Draft version September 30, 2015
Preprint typeset using LaTeX style emulateapj v. 12/16/11

ACTION-BASED DYNAMICAL MODELLING OF THE MILKY WAY DISK AND OUR IMPERFECT KNOWLEDGE OF THE "REAL WORLD"

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

Draft version September 30, 2015

ABSTRACT

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

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

1. INTRODUCTION

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

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

There is a variety of practical approaches to dynamical modelling of discrete collisionless tracers, such as the stars in the Milky Way, e.g., Jeans modelling (Kuijken & Gilmore 1989, Bovy & Tremaine 2012, Garbari et al. 2012, Zhang et al. 2013, 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), 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 distribution function.

Recently, Binney (2012) and Bovy & Rix (2013) [TO DO: are these the correct references???] proposed to constrain the MWs gravitational potential by combining parametrized axisymmetric potentials models with DFs that are simple analytic functions of the three orbital actions (Binney & Tremaine 2008, §3.5 & §4.6, Binney 2011) to model discrete data.

Bovy & Rix (2013) (BR13 hereafter) put this in prac-

[1] Max-Planck-Institut für Astronomie, Königstuhl 17, D-69117 Heidelberg, Germany
[2] Correspondence should be addressed to trick@mpia.de.
[3] University of Toronto [TO DO: What is Jo's current address???]

tice 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 allow simple DFs (Bovy et al. 2012b,c,d). Given an assumed axisymmetric model for the Galaxy 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 among SEGUE G-dwarf stars (Yanny et al. 2009). They also accounted for the complex, but known selection function [TO DO: Reference to SEGUE SF???] of the kinematic tracers. For each MAP the modelling resulted in an independent estimate on the same gravitational potential. Taken as an ensemble, they constrained the disk surface mass density over a wide range of radii ($\sim 4-9$ kpc), and proved to be a powerful constraint on the disk mass scale length and on the disk-to-dark-matter ratio at the Solar radius.

BR13 made however a number of quite severe and idealizing assumptions about potential, DF and the knowledge of observational effects. These idealizations are likely to 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, explicitly developed to deal with large data sets. 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 huge number of stars: (i) We have to make sure that *RoadMapping* is an unbiased estimator (Section 3.1). (ii) Numerical inaccuracies must not be an important source of systematics (Section 2.6). (iii) As parameter estimates become much more precise (Section 3.1, we need more flexibility in the potential and DF model and effective strategies to find the best fit parameters. The improvements made in *RoadMapping* as compared to the machinery used in BR13 are presented in Section 2.7.

We also explore how different aspects of the observational experiment design impact the parameter recovery. (i) It might be worth to explore the importance of the survey volume geometry, size, shape and position within the MW to constrain the potential (Section 3.2). (ii) What if our knowledge of the sample selection function is imperfect, and potentially biased (Section 3.3)? (iii) How to best account for individual and possibly misjudged measurement uncertainties (Section 3.4)?

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

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

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

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[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 $(m-M)$ as proxy for the distance from the Sun, and velocity $\tilde{\boldsymbol{v}} \equiv (\mu_{\mathrm{RA}}, \mu_{\mathrm{DEC}}, v_{\mathrm{los}})$ as proper motions $\boldsymbol{\mu} = (\mu_{\mathrm{RA}}, \mu_{\mathrm{DEC}})$ [TO DO: cos somwhere???] in both RA and DEC direction and line-of-sight velocity v_{los}, the data $(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{v}})$ has to be converted first into the Galactocentric rest-frame coordinates $(\boldsymbol{x}, \boldsymbol{v})$ using the Sun's position and velocity. 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 and Potential Models*

Orbits in axisymmetric potentials are best described and fully specified by the three actions $\boldsymbol{J} \equiv (J_R, J_z, J_\phi = L_z)$ (Binney & Tremaine 2008, §3.5). Their computation from a star's phase-space coordinates, $(\boldsymbol{x}, \boldsymbol{v}) \longrightarrow \boldsymbol{J}$, is typically very expensive and depends on the choice of gravitational potential in which the star moves. The spherical isochrone potential (Henon 1959) and axisymmetric Stäckel potential [TO DO: REF] are the most general Galactic potentials, that allow exact action calculations (Binney & Tremaine 2008, §3.5.2 and [TO DO: REF]). In all other potentials actions have to be numerically estimated, e.g., by using the *Stäckel fudge* by Binney (2012) for axisymmetric potentials and action interpolation grids (Bovy 2015) to speed up the calculation. The latter is one of the improvements employed by *RoadMapping*, which was not used by BR13.

For the gravitational potential in our modelling we assume a family of parametrized potential models. We use: The Milky Way-like potential from BR13 (`MW13-Pot`) with bulge, disk and halo; the spherical isochrone potential (`Iso-Pot`); and the 2-component Kuzmin-Kutuzov Stäckel potential (Batsleer & Dejonghe 1994; `KKS-Pot`), which also displays a disk and halo structure. Table 1 summarizes all reference potentials used in this work

[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/`.

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

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

[(a)] The isochrone potential `Iso-Pot` has one free parameter, the scale length b.
[(b)] The coordinate system of each of the two Stäckel-potential components of the `KKS-Pot` is $\frac{R^2}{\tau_{i,p}+\alpha_p} + \frac{z^2}{\tau_{i,p}+\gamma_p} = 1$ with $p \in \{\text{Disk}, /\text{Halo}\}$ and $\tau_{i,p} \in \{\lambda_p, \nu_p\}$. Both components have the same focal distance $\Delta \equiv \sqrt{\gamma_p - \alpha_p}$, to make sure that the superposition of the two components itself is still a Stäckel potential. The axis ratio of the coordinate surfaces $\left(\frac{a}{c}\right)_p := \sqrt{\frac{\alpha_p}{\gamma_p}}$ describes the flatness of the corresponding Stäckel component. The parameter k describes the relative contribution of the disk mass to the total mass.
[(c)] The free parameters of the `MW13-Pot` are stellar disk scale length R_d and height z_d, as well as the relative halo contribution to $v^2_{\rm circ}(R_\odot)$, f_h, and the slope of the rotation curve, $\frac{{\rm d}\ln(v_{\rm circ}(R_\odot))}{{\rm d}\ln(R)}$.
[(d)] The `MWPotential2014` by Bovy (2015) (see their Table 1) has a circular velocity at the Sun of $v_{\rm circ}(R_\odot) = 220$ km s^{-1}. In this work we use however $v_{\rm circ}(R_\odot) = 230$ km s^{-1} for all potentials.
[(e)] We use a finely spaced action interpolation grid with $R_{\rm max} = 10$ [TO DO: What's that??? units???] and 50 grid points in E and ψ [TO DO: Find out what's that???], and 60 grid points in L_z.

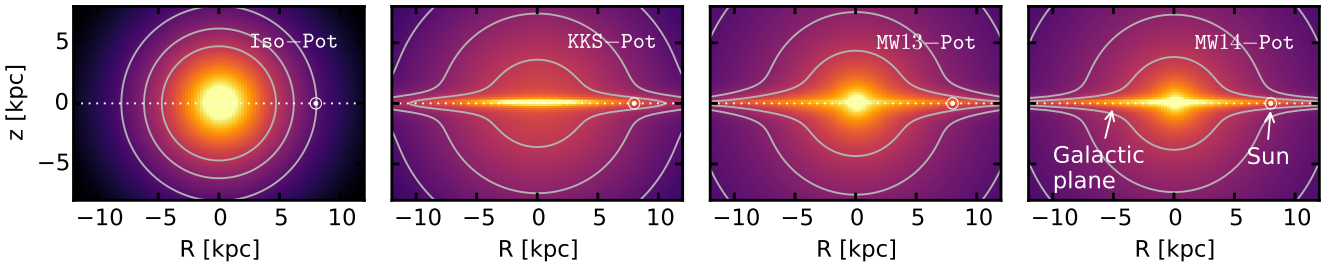


Figure 1. Density distribution of the four reference galaxy potentials in Table 1, for illustration purposes. These potentials are used throughout this work for mock data creation and potential recovery.

together with their free parameters p_Φ. The density distribution of these potentials is illustrated in Figure 1.

2.3. *Stellar Distribution Functions*

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

$$\begin{aligned} &\text{qDF}(\boldsymbol{J} \mid p_{\rm DF}) \\ &= f_{\sigma_R}(J_R, L_z \mid p_{\rm DF}) \times f_{\sigma_z}(J_z, L_z \mid p_{\rm DF}) \end{aligned} \qquad (1)$$

Table 2
Reference distribution-function parameters for the qDF in Equations (1)-(6). These DFs describe the phase-space distribution of stellar populations for which mock data is created and analysed throughout this work for testing purposes. The parameters of the `cooler` & `colder` (`warmer`) qDFs were chosen to have the same $\sigma_{R,0}/\sigma_{z,0}$ ratio as the `hot` (`cool`) qDF. The `colder` and `warmer` qDF have a free parameter X that governs how much colder/warmer they are then the reference `hot` and `cool` qDFs. Hotter populations have shorter tracer scale lengths (Bovy et al. 2012d) and the velocity dispersion scale lengths were fixed according to Bovy et al. (2012c).

name	qDF parameters $p_{\rm DF}$				
	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`	$2+50\%$	$55-50\%$	$66-50\%$	8	7
`colder`	$2+X\%$	$55-X\%$	$66-X\%$	8	7
`warmer`	$3.5-X\%$	$42+X\%$	$32+X\%$	8	7

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

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

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

Here $R_g \equiv R_g(L_z)$ and $\Omega \equiv \Omega(L_z)$ are the (guiding-center) radius and the circular frequency of the circular orbit with angular momentum L_z in a given potential. The radial/epicycle ($\kappa \equiv \kappa(L_z)$) and vertical ($\nu \equiv \nu(L_z)$) frequencies describe how a star would oscillate around the circular orbit when slightly disturbed [TO DO: Ask Jo, how to say this shorter and if definition is correct] (Binney & Tremaine 2008, §3.2.3). The term $[1+\tanh(L_z/L_0)]$ suppresses counter-rotation for orbits in the disk with $L \gg L_0$ (with $L_0 = 10 \times R_\odot/8 \times v_{\rm circ}(R_\odot)/220$ [TO DO: Jo said, galpy default is 10 km/s kpc. But I got the value actually from the code...]). Again following BR13, we choose the functional forms

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

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

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

which indirectly set the stellar number density and radial and vertical velocity dispersion profiles. The qDF has therefore a set of five free parameters $p_{\rm DF}$: 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 in Table 2: Most tests use the `hot` and `cool` qDFs, which correspond to kinematically hot and cool populations, respectively.

