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

W. Trick[1,2], J. Bovy[3,4], and H.-W. Rix[1]

trick@mpia.de

ABSTRACT

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

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

Contents

[1]Max-Planck-Institut für Astronomie, Königstuhl 17, D-69117 Heidelberg, Germany

[2]Correspondence should be addressed to trick@mpia.de.

[3]Institute for Advanced Study, Einstein Drive, Princeton, NJ 08540, USA

[4]Hubble fellow

1. Introduction

[TO DO]

Collection of thoughts for the introduction: *(Text is not yet perfect or concise, but should serve as a starting point to setup a basic structure for the introduction. The text will then have to be shortened, redundant formulations have to be removed, phrasing has to be improved and everything has to be supported with appropriate references.)*

- RoadMapping stands for "Recovery of the Orbit/Action Distribution of Mono-Abundance Populations and Potential INference for our Galaxy".
- **Our modelling method in a nutshell:** We fit simultaneously a model for the Galaxy's gravitational potential and an orbit distribution function (df) to stellar phase-space data. To turn a star's position and velocity into a full orbit, we need the gravitational potential in which the star moves. We assume that we know a family of orbit distribution functions that are close enough to the real distribution of orbits. In this case the stellar orbits calculated within a proposed potential will only follow such a df, if this potential model is close enough to the true potential.
  Or in other words: We need the potential to calculate orbits. At the same time, if we *know* the true orbits, we can deduce the true potential from them. To find the true orbits, we make use of the predictive power of an orbit distribution function.
- **Introducing orbits and actions:** There are different ways to describe stellar orbits. The most obvious is to give the stars position and velocity vector at each point in time, by evaluating the potential forces that act on the star in each time step. Most orbits in realistic galaxy potentials are however not closed, so we would have to integrate the orbit forever. Another, much more convenient way to describe orbits, are so called integrals of motion. These integrals are functions of the star's time-dependent position and velocity, but are themselves constants in time, i.e. conserved quantities. The most obvious integral in static potentials is the energy of the orbit. Symmetries in potentials frequently allow more than one integral: In spherical potentials all three components of the angular momentum are conserved. In many axisymmetric potentials there is, in addition to the energy E and vertical component of the angular momentum L_z, a third non-classical integral of motion I_3, which has however no easy physical meaning. (Binney & Tremaine, Galactic Dynamics)
  Because any function of integrals is an integral of motion itself, it is possible to construct integrals that have both very convenient properties and intuitive physical meanings.

One such a set are the so-called actions. In axisymmetric potentials they are frequently called the radial action J_R, the vertical action J_z and the ϕ-action, which is simply the vertical component of the angular momentum, L_z. The radial action and vertical action quantify the amount of oscillation in radial and vertical direction that the orbit exhibits. Actions are constructed in such a way, that they are not only integrals, but also correspond to the momenta in a set of canonical coordinates. The canonical conjugate positions of the actions are the so-called angles, which have the convenient properties, that they increase strictly linearly in time while the star moves along the orbit. They are periodic in 2π and the frequencies by which they change are functions of the actions. In the action-angle coordinate system, the only thing we need to fully describe an orbit in an axisymmetric potential are therefore just three fixed numbers, the actions.

- **Using actions for distribution functions:** Actions are therefore the natural coordinates of orbits and each point in action space corresponds to one specific orbit in a given potential. It is often used in dynamical modelling, e.g. in the Schwarzschild superposition method (source???), to reconstruct a galaxy by superimposing different orbits and populating them with stars. In this way these kind of methods construct orbit distribution functions for galaxies, which are at the same time distribution functions in action space. Because angles increase linearly in time, when a star moves along its orbit, stars are uniformly distributed in angle space. Therefore a orbit distribution function in terms of actions and a uniform distribution of stars in angle-space can be directly mapped to a distribution of stars in canonical configuration phase-space, measurable stellar positions and velocities. While a stellar distribution in configuration space is six-dimensional, the distribution in action-angle space is effectively three-dimensional, because of the uniformity in angles. (Rewrite, too verbose...)

- **Motivation to use Binney's qDF in the modelling:** Astronomers consider galaxies frequently as a superposition of several components. The stellar component itself is often separated into bulge, halo and several disk sub-populations. Distribution functions in terms of actions have the advantage, that a full distribution function for the whole galaxy can be constructed by a superposition of action-based DFs for each component. [TO DO: This was what Payel Das told me, but I forgot why this is the case and I also didn't find any reference for this. ???] Assuming a DF with a simple form for each galaxy component can give, in superposition, very realistic looking, flexible and successful models for the distribution of stars in galaxies (Bovy & Rix 2013; Sanders & Binney 2015; Piffl et al. 2014) (other references???). Any dynamical modelling approach still depends crucially the assumption one makes about the structure of the galaxy and on the choices for the DFs:

The structure of the MW disk is still under debate. While many still support the thin-thick disk dichotomy in the MW disk (references ???), Bovy et al. (2012b) found indications that the MW disk might actually be a super-position of many stellar subpopluations with a continuous spectrum of scale heights, scale lengths, metallicity and [α/Fe] abundances (dubbed mono-abundance populations (*MAP*s)). Further investigation lead to the findings that *MAP*s in the MW disk have a simple spatial structure that follows an exponential in both radial and vertical direction (Bovy et al. 2012d). The corresponding velocity dispersion profile of the *MAP*s also decreases exponentially with radius and is nearly independent of height above the plane, i.e. quasi-isothermal (Bovy et al. 2012c). The radial decrease in vertical velocity dispersion has, according to Bovy et al. (2012c), a long scale length of $h_{\sigma,z} \sim 7$ kpc for all *MAP*s . Older *MAP*s , which are characterized by lower metallicities and [α/Fe] abundances, have in general shorter density scale lengths, larger scale heights and velocity dispersion (Bovy et al. 2012d). Ting et al. (2013) and Bovy & Rix (2013) finally proposed that these findings could be employed for dynamical modelling techniques using action-based distribution functions. An action-based distribution function, that is flexible enough to describe the spectrum of simple phase-space distributions of different *MAP*s , is the quasi-isothermal distribution function (qDF) by Binney & McMillan (2011), as demonstrated by Ting et al. (2013).

- **Some caveats of DF assumptions as compared to others:** Sanders & Binney (2015) and ??? develop extended distribution functions (EDFs), that extend action-based DFs to also describe the distribution of the star's metallicities. While a full chemo-dynamical modelling, including metallicity as well as α- and other chemical abundances, is ultimately the right way to go, the form of the EDFs still depends on a lot of additional assumptions. By looking at fig. 6 in Bovy & Rix (2013) (other references???) we doubt that a final version of an EDF will have a simple form in action-metallicity space. Motivated by the findings by Bovy et al. 2012, we therefore resign to the simpler approach outlined in Bovy & Rix (2013) and here, were metallicity and α-abundances are implicitly taken into account by describing each MAP separately by one qDF. This procedure could have two caveats:
  First, the binning of the stars according to their abundances could lead to pollution of one MAP, by either choosing the bin sizes too large, or too small compared to the stars' inherent abundance errors.
  Second, while Ting et al. (2013) makes us confident that the qDF is indeed a good functional form to describe each MAP, it could very well be, that the stars' true distribution is close to but not exactly of the family of assumed qDFs.
  Some comments by HWR regarding Sanders & Binneys take on our mod-

elling, should be also included: [TO DO]

- Overall, there is no doubt that making a simultaneous model for the "chemo-orbital-potential" distribution has some advantages over the "orbital-potential" distribution at a given abundance. The main advantage for pursueing MAP modelling at least as a first/intermediate step is: a) it separates out complexity (i.e. it's much easier to "see" what goes right or wrong), b) it provides true cross-checking redundancy w.r.t. to the potential estimates [TO DO: I don't understand the latter]
- "First, choosing bin sizes always requires a compromise between losing the information contained in the position of each datum within its bin and increasing Poisson noise by making the bins small." –¿ That is true for any binning. But with the realistic samples sizes, bins within which the abundances vary "little" have a sensible number of stars (for SEGUE)
- It is true that the MAP approach does not exploit that the abundance space distribution is "smooth"; however, the data show that there is no "simple and large-scale" pattern that lends itself to a simple functional form.
- " Third, we require errors in the ([Fe/H], [alpha/Fe]) space that are much smaller than the bin sizes, otherwise we are neglecting the possibility of contamination on each bin by neighbouring bins." –¿ I don't think the argument is valid; it wouldn't make sense to make the bins SMALLER than the errors, because then you would reduce the samples size and increase the shot-noise, WITHOUT making the approximation to the model DF better. You could make the bin size larger (therefore the error smaller than the bin-size), but would pay the prize of a poorer approximation. So, I actually would think that making the bin size of order of the abundance error is a sensible choice, if that leaves you with "enough" objects in the bin.
- " Additionally, a continuous parametrization allows for a rigorous treatment of the error distributions in ([Fe/H], [alpha/Fe]) and how these errors correlate with the kinematic errors. Hence we believe that it is best to work with an EDF provided we are confident that we have a sufficiently flexible and well-tailored functional form." –¿I would agree with that statement but a) it's not easy to get a simple form for that, see sigma_z(FeH,aFe) Figures in B12; and the redundance argument from above applies...

- **Why should we care about actions in realistic galaxies?** In reality galaxies have rarely perfectly static and axisymmetric potentials, which drastically reduces the number of conserved quantites along orbits. In static non-axisymmetric potentials

there can still be two integrals of motion, angular momentum however is no longer conserved. The Milky Way's disk might have an overall axisymmetric appearance, but is perturbed by spiral arms. The strongest deviation from axisymmetry in the Galaxy is the bar, which also causes the Galactic potential to vary slowly in time. The stirrs up the stars of the disk and the potential and causes radial migration of the orbits (Reference???), orbits change and with them the actions. One could wonder if, under such non-axisymmetric, non-static potential conditions, the assumption and treatment of globally conserved actions in the Milky Way is still a sensible approach. First of all, actions are the natural way to treat orbits and they can be locally defined, even if they might not be globally conserved. As long as we care about orbits, we should care about actions. An orbit carries information about the star's past, about where the star was born and which tidal processes might have carried it away from its inital orbit. Together with the chemistry of the stars, which determined by their place of birth, their current orbits are valuable diagnostics for the evolution and structure of the Milky Way. Secondly, gravitational processes do only in the most extreme cases completely change the actions. In a slowly changing potential, where orbits adapt adiabatically to those changes, actions are conserved (Binney & Tremaine, Galactic Dynamics). And even during bar-induced radial migration at least the vertical actions are conserved and will continue to carry some amount of information about the stars' inital orbit distribution.

[TO DO] (Maybe cite Potzen 2015, who showed that analysing aspherical systems in spherical actions can still be a powerful tool, when used with care...)

- **Why should we care about an axisymmetric "best fit" model for the Milky Way disk?** One of the key assumptions of our modelling technique is the assumed axisymmetry of the Milky Way's gravitational potential, especially its disk. As we discussed already in the previous paragraph, this assumption is indeed only an approximation to the real disk, which has a much richer structure and more complicated potential, with spiral arms and ring-like structures (like the Monocerros ring), with a warp and a flare in the outer disk (references????). Also the Milky Way's halo has substructure, a multitude of streams (references???) and shell-like overdensities (reference???). The ultimate goal will be to find and identify substructures observationally and describe theoretically the structure and evolution of potential perturbations. Our method and efforts to extract information about the axisymmetric Milky Way potential from disk stars aims to create a reliable and well-constraint basis for these endevours: The best possible axisymmetric approximation to the Milky Way's potential could serve as a realistic equilibrium model from which a description of non-axisymmetric tidal perturbations can be theoretically established by perturbation theory. It will

also help a great deal to identify sub-structures, e.g. to find and orbitally connect tidal streams, which in return will then give better constraints on the deviations from axisymmetry. Many modelling and techniques, both purely gravitational, but also chemo-dynamical, can greatly profit from a good axisymmetric model for the galaxy: While we are still far away from knowing the MW's potential all over the place, an axisymmetric model will be the best reference to turn phase-space coordiantes into whole orbits. And orbits are the diagnostics that carry information from everywhere in the galaxy into the solar neighbourhood, where we can hope to exploit them. (Some overlap with section before. How to better structure these two sections and assign the arguments more clearly to "axisymmetric disk" or "actions"?)

