\rm disk}$, the NFW halo scale length $a_{\rm halo}$ and its relative contribution to $v^2_{\rm circ}(R_\odot)$ (with respect to the total disk+halo contribution), $f_{\rm halo}$, and the Hernquist bulge scale length $a_{\rm bulge}$ and its contribution to the total $v^2_{\rm circ}(R_\odot)$, $f_{\rm bulge}$.
[(d)] The `MWPotential2014` by Bovy (2015) (see their Table 1) has $v_{\rm circ}(R_\odot) = 220$ km s^{-1}. We use however $v_{\rm circ}(R_\odot) = 230$ km s^{-1}.

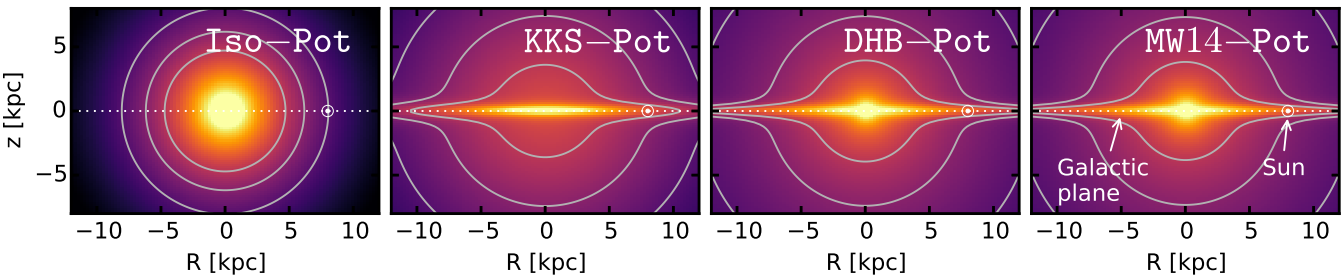

Figure 1. [This is a new version of the original plot - using the DHB-Pot instead of the MW13-Pot. (TO DO: Remove this note for final version.)] Density distribution of the four reference galaxy potentials in Table 1. These potentials are used throughout this work to create and model mock data with *RoadMapping*.

Table 2
Reference parameters for the qDF in Equations (2)-(7), used to create 6D phase-space mock data sets for stellar populations of different kinematic temperature.

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

where $\sigma_R(R \mid p_{\mathrm{DF}})$ and $\sigma_z(R \mid p_{\mathrm{DF}})$ are given by Equations (6) and (7).[5] Each integral is evaluated using a N_v-th order Gauss-Legendre quadrature. For a given p_Φ and p_{DF} we explicitly calculate the density on $N_x \times N_x$ regular grid points in the (R, z) plane and interpolate $\log \rho_{\mathrm{DF}}$ in between using bivariate spline interpolation. The grid is chosen to cover the extent of the observations (for $|z| \geq 0$, because the model is symmetric in z by construction). The total number of actions to be calculated to set up the density interpolation grid is $N_x^2 \times N_v^3$, which is one of the factors limiting the computation speed. To complement the work by BR13, we will specifically work out in Section **??** and Figure 2 how large N_x, N_v and n_σ have to be chosen to get the density with a sufficiently high numerical accuracy.

2.5. *Data likelihood*

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

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

(For simplicity we assume in most tests of this study contiguous SFs of spherical shape around the Sun, which are functions of $\boldsymbol{x}$ only and which we motivate in Appendix .2. The maximum radius of this spherical observed volume is denoted by $r_{\max}$.)

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

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

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

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

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

where N_* is the number of stars in D, and in the last step we used Equation (8).[6] $\prod_i \mathrm{SF}(\boldsymbol{x}_i)$ is independent of p_M, so we treat it as unimportant proportionality factor.

[5] The integration ranges over the velocities are motivated by Figure 19 and n_σ should be chosen as $n_\sigma \sim 5$ (see Figure 2). The integration range $[0, 1.5 v_{\mathrm{circ}}(R_\odot)]$ over v_T is in general sufficient, only for observation volumes with larger mean stellar v_T this upper limit needs to be increased.

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

We find the best fitting p_M by maximizing the posterior probability distribution $pdf(p_M \mid D)$, which is, according to Bayes' theorem

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

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

$$p_{\mathrm{DF}} := \{\ln h_R, \ln \sigma_{R,0}, \ln \sigma_{z,0}, \ln h_{\sigma,R}, \ln h_{\sigma,z}\} \tag{10}$$

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

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

2.6. *Likelihood normalisation*

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

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

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

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

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

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

$$\begin{aligned} &\log \mathscr{L}_{\mathrm{approx}}(D \mid p_M) \\ &= \sum_i^{N_*} \log \mathrm{DF}(\boldsymbol{J_i} \mid p_M) - N_* \log(M_{\mathrm{tot}}) \\ &- N_* \log(1 + \delta M_{\mathrm{tot}}), \end{aligned} \tag{13}$$

with

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

and therefore $\delta M_{\mathrm{tot}} \lesssim 1/N_*$. Otherwise numerical inaccuracies could lead to systematic biases in the potential

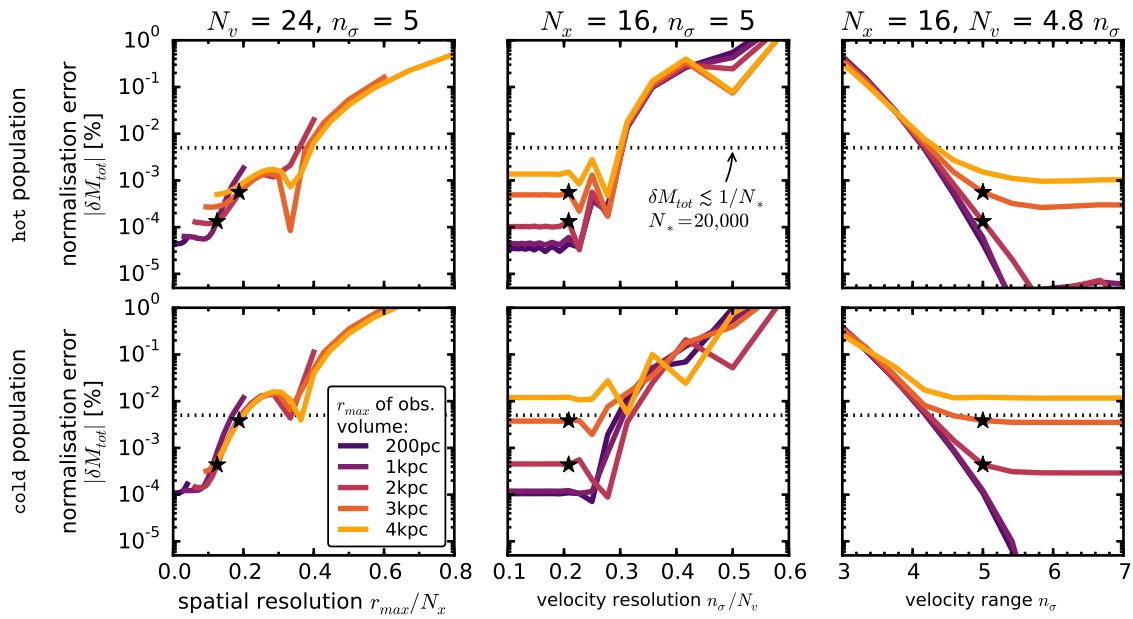

Figure 2. [This is a new version of the original plot - using the DHB-Pot instead of the KKS-Pot, considering a more accurate calculation of the $M_{\mathrm{tot,true}}$, and we also considered a cool population to test the influence of the kinematic temperature. (TO DO: Remove this note for final version.)]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

and DF recovery. For data sets as large as $N_* = 20,000$ stars, which in the age of Gaia could very well be the case, one needs a numerical accuracy of 0.005% in the normalisation. We made sure that this is satisfied for all analyses in this work. Figure 2 demonstrates how the numerical accuracy for analyses with the `DHB-Pot` depends on the spatial and velocity resolution of the grid and that the accuracy we use, $N_x = 16$, $N_v = 24$ and $n_\sigma = 5$, is sufficient.[7] It has to be noted however, that the optimal values for N_x, N_v and n_σ depend not only on N_*, but also on the kinematic temperature of the population (and to a certain degree even on the choice of potential[8]) and it has to be checked on a case-by-case basis what the optimal accuracy is.

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

[7] The accuracy used in this work's analyses is slightly higher than in BR13, where N_* was only a few ~ 100.

[8] In potentials where the Hamilton-Jacobi equation [TO DO: check] is separable (like the `Iso-Pot` and `KKS-Pot`) the DF scaling profiles in Equations (6) and (7) appear to be close to the physical velocity dispersion. It works therefore well to use them together with $n_\sigma \sim 5$ for the integration limits over velocity in Equation (8). For non-separable potentials the DF dispersion parameters and physical velocity dispersion might deviate strongly and n_σ has to be adapted accordingly to get a suitable integration regime.

2.7. *Measurement errors*

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

$$\tilde{p}(\boldsymbol{y}_i \mid p_\Phi, p_{\mathrm{DF}}, \delta\boldsymbol{y}_i) \equiv \frac{\mathrm{SF}(\boldsymbol{y}_i) \cdot \int \mathrm{DF}(\boldsymbol{y}') \cdot N[\boldsymbol{y}_i, \delta\boldsymbol{y}_i]\, \mathrm{d}^6 y'}{\int \left(\mathrm{DF}(\boldsymbol{y}') \cdot \int \mathrm{SF}(\boldsymbol{y}) \cdot N[\boldsymbol{y}', \delta\boldsymbol{y}_i]\, \mathrm{d}^6 y\right) \mathrm{d}^6 y'}. \quad (15)$$

In the case of uncertainties in distance or (RA,Dec) the evaluation of this is computational expensive—especially if the stars have heteroscedastic $\delta\boldsymbol{y}_i$. In practice we compute the convolution using Monte Carlo (MC) integra-

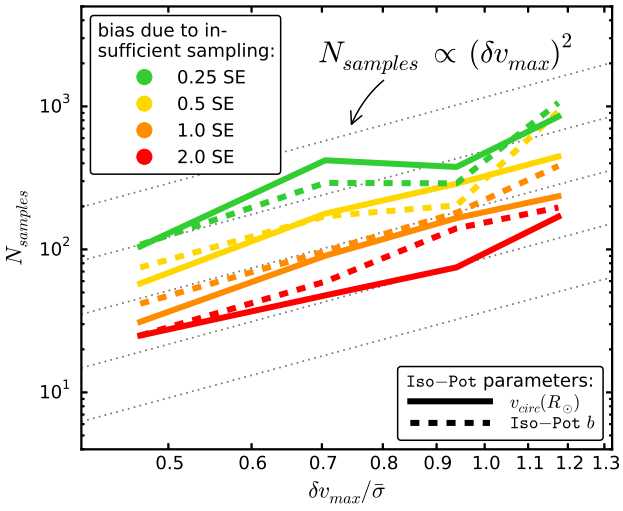

Figure 3. Number of MC samples $N_{\rm samples}$ needed for the numerical convolution of the model probability with the measurement uncertainties in Equation (16), given the maximum velocity uncertainty $\delta v_{\rm max}$ within the stellar sample with respect to the sample's kinematic temperature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

tion with $N_{\rm samples}$ samples,

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

with

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

[TO DO: The referee writes: "Equation (15) is a novelty. It is troubling that the tests that use this approximation all seem to use the isochrone but the approximation is still necessary. Is that because it is computationally awkward to calculate this integral or just very slow?"] [Make clear the otherwise we have to calculate normalization for each star] [Is a reason why we have to restrict the analysis to stars with good distance such that this approximation is not too far off.]