One crucial point in our dynamical modelling technique (Section 2.6), as well as in creating mock data (Section2.5), 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,

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

$$\approx \int_{-n_\sigma\sigma_R(R|p_{\rm DF})}^{n_\sigma\sigma_R(R|p_{\rm DF})} \int_{-n_\sigma\sigma_z(R|p_{\rm DF})}^{n_\sigma\sigma_z(R|p_{\rm DF})} \int_0^{1.5v_{\rm circ}(R_\odot)} {\rm qDF}(J[R, z, \boldsymbol{v} \mid p_\Phi] \mid p_{\rm DF})\, {\rm d}v_T\, {\rm d}v_z\, {\rm d}v_R, \quad (8)$$

where $\sigma_R(R \mid p_{\rm DF})$ and $\sigma_z(R \mid p_{\rm DF})$ are given by Equations 5 and 6.[5] Each integral is evaluated using a N_v-th order Gauss-Legendre quadrature. For a given p_Φ and $p_{\rm DF}$ we explicitly calculate the density on $N_x \times N_x$ regular grid points in the (R, z) plane and interpolate in between using bivariate spline interpolation. The grid is chosen to cover the extent of the observations (for $|z| \geq 0$, 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 speed limiting factors. To complement the work by BR13, we will specifically work out in Section 2.6 and Figure 3 how large N_x, N_v and n_σ have to be chosen to get the density with a sufficiently high numerical accuracy [TO DO: Ask Jo, if he really think that this is difficult to understand here (because we have not yet talked about the normalization)].

2.4. *Selection Functions*

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

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

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

The SF of the SEGUE survey [TO DO: Ask Jo for reference to SEGUE SF] used by BR13 consists of many pencil-beams. In anticipation of [TO DO: What is the word for surveys covering large volumes????] like Gaia, we use SFs that span large observed volumes of simple geometrical shapes: a sphere of radius $r_{\rm max}$ with the Sun at its center; or an angular segment of an cylindrical annulus (wedge), i.e. the volume with $R \in [R_{\rm min}, R_{\rm max}], \phi \in [\phi_{\rm min}, \phi_{\rm max}], z \in [z_{\rm min}, z_{\rm max}]$ within the model Galaxy. The sharp outer edge of the survey volume could be interpreted as a detection limit in apparent brightness in the case where all stars have the same luminosity. Here $0 \leq \text{completeness}(\boldsymbol{x}) \leq 1$ everywhere inside the observed volume, so it can be understood as a position-dependent detection probability. Unless explicitly stated otherwise, we simplify to $\text{completeness}(\boldsymbol{x}) = 1$.

2.5. *Mock Data*

We will rely on mock data as input to explore the limitations of the modelling. We assume that the positions and velocities of our stellar mock sample are indeed drawn from our assumed family of potentials and DFs (with given parameters p_Φ and $p_{\rm DF}$). The DF is in terms of actions, while the transformation $(\boldsymbol{x}_i, \boldsymbol{v}_i) \longrightarrow \boldsymbol{J}_i$ is computationally much less expensive than the inversion. We therefore employ the following effective two-step method for creating mock data, which also accounts for a survey SF.

In the first step we draw stellar positions $\boldsymbol{x}_i$. 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 2.3.[6] 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 sample to follow $\rho(R, |z| \mid p_\Phi, p_{\rm DF})$. To apply a non-uniform $\text{SF}(\boldsymbol{x})$, we use the rejection method a second time. The resulting set of positions $\boldsymbol{x}_i$ follows the distribution$p(\boldsymbol{x}) \propto \rho_{\rm DF}(R, |z| \mid p_\Phi, p_{\rm DF}) \times \text{SF}(\boldsymbol{x})$.

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

Figure 2 shows examples of mock data sets in configuration space $(\boldsymbol{x}, \boldsymbol{v})$ and action space. The mock data from the qDF lead to the expected distributions in configuration space. The distribution in action space illustrates the intuitive physical meaning of actions: The stars of the `cool` population have in general lower radial and vertical actions, as they are on more circular orbits. Circular orbits with $J_R = 0$ and $J_z = 0$ can only be observed in the Galactic mid-plane. The different ranges of angular momentum L_z in the two example observation volumes reflect $L_z \sim R \times v_{\rm circ}$ 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 2.

Measurement uncertainties can be added to the mock data by applying the following modifications to the above procedure. We assume Gaussian errors in the heliocentric phase-space coordinates $\tilde{\boldsymbol{x}} = (\text{RA}, \text{DEC}, (m - M)), \tilde{\boldsymbol{v}} = (\mu_{\rm RA}, \mu_{\rm DEC}, v_{\rm los})$ (see Section 2.1), where we have taken $(m - M)$ as a proxy for distance. In the case of distance uncertainties $\delta(m - M)$, stars virtually scatter in and out of the observed volume. To account for this, we draw the *true* $\boldsymbol{x}_i$ 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.

2.6. *Data Likelihood*

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

$$D = \{\boldsymbol{x}_i, \boldsymbol{v}_i \mid \begin{array}{l}(\text{star } i \text{ in given stellar population}) \\ \wedge\, (\text{SF}(\boldsymbol{x_i}) > 0)\}.\end{array}$$

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

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

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

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

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

where N_* is the number of stars in D, and in the last step we used Equation 8. $\prod_i \text{SF}(\boldsymbol{x}_i)$ is independent of p_M, so we treat it as unimportant proportionality factor. We find the best fitting p_M by maximizing the posterior probability distribution $pdf(p_M \mid D)$, which is, according to Bayes' theorem, proportional to the likelihood $\mathscr{L}(D \mid p_M)$ times a prior $p(p_M)$. We assume flat priors in both p_Φ and

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

(see Section 2.3) throughout this work. Then *pdf* and likelihood can be used interchangeably.

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

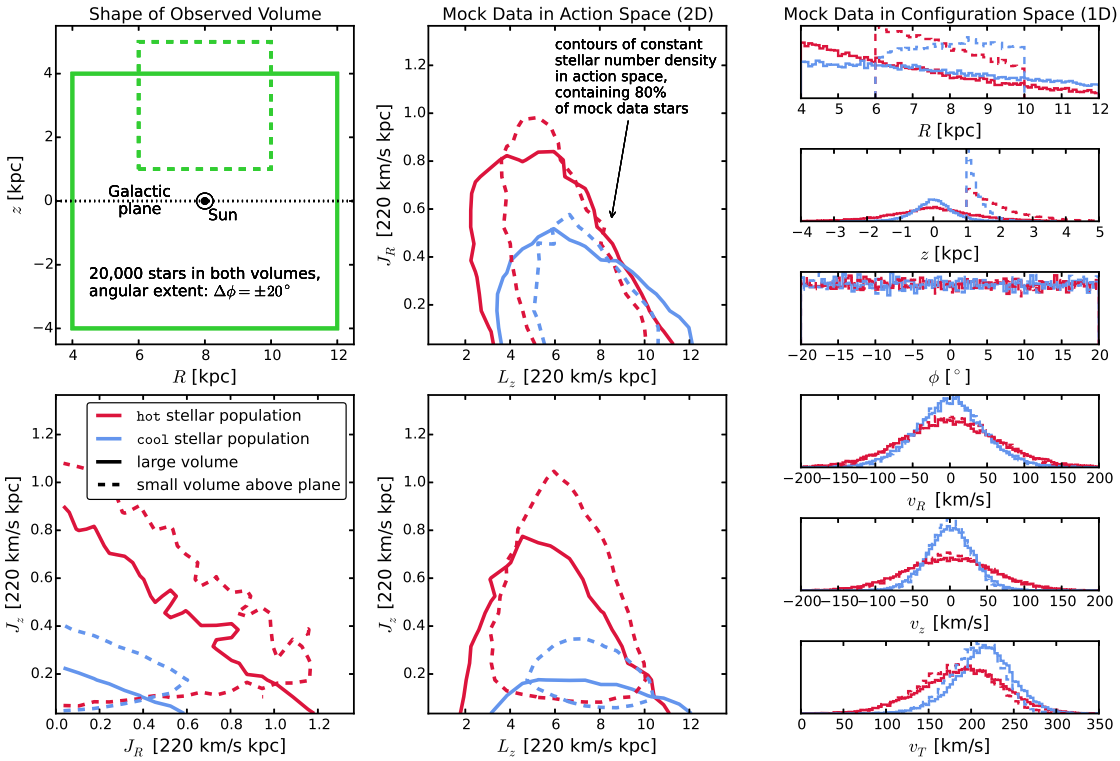


Figure 2. Distribution of mock data in action space (2D iso-density contours, enclosing 80% of the stars, the two central and the lower left panels) and configuration space (1D histograms, right panels), depending on shape and position of a wedge-like survey observation volume (upper right panel) and temperature of the stellar population (indicated in the legend). The p_M of the mock data, created in the `KKS-Pot` potential, are given as Test 1 in Table 3. The distribution in action space visualizes how orbits with different actions reach into different regions within the Galaxy. The 1D histograms on the right illustrate that qDFs generate realistic stellar distributions in Galactocentric coordinates $(R, z, \phi, v_R, v_z, vT)$. [TO DO: Jo suggests to make two or three separate figures out of this. I'm not yet convinced, as I think it is nice and tidy like this.]