- **Previous results with this modelling technique:** Bovy & Rix (2013) ... [TO DO]
  - disk scale length $R_d = 2.15 \pm 0.14$ kpc (Bovy & Rix 2013)
  - disk is maximal (Bovy & Rix 2013)
  - slope of dark matter halo $\alpha < 1.53$ (Bovy & Rix 2013)
- **What do we already know about the axisymmetric MW disk (from other references)?** [TO DO]
  - rotation curve is well-known (reference???)
- **What is there left to learn about the axisymmetric MW disk?** (as Jo asked at the Santa Barbara conference... [TO DO]
  - separation of different MW component is still unclear: individual density profiles, contributions to total pot
  - thin/thick disk vs. continuum of exponential disks
  - dark matter at smaller radii
  - slope & shape of dark matter halo (current state of knowledge?)
- **Other modelling approaches:**
  - Piffl et al. (2014) used a slightly different DF-based modelling approach to constrain the MW's vertical density profile near the sun. They fitted a superposition of "quasi-isothermal" DFs for thick and thin disk, and a DF for the halo to ˜200,000 giant stars from the RAVE survey (RAdial Velocity Experiment, Steinmetz et al. (2006)). They didn't use any chemical information of the stars. To account for different populations within the thin disk, they weighted the corresponding DF's with an assumed star-formation rate instead. To circumvent the

use of RAVE's non-trivial spatial selection function, they separated stars into spatial bins in (R, z) and fitted the velocity distribution predicted by their DF and potential model at the mean (R, z) of each bin to the observed velocities only. They're result for their radial profile of the vertical force within $|z| = 1.1$ kpc and $R > 6.6$ kpc agrees well with the previous results from our method by Bovy & Rix (2013). By not using chemical information and hiding the spatial distribution of stars by binning to circumvent a complicated selection function, Pifffl et al. (2014) is however rejecting a lot of valuable information in the data set. ([TO DO: Look at other useful references in this paper: Bienayme et al. 2014, Zhang et al. 2013, Binney et al. 2014a, Binney 2012b, McMillan & Binney 2013])

- **Motivating this method characterization in anticipation of GAIA:** [TO DO]
- **Ideas how to structure this introduction:**
  - Part I: Basic task is fitting potential and DF at the same time. This is a great, useful and successful way to constrain the galactic gravitational potential. Was already done in Bovy & Rix (2013).
  - Part II: While Bovy & Rix (2013) were successful in their application, they made many approximations / assumptions / idealisations, which were not tested for their valicity and might actually not hold up well. We want to investigate this. We cannot test everythong, but we show some plausible and illustrative examples (often using a spherical isochrone potential for convenience). (Also mention what Sanders & Binney (2015) say about this modelling approach.)

2. Dynamical Modelling

2.1. Model

2.1.1. Actions

[TO DO]

2.1.2. Potential models

[TO DO: Mention different ways to calculate actions in different potentials.] [TO DO: Mention that the potential parameters are denoted by p_Φ]

2.1.3. Distribution function

Motivated by the findings of Bovy et al. 2012??? and Ting et al. (2013) about the simple phase-space structure of *MAP*s (see §1), and following Bovy & Rix (2013) and their successful application, we also assume that each *MAP* follows a single qDF of the form given by Binney & McMillan (2011). This qDF is a function of the actions $\boldsymbol{J} = (J_R, J_z, L_z)$ and

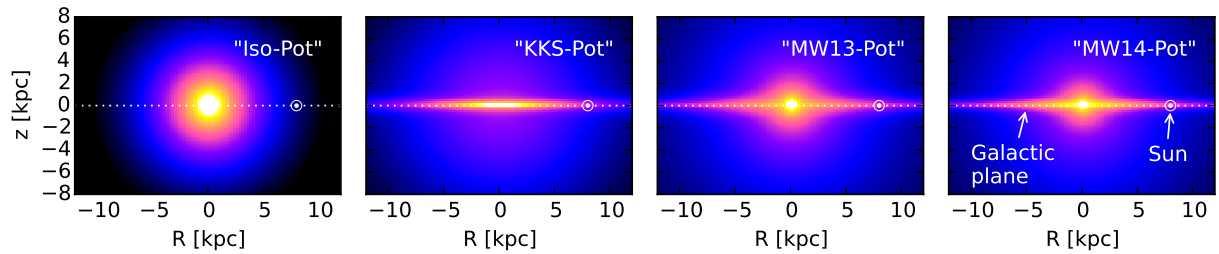

Fig. 1.— Density distribution of the four reference galaxy potentials in table 1, for illustration purposes. These potentials are used throughout this work for mock data creation and potential recovery. [TO DO: Halo sichtbarer machen, evtl. mit isodensity contours]

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. In the subsequent analyses we aim to recover these potential parameters again. All reference potentials assume the sun to be located at $(R_\odot, z_\odot) = (8\ \mathrm{kpc}, 0)$.

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

[a]The coordinate system of each of the two Stäckel-potential components is $\frac{R^2}{\tau_{i,p}+\alpha_p} + \frac{z^2}{\tau_{i,p}+\gamma_p} = 1$ with $p \in \{\mathrm{Disk}, /\mathrm{Halo}\}$ and $\tau_{i,p} \in \{\lambda_p, \nu_p\}$. Both components have the same focal distance $\Delta = \sqrt{\gamma_p - \alpha_p}$, to make sure that the superposition of the two components itself is still a Stäckel potential. The axis ratio of the coordinate surfaces $\left(\frac{a}{c}\right)_p := \sqrt{\frac{\alpha_p}{\gamma_p}}$ describes the flattness of the corresponding Stäckel component.

has the form

$$\begin{aligned}
\mathrm{qDF}(\boldsymbol{J} \mid p_{\mathrm{DF}}) &= f_{\sigma_R}\left(J_R, L_z \mid p_{\mathrm{DF}}\right) \times f_{\sigma_z}\left(J_z, L_z \mid p_{\mathrm{DF}}\right) && (1)\\
\text{with } f_{\sigma_R}\left(J_R, L_z \mid p_{\mathrm{DF}}\right) &= n \times \frac{\Omega}{\pi\sigma_R^2(R_g)\kappa}\left[1+\tanh\left(L_z/L_0\right)\right]\exp\left(-\frac{\kappa J_R}{\sigma_R^2(R_g)}\right) && (2)\\
f_{\sigma_z}\left(J_z, L_z \mid p_{\mathrm{DF}}\right) &= \frac{\nu}{2\pi\sigma_z^2(R_g)}\exp\left(-\frac{\nu J_z}{\sigma_z^2(R_g)}\right) && (3)\\
& && (4)
\end{aligned}$$

Here $R_g \equiv R_g(L_z)$ and $\Omega \equiv \Omega(L_z)$ are the (guidig-center) radius and the circular frequency of the circular orbit with angular momentum L_z in a given potential. $\kappa \equiv \kappa(L_z)$ and $\nu \equiv \nu(L_z)$ are the radial/epicycle (κ) and vertical (ν) frequencies with which the star would oscillate around the circular orbit in R- and z-direction when slightly perturbed (Binney & Tremaine 2008). The term $[1+\tanh(L_z/L_0)]$ suppresses counter-rotation for orbits in the disk with $L \gg L_0$ which we set to a random small value ($L_0 = 10 \times R_\odot/8 \times v_{\mathrm{circ}}(R_\odot)/220$).
For this qDF to be able to incorporate the findings by Bovy et al. 2012??? about the phase-space structure of *MAP*s summarized in §1, we set the functions n, σ_R and σ_z, which indirectly set the stellar number density and radial and vertical velocity dispersion profiles,

$$\begin{aligned}
n(R_g \mid p_{\mathrm{DF}}) &\propto \exp\left(-\frac{R_g}{h_R}\right) && (5)\\
\sigma_R(R_g \mid p_{\mathrm{DF}}) &= \sigma_{R,0} \times \exp\left(-\frac{R_g - R_\odot}{h_{\sigma_R}}\right) && (6)\\
\sigma_z(R_g \mid p_{\mathrm{DF}}) &= \sigma_{z,0} \times \exp\left(-\frac{R_g - R_\odot}{h_{\sigma_z}}\right). && (7)
\end{aligned}$$

The qDF for each *MAP* has therefore a set of five free parameters p_{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_{σ_R} and h_{σ_z}, that describe the radial decrease of the velocity dispersion. The *MAP*s we use for illustration through out this work are summarized in Table 2.

One crucial point in our dynamical modelling technique (§???), as well as in creating mock data (§2.2), is to calculate the (axisymmetric) spatial tracer density $\rho_{\mathrm{DF}}(\boldsymbol{x} \mid p_\Phi, p_{\mathrm{DF}})$ for a given qDF and potential . We do this by integrating the qDF at a given (R, z) over all three

velocity components, using a $N_{\rm velocity}$-th order Gauss-Legendre quadrature for each integral:

$$\rho_{\rm DF}(R, |z| \mid p_\Phi, p_{\rm DF}) = \int_{-\infty}^{\infty} {\rm qDF}(\boldsymbol{J}[R, z, \boldsymbol{v} \mid p_\Phi] \mid p_{\rm DF})\, {\rm d}^3\boldsymbol{v} \tag{8}$$

$$\approx \int_{-N_{\rm sigma}\sigma_R(R|p_{\rm DF})}^{N_{\rm sigma}\sigma_R(R|p_{\rm DF})} \int_{-N_{\rm sigma}\sigma_z(R|p_{\rm DF})}^{N_{\rm sigma}\sigma_z(R|p_{\rm DF})} \int_0^{1.5 v_{\rm circ}(R_\odot)} {\rm qDF}(J[R, z, \boldsymbol{v} \mid p_\Phi] \mid p_{\rm DF})\, {\rm d}v_T\, {\rm d}v_z\, {\rm d}v_R, \tag{9}$$

where $\sigma_R(R \mid p_{\rm DF})$ and $\sigma_z(R \mid p_{\rm DF})$ are given by eq. (6) and (7) and the integration ranges are motivated by Fig. 2. For a given p_Φ and $p_{\rm DF}$ we explicitly calculate the density on $N_{\rm spatial} \times N_{\rm spatial}$ regular grid points in the (R, z) plane; in between grid points the density is evaluated with a bivariate spline interpolation. The grid is chosen to cover the extent of the observations for $z > 0$. The total number of actions that need to be calculated to set up the density interpolation grid is $N_{\rm spatial}^2 \cdot N_{\rm velocity}^3$. Fig. ??? shows the importance of choosing $N_{\rm spatial}$, $N_{\rm velocity}$ and $N_{\rm spatial}$ sufficiently large in order to get the density with an acceptable numerical accuracy.

[TO DO: Rename everywhere $N_{\rm sigma}$ to $n_{\rm interval}$ or something like this.]

2.1.4. Selection function: observed volume and completeness

[TO DO]

Table 2. Reference distribution function parameters for the qDF in eq. (1)-(7). These qDFs describe the phase-space distribution of stellar *MAP*s for which mock data is created and analysed throughout this work for testing purposes. The parameters of the "cooler" ("hotter") *MAP*s were chosen such, that the they have the same σ_R/σ_z ratio as the "hot" ("cool") *MAP* . Hotter populations have shorter tracer scale lengths (Bovy et al. 2012d) and the velocity dispersion scale lengths were fixed according to Bovy et al. (2012c).

name of *MAP*	qDF parameters $p_{\rm DF}$				
	h_R [kpc]	σ_R [km s^{-1}]	σ_z [km s^{-1}]	h_{σ_R} [kpc]	h_{σ_z} [kpc]
"hot"	2	55	66	8	7
"cool"	3.5	42	32	8	7
"cooler"	2 +50%	55-50%	66-50%	8	7
"hotter"	3.5-50%	42+50%	32+50%	8	7

2.2. Mock Data

One goal of this work is to test how the loss of information in the process of measuring stellar phase-space coordinates can affect the outcome of the modelling. To investigate this, we assume first that our measured stars do indeed come from our assumed families of potentials and distribution functions and draw mock data from a given true distribution. In further steps we can manipulate and modify these mock data sets to mimick observational effects.
The distribution function is given in terms of actions and angles. The transformation $(\boldsymbol{J}_i, \boldsymbol{\theta}_i) \longrightarrow (\boldsymbol{x}_i, \boldsymbol{v}_i)$ is however difficult to perform and computationally much more expensive than the transformation $(\boldsymbol{x}_i, \boldsymbol{v}_i) \longrightarrow (\boldsymbol{J}_i, \boldsymbol{\theta}_i)$. We propose a fast and simple two-step method for drawing mock data from an action distribution function, which also accounts effectively for a given survey selection function.

Preparation: Tracer density. We first setup the interpolation grid for the tracer density $\rho(R, |z| \mid p_\Phi, p_{\mathrm{DF}})$ generated by the given qDF and according to §2.1.3 and Eq. 9. For the creation of the mock data we use $N_{\mathrm{spatial}} = 20$, $N_{\mathrm{velocity}} = 40$ and $N_{\mathrm{sigma}} = 5$.