The above approximation also assumes that the star's *position* $\tilde{\boldsymbol{x}}_i$ is perfectly measured. As the SF is also velocity independent, this simplifies the normalisation drastically to Equation (11). Measurement uncertainties in RA and Dec are often negligible anyway. The uncertainties in the Galactocentric *velocities* $\boldsymbol{v}_i = (v_{R,i}, v_{T,i}, v_{z,i})$ depend not only on $\delta\boldsymbol{\mu}$ and $\delta v_{\rm los}$ but also on the distance and its uncertainty, which we do *not* neglect when drawing MC samples $\boldsymbol{y}'_{i,n}$ from the full uncertainty distribution $N[\boldsymbol{y}_i, \delta\boldsymbol{y}_i]$. Figure 3 demonstrates that in the absence of position uncertainties the $N_{\rm samples}$ needed for the convolution integral to converge depends as

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

on the uncertainties in the (1D) velocities. We found that the required $N_{\rm samples}$ to reach a given accuracy does not depend strongly on the number of stars in the sample. But in general we expect that we need higher accuracy and therefore more $N_{\rm samples}$ for larger data sets.

[TO DO: The penultimate sentence of this section contradicts the previous sentence without validation. Why is this?] [Make more clear that we did not see a very pronounced trend but expect it in generall.]

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

2.8. *Fitting procedure*

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

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

- *Initialization.* For N_p free model parameters p_M we start with a sufficiently large grid with 3^{N_p} regular points.
- *Evaluation.* We evaluate the *pdf* at each grid-point similar to BR13 (their Figure 9): An outer loop iterates over the potential parameters p_Φ and precalculates all $N_* \times N_{\rm samples} + N_x^2 \times N_v^3$ actions required for the likelihood calculation (see Equations (8), (9) and (16)). Then an inner loop evaluates Equation (9) (or (16)) for all DF parameters $p_{\rm DF}$ in the given potential.
- *Iteration.* For each of the model parameters p_M we marginalize the *pdf*. A Gaussian is fitted to the marginalized *pdf* and the peak $\pm\ 4\sigma$ become the boundaries of the next grid with 3^{N_p} grid points. The grid might be still too coarse or badly positioned to fit Gaussians. In that case we either zoom into the grid point with the highest probability or shift the current range to find new approximate grid boundaries. We proceed with iteratively evaluating the *pdf* on finer and finer grids, until we have found a reliable 4-sigma fit range in each of the p_M dimensions. The central grid point is then very close to the best fit p_M, and the grid range is of the order of the *pdf* width.

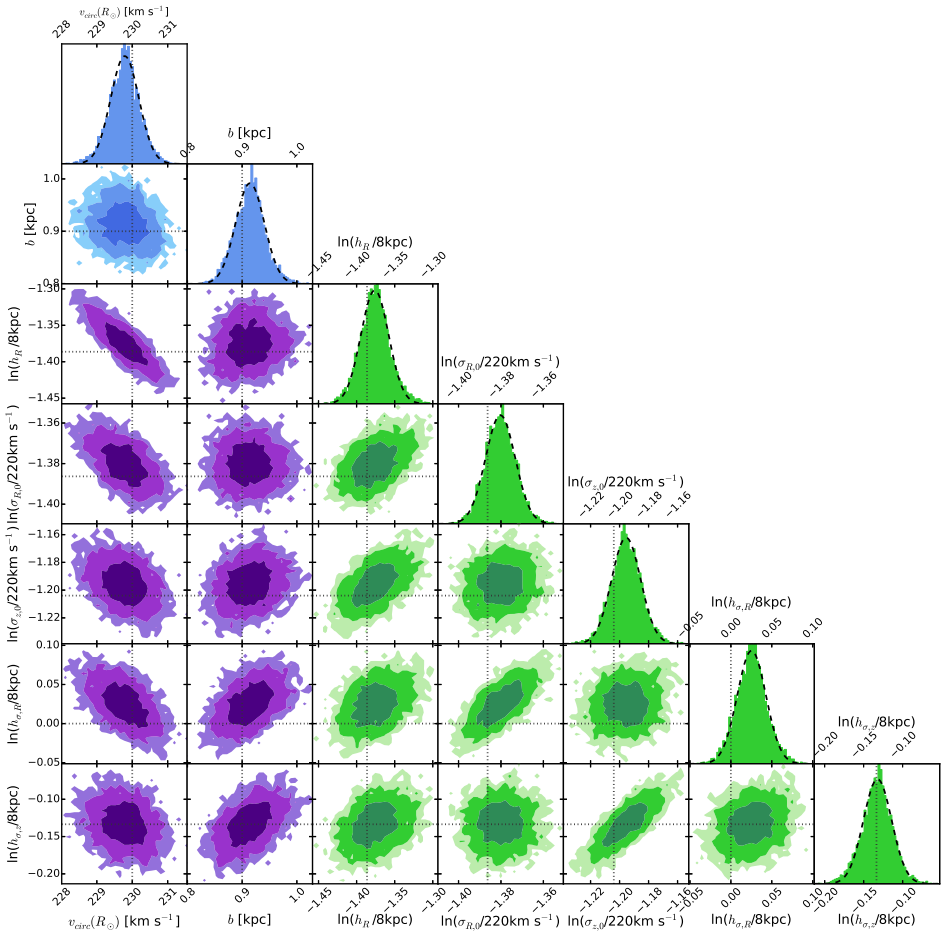

Figure 4. The *pdf* in the parameter space $p_M = \{p_\Phi, p_{\rm DF}\}$ for one example mock data set (see Test 3.1 in Table 3). Blue indicates the *pdf* for the potential parameters p_Φ, green the qDF parameters $p_{\rm DF}$. The true parameters are marked by dotted lines. The dark, medium and bright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

- *The fiducial qDF.* To save time by pre-calculating actions, they have to be independent of the choice of $p_{\rm DF}$. However, the normalisation in Equation (11) requires actions on a $N_x^2 \times N_v^3$ grid and the grid ranges in velocity space *do* depend on the current $p_{\rm DF}$ (see Equation (8)). To relax this, we follow BR13 and use a fixed set of qDF parameters (the *fiducial qDF*) to set the velocity grid boundaries in Equation (8) globally for a given p_Φ. Choosing a fiducial qDF that is very different from the true DF can however lead to large biases in the p_M recovery. BR13 did not account for that. *RoadMapping* avoids this as follows: To get successively closer to the optimal fiducial qDF—with the (yet unknown) best fit $p_{\rm DF}$—we use in each iteration step of the nested-grid search the central grid point of the current p_M grid as the fiducial qDF's $p_{\rm DF}$. As the nested-grid search approaches the best fit values, the fiducial qDF approaches its optimum as well.

- *Computational expense.* Overall the computation speed of this nested-grid approach is dominated (in descending order of importance) by a) the complexity of potential and action calculation, b) the $N_* \times N_{\rm samples} + N_x^2 \times N_v^3$ actions required to be calculated per p_Φ, c) the number of potential parameters and d) the number of DF parameters.

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

[TO DO: Here a fixed sampling is used for the error samples. I think again you should reference McMillan & Binney (2013) as they discussed the numerical stability of this method.]

3. RESULTS

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

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

[TO DO: It is stated that the breakdown of axisymmetry and steady state assumptions is not explored. I wonder as well about the impact of resonances, particularly when the data are very high quality. –¿ pragmatic way look at simulations where that happens –¿ we presume that the data deviation residuals from the best fit axisymmetric model may be a good way to investiagte resonances This cannot be explored in the current setup as the data are generated from an action-based DF but perhaps should be mentioned as a potential limitation of the approach.]

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

The MAPs in BR13 contained between 100 and 800 objects, which implied broad *pdf*s for the model parameters p_M. Several consequences arise in the limit of much larger samples, say $N_* = 20,000$: (i) As outlined in Section **??** and investigated in Figure 2 (Test 1 in Table 3), higher numerical accuracy is needed due to the likelihood normalisation requirement $\delta M_{\rm tot} \lesssim 1/N_*$ (see Equation (14)), which drives the computing time. (ii) The *pdf*s of the p_M become Gaussian, with a *pdf* width (i.e., the standard error (SE) on the parameter estimate) that scales as $1/\sqrt{N_*}$. The former is demonstrated in Figure 4 (Test 3.1 in Table 3) and we also verified that the latter is true. (iii) Any bias in the *pdf* expectation value has to be considerably less than the SE. Figure 5 (Test 3.2 in Table 3) illustrates that *RoadMapping* behaves like an unbiased maximum likelihood estimator: The average parameter estimates from many mock data sets are very close to the input p_M, and the distribution of the actual parameter estimates are Gaussian around it.