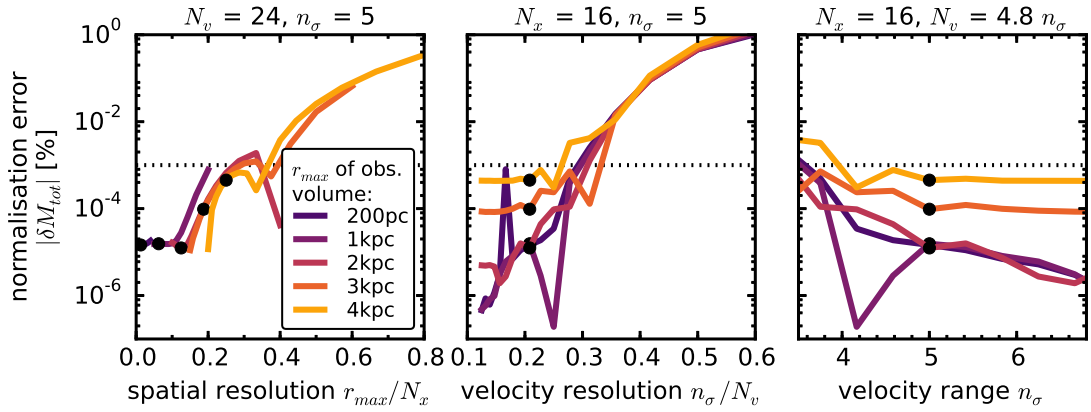
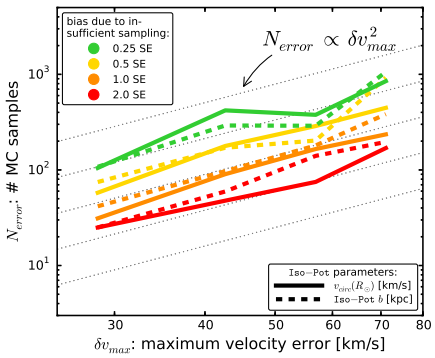


Figure 3. Relative error $\delta M_{\rm tot}$ of the likelihood normalization in Equation 12 depending on the accuracy of the grid-based density calculation in Equation 8 (and surrounding text) in five spherical observation volumes with different radius $r_{\rm max}$. (Test 2 in Table 3 summarizes the model parameters.) The tracer density in Equation 8 is calculated on $N_x \times N_x$ spatial grid points in $R \in [R_\odot \pm r_{\rm max}]$ and $|z| \in [0, r_{\rm max}]$. 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 $[0, 1.5 v_{\rm circ}]$ in v_T). (We vary N_x, N_v and n_σ separately and keep the other two fixed at the values indicated above the columns.) We calculate the "true" normalization in Equation 12 with high accuracy as $M_{\rm tot} \approx M_{\rm tot,approx}(N_x = 20, N_v = 56, n_\sigma = 7)$. The black dots indicate the accuracy used in our analyses: It is better than 0.001% (dotted line). We find that the accuracy depends on the *spatial* resolution of the grid and requires more accurate integrations over the *velocity* for larger volumes within which the density varies more strongly.



(a) $N_* = 10,000$

(b) $N_* = 5,000$

Figure 4. Number of Monte Carlo (MC) samples $N_{\rm error}$ needed for the numerical convolution of the model probability with the measurement uncertainties in Equation 14, given the maximum velocity error $\delta v_{\rm max}$ within the stellar sample. Unsufficent sampling introduces systematic biases in the parameter recovery as indicated in the legend. The relation found here, $N_{\rm error} \propto \delta v^2_{\rm max}$, was distilled from a set of analyses of mock data sets with different proper motion uncertainties $\delta\mu \in [2,5]$ mas yr^{-1} in the absence of distance errors (see Test 6.1 in Table 3). The proper motion error $\delta\mu$ translates to heteroscedastic [TO DO: make sure that this word is written correctly everywhere.] velocity errors according to $\delta v[\mathrm{km\ s^{-1}}] \equiv 4.74047 \cdot r[\mathrm{kpc}] \cdot \delta\mu[\mathrm{mas\ yr^{-1}}]$, with r being the distance of the star from the Sun. Stars with larger δv require more $N_{\rm error}$ for the integral over its measurement uncertainties to converge. We therefore show how the $N_{\rm error}$ needed for the potential *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 $N_{\rm error} = 800$ 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 12). [TO DO: no units in legend] [TO DO: some of the 25 MC sample analyses have to be re-done.] [TO DO: Replace lower plot with new plot with $N_* = 5,000$] [TO DO: Use N_* everywhere where applicable, no $N_{\rm sample}$] [TO DO: Introduce N_* somewhere.] [TO DO: Comment from Jo: I think it is important to test, if the MC vs error plot depends on number pf stars. Maybe test it with less stars (5000), to test this quickly. Naively, I would expect a large depedence on Ndata.]

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

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

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}. 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 the number of stars in the sample (N_* action calculations) and the numerical accuracy of the tracer density grid in Equation 8 needed for the likelihood normalization in Equation 11 ($N_x^2 \times N_v^3$ action calculations). The accuracy has to be chosen high enough, such that the resulting numerical error

$$\delta_{M_{\rm tot}} \equiv \frac{M_{\rm tot,approx}(N_x, N_v, n_\sigma) - M_{\rm tot}}{M_{\rm tot}} \qquad (12)$$

[TO DO: make sure every Mtottrue is replaced by Mtot] does not dominate the log-likelihood, i.e.,

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

with

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

Otherwise numerical inaccuracies could lead to systematic biases in the potential and DF recovery. For data sets as large as $N_* = 20,000$ stars, which in the age of Gaia could very well be the case [TO DO: Really???], one needs a numerical accuracy of 0.005% in the normalisation. Figure 3 demonstrates that the numerical accuracy we use in the analysis, $N_x = 16$, $N_v = 24$ and $n_\sigma = 5$, does satisfy this requirement. This is slightly higher than in BR13, where $N_* \sim 100$ [TO DO: CHECK].

Measurement uncertainties of the data have to be incorporated in the likelihood. We assume Gaussian uncertainties in the observable space $\boldsymbol{y} \equiv (\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{v}}) = (\mathrm{RA}, \mathrm{DEC}, (m-M), \mu_{\rm RA}, \mu_{\rm DEC}, v_{\rm los})$, i.e. the i-th star's observed $\boldsymbol{y}_i$ are drawn from the normal distribution $N[\boldsymbol{y'}_i, \delta\boldsymbol{y}_i]$, with $\boldsymbol{y}_i{}'$ being the star's true phase-space position and $\delta\boldsymbol{y}_i$ its uncertainty . Stars follow the distribution function ($\mathrm{DF}(\boldsymbol{y'}) \equiv \mathrm{DF}(\boldsymbol{J}[\boldsymbol{y'} \mid p_\Phi] \mid p_{\rm DF})$ for short), convolved with the measurement uncertainties $N[0, \delta\boldsymbol{y}]$ [TO DO: CHECK AGAIN]. The selection function $\mathrm{SF}(\boldsymbol{y})$ acts on the space of (error affected) observables. Then the probability of one star becomes

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

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 apply the following approximation,

$$\begin{aligned}&\tilde{p}(\boldsymbol{y}_i \mid p_\Phi, p_{\rm DF}, \delta\boldsymbol{y}_i)\\ &\approx \frac{\mathrm{SF}(\boldsymbol{x}_i)}{M_{\rm tot}} \cdot \frac{1}{N_{\rm error}} \sum_n^{N_{\rm error}} \mathrm{DF}(\boldsymbol{x}_i, \boldsymbol{v}[\boldsymbol{y}'_{i,n}])\end{aligned} \qquad (14)$$

with

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

We calculate the convolution using Monte Carlo (MC) integration with $N_{\rm error}$ samples. The above approximation assumes that the star's *position* $\boldsymbol{x}_i$ is perfectly measured. As the selection function 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 besides on $\delta\boldsymbol{\mu}$ and $\delta v_{\rm los}$ 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 4 demonstrates that in the absence of position uncertainties the $N_{\rm error}$ needed for the convolution integral to converge depends as

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

on the uncertainties in the (1D) velocities.
A similar but only one-dimensional treatment of measurement uncertainties in v_z was already applied by BR13.

2.7. *Fitting Procedure*

To search the $(p_\Phi, p_{\rm DF})$ parameter space for the maximum of the *pdf* in Equation 11, we go beyond the single fixed grid search by BR13 and employ an effective 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 at a low number of likelihood evaluations.

- *Initialization.* For N_p free model parameters p_M, we start with a sufficiently large grid with 3_p^N 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 $N_* \times N_{\rm error} + N_x^2 \times N_v^3$ actions (# stars times # MC samples for error convolution, plus actions required for density interpolation grid in Equation 8). Then an inner loop evaluates Equation 11 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 3^{N_p} grid. 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.
- *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 range 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—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. 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 $N_* \times N_{\rm error} + N_x^2 \times N_v^3$ 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 a Monte-Carlo Markov Chain (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.

3. RESULTS

We are now in a position to examine the limitations of action-based modelling posed in the introduction (see Section 1) using our *RoadMapping* machinery. We explore: (i) unbiased estimates, (ii) the role of the survey volume, (iii) imperfect selection functions, (iv) measurement errors and what happens if the actual (v) DF or (vi) Potential are not spanned by the space of models. We do not explore the breakdown of the assumption that the system is axisymmetric and in steady state. With the exception of the test suite on measurement errors in §3.4, we assume that phase-space errors are negligible. All tests are also summarized in Table 3.

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

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

The individual MAPs in BR13 contained typically between 100 and 800 objects, so that each MAP implied a quite broad *pdf* for the model parameters p_M. Here we explore what happens in the limit of much larger samples, say $N_* = 20,000$ objects. As outlined in §2.6 the immediate consequence of larger samples is given by the likelihood normalization requirement, $\log(1+\delta M_{\rm tot}) \leq 1/N_*$, (see Equation 13)), which is the modelling aspect that drives the computing time. This issues aside, we would, however, expect that in the limit of large data sets with vanishing measurement errors the *pdf*s of the p_M become Gaussian, with a *pdf* width that scales as $1/\sqrt{N_*}$. Further, we must verify that any bias in the *pdf* expectation value is considerably less than the error, even for quite large samples.

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

3.2. *The Role of the Survey Volume Geometry*

[TO DO: Mach einheitlich, wie die Tests in Captions und Text erwhnt werden (in Klammer)]

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 from the same p_M, two different potentials (`Iso-Pot` and `MW13-Pot`, see Test 4 in Table 3), and volume wedges (see Section 2.4) at *different positions within the Galaxy*, illustrated in the right upper panel of Figure 8. To isolate the role of the survey volume geometry, the mock data sets are equally large ($N_* = 20,000$) 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 8.

The second suite of mock data sets was already introduced in Section 3.1 (see also Test 3.3 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 7 and demonstrate the effect of the *size of the survey volume*.

Figure 7 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 [TO DO: Comment from HW: The question of whether a hotter or a colder population gives tighter constraints is an important question, but it seems buries here in a section that is dedicated to another matter, namely the question of volume ... It's OK to leave it here, but somewhere we need to say clearly: whether the population is hot or cold does not make a big and generic difference...]; it depends on the survey volume and the model parameter in question. In Figure 8 the wedges all have the same volume and all give results of similar precision. Minor differences (e.g., 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 8 implies that for these cases volume offsets in the radial or vertical direction have at most a modest impact - even in case of the very large sample size at hand.

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

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

The selection function (see Section 2.4) 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 a spherical survey volume around the Sun with a spatially varying incompleteness function, while assuming constant completeness in the *RoadMapping* analysis. Our test case uses a completeness function which drops linearly with distance r from the Sun (see Test 5 in Table 3 and Figure **??**). This 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 10 demonstrates that the potential recovery with *RoadMapping* is very robust against somewhat wrong assumptions about the radial completeness of the data. The robustness for the `cool` stellar population is even more striking than for the `hot` population. The reason for this robustness could be, that much information about the potential comes from the rotation curve measurements in the plane, which is not affected by the incompleteness of the sample. We test this by not including tangential velocity measurements in the analysis of the data sets from Figure 10 (by marginalizing the likelihood in Equation 9 over v_T). Figure 11 shows that in this case the potential is much less tightly constrained, even for 20,000 stars. For only small deviations of true and assumed completeness ($\lesssim 10\%$) we can however still incorporate the true potential in our fitting result (see Figure **??**).