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

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

Step 2: Drawing velocities according to the distribution function. The velocities are independent of the selection function and observed volume. For each of the positions (R_i, z_i) we now sample velocities directly from the $\mathrm{qDF}(R_i, z_i, \boldsymbol{v} \mid p_{Phi}, p_{\mathrm{DF}})$ using a rejection method. To reduce the number of rejected velocities, we use a Gaussian in velocity space as an envelope function, from which we first randomly sample velocities and then apply the rejection method to shape the Gaussian velocity distribution towards the velocity distribution predicted by the qDF. We now have a mock data set according to the required:

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

[TO DO: mention fig. 2. ???]

Introducing measurement errors. If we want to add measurement errors to the mock data, we need to apply two modifications to the above procedure.
First, measurement errors are best described in the phase-space of observables. We use the heliocentric coordinate system right ascension and declination (α, δ) and distance modulus $(m-M)$ as proxy for the distance from the sun, the proper motion in both α and δ direction (μ_α, μ_δ) and the line-of-sight velocity $v_{\rm los}$. For the conversion between these observables and the Galactocentric cylindrical coordinate system in which the analysis takes place, we need the position and velocity of the sun, which we set for simplicity in this study to be $(R_\odot, z_\odot) = (8, 0)$ kpc and $(v_R, v_T, v_z) = (0, 230, 0)$ km s^{-1}. We assume Gaussian measurement errors in the observables $\tilde{\boldsymbol{x}} = (\alpha, \delta, (m-M)), \tilde{\boldsymbol{v}} = (\mu_\alpha, \mu_\delta, v_{\rm los})$.
Second, in the case of distance errors, stars can virtually scatter in and out of the observed volume. To account for this, we first draw "true" positions from a volume that is larger than the actual observation volume, perturb the stars positions according to the distance errors and then reject all stars that lie now outside of the observed volume. This procedure mirrors the Poisson scatter around the detection threshold for stars whose distances are determined from the apparent brightness and the distance modulus. [TO DO: Can I say it like this???] We then sample velocities (given the "true" positions of the stars) as described above and perturb them according to the measurement errors as well.

[TO DO] Possible plots: *Diagram*: schematic flow chart of how to sample mock data (could be helpful for people, who want to sample mock data in action space and didn't know how to start, like me)

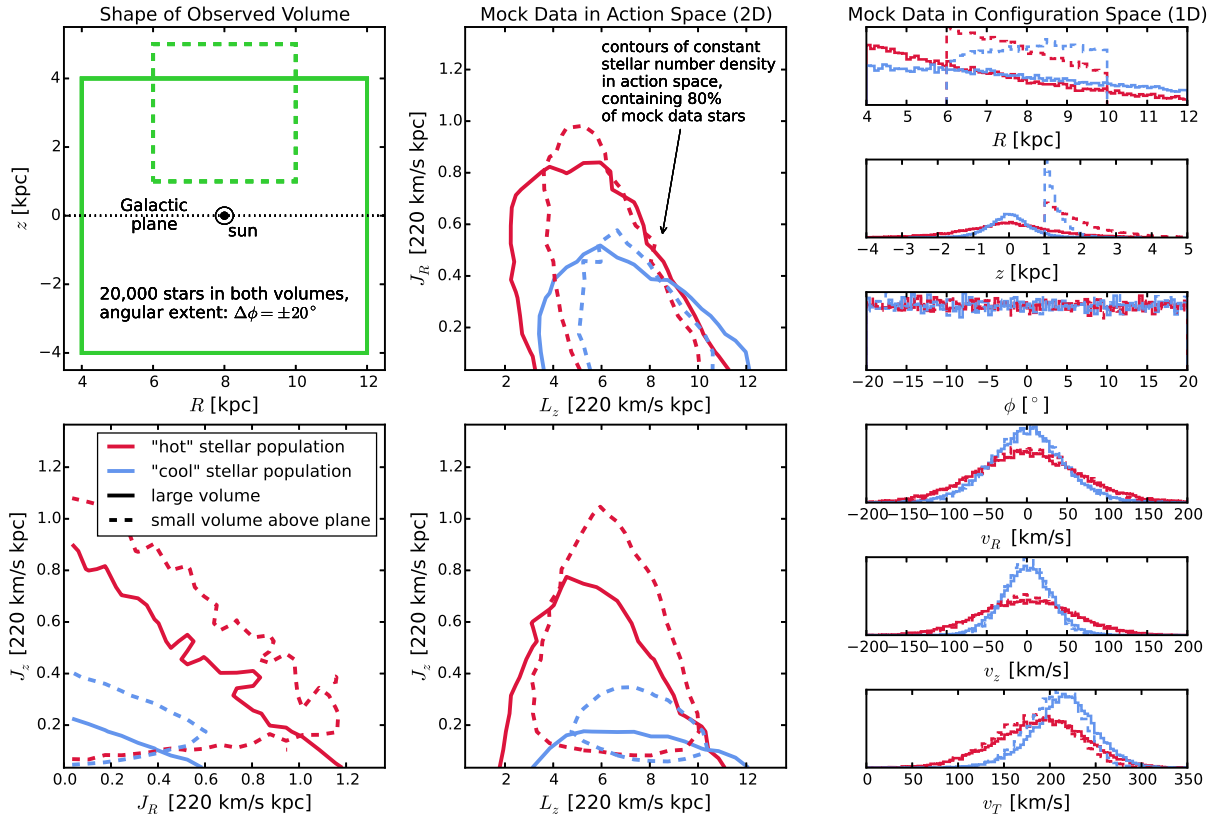

Fig. 2.— Distribution of mock data in action space (2D iso-density contours enclosing 80% of the stars, in the two central and the lower left panel) and configuration space (1D histograms in right panels), depending on shape and position of observation volume and temperature of the stellar population. The parameters of the mock data model is given as Test ① in Table 3. In the upper left panel we demonstrate the shape of the two different observation volumes within which we were creating each a "hot" (red) and "cool" (blue) mock data set: a large volume centered on the Galactic plane (solid lines) and a smaller one above the plane (dashed lines). The stars of the "cool" population have in general lower radial and vertical actions, i.e. are on more circular orbits. The different ranges of L_z's in the two volumes reflect $L_z \sim Rv_{\rm circ}$ and the different radial extent of both volumes.The volume above the plane contains no stars with $J_z = 0$ and more with J_z: The higher a volume is located above the plane, the larger J_z has to be for the star's orbit to cross this volume. Circular orbits with $J_R = 0$ and $J_z = 0$ can obviously only be observed in the Galactic mid-plane. The smaller an orbit's L_z, the smaller also its mean orbital radius. For this orbit to be able to reach into a volume located at larger Galactocentric radius, it needs to be more eccentric and therefore have a larger J_z. This anti-correlation between L_z and J_R can be seen in the top central panel. Orbits with both large J_R and large J_z would be very energetic and are therefore less likely to be observed.

2.3. Likelihood

The idea behind our modeling approach is that the orbits of the stars belonging to one MAP [TO DO: explain MAP???], calculated from a phase-space observation for each star within a proposal potential, will only follow a distribution function from the family of qDFs (cf. §2.1.3) if this propsal potential is (close to) the true potential in which the stars move. This opens up the possibility to fit the qDF and the potential simultaneously to the stellar phase-space data of one MAP, using the orbits of the stars.

2.3.1. Data and Selection Function

We're fitting the potential and the qDF to the data

$$D_j = \{\boldsymbol{x}_i, \boldsymbol{v}_i \mid (\text{star } i \text{ belonging to MAP } j) \wedge (\text{sf}(\boldsymbol{x_i}) > 0)\},$$

where $\boldsymbol{x}_i$ and $\boldsymbol{v}_i$ are the position and velocity of one star. The phase-space volume within which stars are observed by a given survey is defined by the survey's selection function $\text{sf}(\boldsymbol{x}, \boldsymbol{v})$, which is in general a function of the position only, $\text{sf}(\boldsymbol{x})$. To first order the shape of the selection function ("observed volume") is limited by the directions in which the survey is pointed and the sensitivity down to which limiting magnitude it can detect stars. In the simplest case, if all stars had the same brightness, the selection function is 1 everywhere inside the observed volume and 0 outside. Because stars have different brightness the selection function will usually decrease from 1 close to the sun to 0 at the edges of the observed volume ("completeness"). [TO DO: Explain selection function somewhere else????] Only stars for which the selection function is non-zero are contained in the data set D_j.
Our modeling takes place in the Galactocentric rest-frame with cylindrical coordinates $\boldsymbol{x} = (R, \phi, z)$ and velocity components in the corresponding coordinate directions $\boldsymbol{v} = (v_R, v_\phi, v_z)$.[1]

2.3.2. Model Parameters

We fit the five free parameters of the qDF family, h_R, σ_R, σ_z, h_{σ_R} and h_{σ_z}, in logarithmic scale, which corresponds to a logarithmically flat prior in the framework of Bayesian

[1]If the phase-space data is given in observed coordinates, position $\tilde{\boldsymbol{x}} = (\alpha, \delta, m - M)$ in right ascension α, declination δ and distance modulus $m - M$ and velocity $\tilde{\boldsymbol{v}} = (\mu_\alpha, \mu_\delta, v_{\text{los}})$ as proper motions $\boldsymbol{\mu} = (\mu_\alpha, \mu_\delta)$ [TO DO: cos somwhere???] 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 (cf. §???).

statistics. The set of qDF fit parameters is therefore

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

To be able to control the number of degrees of freedom in the potential fit, we have to assume a certain family of potential models, parametrized by the parameters p_Φ (cf. §2.1.2).
The total set of model parameters to fit is then

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

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 qDF evaluated for the i-th star is then $\mathrm{qDF}(\boldsymbol{J}_i \mid M) := \mathrm{qDF}(\boldsymbol{J}[\boldsymbol{x}_i, \boldsymbol{v}_i \mid p_\Phi] \mid p_{\rm DF})$.

2.3.3. Form of the Likelihood

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

$$\mathscr{L}(M \mid D_j) = \prod_i^{N_j} P(\boldsymbol{x}_i, \boldsymbol{v}_i \mid M), \tag{10}$$

where N_j is the number of stars in the data set D_j. This probability is, properly normalized and in the correct units,

$$\begin{aligned} P(\boldsymbol{x}_i, \boldsymbol{v}_i \mid M) &= \frac{1}{\left(r_o v_o\right)^3} \cdot \frac{\mathrm{qDF}(\boldsymbol{J}_i \mid M) \cdot \mathrm{sf}(\boldsymbol{x}_i)}{\int \mathrm{d}^3x\, \mathrm{d}^3v\ \mathrm{qDF}(\boldsymbol{J} \mid M) \cdot \mathrm{sf}(\boldsymbol{x})} \\ &\propto \frac{1}{\left(r_o v_o\right)^3} \cdot \frac{\mathrm{qDF}(\boldsymbol{J}_i \mid M)}{\int \mathrm{d}^3x\ \rho_{\rm DF}(R, |z| \mid M) \cdot \mathrm{sf}(\boldsymbol{x})}. \end{aligned} \tag{11}$$

In the second step we used eq. (8). The factor $\prod_i \mathrm{sf}(\boldsymbol{x}_i)$ is independent of the model parameters, so we use simply eq. (11) in the likelihood calculation. We find the best set of model parameters by maximising the likelihood.

2.3.4. A word on units

We evaluate the likelihood in a scale-free potential within a Galactocentric coordinate system which is defined as $v_{\rm circ}(R = 1) = 1$. $v_{\rm circ}(R_\odot = 8.{\rm kpc}) \sim 230{\rm km\ s^{-1}}$ is the Galaxy potential parameter that determines the total Galaxy mass / amplitude of the potential.

To switch into our modelling coordinate frame, we first have to re-scale the data and the model parameters: all spatial coordinates to units of $r_o := R_\odot$ and all velocities to units of $v_o := v_{\rm circ}(R_\odot)$. The prefactor $1/\left(r_o v_o\right)^3$ in eq. (11) makes sure that the likelihood has the correct units to satisfy:

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

Including this prefactor is crucial when $v_{\rm circ}(R_\odot)$ is a free fitting parameter.

2.3.5. Numerical accuracy in calculating the likelihood

[TO DO: Consistent capitals in section titles. ????]