3.2. *The role of the survey volume geometry*

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

The first suite draws mock data for two different potentials (`Iso-Pot` and `MW13-Pot`) and four volume wedges (see Section **??**) with different extent and at *different positions within the Galaxy*, illustrated in the upper panel of Figure 6. Otherwise the data sets are generated from the same p_M (see Test 4 in Table 3). To isolate the role of the survey volume geometry, the mock data sets all have

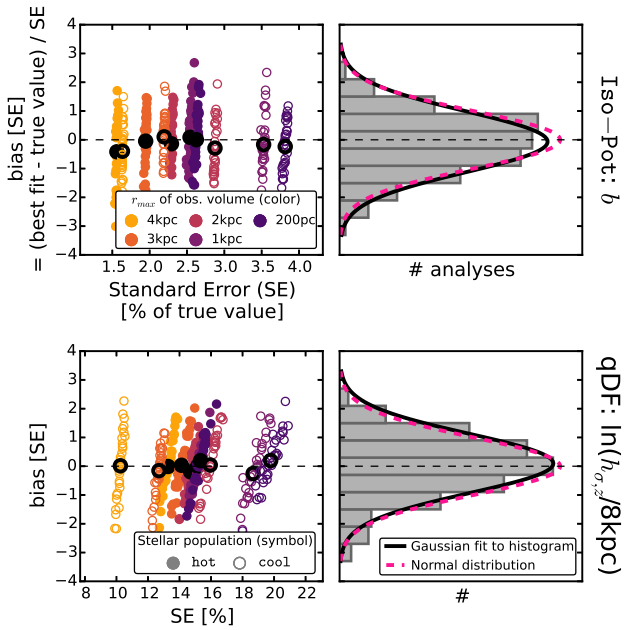

Figure 5. Lack of bias in the parameter estimates. Maximum likelihood estimators converge to the true parameter values for large numbers of data points and have a Gaussian spread—if the model assumptions are fulfilled. To test that these conditions are satisfied for *RoadMapping*, we create 320 mock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

the same number of stars ($N_* = 20,000$) in all cases, and are drawn from identical total survey volumes (4.5 kpc^3, achieved by adjusting the angular width of the wedges). The results are shown in Figure 6.

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

Figure 5 demonstrates that, given a choice of $p_{\rm DF}$, a larger volume always results in tighter constraints. There is no obvious trend that a hotter or cooler population will always give better results; it depends on the survey volume and the model parameter in question. In Figure 6 the wedges all have the same volume and all give results of similar precision. Minor differences (e.g., the `Iso-Pot` potential being less constrained in the wedge with large vertical but small radial extent) are a special property of the considered potential and parameters, and not a global property of the corresponding survey volume. In the case of an axisymmetric model galaxy, the extent in ϕ direction is not expected to matter. Overall radial extent and vertical extent seem to be equally important to constrain the potential. In addition, Figure 6 implies that volume offsets in the radial or vertical direction have

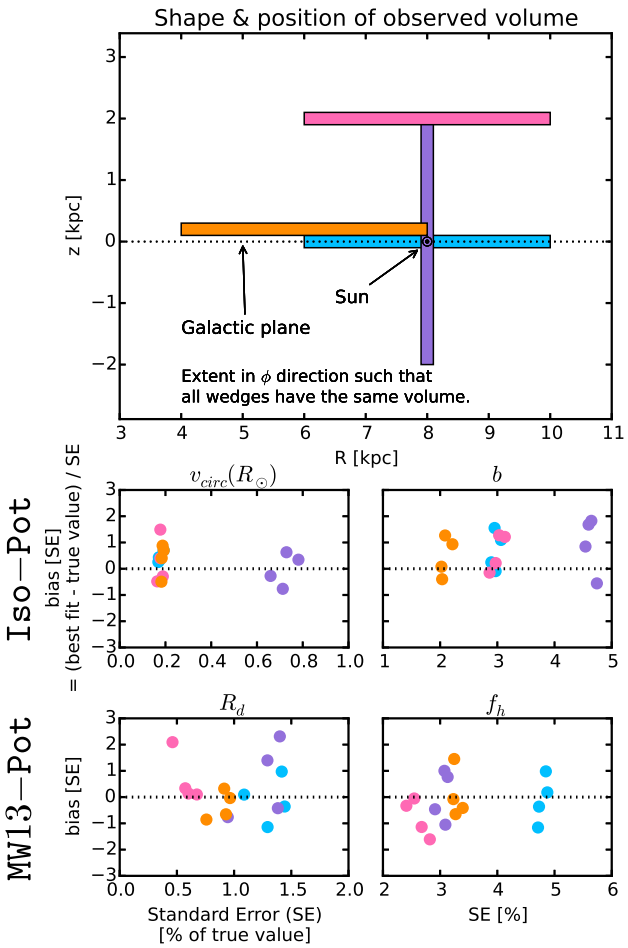

Figure 6. Bias vs. standard error in recovering the potential parameters for mock data sets drawn from four different wedge-shaped test observation volumes within the Galaxy (illustrated in the upper panel; the corresponding analyses are colour-coded) and two different potentials (`Iso-Pot` and `MW13-Pot` from Table 1; see also Test 4 in Table 3 for all model parameters used). Standard error and offset were determined from a Gaussian fit to the marginalized *pdf*. The angular extent of each wedge-shaped observation volume was adapted such that all have a volume of 4.5 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: the referee writes: I understand that the selections used in Fig 9 are illustrative but the pink selection just doesn't seem realistic. I think Fig 8. is a sufficient demonstration of the difference between different selections. Fig. 9 doesn't add anything and is barely discussed in the text. Also, without observational uncertainties (which will be greater for the more distant boxes) the discussion seems superficial. I would consider removing this.] [TO DO: I will keep this figure but explain it more in the text and what the idea behind this figure was.] [TO DO: we have chosen extreme cases to see the effects most dramatic, but we don't claim that all of them are realistic] [TO DO: To be consistent I probably have to replace the MW13 analyses with DHB analyses...]

at most a modest impact—even in case of the very large sample size at hand.

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

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

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

The survey SF (see also Appendix .2) can be very complex and is therefore sometimes not perfectly known. Here we investigate how much this could affect the recovery of the potential. We do this by creating mock data in in the `DHB-Pot` within a spherical survey volume with radius $r_{\max}$ around the Sun (see Test 5 in Table 3) and a spatially varying completeness function

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

which drops linearly with distance r from the Sun. The completeness function can be understood as the probability of a star at distance r to be detected (see also Equation 1). In the *RoadMapping* analysis on the other hand, we assume constant completeness ($\epsilon_r = 0$). The incompleteness parameter ϵ_r of the mock data quantifies therefore by how much we misjudge the SF. This mock test captures the relevant case of stars being less likely to be observed (than assumed) the further away they are (e.g., due to unknown dust obscuration).

Figure 7 demonstrates that the potential recovery with *RoadMapping* is quite robust against somewhat wrong assumptions about the completeness of the data, i.e., $\epsilon_r \lesssim 0.15$ for the `hot` and $\epsilon_r \lesssim 0.2$ for the `cool` population. The `cool` population is more robust, because it is less affected by the SF misjudgement at high $|z|$ than the `hot` population. Our simple model SF affects stars at large and small radii in equal proportion. As long as the misjudgement is small, the tracer scale length parameter h_R can still be reliable recovered, and with it the potential.

We have also investigated several test suites using the `Iso-Pot`. The recovery of $v_{\rm circ}(R_\odot)$ and the qDF parameters at different ϵ_r is qualitatively and quantitatively similar to Figure 7 for the `DHB-Pot`. The isochrone scale length b however is recovered independently of ϵ_r—probably because rotation curve measurements in the plane alone, which are not affected by the SF cuts, give reliable constraints on b. When not including tangential velocity measurements in the analysis (which is done by marginalizing the likelihood in Equation (9) over v_T), the parameters are well recovered only for $\epsilon_r \lesssim 0.15$ and $\epsilon_r \lesssim 0.2$ for the `hot` and `cool` population respectively. As this is in concordance with our findings for the `DHB-Pot`, this result seems to be valid for different choices of potentials.

For spatial completeness functions varying with the distance from the plane $|z|$ only, the `Iso-Pot` potential recovery is similarly robust as long as v_T measurements are included.

[TO DO: Hans-Walter said: "In general the SF is not so important for cool populatoins, while for hot populations getting the radial profile right matters more." - Not sure if this is right and how to fit this in the text.]

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

Measurement uncertainties in proper motions and distance dominate over uncertainties in position on the sky

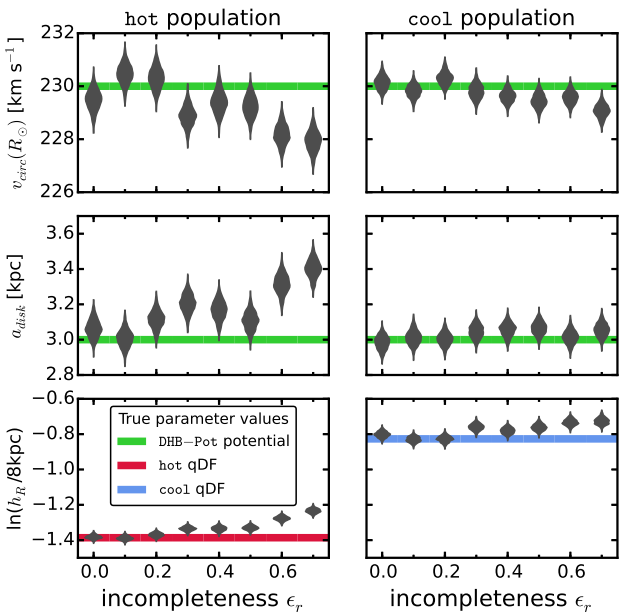

Figure 7. [This is a new version of the original plot (Figure 22) - using the DHB-Pot instead of the Iso-Pot. (TO DO: Remove this note for final version.)] Impact of misjudging the completeness of the data on the parameter recovery with *RoadMapping*. Each mock data set was created with a different incompleteness parameter ϵ_r (shown on the x-axis, see Equation (17)). (The model parameters are given as Test 5 in Table 3.) The analysis however assumed that all data sets had constant completeness within the survey volume ($\epsilon_r = 0$). The violins show the full shape of the projected *pdf*s for each model parameter, and the solid lines their true values. The *RoadMapping* method seems to be very robust against small to intermediate deviations between the true and the assumed data incompleteness. (The potential parameter $f_{\rm halo}$ and the other qDF parameters are recovered to a comparable accuracy and are therefore not shown here.)

(RA, Dec) and line-of-sight velocity, which can be more accurately determined.

The range of proper motion uncertainties we will investigate in this section, $1-5$ mas yr^{-1}, is the approximate measurement accuracy that can be achieved by combining catalogues from ground-based surveys like the Sloan Digital Sky Survey (SDSS; Abazajian et al. 2003), the USNO-B catalog (Monet et al. 2003), 2MASS (Skrutskie et al. 2006) and Pan-STARRS1 photometric catalog (PS1; Kaiser et al. 2010).[9] Space-based surveys can do even better: The Hipparcos (ESA 1997) and Tycho-2 (Høg et al. 2000) catalogues achieve $\delta\mu \sim 2.5$ mas yr^{-1} (and even $\delta\mu \lesssim 1$ mas yr^{-1} for all stars with $V < 12$), which will be soon superseded by Gaia with only $\delta\mu \sim 0.3$ at its faint end at magnitude $G \sim 20$ (de Bruijne et al. 2014).