We found similar results also for spatial incompleteness functions varying with z

[TO DO: Comment by HW: I don't have an immediate solution for this, but again, it seems the interesting question of "how much of the information is in the rotation curve" is 'hidden' in the section on selection functions... - What does he mean by that?]

[TO DO: Remove vertical incompleteness from test table]

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

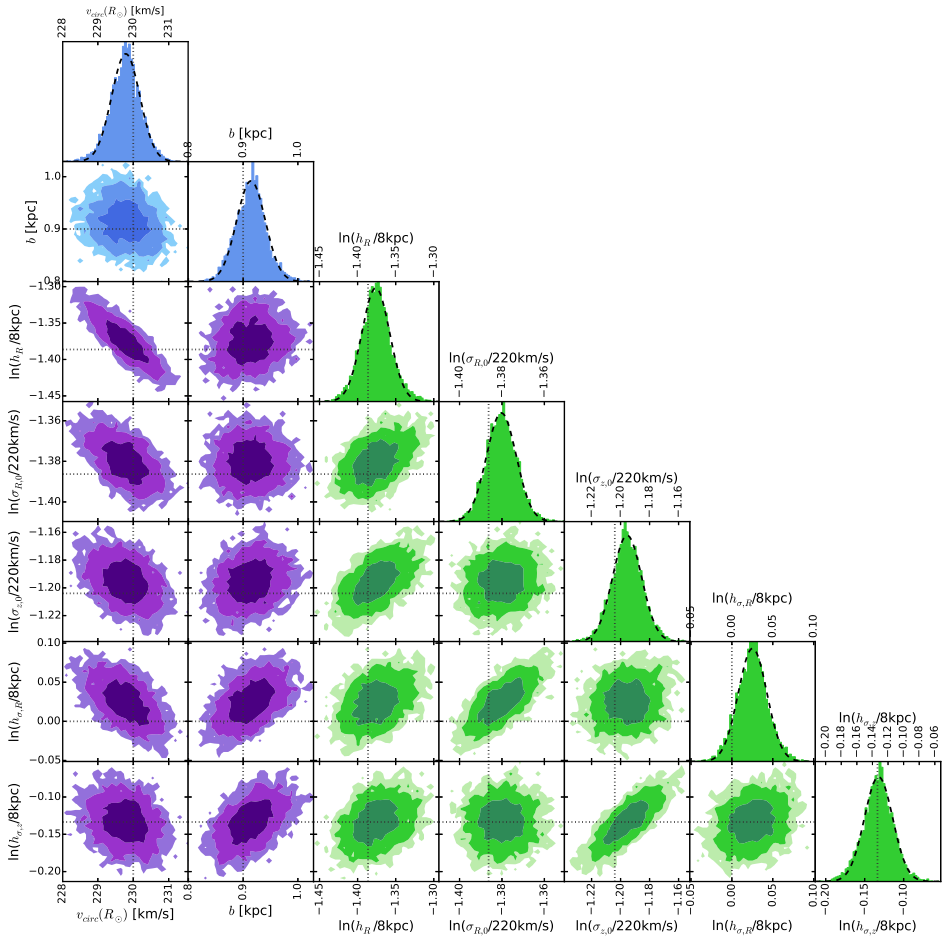


Figure 5. The *pdf* (proportional to the likelihood in Equation 9) in the parameter space $p_M = \{p_\Phi, p_{\rm DF}\}$ for one example mock data set created according to Test 3.1 in Table 3. Blue indicates the *pdf* for the potential parameters, green the qDF parameters. The true parameters are marked by dotted lines. The dark, medium and bright contours in the 2D distributions represent 1, 2 and 3 sigma confidence regions [TO DO: HW: "likelihood vs. pdf - This is where this matters: is this a confidence on the data or on the parameters?" Don't understand, what he means...], respectively. The parameters are weakly to moderately covariant, but their level of covariance depends on the actual choice of the mock data's p_M. The *pdf* here was sampled using MCMC. The dashed lines in the 1D distributions are Gaussian fits to the histogram of MCMC samples. This demonstrates very well that for such a large number of stars, the *pdf* approaches the shape of a multi-variate Gaussian, as expected from the central limit theorem [TO DO: Jo wrote, that he is not sure if the central limit theorem is directly relevant here].

[TO DO: Comment from HW: This Section has three parts:
– convergence of the integral (ALREADY REMOVED)
– testing the approximation
– underestimating errors
It seems to me that the basic Section: What is the impact of the errors? Is missing. That should be the center piece, and the other three aspects should be quick summary notes, only 1-2 sentences long.] [I'll try to address this with a plot mean(SE) vs. proper motion error - also for cold population (currently running on wolf).]

In absence of distance uncertainties the error convolved likelihood given in Equation 14 is unbiased. When including distance (modulus) errors, Equation 14 is just an approximation for the true likelihood. The systematic bias thus introduced in the parameter recovery gets larger with the size of the error. This is demonstrated in Figure 6.2. We find however that in case of $\delta(m-M) \lesssim 0.3$ mag (if also $\delta\mu \leq 2$ 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 2 sigma [TO DO: Make sure this is what I claim in abstract and discussion.]. This corresponds to a relative distance error of $\sim 10\%$.

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

We found that in case we perfectly knew the measurement errors (and the distance error is negligible), the convolution of the model probability with the measurement uncertainties gives precise and accurate constraints on the model parameters - even if the error itself is quite large.

Figure 14 now investigates the effect of a systematic

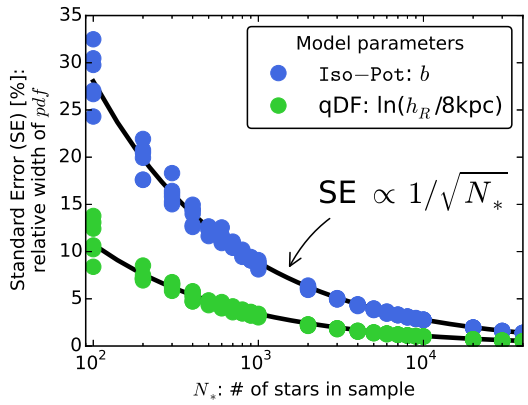


Figure 6. The width of the *pdf* (see Equation 9) for two fit parameters found from analyses of 132 mock data sets vs. the number of stars in each data set. (The mock data was created according to the model parameters given in Test 3.2 in Table 3.) The relative standard error (SE) was found from a Gaussian fit to the marginalized *pdf* for each model parameter. As can be seen, for large data samples the width of the *pdf* scales with $1/\sqrt{N_*}$ as predicted by the central limit theorem.

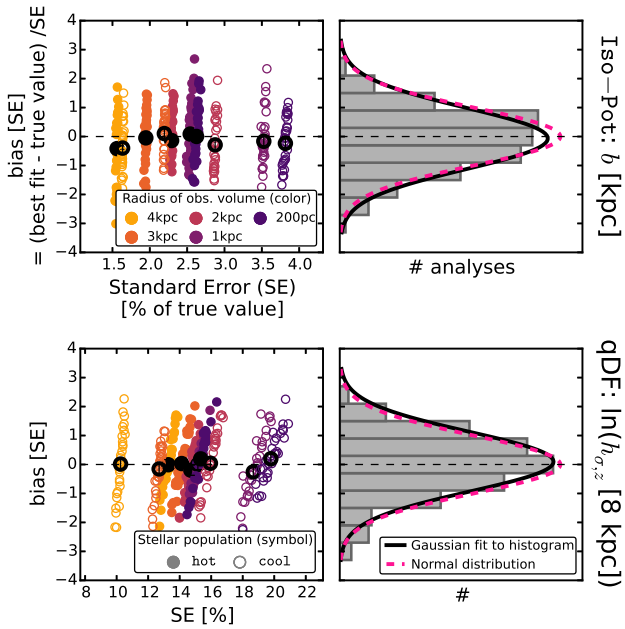


Figure 7. (Un-)bias of the parameter estimates: According to the central limit theorem the the best fit values for a large number of data sets, each containing a large number of stars, will follow the Normal distribution. To test this, we create 320 mock data sets, which come from two different stellar populations and five spherical observation volumes (see legends). (All model parameters are summarized in Table 3 as Test 3.3.) Bias and relative standard error (SE) are derived from the marginalized *pdf* for one potential parameter (isochrone scale length b in first row) and one qDF parameter ($h_{\sigma,z}$ in second row). The second column displays a histogram of the 320 offsets. As it closely follows a Normal distribution, our modelling method is therefore well-behaved and unbiased. The black dots show the mean offset and SE for the 32 analyses belonging to one model. [TO DO: $r_{\rm max}$ instead of radius in legend] [TO DO: Leerzeichen fehlt in y-achsenbeschriftung] [TO DO: no kpc at b, and / 8kpc at hsz]

underestimation of the true proper motion uncertainties $\delta\mu$ by 10% and 50%. We find that this causes a bias in the parameter recovery that grows seemingly linear with $\delta\mu$. For an underestimation of only 10% however, the bias is still $\lesssim 2$ sigma for 10,000 stars [TO DO: Check] - even for $\delta\mu \sim 3$ 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 14). The qDF parameters are for example better recovered by hotter populations. This is, because the *relative* difference between the true $\sigma_i(R)$ (with $i \in \{R, z\}$) and measured $\sigma_i(R)$ (which comes from the deconvolution with an underestimated velocity uncertainty) is smaller for hotter populations.

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

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

Our modelling approach assumes that each stellar population follows a simple DF (here: 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[7] (see Table 2 and Test 7 in Table 3), and analyze the composite mock data set by fitting a single qDF to it. The mock data sets and best fit qDF are illustrated in Figure 15, and the comparison of input and best fit parameters in Figures 16 and 17. In *Example 1* we choose qDFs of widely different temperatures and vary their relative fraction (Figure 16); in *Example 2* we always mix mock data stars from two different qDFs in equal proportion, but vary by how much the qDF's temperatures differ (Figure 17).

The first set of tests mimics a DF that has wider wings or a sharper core in velocity space than a qDF (see Figure 15). 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 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 only small, error in recovering the potential parameters, monotonically increasing with the qDF parameter difference. In particular for fractional differences in the qDF parameters of $\lesssim 20\%$ the systematics are insignificant even for sample sizes of 20,000, as used in the mock data.