To evaluate the likelihood at a given set of $(p_\Phi, p_{\rm DF})$ we proceed in principle in the following way: The numerator in eq. (11) can be calculated straightforward by calculating the actions of each star in the given potential (cf. §???) and then evaluating the qDF at each action. For the normalisation of the likelihood we first have to calculate the density $\rho_{\rm DF}(R, |z| \mid M)$ on a grid as described in §**??**. The density is then interpolated using bivariate spline interpolation. In the case of $\mathrm{sf}(\boldsymbol{x}) = 1$ everywhere inside the observed volume and $\mathrm{sf}(\boldsymbol{x}) = 0$ outside, i.e. for a complete sample, the integral in the normalisation in eq. (11) is essentially two-dimensional in R and z and we can use the shape of the observed volume to set finite integration limits. We perform this integral over the interpolated tracer density by using Gauss Legendre integration of order 40 in each R and z direction. The integration over ϕ is done analytically.
Unfortunately the evaluation of the likelihood for only one set of model parameters is already very computationall expensive. The computation speed is set by the number of action calculations needed, i.e. the number of stars and the numerical accuracy of the integrals in the normalisation, which requires $N^2_{\rm spatial} \times N^3_{\rm velocity}$ action calculations. The numerical accuracy has to be choosen high enough, such that the integrals in the normalisation are mostly converged and the error introduced by this does not dominate in the likelihood, i.e.

$$\begin{aligned} \log \mathscr{L}(M \mid D_j) \;&=\; \sum_i^{N_j} \log \mathrm{qDF}(\boldsymbol{J_i} \mid M) \\ &\quad -N_j \log(\text{true normalisation}) - N_j \log(1 + \text{rel. error}), \qquad (12) \\ \text{with} \quad &\quad N_j \log(1 + \text{rel. error}) \lesssim 1. \end{aligned}$$

[TO DO: Don't understand why 1 is the threshold here. ???] For data sets as large as $N_j =$ 20,000 stars in one MAP, which in the age of GAIA could very well be the case [TO DO:

Really???], we would need a numerical accuracy of 0.005% in the normalisation. Fig. 3 demonstrates that the numerical accuracy we use in the analysis, $N_{\rm spatial} = 16$, $N_{\rm velocity} = 24$ and $N_{sigma} = 5$, does satisfy this requirement.[2] [TO DO: Should we also show that 40th order GL integraiton over interpolated density is enough? as this is really a lot and well converged, I would simply state that this is enough, but not show anythin.????] [TO DO: Look up what McMillan & Binney 2013 have to say about the numerical accuracy of the normalisation. Sanders & Binney (2015) are quoting them on that matter.]

[TO DO] Stuff to explain about fig. 3: When fitting a potential and DF model to stellar data, the numerical accuracy of the normalisation is very important. The tracer density changes smoothly with the potential and DF parameters. Systematic errors in the density calculation can therefore lead to systematic over- or underestimation of the true potential parameters. This effect is less severe for larger volumes: The density gradient within a large volume is larger and therefore the relative change of the normalisation with the model parameters is also larger as for smaller observation volumes. For small volumes it is therefore more important to get the density right, i.e. having high $N_{\rm velocity}$ and $N_{\rm sigma}$. For larger volumes it is also important to pre-calculate the density at enough spatial points, i.e. having high $N_{\rm spatial}$, while at the same time smaller inaccuracies in the density calculation do not have a comparable severe effect. This could be also the reason, why the normalisation calculation for the smallest volume in fig. 3 is much more well behaved as long as we can calculate exact actions (isochrone and Staeckel potential) [TO DO: Is this really the reason????]. [TO DO: Should we demonstrate how a wrong accuracy introduces biases???] Using a fiducial qDF to fix the integration range over the velocity in the analysis (cf. §???) can help to make the normalisation vary in a more controlled and smooth way. If the fiducial qDF is close to the true qDF parameters, we get already un-biased and well-behaved results for a data set with 20,000 stars in the isochrone potential and an accuracy of $N_{\rm sigma} = 4$ and $N_{\rm velocity} = 20$, as demonstrated in fig. **??**. This lower accuracy leads to an relative error of 0.005% in the normalisation and therefore to an error in the log-likelihood of ~ 1 (cf. §sec:numaccuracynormalisation). [TO DO: Is this correct? $N\log(n+np) = N\log(n) + N\log(1+p)$ with $N\log(1+p) \sim 1$ for N=20000, p=0.00005. But why do we know that 1 is small enough? Argument is also different to Jo's argument with N_{stars}*precision=1, but I don't get that.???] [TO DO: Does this higher accuracy make the biases in MW potential analyses smaller????]
If two hypotheses have a $\Delta \log \mathscr{L} = 1$, one hypothesis is 10 times more likely. Below this we

[2]In case of the isochrone potential we already have high enough accuracy for $N_{\rm spatial} = 16$, $N_{\rm velocity} = 20$ and $N_{sigma} = 4$.

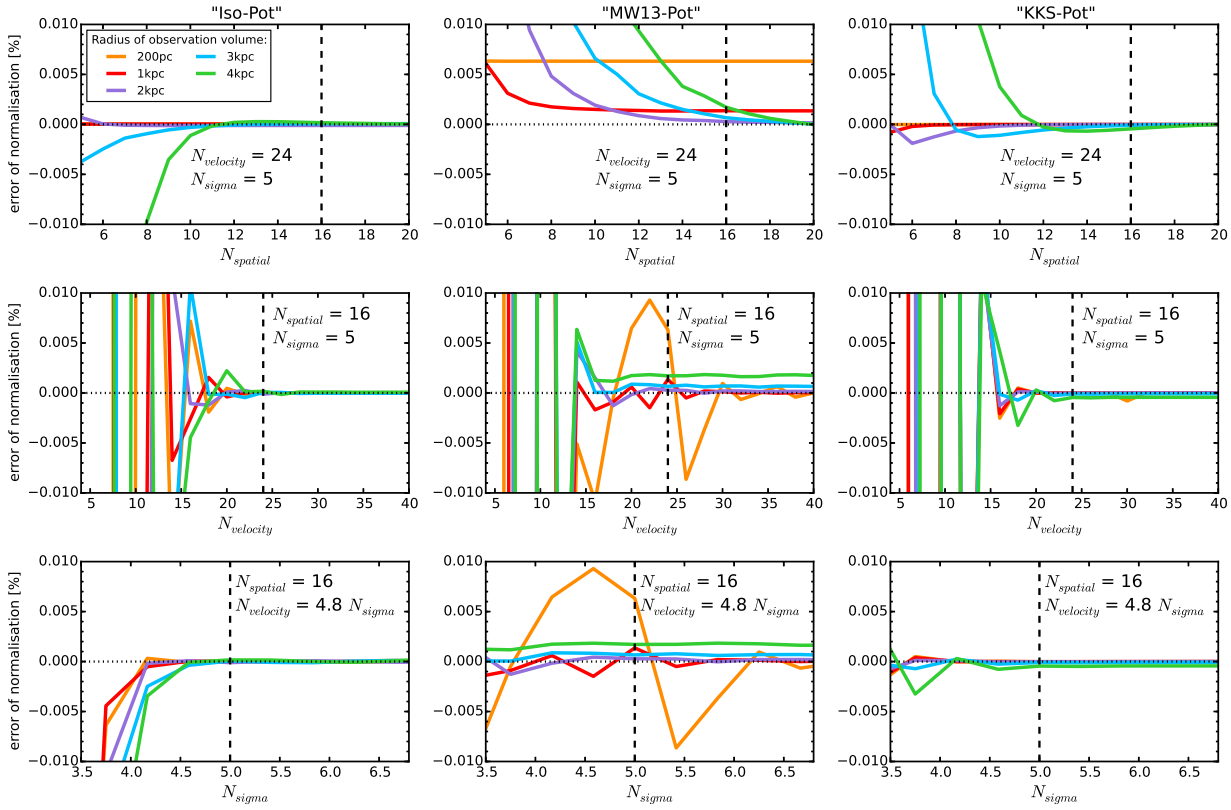

Fig. 3.— [TO DO: Re-write caption. The second potential is not the potA-MW potential anymore but the reference potential "MW13-Pot".] Relative error of the likelihood normalization in eq. (11) and (12) depending on the accuracy of the density calculation in §**??**. The different colors represent calculations for different radii of the spherical observation volume around the sun, as indicated in the legend. $N_{\rm spatial}$ is the number of regular grid points in each R and $z > 0$ within the observed volume on which the tracer density is evaluated according to eq. (9). At each (R, z) a Gauss-Legendre integration of order $N_{velocity}$ is performed over an integration range of $\pm N_{\rm spatial}$ times the dispersion in v_R and v_z and $[0, 1.5 v_{\rm circ}(R_\odot)]$ in v_T. To integrate the interpolated density over the observed volume to arrive at the likelihood normalization in eq. (???), we perform a 40th-order Gauss-Legendre integration in each R and z direction. The distribution function that was evaluated for these plots has the parameters $p_{\rm DF} = \{h_R, \sigma_R, \sigma_z, h_{\sigma_R}, h_{\sigma_z}\} = \{2\ {\rm kpc}, 55\ {\rm km\ s^{-1}}, 66\ {\rm km\ s^{-1}}, 8\ {\rm kpc}, 7\ {\rm kpc}\ \}$. We show the results for three different potentials, an isochrone potential with parameters $p_\Phi = \{v_{\rm circ}, b\} = \{230\ {\rm km\ s^{-1}}, 0.9\ {\rm kpc}\ \}$, a MW-like potential (cf. ???) with parameters $p_\Phi = \{v_{\rm circ}, R_d, z_h, f_h, \frac{{\rm d}\ln v_c}{{\rm d}\ln R}]\} = \{230\ {\rm km\ s^{-1}}, 2.5\ {\rm kpc}, 400\ {\rm pc}\ , 0.8, 0\}$ and a 2-component KK-Staeckel potential with parameters $p_\Phi = \{v_{\rm circ}, \Delta, (a/c)_{\rm disk}, (a/c)_{halo}, k)\} = \{230\ {\rm km\ s^{-1}}, 0.3, 20., 1.07, 0.28\}$. (Caption continues on next page.)

Fig. 3.— (Continued.) We calculate the true normalization with high accuracy as $M_{\rm tot,true} \approx M_{\rm tot}(N_{\rm spatial} = 20, N_{\rm velocity} = 56, N_{\rm sigma} = 7)$. [TO DO: Introduce M_{tot} as the likelihood normalization somewhere as formula... ???] The relative error of the normalization is then calculated as $(M_{\rm tot}[N_{\rm spatial}, N_{\rm velocity}, N_{\rm sigma}] - M_{\rm tot,true})/M_{\rm tot,true}$. The dashed lines indicate the accuracy used in our analyses: it is better than 0.001% for all three potential types. Only for the smallest volume in the MW potential (yellow line) the error is only $\sim 0.005\%$. This could be due to the fact, that, while we have analytical formulas to calculate the actions for the isochrone and the Staeckel potential exactly, we have to resort to an approximate action calculation (Staeckel Fudge by Binney) for the MW-like potential (cf. §???). [TO DO: larger labels???] [TO DO: Try to redo yellow curve in MW. Weird, that it does not depend on $N_{spatial}$.???]

cannot properly decide which hypothesis is better. [TO DO: This limit should be satisfied for the differential log likelihood error (i.e. the derivative of the likelihood w.r.t. the parameters.) than only the likelihood. How to do this? Probably not doing this....]

2.3.6. Marginalization over coordinates

[TO DO]

2.3.7. Measurement Errors

We assume Gaussian errors in the observable space $\boldsymbol{y}_i \equiv (\tilde{\boldsymbol{x}}_i, \tilde{\boldsymbol{v}}_i) = (\alpha, \delta, (m-M), \mu_\alpha, \mu_\delta, v_{\rm los})$,

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

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

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

In case of errors in distance modulus $\mu \equiv (m - M)$, but not in position on the sky (i.e. $\sigma_\alpha = 0$ and $\sigma_\delta = 0$), and a purely spatial selection function, this reduces to

$$\begin{aligned}
\tilde{P}(\boldsymbol{y}_i \mid p_\Phi, p_{\mathrm{DF}}, \sigma_{\boldsymbol{y},i}) &\equiv \frac{\mathrm{sf}(\tilde{\boldsymbol{x}}_i) \cdot \int \mathrm{d}^6 y' \ \mathrm{DF}(\boldsymbol{y}') \cdot N[\boldsymbol{y}_i, \sigma_{\boldsymbol{y},i}](\boldsymbol{y}')}{\int \mathrm{d}^6 y \ \mathrm{DF}(\boldsymbol{y}) \cdot \int \mathrm{d}\mu' \ \mathrm{sf}(\tilde{\boldsymbol{x}}') \cdot N[\mu, \sigma_{\mu,i}](\mu')}, \\
&\approx \frac{\mathrm{sf}(\tilde{\boldsymbol{x}}_i) \cdot \int \mathrm{d}^6 y' \ \mathrm{DF}(\boldsymbol{y}') \cdot N[\boldsymbol{y}_i, \sigma_{\boldsymbol{y},i}](\boldsymbol{y}')}{\int \mathrm{d}^6 y \ \mathrm{DF}(\boldsymbol{y}) \cdot \mathrm{sf}(\tilde{\boldsymbol{x}})} \qquad (13) \\
&\approx \frac{\mathrm{sf}(\tilde{\boldsymbol{x}}_i)}{\int \mathrm{d}^6 y \ \mathrm{DF}(\boldsymbol{y}) \cdot \mathrm{sf}(\tilde{\boldsymbol{x}})} \cdot \frac{1}{N_{\mathrm{error}}} \sum_n^{N_{\mathrm{error}}} \mathrm{DF}(\boldsymbol{y}'_{i,n}) \qquad (14)
\end{aligned}$$

The first approximation, Eq. 13, is valid only in the case of small σ_μ [TO DO: check], but makes the normalisation much less computational expensive - especially in the case of heteroscedastic errors $\sigma_{\mu,i}$, for which the normalisation would have been calculated for each star separately. The second approximation, Eq. 14, is how we compute the convolution using Monte Carlo integration with N_{error} samples drawn from the error Gaussian, $y'_{i,n} \sim N[\boldsymbol{y}_i, \sigma_{\boldsymbol{y},i}](\boldsymbol{y}')$.