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

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

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

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

How does this compare with the distance uncertainties expected for Gaia? For a typical red clump giant star with $M_I \sim 0$ mag and $(V-I) \sim 1$ mag at a distance of $r = 3$ kpc we estimate (using the magnitude transformation by Jordi et al. (2010) and the uncertainty parametrization by de Bruijne et al. (2014)) a parallax uncertainty of $\delta\pi \sim 11$ μas, which is consistent with a distance uncertainty of less than 5%, and a proper motion uncertainty of $\delta\mu \sim 6$ μas yr^{-1}, which is negligible. When not restricting the modelling to giant stars only, a quick investigation by Rene Andrae of stars at $r = 3$ kpc $\pm$ 5 pc from the Gaia Universe model snapshot catalogue (GUMS; Robin et al. 2012) revealed that a magnitude cut at $G \sim 15$ mag in the overall Gaia data set should keep all distance uncertainties within 3 kpc below $\sim 10\%$, while also preserving Gaia's simple SF.

We therefore found that in case we perfectly know the measurement uncertainties (and the distance uncertainty is negligible or of the order of the uncertainties expected from Gaia within $\sim$ 3kpc), the convolution of the model probability with the measurement uncertainties gives *precise and accurate* constraints on the model parameters—even if the measurement uncertainty itself is quite large.

Lastly, Figure 10 investigates the effect of a systematic *underestimation* of the true proper motion uncertainties $\delta\mu$ by 10% and 50% (see also Test 6.3 in Table 3). We find that this causes a bias in the parameter recovery that

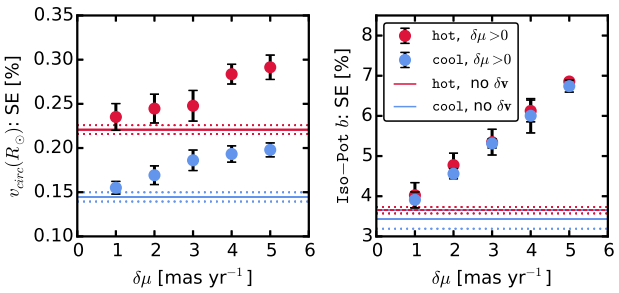

Figure 8. Effect of proper motion uncertainties $\delta\mu$ on the precision of potential parameter recovery for two stellar populations of different kinematic temperature (see Test 6.1 in Table 3 for all model parameters). The relative standard error (SE) derived from the marginalized *pdf* for each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

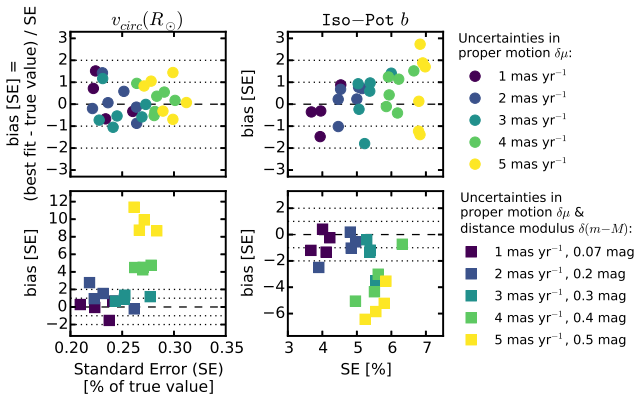

Figure 9. Potential parameter recovery using the approximation for the model probability convolved with measurement uncertainties in Equation (16). We show *pdf* offset and relative width (i.e., standard error SE) for potential parameters recovered from mock data sets (which were created according to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

grows seemingly linear with $\delta\mu$. For an underestimation of only 10% however, the bias is still $\lesssim 2\sigma$ for 10,000 stars—even for $\delta\mu \sim 3$ mas yr^{-1}.

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

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

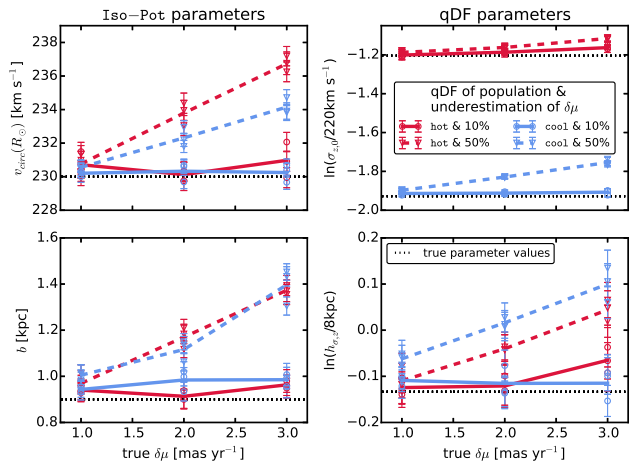

Figure 10. Effect of a systematic underestimation of proper motion uncertainties $\delta\mu$ on the recovery of the model parameters. (The true model parameters used to create the mock data are summarized as Test 6.3 in Table 3, four of them are indicated as black dotted lines in this figure.) The mock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

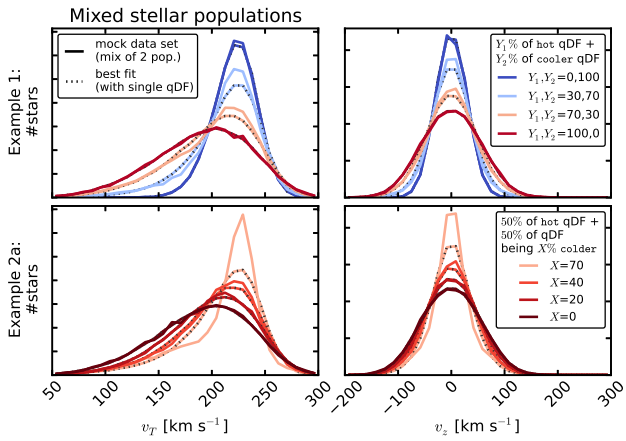

Figure 11. [This is a new version of the original plot - using the DHB-Pot instead of the Iso-Pot. (TO DO: Remove this note for final version.)] Distribution of mock data v_T and v_z created by mixing stars drawn from two different qDFs (solid lines), and the distribution predicted by the best fit of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

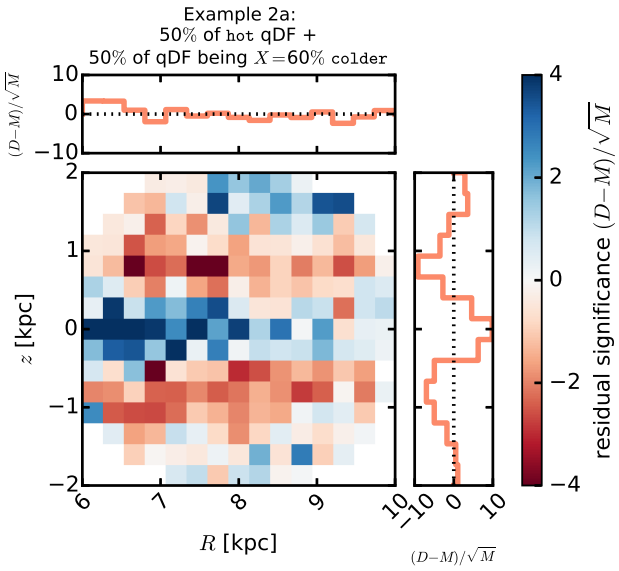

Figure 12. [This is a completely new plot. (TO DO: Remove this note for final version.)] Residual significance $(D-M)/\sqrt{M}$ of one example mock data set D and its best fit single qDF model M in the (R, z) plane. The mock data set D was created by mixing a `hot` and a X% `colder` population (see also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

Our modelling approach assumes that each stellar population follows a simple DF; here we use the qDF. In this section we explore what happens if this idealization does not hold. We investigate this issue by creating mock data sets that are drawn from *two* distinct qDFs of different temperature[10] (see Table 2 and Test 7 in Table 3)in the `DHB-Pot`, and analyse the composite mock data set by fitting a *single* qDF to it. The velocity distribution of some mock data sets and their best fit qDFs are illustrated in Figure 11, while Figure 12 shows the tracer density residuals between data and best fit in the (R, z) plane. Figures 13 and 14 compare the input and best fit parameters. In *Example 1* we choose qDFs of widely different temperature and vary their relative fraction of stars in the composite mock data set (Figure 13); in *Example 2* we always mix mock data stars from two different qDFs in equal proportion, but vary by how much the qDFs' temperatures differ (Figure 14).

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

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

We find from *Example 1* that the potential parameters

[10] Following the observational evidence, our mock data populations with cooler qDFs also have longer tracer scale lengths.

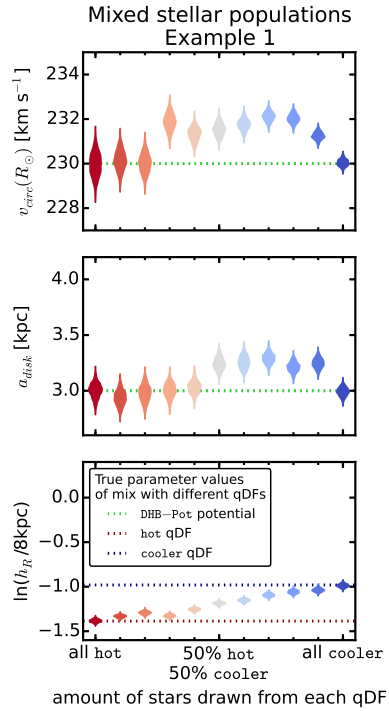

Figure 13. [This is a new version of the original plot (Figure 20)-using the DHB-Pot instead of the Iso-Pot. (TO DO: Remove this note for final version.)] The dependence of the parameter recovery on degree of pollution and temperature of the stellar population. We mix (i.e., "pollute") varying amounts of stars from a `hot` stellar population with stars from a very different `cooler` population (see Table 2), as indicated on the x-axis. (All model parameters used to create the mock data are given as Test 7, *Example 1*, in Table 3.) The composite polluted mock data set follows a true DF that has a slightly different shape than the qDF. We then analyse it using *RoadMapping* and fit a *single* qDF only. The violins represent the marginalized *pdf*s for the best fit model parameters. Some mock data sets are shown in Figure 11, first row, in the same colours as the violins here. We find that a hot population is much less affected by pollution with stars from a cooler population than vice versa.

can be more robustly recovered, if a mock data population is polluted by a modest fraction ($\lesssim 30$%) of stars drawn from a much cooler qDF, as opposed to the same pollution of stars from a hotter qDF. When considering the case of a 50/50 mix of contributions from different qDFs in *Example 2*, there is a systematic, but mostly small, bias in recovering the potential parameters, monotonically increasing with the qDF parameter difference. In particular for fractional differences in the qDF parameters of $\lesssim 20$% the systematics are insignificant even for sample sizes of $N_* = 20,000$, as used in the mock data.