Overall, mock data drawn from a cooler DF always seem to give tighter constraints on the circular velocity at the Sun, because the rotation curve can be constrained easier if more stars are on near-circular orbits. But we

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

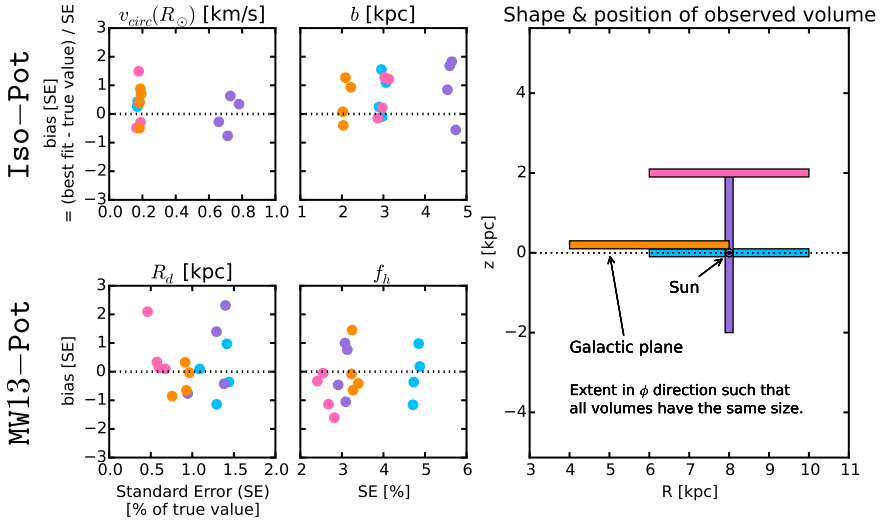


Figure 8. Bias vs. standard error in recovering the potential parameters for mock data stars drawn from four different wedge-shaped test observation volumes within the Galaxy (illustrated in the right panel; the corresponding analyses are colour-coded) and two different potentials (`Iso-Pot` and `MW13-Pot` 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 the volume of 4.5 kpc^3, even though their extent in (R, z) is different. Overall there is no clear trend, that an observation volume around the Sun, above the disk or at smaller Galactocentric radii should give remarkably better constraints on the potential than the other volumes. [TO DO: Write in Plot "... that all wedges have the same volume".] [TO DO: No units in parameters in titles]

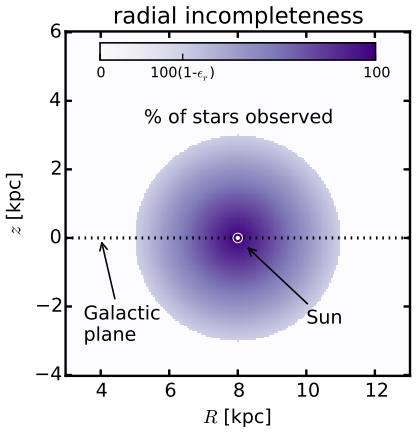


Figure 9. Illustration of the selection function used to investigate the impact of misjudgements of the radial incompleteness of the data in Figure 10. The survey volume is a sphere around the Sun and the percentage of observed stars is decreasing linearly with radius from the Sun. How fast this detection/incompleteness rate drops is quantified by the factor ϵ_r.

found the recovered $v_{\rm circ}(R_\odot)$ not always to be unbiased at the implied precision.

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

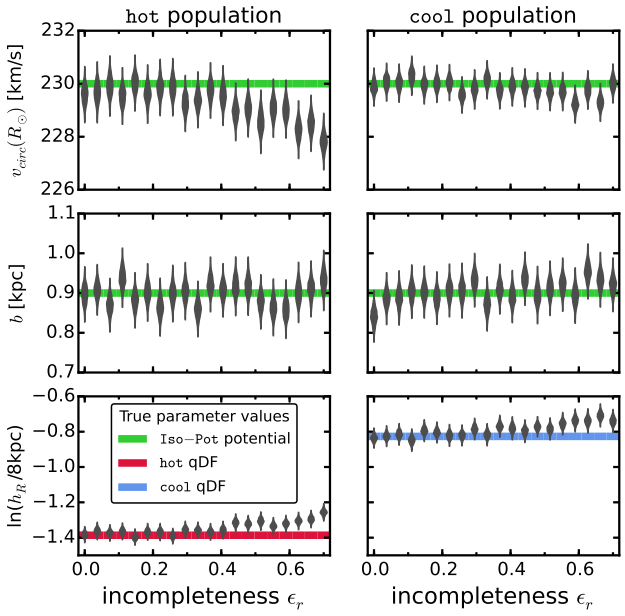


Figure 10. Influence of wrong assumptions about the radial incompleteness of the data on the parameter recovery with *RoadMapping*. Each mock data set was created with different incompleteness parameters ϵ_r (shown on the x-axis and illustrated in Figure 9). (The model parameters are given as Test 5 in Table 3.) The analysis however did not know about the incompleteness and assumed that all data sets had constant completeness within the survey volume ($\epsilon_r = 0$). The 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 qDF parameters not shown here exhibit an even better robustness than h_R.) [TO DO: Jo suggested to also remove the h_R panel, but I like, that one can see that it is the spatial tracer distribution that drives the little degradation of the recovery.]

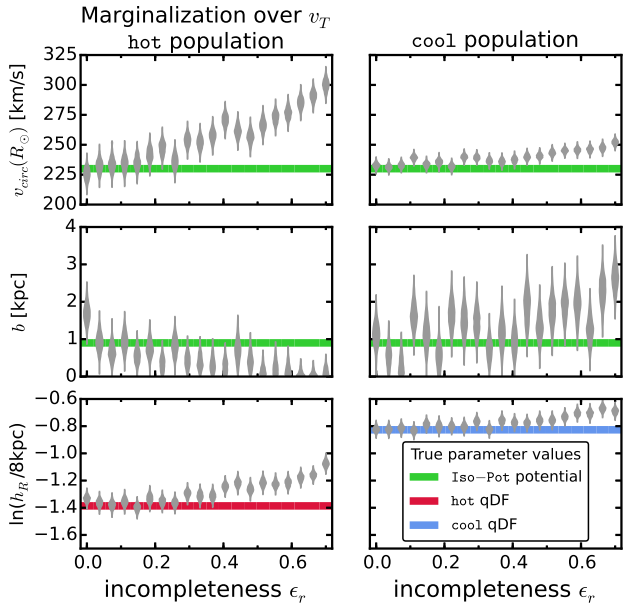


Figure 11. Same as Figure 10, but without including information about the tangential velocities in the analysis. This was done by marginalizing the likelihood in Equation 9 over v_T. The parameter recovery is much worse than in Figure 10 (as can be seen from comparing the parameter ranges on the y-axis). This could indicate that much of the information about the potential is actually stored in the rotation curve, i.e. $v_T(R)$, which is not affected by removing stars from the data set. But even if we do not include v_T we can still recover the potential within the errors, at least for small ($\epsilon_z \lesssim 10\%$).

We note, that in the cases where the systematic bias in the potential parameter recovery becomes several sigma large, a direct comparison of the true mock data set and best fit distribution (see Figure 15) already reveals that the assumed DF is not a good model for the data.

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

[TO DO: Make sure that the vcirc is better recovered from cooler qDF is mentioned here for the first time] [TO DO: Mention somewhere but only once that in case the offsets are big, one can already see in the mock data distribution, that the assumed DF is not the correct DF.]

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

We now explore what happens when the mock data were drawn from one axisymmetric potential family, here `MW14-Pot`, and is then modelled considering potentials from only another axisymmetric family, here `KKS-Pot` (see Table 1 and Figure 1). In the analysis we assume the circular velocity at the Sun to be fixed and known [TO DO: Comment from Hans-Walter: Do we have reason to believe that this very restrictive assumption does not qualitatively impact our upshot (quantitative differences are OK).] and only fit the parametric potential form.

We analyse a mock data set from each a `hot` and `cool` stellar population (see Table 2). The distributions generated from the best fit parameters reproduce the data in configuration space very well (see Figure 18).

The results for the potential are shown in Figure 19. We find that the potential recovered by *RoadMapping* is

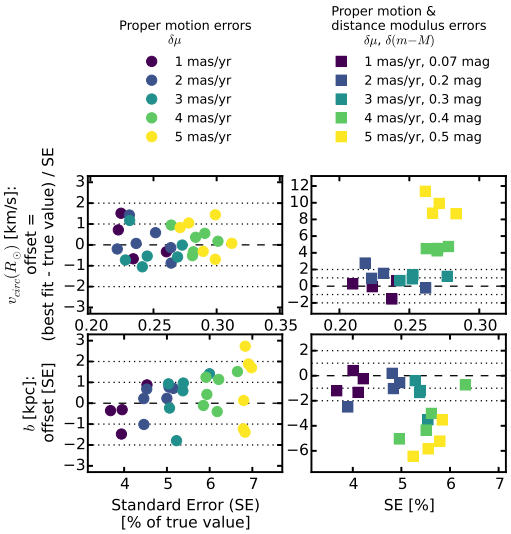


Figure 12. Potential parameter recovery using the approximation for the model probability convolved with measurement uncertainties in Equation 14. We show *pdf* offset and relative width (i.e., standard error SE) for the potential parameters derived from mock data sets, which were created according to Test 6.2 in Table 3). The data sets in the left panels have only uncertainties in line-of-sight velocity and proper motions, while the data sets in the right panels also have distance (modulus) uncertainties, as indicated in the legends in the first row. For data sets with proper motion error errors $\delta(m-M) \leq 3$ mas yr^{-1} Equation 14 was evaluated with $N_{\rm error} = 800$, for $\delta(m-M) > 3$ mas yr^{-1} we used $N_{\rm error} = 1200$. In absence of distance uncertainties Equation 14 gives unbiased results. For $\delta(m-M) \geq 3$mas yr^{-1} (which corresponds in this test to $\delta v_{\rm max} \gtrsim 43$ km s^{-1}, see Equation **??**) however biases of several sigma are introduced as Equation 14 is only an approximation for the true likelihood in this case. [TO DO: No units at b and vcirc on y axis]

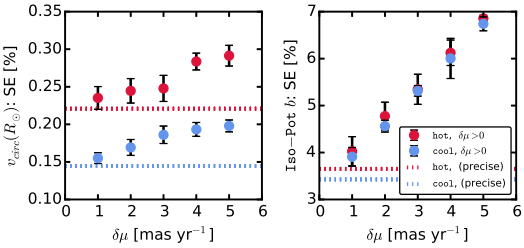


Figure 13. [TO DO: This should be a figure that plots precision (SE) vs. proper motion error for a hot and a cool population (for no distance error). This is to demonstrate the effect of measurement errors in general. Currently running on cluster....]