2.4. Fitting Procedure

We search the $(p_\Phi, p_{\rm DF})$ parameter space for the maximum of the likelihood in eq. (10). The most crucial part of our fitting procedure for finding the peak and width of the likelihood in the $(p_\Phi, p_{\rm DF})$ parameter space is therefore the reduction of computational costs while not introducing systematic errors due to numerical inaccuracies. We do this by a two-step procedure: The first step finds the approximate peak and width of the likelihood using a nested-grid search, while the second step will either sample the shape of the likelihood (or rather the posterior probability distribution) using a Monte-Carlo Markov Chain (MCMC) or calculate the likelihood on a much finer grid.

2.4.1. Fitting Step 1: Finding the likelihood peak with a Nested-grid search

[TO DO: Make consistent: use of $\sigma_{R,0}$ and σ_R as profile or dispersion at sun. ???]

The $(p_\Phi, p_{\rm DF})$ parameter space can be high-dimensional and we do not necessarily have a good notion where to look for the likelihood peak initially. We use a nested-grid approach to find the peak and to minimize effectively the number of models for which we have to evaluate the likelihood.[3]
The nested-grid search works in the following way:

- *Initialization.* We set up an initial grid with 3^N regular grid points, where N is the number of free model parameters M (cf. §2.3.2. The range of this initial grid is chosen sufficiently large and should encompass all reasonable[4] values for the parameters.

- *Evaluation.* Then we evaluate the likelihood at each grid-point. Stepping through different p_Φ parameters is much more computationally expensive than stepping through different DF parameter sets, because of the many $\boldsymbol{x}, \boldsymbol{v} \xrightarrow{p_\Phi} \boldsymbol{J}$ transformations that have to be performed for each new potential. Evaluation on a grid allows us to have an outer loop that iterates over the potential parameters p_Φ and pre-calculates the actions and

[3]The nested-grid approach is preferable to other optimizing methods, because it can be effectively parallelized on multiple computer cores, while methods like ??????? work linearly and would therefore take longer.

[4]To get a better feeling where in parameter space the true $p_{\rm DF}$ parameters lie, we fit eq. (???) directly to the data. This gives a very good inital guess for $\sigma_{R,0}$ and $\sigma_{z,0}$. To improve the estimate for h_R, we fit eq. (???) only to stars within a thin wedge around $(R = 0, z = 0)$ and then apply the relation in fig. 5 in Bovy & Rix (2013) between the stars' measured scale length $h_R^{\rm out}$ and the qDF tracer scale length $h_R^{\rm in} = h_R$.

an inner loop which, for a given potential, goes over the qDF parameters $p_{\rm DF}$ and uses these pre-calculated actions to evaluate the likelihood (analogously to fig. 9 in Bovy & Rix (2013)).
Both, the pre-calculation of actions and the likelihood calculations for all $p_{\rm DF}$s, can be easily sped up by distributing them over many computer cores.

- *Iteration.* To find from the very sparse 3^N likelihood grid a new and better grid, that is more centered on the likelihood and has a width that, in the optimal case, is of order of the width of the likelihood, we proceed in the following: For each of the model parameters M the likelihood is marginalized over all the other dimensions. From the resulting three grid points, the fraction of second highest and highest likelihood is compared with e^{-8}: If the fraction is larger than that, the range of the grid is still larger than a ˜4-sigma likelihood environment around the peak. In this case we simply choose the grid point with the highest likelihood as the new grid range. Otherwise, if the width of the grid is already small enough, we can fit a Gaussian to the three grid points and determine a new and better 4-sigma fitting grid range from it, with the best-fit Gaussian mean as the new central grid point.
We proceed with iteratively evaluating the likelihood on finer and finer grids, until we have found a 4-sigma fit range in each of the model parameter dimensions.

- *The fiducial qDF.* For the above strategy to work properly, the action pre-calculations have to be independent of the choice of qDF parameters. This is clearly the case for the $N_j \times N_{\rm error}$ [TO DO: explain $N_{\rm error}$???] stellar data actions $\boldsymbol{J}_i$. To calculate the normalisation in eq. (11), $N^2_{\rm spatial} \times N^3_{\rm velocity}$ actions $\boldsymbol{J}_n$ are needed. Formally the spatial coordinates at which the $\boldsymbol{J}_n$ are calculated depend on the $p_{\rm DF}$ parameters via the integration ranges in eq. (9). To relax this dependence we instead use the same velocity integration limits in the likelihood calculations for all $p_{\rm DF}$s in a given potential. This set of parameters, that sets the velocity integration range globally, $(\sigma_{R,0}, \sigma_{z,0}, h_{\sigma_R}, h_{\sigma_z})$ in eq. (???), is referred to as the "fiducial qDF". Using the same integration range in the density calculation for all qDFs at a given p_Φ makes the normalisation vary smoothly with different $p_{\rm DF}$. Choosing a fiducial qDF that is very off from the true qDF can however lead to large biases. The optimal values for the fiducial qDF are the (yet unknown) best fit $p_{\rm DF}$ parameters. We take care of this by setting, in each iteration step of the nested-grid search, the fiducial qDF simply to the $p_{\rm DF}$ parameters of the central grid point. As the nested-grid search approaches the best fit values, the fiducial qDF approaches automatically the optimal values as well. This is another advantage of the nested-grid search, because the result will not be biased by a poor choice of the fiducial qDF.

- *Speed Limitations.* Overall the computation speed of this nested-grid approach is dominated (in descending order of importance) by a) the complexity of potential and action calculation, b) the number $N_j \times N_{\rm error} + N^2_{\rm spatial} \times N^3_{\rm velocity}$ of actions to calculate, i.e. the number of stars, error samples and numerical accuracy of the normalisation calculations, c) the number of different potentials to investigate (i.e. the number of free potential parameters and number of grid points in each dimension) and d) the number of qDFs to investigate. The latter is also non-negligible, because for such a large number of actions the number of qDF-function evaluations also take some time. We therefore restrict the nested grid search to just three points in each dimension of potential and qDF parameters.

2.4.2. Fitting Step 2: Sampling the shape of the likelihood with MCMC

After the nested-grid search is converged, we already have a very good feeling for where the peak of the likelihood is and how large the approximate 4-sigma likelihood environment is. In the next step we also want to sample the shape of the likelihood. We can either do this by a grid search as well, simply using $K > 3$ grid points in each dimension. The number of grid points scales exponentially with N and it might be, that some of the grid points have very low likelihood and we would waste time on calculating them anyway. In this case it could be a better idea to sample the likelihood (or rather the posterior probability distribution, which is the likelihood times some priors, cf. §????) using a Monte-Carlo Markov Chain (MCMC). Launching the walkers close to the already known peak could lead to a convergence of the MCMC in much less than K^N likelihood evaluations.

[TO DO]

3. Results

We are now in a position to explore the questions about the ultimate limitations of action based modelling, posed in the introduction:

- Can we still retrieve unbiased model parameter estimates p_M in the limit of large sample sizes?
- What role does the survey volume and geometry play, at given sample size?
- What if our knowledge of the sample selection function is imperfect, and potentially biased?
- How do the parameter estimates deteriorate if the individual errors on the phase-space coordinates become significant?

But we also consider the more fundamental limitations:

- What if the observed stars are not extacly drawn from the family of model distribution functions?
- What happens to the estimate of the potential and the DF, if the actual potential is not contained in the family of model potentials?

We do not explore the breakdown of the assumption that the system is axisymmetric and in steady state. [**hat shouldl also be at the end of the introduction..** [say: except for the case of "errors" we assume that thne phase-space errors are negligible..]

[TO DO: Make consistent h_{σ_R} –¿ $h_{\sigma,R}$]

Table 3. [TO DO: Caption]

Test		Model for Mock Data	Model in Analysis	Figures
①	*Potential:*	"KKS-Pot"	-	Fig. 2
Influence of	*MAP :*	"hot" & "cold" qDF		
survey volume on	*Survey volume:*	a) $R \in [4, 12]$ kpc,		
mock data distribution,		$z \in [-4, 4]$ kpc,		
also in action space		$\phi \in [-20°, 20°]$.		
		b) $R \in [6, 10]$ kpc,		
		$z \in [1, 5]$ kpc,		
		$\phi \in [-20°, 20°]$.		
	# stars per data set:	20,000		
	# data sets:	4		
②	*Potential:*	"Iso-Pot"	"Iso-Pot", free parameter: b	Fig. ???
Width of the	*MAP :*	"hot" qDF	"hot" qDF, free parameters:	
likelihood scales			$\ln\left(\frac{h_R}{8\text{kpc}}\right), \ln\left(\frac{\sigma_R}{230\text{km s}^{-1}}\right), \ln\left(\frac{h_{\sigma,R}}{8\text{kpc}}\right)$	
with number of stars	*Survey volume:*	sphere around sun,		
by $\propto 1/\sqrt{N}$		$r_{\max} = 3$ kpc		
	# stars per data set:	between 100 and 40,000		
	# data sets:	132		

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

The individual *MAP* in Bovy & Rix (2013) contained typically 200 [CHECK] objects, so that each *MAP* implied a quite broad *pdf* for the p_M. Here we explore what happens in the limit of very much larger samples for each *MAP* , say 20,000 objects. As outlined in §[TO DO CHECK] the immediate consequence of larger samples is given by the likelihood normalization requirement, $\log(1 + rel.error) \leq 1/N_{sample}$, (see Eq. 5 [TO DO CHECK]), which is the modelling aspect that drives the computing time. This issues aside, we would, however, expect that in the limit of large data sets with vanishing measurement errors the *pdf*s of the p_M become Gaussian, with a *pdf* width, σ_p that scales as $1/N_{sample}$. Further, we must verify that any bias in the *pdf* expectation value is far less than σ_p, even for quite large samples.

Using sets of mock data ([TO DO: describe by referencing to Section]) and our fiducial model for p_M, we verified that the *RoadMapping* satisfies all these conditions and expectations. Fig. 4 illustrates the joint *pdf*'s of all p_M. This figure illustrates that the *pdf*'s are multivariate Gaussians that project into Gaussians when considering the marginalized *pdf* for all the individual p_M. Note that some of the parameters are quite covariant, but the level of their actual covariance depends on the of the p_M from with the mock data were drawn. Figure5 then illustrates that the *pdf* width, σ_p indeed scales as $1/N_{sample}$. Fig.6 illustrates even more, that the *RoadMapping* satisfies the central limit theorem. The average parameter estimates from many mock samples with identical underlying p_M are very close to the input p_M, and the distibution of the actual parameter estimates are a Gaussian around it.