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

The recovery of the effective qDF parameters, in light of non-qDF mock data, is quite intuitive (in Figures 13 and 14 we therefore show only h_R): the effective qDF temperature lies between the two temperatures from which the mixed DF of the mock data was drawn; in all cases the scale lengths of the velocity dispersion fall-off, $h_{\sigma,R}$ and $h_{\sigma,z}$, are shorter than the true scale lengths, because the stars drawn form the hotter qDF dominate at small radii, while stars from the cooler qDF (with its longer tracer scale length) dominate at large radii;

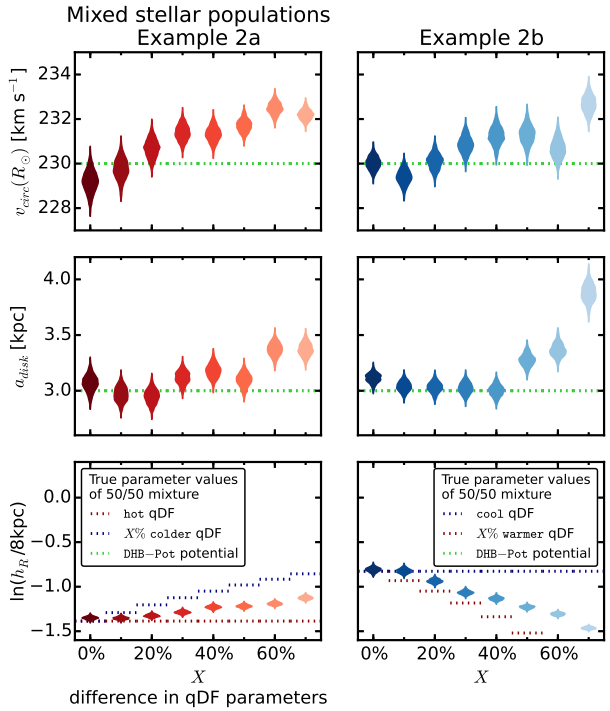

Figure 14. [This is a new version of the original plot (Figure 21) - using the DHB-Pot instead of the Iso-Pot. (TO DO: Remove this note for final version.)] The dependence of the parameter recovery on the difference in qDF parameters of a 50/50 mixture of two stellar populations and their temperature. The two qDFs from which the stars in each mock data set were drawn are indicated in the legend, with the qDF parameters $\sigma_{R,0}, \sigma_{z,0}$ and h_R differing by X% (see also Table 2 and Section 2.4), as indicated on the x-axis. (The model parameters used for the mock data creation are given as Test 7, *Example 2a & b*, in Table 3.) Each composite mock data set is fitted with a *single* qDF and the marginalized *pdf*s are shown as violins. Some mock data sets of *Example 2a* and their best fit distributions are shown in Figure 11, last row (colour-coded analogous to the violins here), and Figure 12 shows the corresponding residuals in the (R, z) plane. By mixing populations with varying difference in their qDF parameters, we model the effect of finite bin size or abundance errors when sorting stars into different MAPs in the [α/Fe]-vs.-[Fe/H] plane and assuming they follow single qDFs (cf. BR13). We find that the bin sizes should be chosen such that the difference in qDF parameters between neighbouring MAPs is less than 20%.

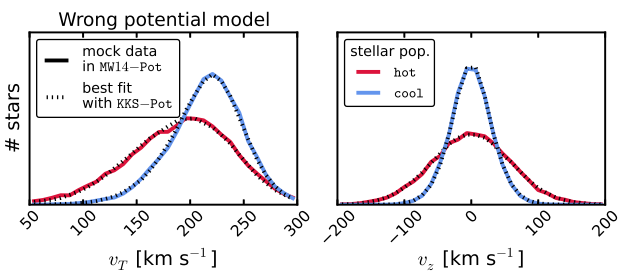

Figure 15. Comparison of the distribution of mock data v_T and v_z created in the `MW14-Pot` potential and with two different stellar populations (see Test 8 in Table 3 for all mock data model parameters), and the best fit distribution recovered by fitting the family of `KKS-Pot` potentials to the data. The best fit potentials are shown in Figure 16 and the corresponding best fit qDF parameters in Figure 17. The data is very well recovered, even though the fitted potential family did not incorporate the true potential. [TO DO: Update with new analysis]

the recovered tracer scale lengths, h_R, vary smoothly between the input values of the two qDFs that entered the mix of mock data. The latter is demonstrated in Figure 12: The radial tracer density profile of the mock data is steeper than a single qDF in the mid-plane and more shallow at higher $|z|$; overall the best fit h_R lies therefore in between.

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

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

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

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

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

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

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

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

The density contours are less tightly constrained than the forces, but we still capture the essentials. Overall the best fit disk is less dense in the mid-plane than the true disk. While it is in general possible to generate very flattened density distributions from Stäckel potentials, it might be difficult to simultaneously have a roundish halo and to require that both Stäckel components have

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

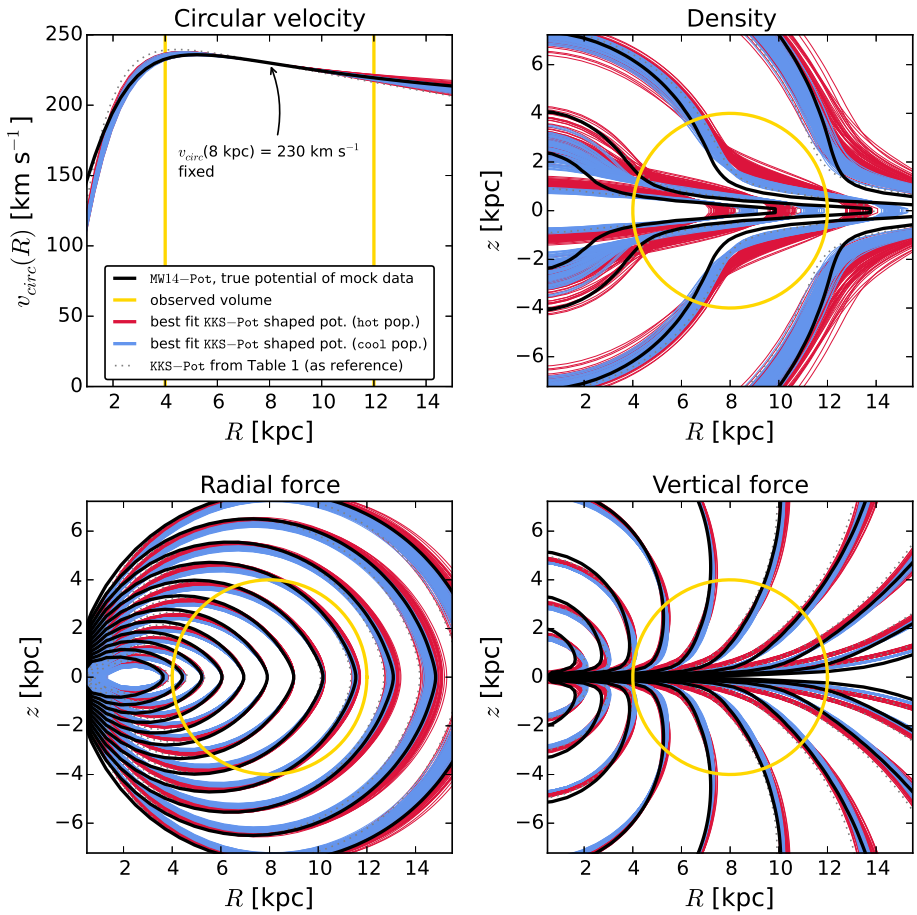

Figure 16. Recovery of the gravitational potential if the assumed potential model family (`KKS-Pot` with fixed $v_{\rm circ}(R_\odot)$) and the true potential of the (mock data) stars (`MW14-Pot` in Table 1) have slightly different parametric forms. We show the circular velocity curve, as well as contours of equal density, radial and vertical force in the R-z-plane, and compare the true potential (black lines) with 100 sample potentials (red and blue lines) drawn from the *pdf* found with MCMC for a `hot` (red) and a `cool` (blue) stellar population. (All mock data model parameters are given as Test 8 in Table 3.) Overall, the true potential is well recovered. [TO DO: The referee writes: "Fig 19 is difficult to interpret. Is it possible to display the difference?"] [TO DO: Make frame lines and ticks thick.] [TO DO: replace with analyses with Delta=0.45 and proper fiducial qDF (This has to be re-run. I do not yet have those analyses.).] [TO DO: make sure you use physical units. Done. Is already in plotting code.] [TO DO: Add in the plot titles of the contour plots $\rho(R,z)$ and $F_R(R,z)$ etc. Done. Is already in plotting code.] [TO DO: Noch zwei panels with density comparison along R and z to show the quantitative difference better. Done. Is already in plotting code.]

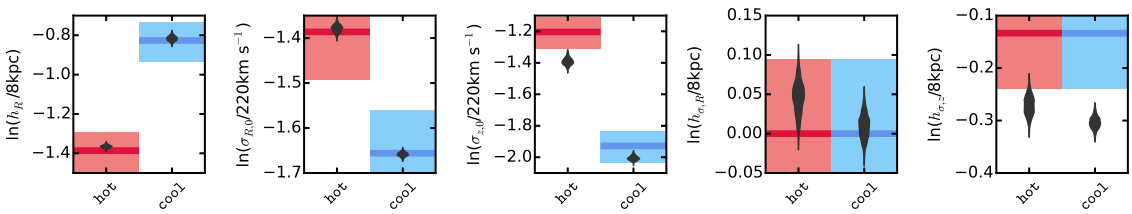

Figure 17. Recovery of the qDF parameters for the case where the true and assumed potential deviate from each other (see Test 8 in Table 3). The thick red (blue) lines represent the true qDF parameters of the `hot` (`cool`) qDF in Table 2 used to create the mock data, surrounded by a 10% error region. The grey violins are the marginalized *pdf*s for the qDF parameters found simultaneously with the potential constraints shown in Figure 16. [TO DO: I think Fig. 20 could be removed. As mentioned it doesn't make sense to compare the DF parameters between different potentials so I am not sure what Fig 20 is telling us.] [TO DO: keep this figure but make more clear why it is "important". e.g. EDF fitting as in sanders and binney in a wrong potential –¿ not so much importance to the actual best fit df parameters.] [TO DO: overplott true Gaussian velocity dispersion at sun.]

the same focal distance (see Table 1).

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

(cf. BR13) and how close they are to the actual velocity profile depends on the choice of potential; that is, the *physical* velocity dispersion is well recovered, even if the velocity-dispersion parameter is not.