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 19 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 $r_{\rm max} = 4$ 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

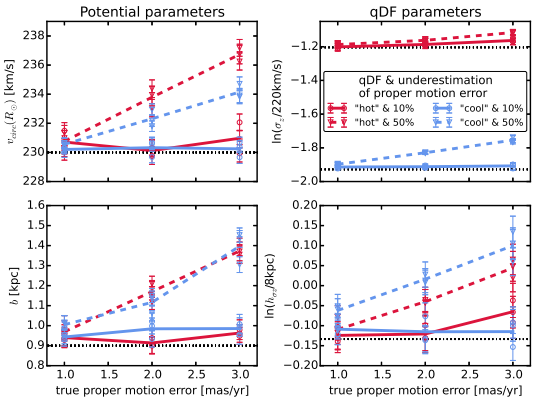


Figure 14. Effect of a systematic underestimation of proper motion errors in the recovery of the model parameters. The true model parameters used to create the mock data are summarized as Test 6.3 in Table 3, four of them are given on the y-axes and the true values are indicated as black dashed lines. The velocities of the mock data were perturbed according to Gaussian errors in the RA and DEC proper motions as indicated on the x-axis. The circles and triangles are the best fit parameters of several mock data sets assuming the proper motion uncertainty, with which the model probability was convolved, was underestimated in the analysis by 10% or 50%, respectively. The error bars correspond to 1 sigma confidence. The lines connect the mean of each two data realisations and are just to guide the eye. [TO DO: rename $h_{\sigma z}$ to $h_{\sigma,z}$, σ_z to $\sigma_{z,0}$] [TO DO: Potential and/or population names in typewriter font] [TO DO: Iso-Pot in Title] [TO DO: Delta mu on x-axis] [TO DO: legend with true values]

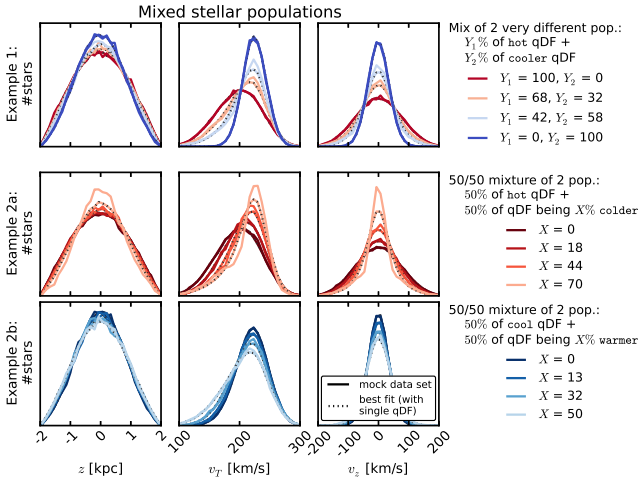


Figure 15. Distribution of mock data created by mixing stars drawn from two different qDFs (solid lines), and the distribution predicted by the best fit of a single qDF and potential to the data (dotted lines). The model parameters used to create the mock data are given in Table 3 as Test 7, with the qDF parameters refereed 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 16 and 17. (The data sets are color-coded in the same way as the corresponding analyses in Figures 16 and 17.) 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 intermediate to large deviations it becomes already obvious from directly comparing the mock data and best fit distribution, that a single qDF is a bad assumption for the star's true DF. [TO DO: Replace Example 2b analyses.]

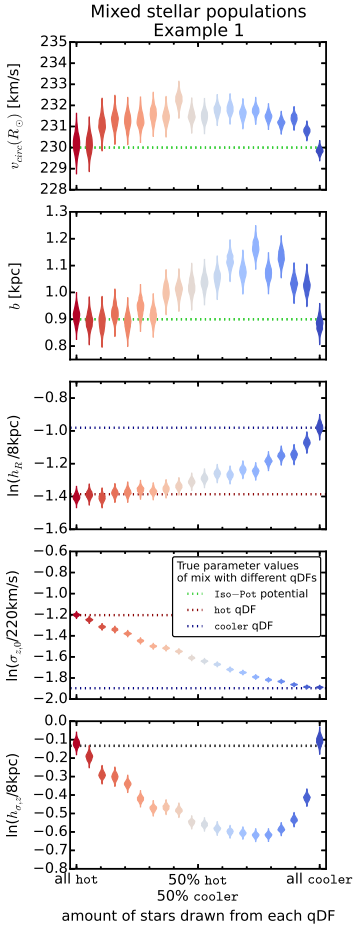


Figure 16. The dependence of the parameter recovery on degree of pollution and temperature of the stellar population. We mix (i.e., pollute) varying amounts of stars from a `hot` stellar population with stars from a very different `cooler` population (see Table 2), as indicated on the x-axis. (All model 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 15, first row, in the same colors as the violins here. We find that a hot population is much less affected by pollution with stars from a cooler population than vice versa. [TO DO: Replace Example 2b analyses.]

in functional forms.

The density contours are less tightly constrained than the forces, but we still capture the essentials: The `hot` stellar population constrains the halo—especially at smaller radii it is equally good or better than the `cool` population; the `cool` population gives tighter constraints on the halo in the outer region and recovers the disk better than the `hot` population. This is in concordance with expectations as the `cool` population has a longer tracer scale length and is more confined to the disk than the `hot` population and therefore also probes the Galaxy in these regions better.

Overall the best fit disk is less dense in the midplane than the true disk, because the generation of very flattened components like exponential disks with Stäckel potentials is not possible [TO DO: Ask Glenn, if this is true].

Figure 20 compares the true qDF parameters with the best fit qDF parameters belonging to the potentials in

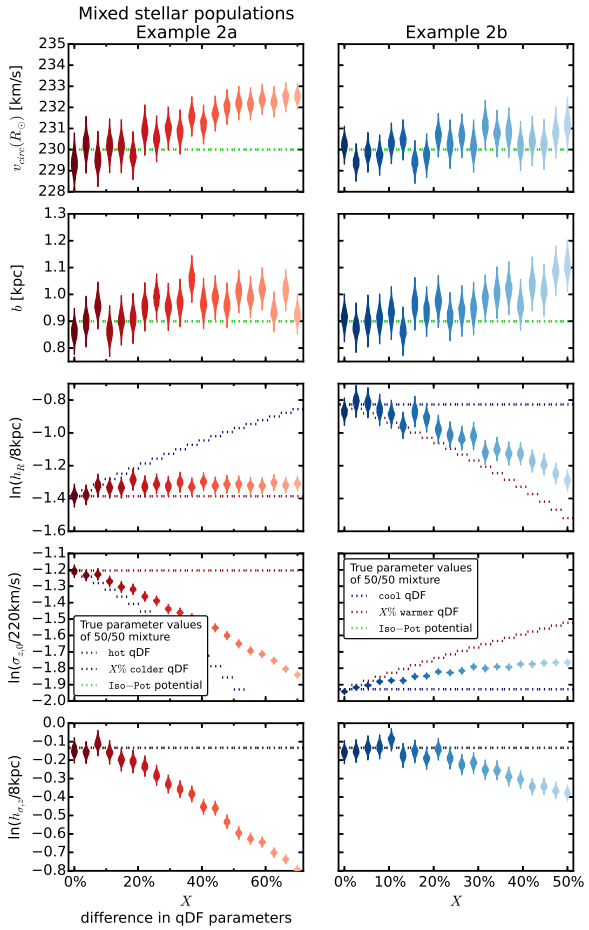


Figure 17. The dependence of the parameter recovery on the difference in qDF parameters of a 50%/50% mixture of two stellar populations and their temperature. The two qDFs from which the stars in each mock data set were drawn are indicated in the legend, with the qDF parameters $\sigma_{R,0}, \sigma_{z,0}$ and h_R differing by X% (see also Table 2), 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 are shown in figure 15, last two rows (color-coded analogous to the violins here). 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%. [TO DO: Replace Example 2b analyses.]

Figure 19. 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 produced. This discrepancy could be connected to the `KKS-Pot` not being able to reproduce the flattness of the disk. Also, σ_z and σ_R in Equations 5-6 are scaling profiles for the qDF (cf. BR13) and how close they are to the actual velocity profile depends on the choice of potential.

4. SUMMARY AND DISCUSSION

[TO DO: Introduce DF somewhere - use DF wherever we don't need qDF.] [TO DO: Introduce MW somewhere.]
[TO DO: Compare these sections with the results. Points should be made detailed in the results section and short here in the discussion. Says Hans-Walter.]
[TO DO: Abstze mit Indent. Keine Leerzeilen.]

Recently, implementations of action DF-based modelling of 6D data in the Galactic disk have been put forth, in part to lay the ground-work fo Gaia (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 distribution function 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 reliable, even when there are small deviations of the model assumptions from the real world Galaxy.

RoadMapping applies a full likelihood analysis and is statistically well-behaved. It 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 2012) 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 the limit of large datasets, as long as the modelling assumptions are fulfilled.

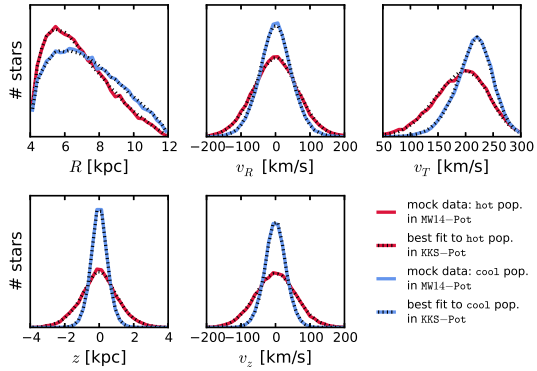


Figure 18. Comparison of the distribution of mock data in configuration space created in the `MW14-Pot` potential and with two different stellar populations (see Test 8 in Table 3 for all mock data model parameters), and the best fit distribution recovered by fitting the family of `KKS-Pot` potentials to the data. The best fit potentials are shown in Figure 19 and the corresponding best fit qDF parameters in Figure 20. The data is very well recovered, even though the fitted potential family did not incorporate the true potential.