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

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

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

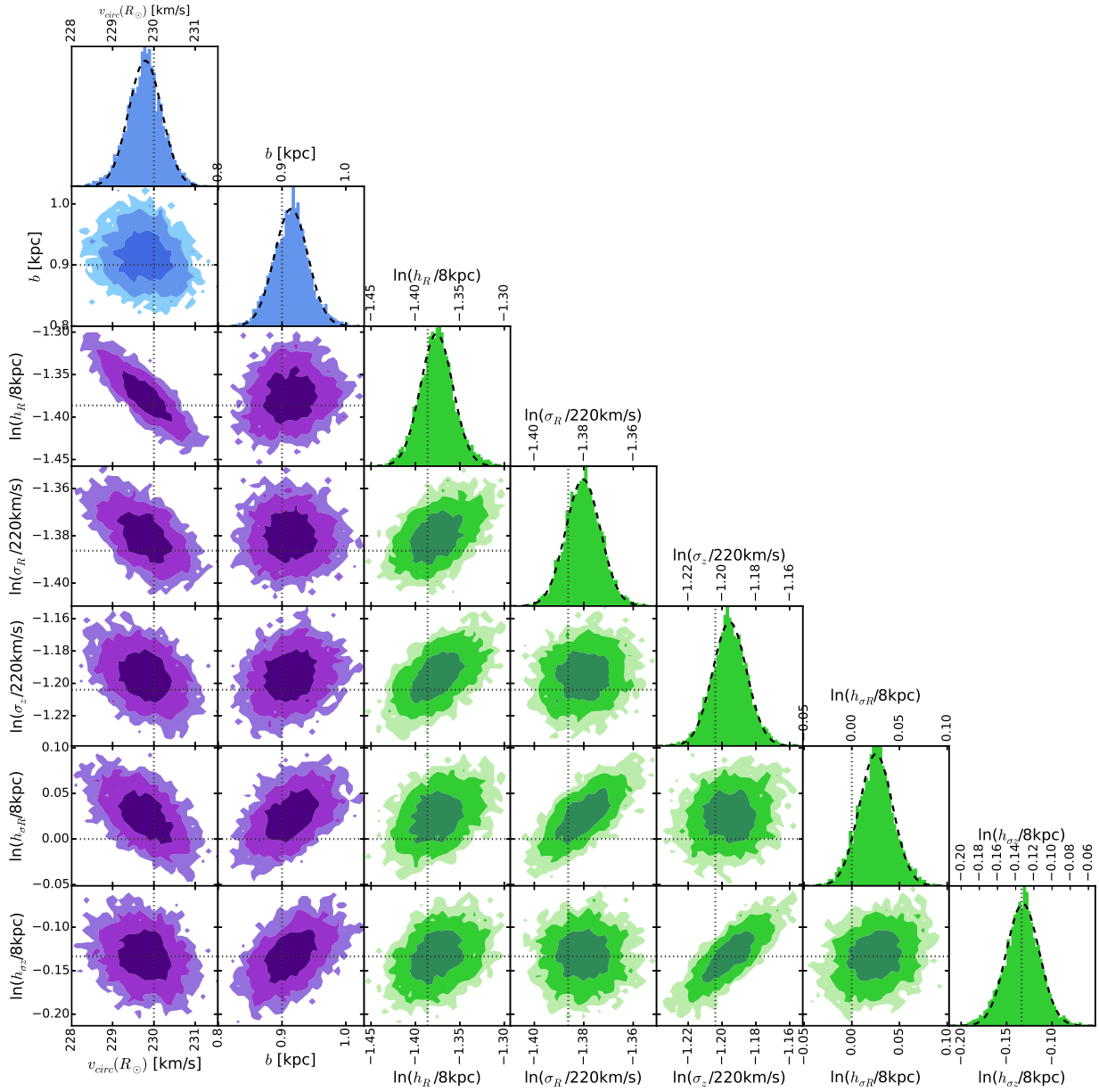

Fig. 4.— The likelihood in eq. (???) in the parameter space $\{p_\Phi, \ln(p_{\rm DF})\}$ for one example mock data set. This mock data set has 20,000 stars and was created in the potential "IsoPot" and from the "hot" qDF, and was observed within a spherical volume around the sun of radius $r = 2$ kpc . The true parameters are marked by dotted lines. The dark, medium and bright contours in the 2D distributions represent 1, 2 and 3 sigma confidence regions, respectively, and show weak or moderate covariances. The likelihood here was sampled using MCMC (with flat priors in p_Φ and $\ln(p_{\rm DF})$ to turn the likelihood into a full posterior distribution function). Because only 10,000 MCMC samples were used to create the histograms shown, the 2D distribution has noisy contours. The dashed lines in the 1D distributions are Gaussian fits to the histogram of MCMC samples. This demonstrates very well that for such a large number of stars, the likelihood approaches the shape of a multivariate Gaussian, as expected from the central limit theorem. [TO DO: Maybe re-do with higher accuracy??? This was done with $N_{sigma} = 4$.] [TO DO: Mention "Note: this was picked among 5 to have all 1sigma contours encompass the input values." ???] [TO DO: it's the cold population, not the hot one??? I'm not sure]

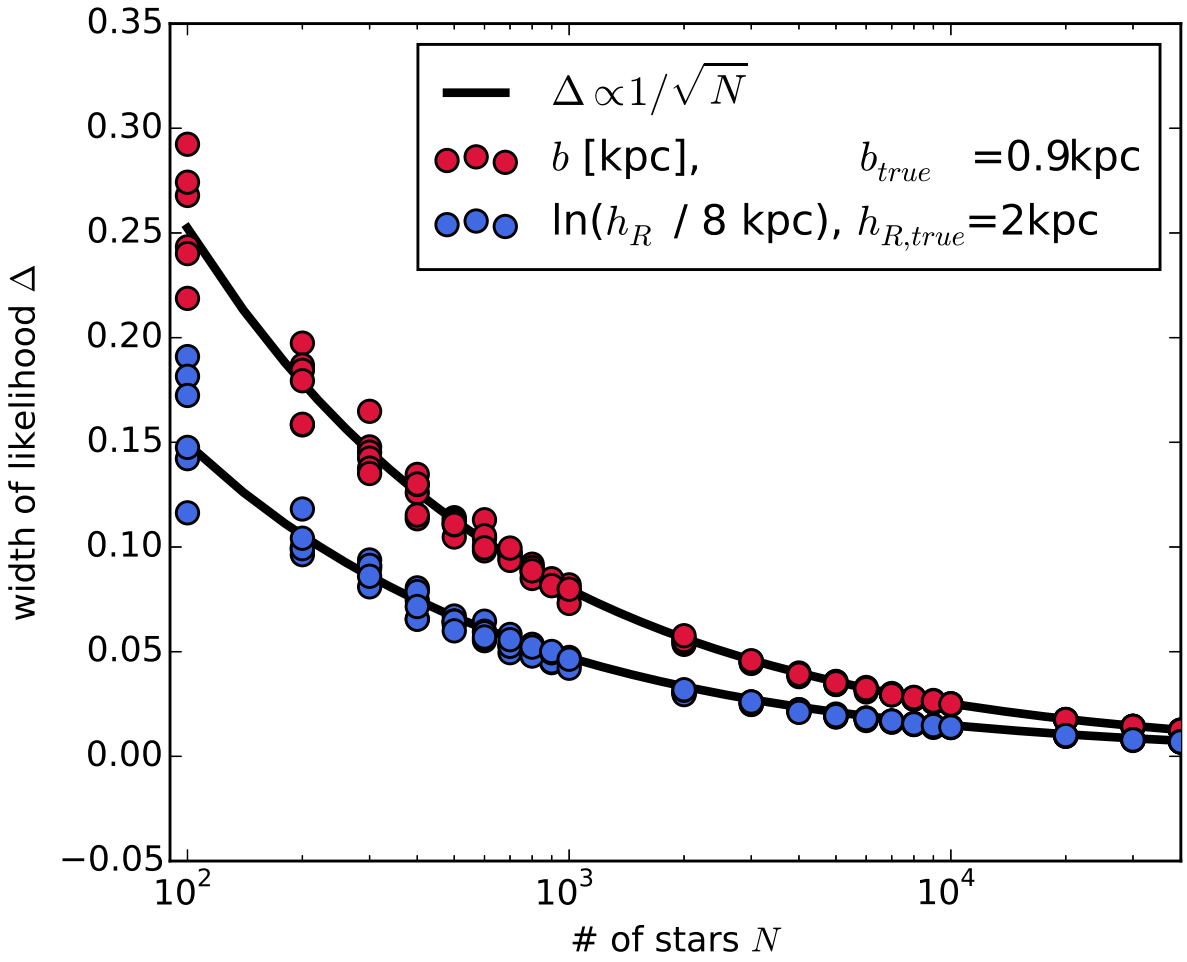

Fig. 5.— The width of the likelihood for two fit parameters found from analyses of 132 mock data sets vs. the number of stars in each data set. The parameters of the mock data model are given as Test ② in Table **??**. The likelihood in Eq. ??? was evaluated on a grid in the free parameters. We then fitted a Gaussian to the marginalized likelihoods of each free fit parameter. The standard error (SE) of these best fit Gaussians is shown for the potential parameter b in kpc (red dots) and for the qDF parameter $\ln(h_R/8\text{kpc})$ in dimensionless units (blue). The black lines are fits of the functional form $\Delta(N) \propto 1/\sqrt{N}$ to the data points of both shown parameters. As can be seen, for large data samples the width of the likelihood behaves as expected and scales with $1/\sqrt{N}$ as predicted by the central limit theorem. [TO DO: Maybe re-do with higher accuracy??? This was done with $N_{sigma} = 4$.] [TO DO: rename width of likelihood into Standard Error (SE). Also x-axis: N (# of stars in data set)???]

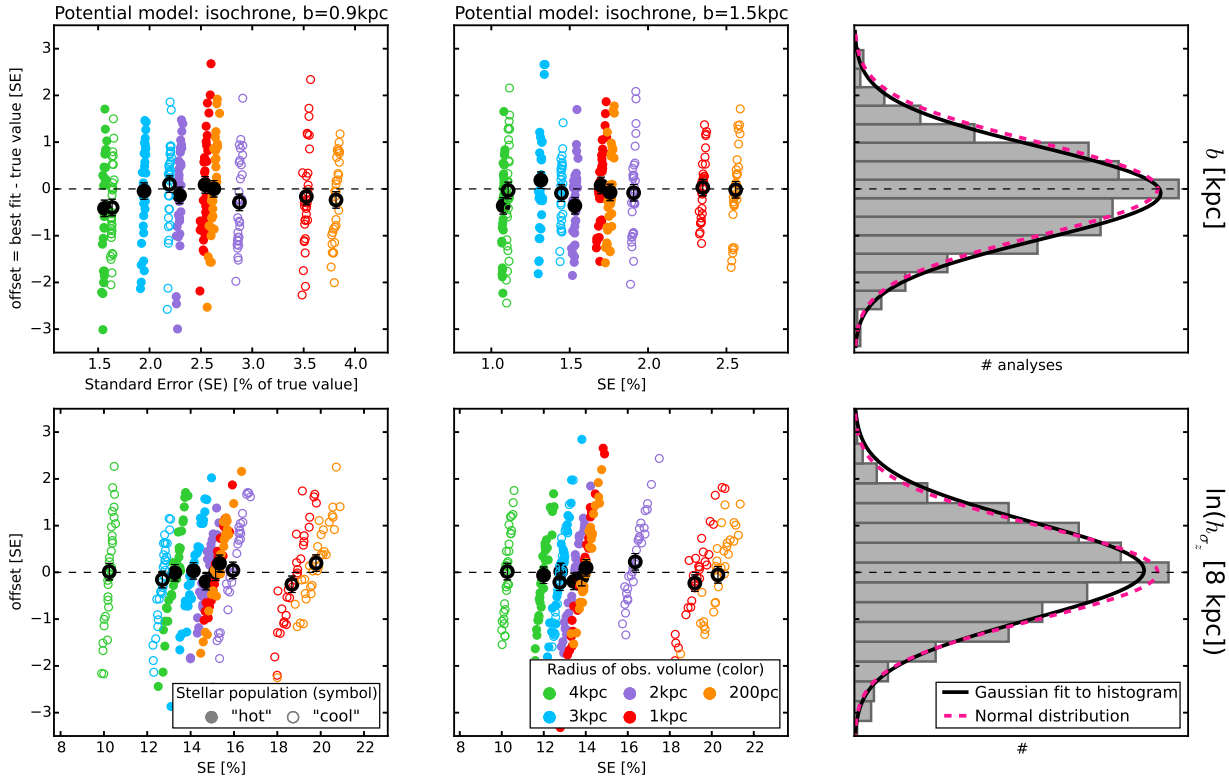

Fig. 6.— (Un-)bias of the parameter estimate: According to the central limit theorem the likelihood will follow a Gaussian distribution for a large number of stars. From this follows that also for a large number of data sets the corresponding best fit values for the model parameters have to follow a Gaussian distribution, centered on the true model parameters. That our method satisfies this and is therefore an unbiased estimator [TO DO: can I say that????] is demonstrated here. We create 640 mock data sets. They come from two different isochrone potentials ($p_\Phi = \{v_{\rm circ}, b\} = \{230 \text{ km s}^{-1}, b\}$ with $b = 0.9$ kpc (first column) and $b = 1.5$ kpc (second column)), two different stellar populations ('hot' with $p_{DF,hot} = \{h_R, \sigma_R, \sigma_z, h_{\sigma_R}, h_{\sigma_z}\} = \{2 \text{ kpc}, 55 \text{ km s}^{-1}, 66 \text{ km s}^{-1}, 8 \text{ kpc}, 7 \text{ kpc}\}$ (solid symbols) and 'cool' with $p_{DF,cool} = \{h_R, \sigma_R, \sigma_z, h_{\sigma_R}, h_{\sigma_z}\} = \{3.5 \text{ kpc}, 42 \text{ km s}^{-1}, 32 \text{ km s}^{-1}, 8 \text{ kpc}, 7 \text{ kpc}\}$ (open symbols)) and five spherical observation volumes of different sizes (color coded, see legend). For each parameter set we therefore sample 32 mock data realisations and analyse them by evaluating the likelihood ??? on a grid. As numerical accuracy we use $N_{\rm velocity} = 20$ and $N_{\rm sigma} = 4$. The fit parameters are $\{b, \ln(h_R/8\text{kpc}), \ln(\sigma_R/230\text{km s}^{-1}), \ln(h_{\sigma_R}/8\text{kpc})\}$. All other model parameters are kept at their true value in the modelling. We determine the best fit value and the standard error (SE) for each fit parameter by fitting a Gaussian to the marginalized likelihood. The offset is the difference between the best fit and the true value of each model parameter. In the first two columns the offset in units of the SE is plotted vs. the SE in % of the true model parameter. The first row shows the results for the isochrone scale length b and the second row the qDF parameter h_{σ_z}, which corresponds to the scale length of the vertical velocity distribution.