[TO DO: The fact the density is not well recovered seems interesting as it points to possible biases in the surface density of the disc/dark matter measurements if one uses the wrong potential. It would be good to have the discreapncy quantized in the text.]

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

[TO DO: The referee writes: "Should this section be moved to the discussion section?"]

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

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

First, the circular velocity at the Sun, $v_{\rm circ}(R_\odot)$, is always best recovered with cooler populations (see Figures 8, 10, 13, 14 and 16), because more stars are on near-circular orbits (see Figure 18). As cooler populations probe the rotation curve better, which in turn probes the gravitational potential, the potential recovery using cool stellar populations is less sensitive to misjudgements of (spatial) selection functions (see Figures 7 and 23). There is however the caveat, that cool populations are more susceptible to non-axisymmetric streaming motions in the disk.

Second, hotter populations seem to be less sensitive to misjudgements of proper motion measurement uncertainties (see Figure 10) and pollution with stars from a cooler population (see Figures 13 and 14), because of their higher intrinsic velocity dispersion (see Figure 19).

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

4. SUMMARY AND DISCUSSION

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

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

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

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

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

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

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

[TO DO: Perhaps add statements comparing the errors explored to those anticipated from Gaia.]

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

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

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

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

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

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

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

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

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

For as many as 20,000 stars constraints become already so tight that it should presumably be possible to distinguish between different parametric MW potential models (e.g., the `MW13-Pot` used by BR13 and the `KKS-Pot`).

BR13 fitted a MW-like model potential and calculated actions using the Stäckel approximation (Binney 2012a); in this case study we directly fitted a Stäckel potential to the data, with exact actions in the model potential. The latter is computationally much less expensive due to the simple analytic form of the potential. It would also allow flexibility by expressing the MW potential as a superposition of many more simple Kuzmin-Kutuzov Stäckel components (Famaey & Dejonghe (2003) used for example 3 components). The former approach by BR13 however allows to parametrize the potential with intuitive and physically motivated building blocks (exponential disks, power-law dark matter halo etc.). While both approaches are formally similar, it remains to decide which is better.

[TO DO: The two approaches mentioned at the end of 'Gravitational potential beyond the...' are stated as formally similar but I think it is clear that one is better than the other. The true Staeckel approach limits you to potentials with the same foci. This is an obvious limitation and has been discussed before.]

[TO DO: We use the Staeckel Fudge with fixed focal length for all orbits. Or am I mistaken? The Staeckel Fudge as presented in Sanders & Binney (2015) calculates a good focus for each orbit separately. Or is the galpy code doing that as well?]

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

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

[TO DO: The referee writes: "The definition of X in f(J,[X/H]) doesn't seem to make sense."]

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

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

The vertical force profile of their best fit mass model nicely agrees with the results from BR13 for $R > 6.6$ kpc. The disadvantage of their approach is, that by binning the stars spatially, a lot of information is not used.

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

Future work: We know that real galaxies, including the MW, are not axisymmetric. Using N-body models, we will explore in a subsequent paper how the recovery of the gravitational potential with *RoadMapping* will be affected when data from a non-axisymmetric system get interpreted through axisymmetric models. In this context further investigations of questions that came up in Section 3.3 ("How much of the information on the potential is stored in the rotation curve?") and earlier in Section 4 ("What is better: fitting a MW-like potential using approximate Stäckel actions, or fitting a Stäckel potential to the MW using exact actions?") should also be conducted.

[TO DO: The first section in future work is very interesting. Use of different DFs and potentials as explored in this paper is interesting but a true test of the apparatus on a more realistic galaxy would make the 'RoadMapping' tool much more attractive.]

[TO DO: I think that the final two questions of the future work section are weak. Clearly the rotation curve is only describing the in-plane force not the force everywhere. Parametrizations will naturally convince you that the rotation curve is well measured but I think there is a lot more flexibility. Also, the advantage of using the approximate actions is that more realistic potentials can be considered.]

[TO DO: Yes, the final questions are weak. Remove them. Better questions: Stress the upcoming test with a simulation and spiral arms. also: Do we get potential good enough to calculate actions accurate enough to identify clumps? To compare it to clumps in abundance space vs. clustering of stars in action space. Also: How do results from RM (potential and DF) compare with results from Jeans models?]

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

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

[TO DO: mention how long new tests took on how many cores]

5. ACKNOWLEDGEMENTS

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

[TO DO: Use appropriate journal abbreviations in the reference list.]

REFERENCES

Abazajian, K., Adelman-McCarthy, J. K., Agüeros, M. A., et al. 2003, AJ, 126, 2081
Ahn, C. P., Alexandroff, R., Allende Prieto, C., et al. 2014, ApJS, 211, 17
Batsleer, P., & Dejonghe, H. 1994, A&A, 287, 43
Binney, J., & Tremaine, S. 2008, Galactic Dynamics: Second Edition
Binney, J. 2010, MNRAS, 401, 2318
Binney, J. 2011, Pramana, 77, 39
Binney, J., & McMillan, P. 2011, MNRAS, 413, 1889
Binney, J. 2012a, MNRAS, 426, 1324
Binney, J. 2012b, MNRAS, 426, 1328
Binney, J. 2013, NAR, 57, 29
Bovy, J., Rix, H.-W., & Hogg, D. W. 2012b, ApJ, 751, 131
Bovy, J., Rix, H.-W., Hogg, D. W., et al. 2012c, ApJ, 755, 115
Bovy, J., Rix, H.-W., Liu, C., et al. 2012d, ApJ, 753, 148
Bovy, J., & Tremaine, S. 2012, ApJ, 756, 89
Bovy, J., & Rix, H.-W. 2013, ApJ, 779, 115 (BR13)
Bovy, J. 2015, ApJS, 216, 29
Bovy, J., Rix, H.-W., Green, G. M., Schlafly, E. F., & Finkbeiner, D. P. 2016, ApJ, 818, 130
Büdenbender, A., van de Ven, G., & Watkins, L. L. 2015, MNRAS, 452, 956
Casagrande, L., Schönrich, R., Asplund, M., et al. 2011, A&A, 530, A138
de Bruijne, J. H. J., Rygl, K. L. J., & Antoja, T. 2014, EAS Publications Series, 67, 23
Dehnen, W., & Binney, J. 1998, MNRAS, 294, 429
de Lorenzi, F., Debattista, V. P., Gerhard, O., & Sambhus, N. 2007, MNRAS, 376, 71
de Zeeuw, T. 1985, MNRAS, 216, 273
ESA 1997, ESA Special Publication, 1200,
Famaey, B., & Dejonghe, H. 2003, MNRAS, 340, 752
Foreman-Mackey, D., Hogg, D. W., Lang, D., & Goodman, J. 2013, PASP, 125, 306
Garbari, S., Liu, C., Read, J. I., & Lake, G. 2012, MNRAS, 425, 1445
Gilmore, G., & Reid, N. 1983, MNRAS, 202, 1025
Gould, A., & Kollmeier, J. A. 2004, ApJS, 152, 103
Henon, M. 1959, Annales d'Astrophysique, 22, 126
Høg, E., Fabricius, C., Makarov, V. V., et al. 2000, A&A, 355, L27
Holmberg, J., Nordström, B., & Andersen, J. 2009, A&A, 501, 941

Hunt, J. A. S., & Kawata, D. 2014, MNRAS, 443, 2112
Ivezic, Z., Tyson, J. A., Abel, B., et al. 2008, arXiv:0805.2366
Ivezić, Ž., Monet, D. G., Bond, N., et al. 2008, A Giant Step: from Milli- to Micro-arcsecond Astrometry, 248, 537
Jordi, C., Gebran, M., Carrasco, J. M., et al. 2010, A&A, 523, A48
Jurić, M., Ivezić, Ž., Brooks, A., et al. 2008, ApJ, 673, 864
Kaiser, N., Burgett, W., Chambers, K., et al. 2010, Proc. SPIE, 7733, 77330E
Kuijken, K., & Gilmore, G. 1989, MNRAS, 239, 605
McMillan, P. J., & Binney, J. J. 2008, MNRAS, 390, 429
McMillan, P. J. 2012, European Physical Journal Web of Conferences, 19, 10002
McMillan, P. J., & Binney, J. 2012, MNRAS, 419, 2251
McMillan, P. J., & Binney, J. J. 2013, MNRAS, 433, 1411
Monet, D. G., Levine, S. E., Canzian, B., et al. 2003, AJ, 125, 984
Munn, J. A., Monet, D. G., Levine, S. E., et al. 2004, AJ, 127, 3034
Munn, J. A., Monet, D. G., Levine, S. E., et al. 2008, AJ, 136, 895
Nordström, B., Mayor, M., Andersen, J., et al. 2004, A&A, 418, 989
Perryman, M. A. C., de Boer, K. S., Gilmore, G., et al. 2001, A&A, 369, 339
Piffl, T., Binney, J., McMillan, P. J., et al. 2014, MNRAS, 445, 3133
Read, J. I. 2014, Journal of Physics G Nuclear Physics, 41, 063101
Reid, M. J., Menten, K. M., Zheng, X. W., et al. 2009, ApJ, 700, 137-148
Rix, H.-W., & Bovy, J. 2013, A&A Rev., 21, 61
Robin, A. C., Luri, X., Reylé, C., et al. 2012, A&A, 543, A100
Sanders, J. L., & Binney, J. 2015, MNRAS, 449, 3479
Sanders, J. L., & Binney, J. 2016, MNRAS, 457, 2107
Sesar, B., Bovy, J., Bernard, E. J., et al. 2015, ApJ, 809, 59
Skrutskie, M. F., Cutri, R. M., Stiening, R., et al. 2006, AJ, 131, 1163
Steinmetz, M., Zwitter, T., Siebert, A., et al. 2006, AJ, 132, 1645
Strigari, L. E. 2013, Phys. Rep., 531, 1
Syer, D., & Tremaine, S. 1996, MNRAS, 282, 223
Ting, Y.-S., Rix, H.-W., Bovy, J., & van de Ven, G. 2013, MNRAS, 434, 652
Yanny, B., Rockosi, C., Newberg, H. J., et al. 2009, AJ, 137, 4377
Zhang, L., Rix, H.-W., van de Ven, G., et al. 2013, ApJ, 772, 108