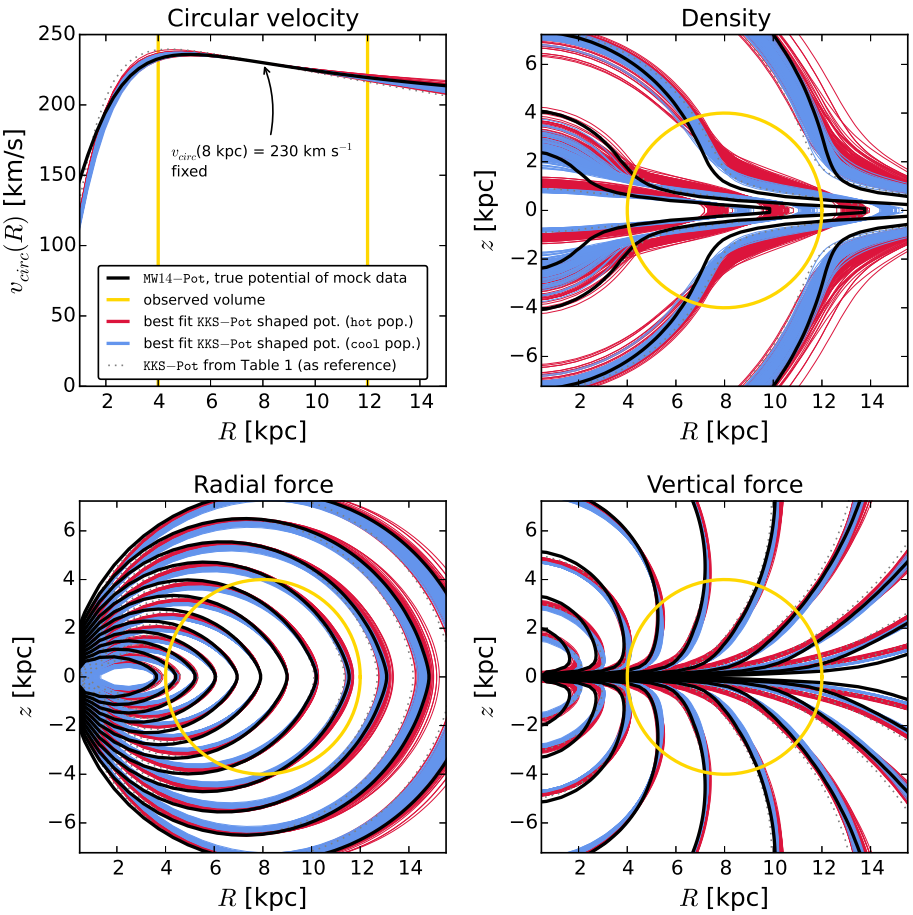


Figure 19. Recovery of the gravitational potential if the assumed potential model (`KKS-Pot` with fixed $v_{\rm circ}(R_\odot)$) and the true potential of the (mock data) stars (`MW14-Pot` in Table 1) are slightly different. We show the circular velocity curve, as well as contours of equal density, radial and vertical force in the R-z-plane, and compare the true potential with 100 sample potentials drawn from the posterior distribution function found with the MCMC for a `hot` (red) and a `cool` stellar population (blue). All model parameters are given as Test 8 in Table 3. [TO DO: Do more analyses???] [TO DO: Reference correct Table in Plot - don't forget!] [TO DO: Redo whole analysis with vcirc not being fixed (HW is not sure if this really doesn't make a difference.]

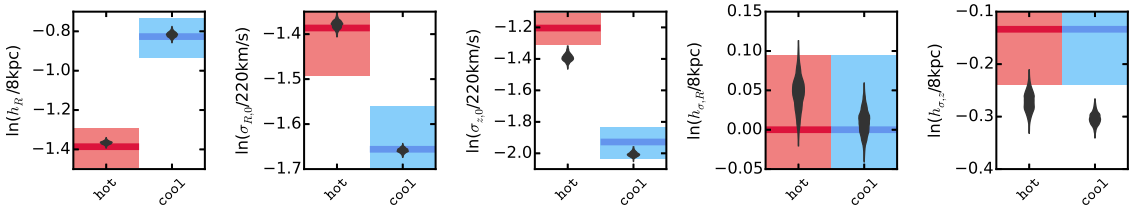


Figure 20. Recovery of the qDF parameters for the case where the true and assumed potential deviate from each other (Test 8 in Table 3). The thick red (blue) lines represent the true qDF parameters of the `hot` (`cool`) qDF in Table 2 used to create the mock data, surrounded by a 10% error region. The grey violins are the marginalized *pdf*s for the qDF parameters found simultaneously with the potential constraints shown in Figure 19.

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.

Stellar populations of different scale length and temperature probe different regions of the Galaxy (BR13). But there is no easy rule of thumb for which survey volume and stellar population which potential and DF parameter is constrained best.[TO DO: Ask HW, if he wants this to be discussed somewhere in detail.]

Surprisingly, (cf. Rix & Bovy 2013) *RoadMapping* seems to be very robust against misjudgements of the spatial data selection function. 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.

[TO DO: Finished up to here. Continue here.] [TO DO: Comment from HW: Author: rix Subject: This paragraph shouldbe 1 or 2 sentences, following the first

paragraph on "Sample/Data Properties". This – at the moments –reads to bequite confusing. I don't quite get whatthe "upshot" is; there is technical detai on N_{error} [enought to say it's expensive]; and, as noted earlier; I don't understand why theerror convolution for a nearbydata point needs to know about $\delta v_{\rm max}$] Properly convolving the likelihood with measurement errors is computational very expensive. By ignoring positional errors and only including distance errors as part of the velocity error, we can drastically reduce the computational costs. For stars within 3 kpc from the Sun this approximation works well for distance errors of $\sim 10\%$ or smaller. The number of MC samples needed for the error convolution using MC integration scales by $N_{\rm error} \propto (\delta v_{\rm max})^2$ with the maximum velocity error at the edge of the sample. If we did not know the true size of the proper motion measurement errors perfectly, we can only reproduce the true model parameters to within $\lesssim 2$ sigma [TO DO: Check???] as long as we do not underestimate it by more than 10% and for proper motion errors $\lesssim 2$ mas yr^{-1}.

Deviations from the qDF Assumption: *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 star's true DF.

In Example 1 in Section 3.5 we considered true stellar DFs being i) hot with less stars at small radii and more stars with low velocities than assumed (reddish data sets in Figure 15), or ii) cool with broader velocity dispersion wings and less stars at large radii than assumed (bluish data sets). [TO DO: I removed the radius mock data distribution from the figure, because you couldn't see differences. Should I put it back in, so people understand this comment better?] We find that case i) would give more reliable results for the potential parameter recovery, but in both cases biases are small if the contamination is less than [TO DO: CHECK] [TO DO: Jo suggested this last part of the sentence, but I'm not sure, this is the case]. In other words, hotter stellar populations appear to be more robust against pollution of stars coming from a cooler population than vice versa.

Binning of stars into MAPs in [α/Fe] and [Fe/H], as done by BR13, could introduce biases due to abundance errors 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 20% in the qDF parameters of two neighbouring bins can still give quite good constraints on the potential parameters.

The relative difference 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 16 and 17 suggest that especially the tracer scale length h_R needs to be recovered to get the potential right. For this parameter however the bin sizes in Figure 6 of 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 Parameterized 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 so tight, that it should be already possible to distinguish between different parametric MW potential models (e.g. the `MW13-Pot` used by BR13 and the `KKS-Pot`). [TO DO: I did not really do the test with the MW13-Pot, can I still claim this???]

We also found indications that populations of different scale lengths and temperature indeed probe different regions of the Galaxy best. [TO DO: Check that this is indeed the case - it is not clear to me from the plot. ???] 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.

BR13 fitted a MW-like model potential and calculated actions using the Stäckel approximation (Binney 2012); 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 for 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.

Different Modelling Approaches using Action-based Distribution Functions: BR13 have 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.

The main drawback is that—for reasons of galaxy and chemical evolution—the DF properties are astrophysically linked between different MAPs. Ultimately, the goal is to do a consistent chemodynamical model that si-

multaneously fits the potential and DF($\boldsymbol{J}$, [X/H]) (where [X/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.

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

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 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 thisk disk, each consisting of several cohorts described by qDFs, a DF for the halo, a functional form of the metallicity of the interstellar medium at the time of birth, and a simple prescription for radial migration. They applied a full likelihood analysis accounting for selection effects and found a best fit for the eDF in 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.

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.

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

[TO DO: Make sure that MW is introduced once and Milky way is always abbreviated]

[TO DO: Somewhere a footnote with code reference to galpy]

5. ACKNOWLEDGMENTS

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

REFERENCES

Ahn, C. P., Alexandroff, R., Allende Prieto, C., et al. 2014, ApJS, 211, 17
Batsleer, P., & Dejonghe, H. 1994, A&A, 287, 43
Binney, J. 2010, MNRAS, 401, 2318
Binney, J., & McMillan, P. 2011, MNRAS, 413, 1889
Binney, J. 2011, Pramana, 77, 39
Binney, J. 2012, MNRAS, 426, 1324
Binney, J. 2012, MNRAS, 426, 1328
Binney, J. 2013, NAR [TO DO: emulateapj doesn't know NAR], 57, 29
Binney, J., & Tremaine, S. 2008, Galactic Dynamics: Second Edition, by James Binney and Scott Tremaine. ISBN 978-0-691-13026-2 (HB). Published by Princeton University Press, Princeton, NJ USA, 2008.
Bovy, J., & Tremaine, S. 2012, ApJ, 756, 89
Bovy, J., Rix, H.-W., & Hogg, D. W. 2012b, ApJ, 751, 131
Bovy, J., Rix, H.-W., Hogg, D. W. et al., 2012c, ApJ, 755,115
Bovy, J., Rix, H.-W., Liu, C., et al. 2012, ApJ, 753, 148
Bovy, J., & Rix, H.-W. 2013, ApJ, 779, 115 (BR13)
Bovy, J. 2015, ApJS, 216, 29
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
Dehnen, W., & Binney, J. 1998, MNRAS, 294, 429
de Lorenzi, F., Debattista, V. P., Gerhard, O., & Sambhus, N. 2007, MNRAS, 376, 71
Famaey, B., & Dejonghe, H. 2003, MNRAS, 340, 752
Foreman-Mackey, D., Hogg, D. W., Lang, D., & Goodman, J. 2013, PASP, 125, 306
Garbari, S., Liu, C., Read, J. I., & Lake, G. 2012, MNRAS, 425, 1445
Gilmore, G., & Reid, N. 1983, MNRAS, 202, 1025
Henon, M. 1959, Annales d'Astrophysique, 22, 126
Holmberg, J., Nordström, B., & Andersen, J. 2009, A&A, 501, 941
Hunt, J. A. S., & Kawata, D. 2014, MNRAS, 443, 2112
Hunt, J. A. S., & Kawata, D. 2014, MNRAS, 443, 2112
Jurić, M., Ivezić, Ž., Brooks, A., et al. 2008, ApJ, 673, 864
Kawata, D., Hunt, J. A. S., Grand, R. J. J., Pasetto, S., & Cropper, M. 2014, MNRAS, 443, 2757
Klement, R., Fuchs, B., & Rix, H.-W. 2008, ApJ, 685, 261
Kuijken, K., & Gilmore, G. 1989, MNRAS, 239, 605
McMillan, P. J. 2011, MNRAS, 414, 2446
McMillan, P. J. 2012, European Physical Journal Web of Conferences, 19, 10002
McMillan, P. J., & Binney, J. J. 2008, MNRAS, 390, 429
McMillan, P. J., & Binney, J. 2012, MNRAS, 419, 2251
McMillan, P. J., & Binney, J. J. 2013, MNRAS, 433, 1411
Nordström, B., Mayor, M., Andersen, J., et al. 2004, A&A, 418, 989
Perryman, M. A. C., de Boer, K. S., Gilmore, G., et al. 2001, A&A, 369, 339
Piffl, T., Binney, J., McMillan, P. J., et al. 2014, MNRAS, 445, 3133
Read, J. I. 2014, Journal of Physics G Nuclear Physics, 41, 063101
Rix, H.-W., & Bovy, J. 2013, A&A Rev., 21, 61
Sanders, J. L., & Binney, J. 2015, MNRAS, 449, 3479
Sellwood, J. A. 2010, MNRAS, 409, 145
Steinmetz, M., Zwitter, T., Siebert, A., et al. 2006, AJ, 132, 1645
Strigari, L. E. 2013, Phys. Rep., 531, 1
Syer D., Tremaine S. 1996, MNRAS, 282, 223
Ting, Y.-S., Rix, H.-W., Bovy, J., & van de Ven, G. 2013, MNRAS, 434, 652
Yanny, B., Rockosi, C., Newberg, H. J., et al. 2009, AJ, 137, 4377
Zhang, L., Rix, H.-W., van de Ven, G., et al. 2013, ApJ, 772, 108