Fig. 6.— (Continued.) The last column finally displays a histogram of the 640 offsets (in units of the corresponding SE). The black solid line is a Gaussian fit to a histogram. The dashed pink line is a normal distribution $\mathcal{N}(0, 1)$. As they agree very well, our modelling method is therefore well-behaved and unbiased. For the 32 analyses belonging to one model we also determine the mean offset and SE, which are overplotted in black in the first two columns (with $1/\sqrt{32}$ as error). [TO DO: Is the scatter of the black symbols too large??? Is the reason for this numerical inaccuracies???] [TO DO: units of b in title????????]

3.2. The Role of the Survey Volume Geometry

Beyond the sample size, the survey volume *per se* must play a role; clearly, even a vast and perfect data set of stars within 100 pc of the Sun, has limited power to tell us about the potential at very different R. Intuitively, having dynamical tracers over a wide range in R suggests to allow tighter contraints on the radial dependence of the potential. To this end, we devise a number mock data sets, drawn from a one single p_M, but drawn from six different volume wedges (see §[TO DO CHECK]), as illustrated in the left panels of fig. 7. To make the parameter inference comparison very differential, the mock data sets are equally large (20,000) in all cases, and are drawn from identical total survey volumes (4.5 kpc^3, achieved by adjusting the angular width of the edges). The right panels of Fig.7 the illustrate the ability of *RoadMapping* to constrain model parameters (in this case two p_Φ parameters). The two top right panels of Fig.7 illustrate that the radial extent and the maximal height above the mid-plane matter. In the case shown, the standard error of the estimated parameters in twice as large for the volume with small ΔR and $\Delta|z|$; unsurprisingly, in the axisymmetic context the larger $\Delta\phi$ extent of that volume does not help to constrain the parameters. The panels in the bottom row explore whether the radial or vertical extent plays a dominant role: it appears that substantive radial and vertical extent are comparably important to constrain the parameters.

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

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

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

TO DO As transparency doesn't work in eps, the orange volume looks smaller than the blue one.

TO DO Maybe skip first row of plots?

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

TO DO We could compare these results with similar results for KKS pot. If the latter has no biases, we can state that to avoid biases when using an non-Staeckel potential, one

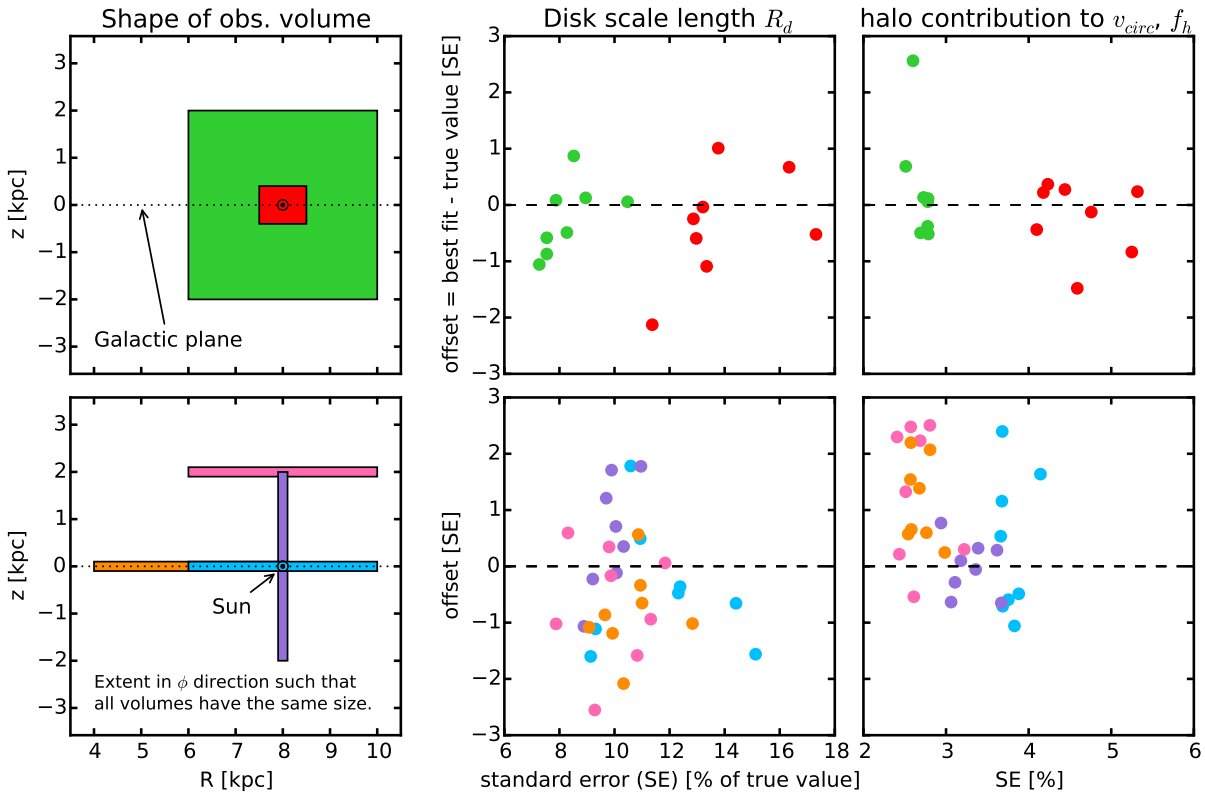

Fig. 7.— We demonstrate that for a given size of the observation volume the shape and position of the volume does not matter much as long as we have both large radial and/or vertical coverage. The left column shows the position of our test observation volumes within the Galaxy with respect to the Galactic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

should use a volume with comparable R *and* z coverage, because for this the biases seem to be smallest.

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

TO DO Do we explicitely want to test, if it matters, if the radial coverage is larger or smaller the disk scale length, and the vertical coverage is larger or smaller than the disk scale height?

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

The selection function of a survey is described by a spatial survey volume and a completeness function, which determines the fraction of stars observed at a given location within the Galaxy with a given brightness, metallicity etc (see §[TO DO CHECK]). The completeness function depends on the characteristics and mode of the survey, can be very complex and is therefore sometimes not perfectly known. We investigate how much an imperfect knowledge of the selection function can affect the recovery of the potential. We model this by creating mock data with varying incompleteness, while assuming constant completeness in the analysis. The mock data comes from a sphere of $r_{\max} = 3$ kpc around the sun and an incompleteness function that drops linearly with distance r from the sun (fig. **??**):

$$\text{completeness}(r) = 1\epsilon_r \cdot \frac{r}{r_{\max}}$$

This could be understood as a model for the important effect of stars being less likely to be observed the further away they are. We demonstrate that the potential recovery with *RoadMapping* is very robust against somewhat wrong assumptions about the (in-)completeness of the data (see fig. **??**). A lot of information about the potential comes from the rotation curve measurements in the plane, which is not affected by applying an incompleteness function. In Appendix ??? we also show that the robustness is somewhat less striking but still given for small misjudgements of the incompleteness in vertical direction, parallel to the disk plane (fig. **??** and **??**). This could model the effect of wrong corrections for dust obscurement in the plane. We also investigate in Appendix ??? if indeed most of the information is stored in the rotation curve. For this we use the same mock data sets as in fig. **??** and **??**, but this time were not including the tangential velocities in the modelling, rather marginalizing the likelihood over v_T. In this case the potential is much less tightly constrained, even for 20,000 stars. For only small deviations of true and assumed completeness ($\lesssim 10\%$) we can however still incorporate the true potential in our fitting result (see fig. ???).

Stuff that needs to be further examined:

[TO DO] Maybe instead of decreasing completeness with height above the plane, a completeness that INcreases with height above the plan, to model e.g. obscuration due to dust.

[TO DO] Make similar test as isoSphFlexIncompR, but with KKS potential, to test, if this robustness is a special case for the isochrone potential.

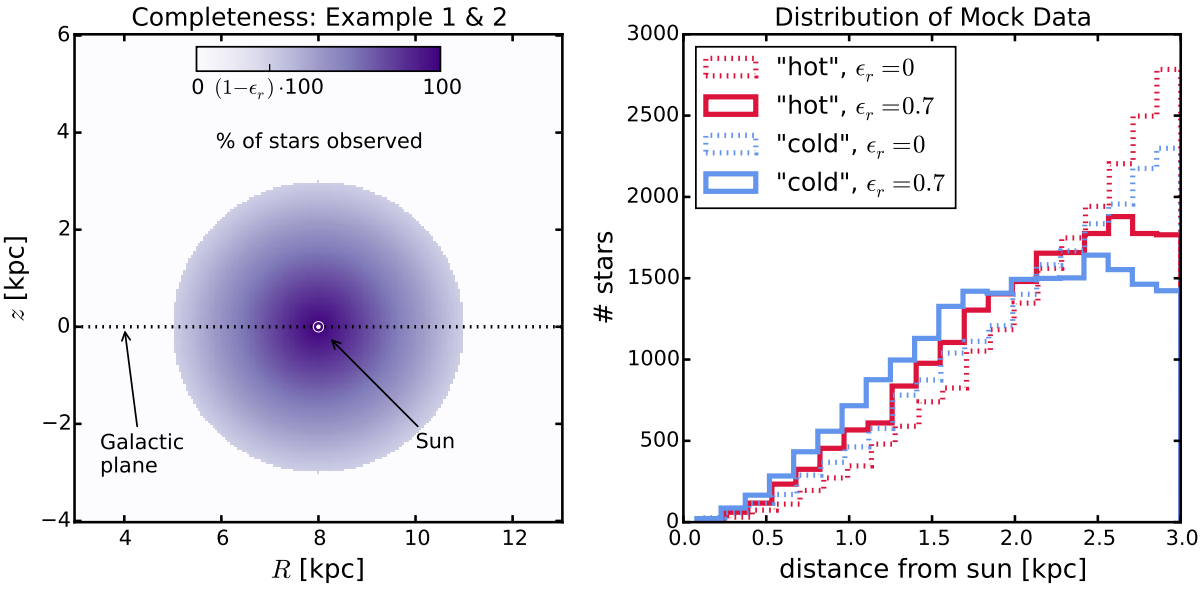

Fig. 8.— [TO DO: Rewrite caption] Selection function and mock data distribution for investigating radial (Example 1 & 2, left) and vertical (Example 3 & 4, right) incompleteness of the data. The survey volume is a sphere around the sun with $r_{\rm max} = 3$ kpc. In Example 1 & 2 (Example 3 & 4) the percentage of observed stars is decreasing linearly with radius from the sun (height above the Galactic plane), as demonstrated in the first row of panels. How fast this detection rate drops is quantized by the factor ϵ_r (ϵ_z) in eq. (**??**) (eq. (**??**)). Different mock data sets have different ϵ_r (ϵ_z). Histograms for four data sets, each with 20,000 and drawn from two *MAP*s ("hot" in red and "cool" in blue, see table 2) and with two different ϵ_r (ϵ_z), 0 and 0.7, are shown in the lower two panels for illustration purposes.[TO DO: Re-do, if new analyses are in violin plot.]