Table 3

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

Test		Model for Mock Data	Model in Analysis	Figures & Results
Test 1 : Numerical accuracy in calculating the likelihood normalisation	*Potential:* *DF:* *Survey volume:* *Numerical accuracy:*	`DHB-Pot` `hot` or `cool` qDF sphere around Sun, $r_{\max} = 0.2, 1, 2, 3$ or 4 kpc $N_x \in [5, 32]$, $N_v \in [4, 48]$, $n_\sigma \in [3, 7]$	-	Figure 2 Suitable accuracy for our tests: $N_x = 16, N_v = 24, n_\sigma = 5$. Higher spatial resolution is required for cooler populations.
Test 2 : Numerical convergence of convolution with measurement uncertainties	*Potential:* *DF:* *Survey Volume:* *Uncertainties:* *Numerical Accuracy:* N_*:	`Iso-Pot` `hot` qDF sphere around Sun, $r_{\max} = 3$ kpc $\delta\mathrm{RA} = \delta\mathrm{Dec} = \delta(m - M) = 0$ $\delta v_{\mathrm{los}} = 2$ km s^{-1} $\delta\mu = 2,3,4$ or 5 mas yr^{-1} 10,000	`Iso-Pot`, all parameters free qDF, all parameters free (fixed & known) (fixed & known) $N_{\mathrm{samples}} \in [25, 1200]$	Figure 3
Test 3.1 : Shape of the model parameters' *pdf* for large data sets	*Potential:* *DF:* *Survey Volume:* N_*:	`Iso-Pot` `hot` qDF sphere around Sun, $r_{\max} = 2$ kpc 20,000	`Iso-Pot`, all parameters free qDF, all parameters free (fixed & known)	Figure 4 The *pdf* becomes a multivariate Gaussian in the limit of large data. The width of the *pdf* scales as $1/\sqrt{N_*}$.
Test 3.2 : Parameter estimates are unbiased; Influence of survey volume size	*Potential:* *DF:* *Survey volume:* N_*:	`Iso-Pot` `hot` or `cool` qDF sphere around Sun, $r_{\max} = 0.2, 1, 2, 3$ or 4 kpc 20,000	`Iso-Pot`, free parameter: b qDF, free parameters: $\ln h_R, \ln \sigma_{R,0}, \ln h_{\sigma,R}$ (fixed & known)	Figure 5 *RoadMapping* behaves like an unbiased maximum likelihood estimator. [TO DO: Volume result]
Test 4 : Influence of position & shape of survey volume on parameter recovery	*Potential:* *DF:* *Survey volume:* N_*:	(i) `Iso-Pot` or (ii) `MW13-Pot` `hot` qDF 4 different wedges, see Figure 6, upper panel 20,000	(i) `Iso-Pot`, all parameters free (ii) `MW13-Pot`, R_d and f_h free (i) qDF, all parameters free (ii) qDF, only h_R, $\sigma_{z,0}$ and $h_{\sigma,R}$ free (fixed & known)	Figure 6
Test 5 : Influence of wrong assumptions about the spatial SF on parameter recovery	*Potential:* *DF:* *Survey volume:* *Completeness:* N_*:	`DHB-Pot` `hot` or `cool` qDF sphere around Sun, $r_{\max} = 3$ kpc Equation (17) with $\epsilon_r \in [0, 0.7]$ 20,000	`DHB-Pot`, free parameters: $v_{\mathrm{circ}}(R_\odot), a_{\mathrm{disk}}, f_{\mathrm{halo}}$ qDF, all parameters free (fixed & known) completeness($\boldsymbol{x}$) = 1, i.e., $\epsilon_r = 0$	Figures 7 For minor misjudgements of a radially symmetric SF (i.e., $\epsilon_r \lesssim 0.15$ for `hot` and $\epsilon_r \lesssim 0.2$ for `cool` populations) the potential recovery is still robust.
Test 6.1 : Effect of proper motion uncertainties on precision of potential recovery	*Potential:* *DF:* *Survey volume:* *Uncertainties:* N_*:	`Iso-Pot` `hot` or `cool` qDF sphere around Sun, $r_{\max} = 3$ kpc (i) $\delta\mathrm{RA} = \delta\mathrm{Dec} = \delta(m - M) = 0$, $\delta v_{\mathrm{los}} = 2$ km s^{-1}, $\delta\mu = 1, 2, 3, 4$ or 5 mas yr^{-1} (ii) no measurement uncertainties 10,000	`Iso-Pot`, all parameters free qDF, all parameters free (fixed & known) (fixed & known)	Figure 8 If $\delta\mu$ is perfectly known, the precision of the potential para- meter recovery is only a factor $\sim 1.15 - 2$ worse for $\delta\mu \sim 1 - 5$ mas yr^{-1} than for a data set without proper motion uncertainties.
Test 6.2 :	*Potential:*	`Iso-Pot`	`Iso-Pot`, all parameters free	Figure 9

Table 3 — *Continued*

Test		Model for Mock Data	Model in Analysis	Figures & Results
Testing the convolution with measurement uncertainties in Equation (16) with & without distance uncertainties	*DF:*	`hot` qDF	qDF, all parameters free	The approximate likelihood in Equation (16) is the true likelihood in the absence of position errors. For distance uncertainties $\lesssim 10\%$ at $r_{\max} \sim 3$ kpc, which is the case for Gaia and $G \lesssim 15$, the introduced bias is still less than 2σ.
	Survey Volume:	sphere around Sun, $r_{\max} = 3$ kpc	(fixed & known)	
	Uncertainties:	$\delta\mathrm{RA} = \delta\mathrm{Dec} = 0$, $\delta v_{\mathrm{los}} = 2$ km s^{-1}, $\delta\mu = 1,2,3,4$ or 5 mas yr^{-1}, (i) $\delta(m-M) = 0$ or (ii) $\delta(m-M) \neq 0$ (see Figure 9)	(fixed & known)	
	N_*:	10,000		
Test 6.3 : Underestimation of proper motion uncertainties	*Potential:*	`Iso-Pot`	`Iso-Pot`, all parameters free	Figure 10
	DF:	`hot` or `cool` qDF	qDF, all parameters free	The bias introduced by underestimating $\delta\mu$ grows with $\delta\mu$. If $\delta\mu \lesssim 3$ mas yr^{-1} and $\delta\mu$ is underestimated by 10% the bias is only $\sim 2\sigma$ or less.
	Survey volume:	sphere around Sun, $r_{\max} = 3$ kpc	(fixed & known)	
	Uncertainties:	only proper motion uncertainties 1, 2 or 3 mas yr^{-1}	proper motion uncertainties 10% or 50% underestimated	
	N_*:	10,000		
Test 7 : Deviations of the assumed DF from the stars' true DF	*Potential:*	`DHB-Pot`	`DHB-Pot`, free parameters: $v_{\mathrm{circ}}(R_\odot), a_{\mathrm{disk}}, f_{\mathrm{halo}}$	Figures 11, 12, 13 & 14
	DF:	mix of two qDFs... (i) *Example 1:* ... with different mixing rates and fixed qDF parameters (`hot` & `cooler` qDF from Table 2) (ii) *Example 2:* ... with 50/50 mixing rate and varying qDF parameters (by X%): *a)* `hot` & `colder` qDF or *b)* `cool` & `warmer` qDF (see Table 2)	single qDF, all parameters free	*Example 1:* Hot qDF-like populations polluted by up to $\sim 30\%$ of stars from a much cooler population give still reliable potential constraints. This is not true for the opposite case. *Example 2:* Differences of $\lesssim 20\%$ in qDF parameters do not matter when mixing sub-populations, e.g., due to finite binning in abundance space and abundance errors.
	Survey volume:	sphere around Sun, $r_{\max} = 2$ kpc	(fixed & known)	
	N_*:	20,000		
Test 8 : Deviations of the assumed potential model from the stars' true potential	*Potential:*	`MW14-Pot`	`KKS-Pot`, all parameters free, only $v_{\mathrm{circ}}(R_\odot) = 230$ km s^{-1} fixed	Figures 15, 16 & 17
	DF:	`hot` or `cool` qDF	qDF, all parameters free	
	Survey volume:	sphere around Sun, $r_{\max} = 4$ kpc	(fixed & known)	
	N_*:	20,000		
	Numerical accuracy:		$N_x = 20, N_v = 28, n_\sigma = 5.5$	

APPENDIX

Mock data

The mock data in this work is generated according to the following procedure:

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

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

In the second step we draw velocities $\boldsymbol{v}_i$. 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 $\mathrm{DF}(\boldsymbol{J}[R_i, z_i, \boldsymbol{v} \mid p_\Phi] \mid p_{\rm DF})$ using a rejection method. We now have a mock data set satisfying $(\boldsymbol{x}_i, \boldsymbol{v}_i) \longrightarrow p(\boldsymbol{x}, \boldsymbol{v}) \propto \mathrm{DF}(\boldsymbol{J}[\boldsymbol{x}, \boldsymbol{v} \mid p_\Phi] \mid p_{\rm DF}) \times \mathrm{SF}(\boldsymbol{x})$.

[TO DO: The referee wrote: "The discussion of selection on very errorneous x coordinates is interesting but surely this isn't the way the data will actually be handled?" - What does he mean? What's wrong? –¿ we are not 100% sure that we understand the referee's main point but given 10data can get in practice it's likeliy to be a magnitude cut but for intrisnic standard candles like red clump stars a magnitude cut translates to a distance cut we realise it's idealized and there won't be a sharp distance cut scatter in and out of survey volume will play a role and we explore an ideal case]

Measurement uncertainties can be added to the mock data by applying the following modifications to the above procedure. We assume Gaussian uncertainties in the heliocentric phase-space coordinates $\tilde{\boldsymbol{x}} = (\mathrm{RA}, \mathrm{Dec}, (m-M)), \tilde{\boldsymbol{v}} = (\mu_{\rm RA} \cdot \cos \mathrm{Dec}, \mu_{\rm Dec}, v_{\rm los})$ (see Section 2.1). In the case of distance and position uncertainties stars virtually scatter in and out of the observed volume. To account for this, we draw the *true* $\boldsymbol{x}_i$ from a volume that is larger than the actual observation volume, perturb the $\boldsymbol{x}_i$ 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* $\boldsymbol{v}_i$ (given the *true* $\boldsymbol{x}_i$) as described above and perturb them according to the velocity uncertainties.

We show examples of mock data sets (without measurement uncertainties) in action space (Figure 18) and configuration space $(\boldsymbol{x}, \boldsymbol{v})$ (Figure 19). The mock data generated from the qDF follow the expected distributions in configuration space. The distribution in action space illustrates the intuitive physical meaning of actions: The stars of the `cool` population have in general lower radial and vertical actions, as they are on more circular orbits. Circular orbits with $J_R = 0$ and $J_z = 0$ can only be observed in the Galactic mid-plane. The different ranges of angular momentum L_z in the two example observation volumes reflect $L_z \sim R \times v_{\rm circ}$ 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 18.