[TO DO: In which order should I give the references????]
[TO DO: replace the references which I typed myself with the ones from ADS.]
[TO DO: Check if all references are actually used in paper. ???]

Table 3
Summary of test suites in this work: The first column indicates the test suite, the second column the potential, DF and selection function model etc. used for the mock data creation, the third model the corresponding model assumed in the analysis, and the last column lists the figures belonging to the test suite. Parameters that are not left free in the analyis, are always fixed to their true value. Unless otherwise stated we calculate the likelihood by the nested-grid and MCMC approach outlined in §2.7 and use $N_x = 16$, $N_v = 24$, $n_\sigma = 5$ as numerical accuracy for the likelihood normalisation in Equations (9) and (8). N_* is the number of stars per mock data set. Unless stated otherwise, the all selection functions have completeness($\boldsymbol{x}$) = 1. [TO DO: Jo suggested to make many tables from this. But I actually like one big table at the end of the paper. Otherwise we had 6 additional tables interrupting the flow of text and figures all the time. And the parameters in the table are really just for reference.] [TO DO: Overall, this table could do with a little less information.]

Test		Model for Mock Data	Model in Analysis	Figure
Test 1 :	*Potential:*	`KKS-Pot`	-	Figure 2
Influence of	*DF:*	`hot` or `cool` qDF		
survey volume on	*Survey volume:*	a) $R \in [4, 12]$ kpc,$z \in [-4, 4]$ kpc, $\phi \in [-20°, 20°]$.		
mock data distribution,		b) $R \in [6, 10]$ kpc,$z \in [1, 5]$ kpc, $\phi \in [-20°, 20°]$.		
also in action space	N_*:	20,000		
Test 2 :	*Potential:*	`KKS-Pot`	-	Figure 3
Numerical accuracy	*DF:*	`hot` qDF		
in calculation	*Survey volume:*	sphere around Sun, $r_{\max} = 0.2, 1, 2, 3$ or 4 kpc		
of the likelihood	*Numerical accuracy:*	$N_x \in [5, 20]$, $N_v \in [6, 40]$, $n_\sigma \in [3.5, 7]$		
normalisation				
Test 3.1 :	*Potential:*	`Iso-Pot`	`Iso-Pot`, all parameters free	Figure 5
pdf is a	*DF:*	`hot` qDF	qDF, all parameters free	
multivariate	*Survey Volume:*	sphere around Sun, $r_{\max} = 2$ kpc	(fixed & known)	
Gaussian	N_*:	20,000		
for large data sets.	*Numerical accuracy:*		$N_v = 20$ and $n_\sigma = 4$	
Test 3.2 :	*Potential:*	`Iso-Pot`	`Iso-Pot`, free parameter: b	Figure 6
Width of the	*DF:*	`hot` qDF	`hot` qDF, free parameters:	
likelihood scales			$\ln h_R, \ln \sigma_{R,0}, \ln h_{\sigma,R}$	
with number of stars	*Survey volume:*	sphere around Sun, $r_{\max} = 3$ kpc	(fixed & known)	
by $\propto 1/\sqrt{N_*}$.	N_*:	between 100 and 40,000		
	Analysis method:		likelihood on grid	
	Numerical accuracy:		$N_v = 20$ and $n_\sigma = 4$ (for speed)	
Test 3.3 :	*Potential:*	`Iso-Pot`	`Iso-Pot`, free parameter: b	Figure 7
Parameter estimates	*DF:*	`hot` or `cool` qDF	`hot/cool` qDF, free parameters:	
are unbiased.			$\ln h_R, \ln \sigma_{R,0}, \ln h_{\sigma,R}$	
	Survey volume:	5 spheres around Sun, $r_{\max} = 0.2, 1, 2, 3$ or 4 kpc	(fixed & known)	
	N_*:	20,000		
	Analysis method:		likelihood on grid	
	Numerical accuracy:		$N_v = 20$ and $N_\sigma = 4$ (for speed)	
Test 4 :	*Potential:*	i) `Iso-Pot` or ii) `MW13-Pot`	i) `Iso-Pot`, all parameters free	Figure 8
Influence of			ii) `MW13-Pot`, R_d and f_h free	
position & shape	*DF:*	`hot` qDF	i) qDF, all parameters free	
of survey volume			ii) qDF, only h_R, $\sigma_{z,0}$ and $h_{\sigma,R}$ free	
on parameter recovery	*Survey volume:*	4 different wedges, see Figure 8, upper right panel	(fixed & known)	
	N_*:	20,000		
	Analysis method:		i) MCMC, ii) likelihood on grid	
Test 5 :	*Potential:*	`Iso-Pot`	`Iso-Pot`, all parameters free	Figure 10 & 11
Influence of	*DF:*	`hot` or `cool` qDF	qDF, all parameters free	
wrong assumptions	*Survey volume:*	sphere around Sun, $r_{\max} = 3$ kpc	(fixed & known)	
about the data set	*Completeness:*	radial incompleteness (see Figure 9),	data set complete,	
(in-)completeness		completeness(r) $= 1 - \epsilon_r \frac{r}{r_{\max}}$, $\epsilon_r \in [0, 0.7]$	completeness(r) = 1, $\epsilon_r = 0$	

Table 3 — *Continued*

Test		Model for Mock Data	Model in Analysis	Figure
on parameter recovery		$r \equiv$ distance from Sun,		
	N_*:	20,000		
Test 6.1 :	*Potential:*	`Iso-Pot`	`Iso-Pot`, all parameters free"	Figure 4
Numerical convergence	*DF:*	`hot` qDF	qDF, all parameters free	
of convolution	*Survey Volume:*	sphere around Sun, $r_{\rm max} = 3$ kpc	(fixed & known)	
with measurement	*Uncertainties:*	δRA = δDEC = $\delta(m-M) = 0$	Convolution with	
uncertainties		$\delta v_{\rm los} = 2$ km/s	perfectly known	
		$\delta\mu_{\rm RA} = \delta\mu_{\rm DEC}$ = 2,3,4 or 5 mas/yr	measurement uncertainties	
	Numerical Accuracy:		$N_{\rm error} \in [25, 1200]$	
	N_*:	10,000		
Test 6.2 :	*Potential:*	`Iso-Pot`	`Iso-Pot`, all parameters free	Figure 12
Testing the	*DF:*	`hot` qDF	qDF, all parameters free	
convolution	*Survey Volume:*	sphere around Sun, $r_{\rm max} = 3$ kpc	(fixed & known)	
with measurement & without	*Uncertainties:*	δRA = δDEC = 0,	Convolution with	
uncertainties with		$\delta v_{\rm los} = 2$ km s^{-1},	measurement uncertainties,	
distance errors		$\delta\mu_{\rm RA} = \delta\mu_{\rm DEC}$ = 1,2,3,4 or 5 mas/yr,	ignoring distance errors in position	
		a) $\delta(m-M) = 0$ or	(see Section 2.6)	
		b) $\delta(m-M) \neq 0$.		
		(see Figure 12)		
	Numerical Accuracy:		$N_{\rm error} = 800$ for $\delta\mu \leq 3$mas yr^{-1},	
			$N_{\rm error} = 1200$ for $\delta\mu > 3$mas yr^{-1}	
	N_*:	10,000		
Test 6.3 :	*Potential:*	`Iso-Pot`	`Iso-Pot`, all parameters free	Figure 14
Underestimation	*DF:*	`hot` or `cool` qDF	qDF, all parameters free	
of proper motion	*Survey volume:*	sphere around Sun, $r_{\rm max} = 3$ kpc	(fixed & known)	
errors	*Uncertainties:*	only proper motion errors	Convolution with proper motion errors	
		1, 2 or 3 mas/yr	10% or 50% underestimated	
	N_*:	10,000		
Test 7 :	*Potential:*	`Iso-Pot`	`Iso-Pot`, all parameters free	Figures 16 & 17
Deviations in the	*DF:*	mix of two qDFs (see Figure 15)	single qDF, all parameters free	
assumed DF		*Example 1:*		
from the		fixed qDF parameters		
star's true DF		(`hot` & `cooler` qDF from Table 2)		
		but different mixing rates.		
		Example 2:		
		50/50 mixture		
		with varying qDF parameters (by X%):		
		a) `hot` & `colder` qDF or		
		b) `cool` & `warmer` qDF (see Table 2)		
	Survey volume:	sphere around Sun, $r_{\rm max} = 2$ kpc	(fixed & known)	
	N_*:	20,000		
Test 8 :	*Potential:*	`MW14-Pot`	`KKS-Pot`, all parameters free,	potential contours:
Deviations of the			only $v_{\rm circ}(R_\odot) = 230$km s^{-1} fixed	Figure 19
assumed potential model	*DF:*	`hot` or `cool` qDF	qDF, all parameters free	qDF recovery:
from the star's	*Survey volume:*	sphere around Sun, $r_{\rm max} = 4$ kpc	(fixed & known)	Figure 20
true potential	N_*:	20,000		