Selection functions

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

$$\mathrm{SF}(\boldsymbol{x}) \equiv \begin{cases} \mathrm{completeness}(\boldsymbol{x}) & \text{if } |\boldsymbol{x} - \boldsymbol{x}_\odot| \leq r_{\rm max}, \\ 0 & \text{otherwise}, \end{cases} \tag{1}$$

We set $0 \leq \mathrm{completeness}(\boldsymbol{x}) \leq 1$ everywhere inside the observed volume, so it can be understood as a position-dependent detection probability for a star at $\boldsymbol{x}$. Unless explicitly stated otherwise, we simplify to $\mathrm{completeness}(\boldsymbol{x}) =$

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

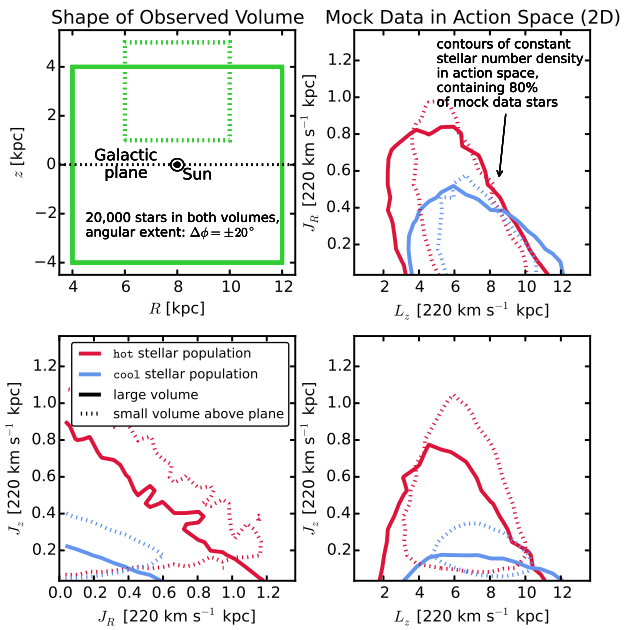

Figure 18. Distribution of mock data in action space (2D iso-density contours, enclosing 80% of the stars), depending on shape and position of a wedge-like survey observation volume (upper left panel) and temperature of the stellar population (indicated in the legend). The four mock data sets are generated in the `KKS-Pot` from Table 1 from either the `hot` or `cool` DF in Table 2. The distribution in action space visualizes how orbits with different actions reach into different regions within the Galaxy. The corresponding mock data in configuration space is shown in Figure 19.

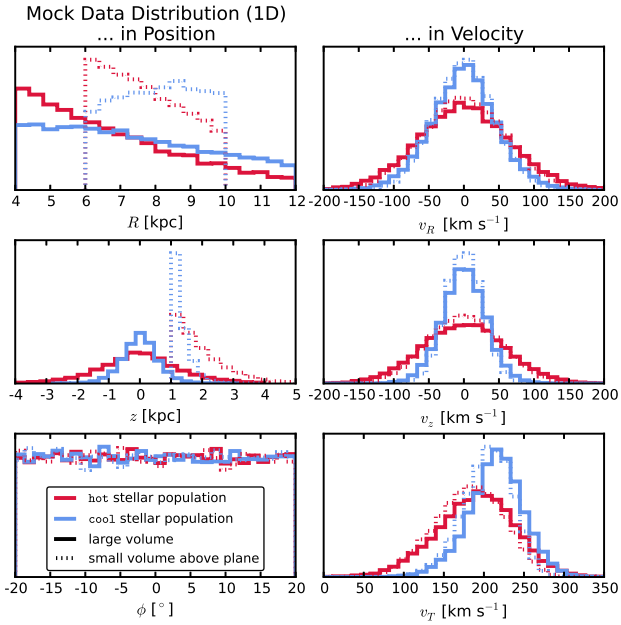

Figure 19. Distribution of the mock data from Figure 18 in configuration space. The corresponding observation volumes (as indicated in the legend) are shown in Figure 18, upper left panel. The 1D histograms illustrate that qDFs generate realistic stellar distributions in Galactocentric coordinates $(R, z, \phi, v_R, v_z, vT)$: More stars are found at smaller R and $|z|$, and are distributed uniformly in ϕ according to our assumption of axisymmetry. The distribution in radial and vertical velocities, v_R and v_z, is approximately Gaussian with the (total projected) velocity dispersion being of the order of $\sim \sigma_{R,0}$ and $\sim \sigma_{z,0}$ (see Table 2). The distribution of tangential velocities v_T is skewed because of asymmetric drift.

1. Additionally, we use in Figures 6 (Test 4), 18 and 19 for illustrative purposes some rather unrealistic survey volumes which are angular segments of a cylindrical annulus (wedge), i.e., the volume with $R \in [R_{\min}, R_{\max}], \phi \in [\phi_{\min}, \phi_{\max}], z \in [z_{\min}, z_{\max}]$ within the model Galaxy.

[TO DO: Make sure numbering in Table 3 is correct.] [TO DO: Remove latex comments] [TO DO: Make sure SF is still somewhere properly introduced.] [TO DO: How to have subchapters with numbers in the appendix? Otherwise referencing doesn't work.] [TO DO: Reference on Staeckel pancakes] [TO DO: Equation numbering in Appendix starts again from (1) - which is stupid.] [TO DO: Write in discussion that hopefully isochrone potential in error section does not affect result, as qualitatively and quantitatively we got very similar results independent of potential in incompleteness and mix section.]

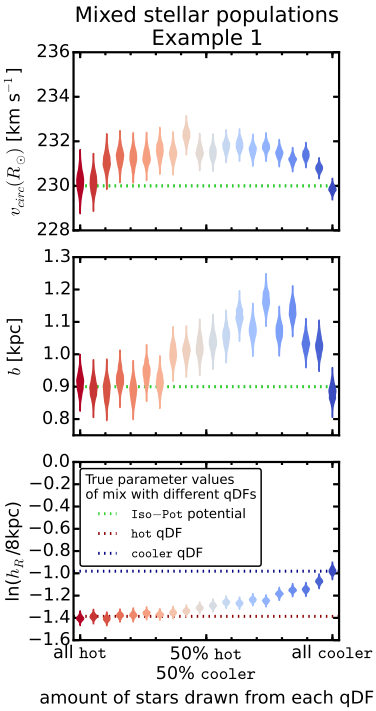

Figure 20. [TO DO: This is an old plot, with analyses using the isochrone potential. Delete it later. The new plot in shown in Figure 13.]

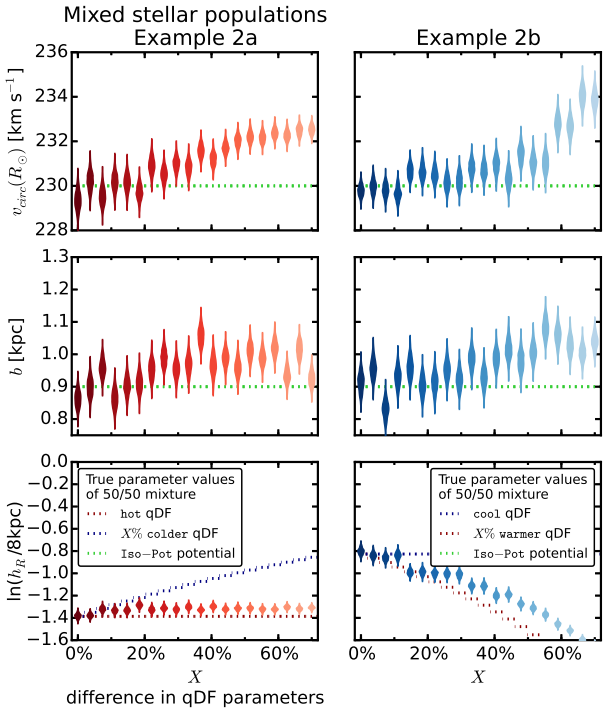

Figure 21. [TO DO: This is an old plot, with analyses using the isochrone potential. Delete it later. The new plot in shown in Figure 14.]

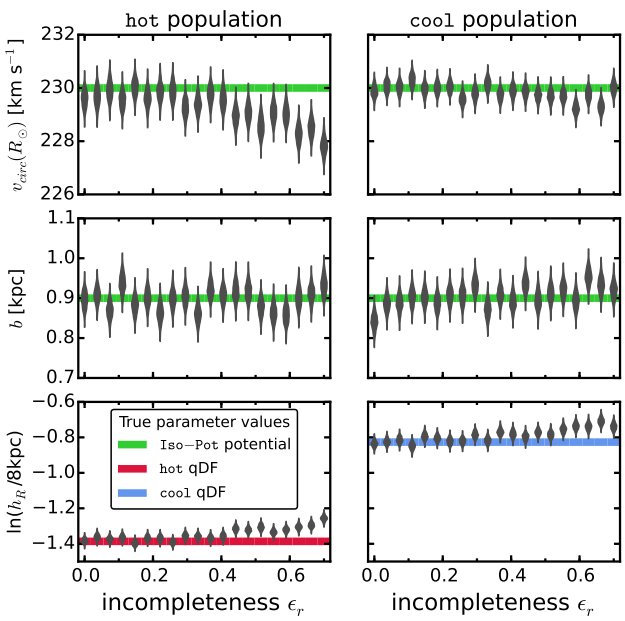

Figure 22. [TO DO: This is an old plot, with analyses using the isochrone potential. Delete it later. The new version of this plot is Figure 7.]

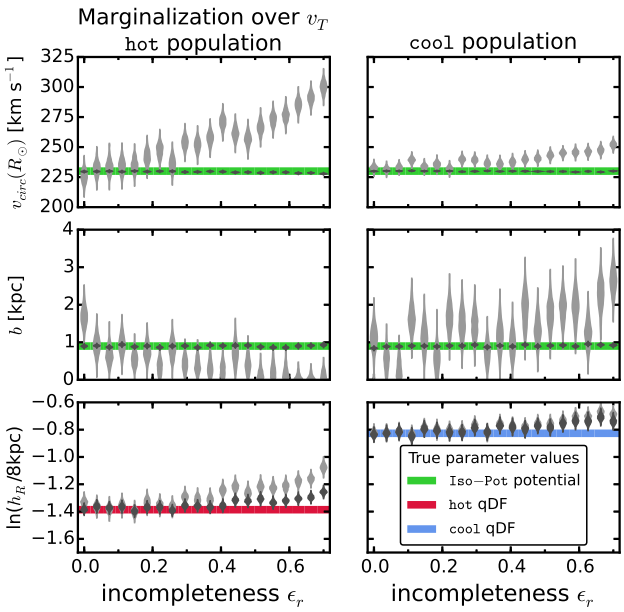

Figure 23. [TO DO: This is an old plot, with analyses using the isochrone potential. I do not need it in the new version of the paper. Delete it later.]