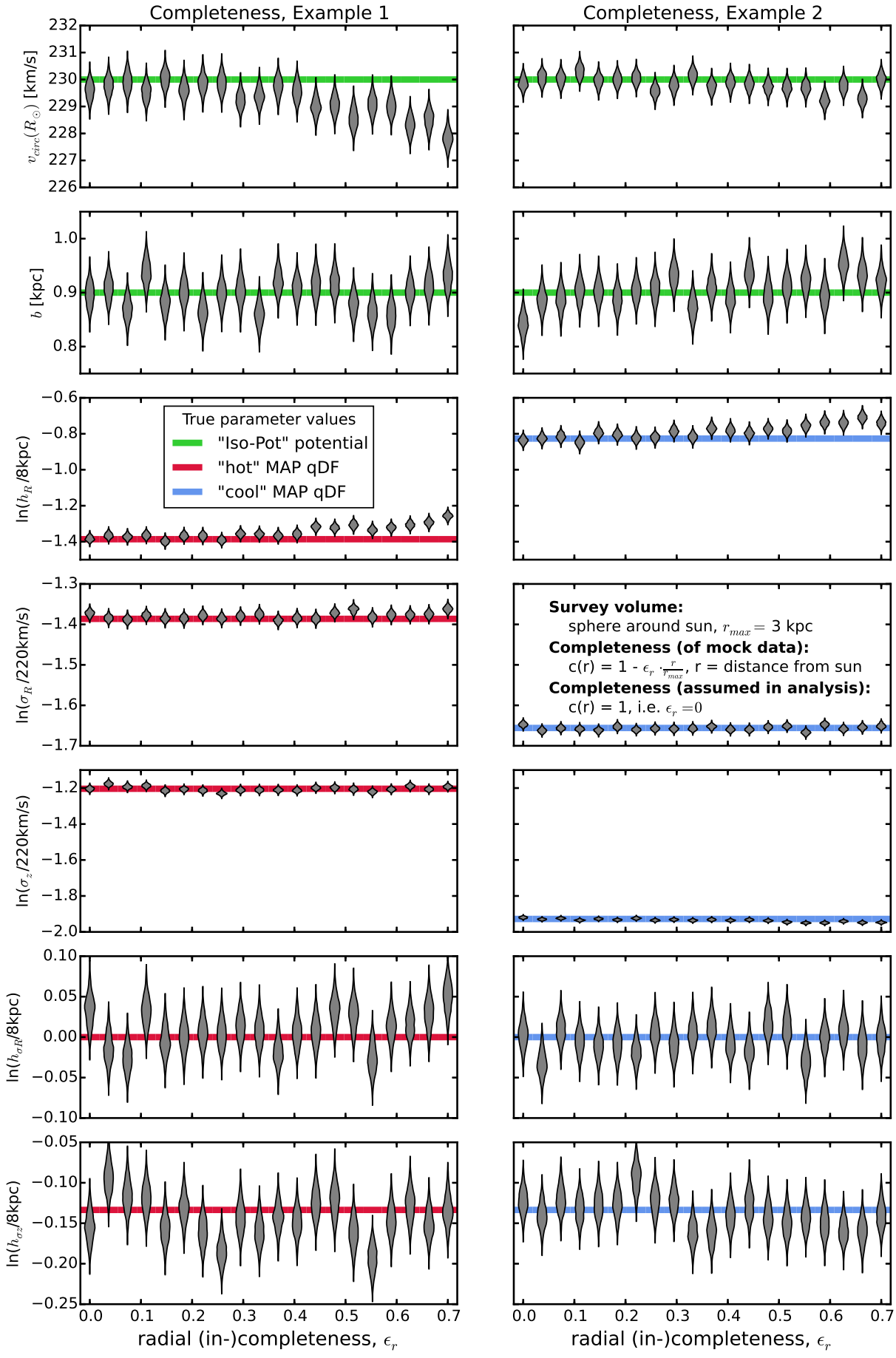

Fig. 9.— Caption [TO DO] (This was done using the current qDF to set the fitting range.

3.4. Effect of measurement errors on recovery of potential?

Collection of possible tests and plots

- *Plot 1:* The plot I had on the poster, which shows the number of MC samples needed for given maximum error. However, we still haven't tested, if this plot depends on: * hotness of stars * number of stars
- *Plot 2:* Some plot that shows, that our approximation of ignoring distance errors works. Any ideas?
- *Test 1:* One selection function, one population, vary the size of the proper motion error (don't forget to adapt the number of MC samples needed)
  Plot 3: (width of pdf) vs. (maximum velocity error / temperature parameter)

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

Our modelling approach assumes that each *MAP* follows a quasi-isothermal distribution function, qDF. In this Section we explore what happens if this idealization does not hold. This could be, because even in the limit of perfectly measured abundances, MAPs do not follow a qDF. Or, even if they did do that, because the finite abundance errors effectively mix different MAPs. We investigate both these issues by creating mock data sets (Fig. 10) that are drawn from two distinct qDFs of different temperature, and analyze the composite mock data set by fitting a single qDF to it. These results are illustrated in Figs. 11 and 12. Following the observational evidence, MAPs with cooler qDFs also have longer tracer scale lengths. In the first set of test, we choose qDFs of widely different temperatures and vary their relative fraction (dubbed "examples 1a/b", Fig. 11) ; in the second set of tests ("examples 2a/b", Fig. 12), we always mix mock data points from two different qDFs in equal proportion, but vary by how much the qDF's temperatures differ.
The first set of tests mimicks a DF that has wider wings or a sharper core in velocity space than a qDF (Fig. 10). The second test could be understood by mixing neighbouring *MAP*s due to too large bin sizes or abundance measurement errors.

It is worth considering separately the impact of the DF deviations on the recovery of the potential and of the qDF parameters.
We find from example 1 that the potential parameters can be better and more robustly recovered, if a mock-data *MAP* is polluted by a modest fraction ($\lesssim 30\%$) of stars drawn from a cooler qDF with a longer scale length, as opposed to the same pollution of stars drawn from a hotter qDF with a shorter scale length.
When considering the case of a 50/50 mix of contributions from different qDFs , there is a systematic, but only small, error in recovering the potential parameters, monotonically increasing with the qDF parameter difference (example 2); in particular for fractional differences in the qDF parameters of $\lesssim 20\%$ the systematics are insignificant even for samples sizes of 20,000, as used in the mock data.
The recovery of the effective qDF parameters, in light of non qDF mock data is quite intuitive: the effective qDF temperature lies between the two temperatures from which the mixed DF of the mock data was drawn; in all cases the scale length of the velocity dispersion fall-off, $h_{\sigma R}$ and $h_{\sigma,z}$, is shorter, because the stars drawn form the hotter qDF dominate at small radii, while stars form the cooler qDF (with its longer tracer scae length) dominate at large radii. The recovered tracer scale lengths, h_R vary smoothly between the input values of the two qDFs that enetered 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.

We interpret the results in example 1 as recovering the potential from a DF, whose

velocity dispersion has a steeper core and more stars at larger radii than expected (bluish data sets in Fig. 10), or a DF that has broader velocity dispersion wings and more stars at small radii than predicted by the qDF (reddish data sets). We find that the latter would give more reliable results for the potential parameter recovery. At the same time, if we assume that the distribution of stars from one *MAP* is caused by radial migration away from the initial location of star formation, it is more likely that the qDF overestimates the true number of stars at smaller radii. [TO DO: Is this actually a sensible argument???] This could be remedied by focusing the analysis especially on hotter *MAP*s with shorter scale length, for which pollution by colder stars is also much less a problem.
Example 2 could be understood as a model scenario for decreasing bin sizes in the metallicity-α plane when sorting stars in different *MAP*s , assuming that there is a smooth variation of qDF within the metallicity-α plane and each *MAP* indeed follows a qDF. We find that, in the case of 20,000 stars in each bin, differences of 20% in the qDF parameters of two neighbouring bins can still give quite good constraints on the potential parameters. We compare this with the relative difference in the qDF parameters in the bins in fig. 6 of Bovy & Rix (2013), which have sizes of $[Fe/H] = 0.1$ dex and $\Delta[\alpha/Fe] = 0.05$ dex. It seems that these bin sizes are large enough to make sure that $\sigma_{R,0}$ and $\sigma_{z,0}$ of neighbouring *MAP*s do not differ more than 20%. As fig. 11 and 12 suggests especially the tracer scale length h_R needs to be recovered to get the potential right. For this parameter however the bin sizes in fig. 6 of Bovy & Rix (2013) might not yet be small enough to ensure no more than 20% of difference in neighbouring h_R, especially in the low-α ($[\alpha/Fe] \lesssim 0.2$), intermediate-metallicity ($[Fe/H] \sim -0.5$) *MAP*s - provided of course, that each bin contains 20,000 stars. In case there are less than 20,000 stars in each bin the constraints are less tight and due to Poisson noise one could also allow larger differences in neighbouring *MAP*s while still getting reliable results.

Additional test: [TO DO] Draw 200,000 stars from best fit qDF, normalize, compare (residuals?) to mock data set

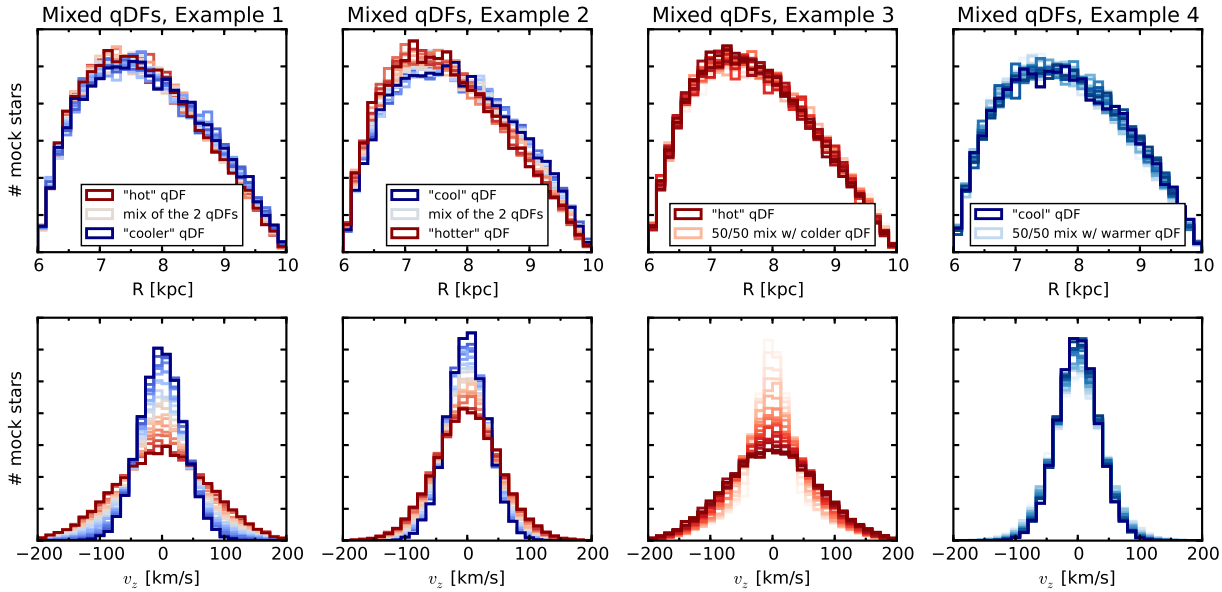

Fig. 10.— Distribution of mock data in two coordinates (R and v_z), created by mixing stars drawn from two different qDFs. This demonstrates how mixing two qDFs can be used as a test case for changing the shape of the DF to not follow a pure qDF anymore, e.g. by adding wings or slightly changing the radial density profile. The distribution in R is also strongly shaped by the selection function, which is, in this case, a sphere around the sun with $r_{\rm max} = 2$ kpc. In total there are always 20,000 stars in each data set and all of them were created in the same potential, the isochrone potential "Iso-Pot" from table 1. The dark red and dark blue histograms show data sets drawn from a single qDF only: the "hot" and "cooler" *MAP*s (Example 1, first column), the "cool" and "hotter" *MAP*s (Example 2, second column), the "hot" (Example 3, third column) and the "cool" *MAP*s (Example 4, fourth column) from table 2. *Example 1 & 2:* The other histograms show data drawn from a superposition of the two reference qDFs. The color coding represents the different mixing rates (reddish: more hot stars, bluish: more cool stars, white: half/half) and is the same as in figure 11, where the corresponding modelling results for each data set are depicted in the same color. *Example 3 & 4:* In this test suite the mixing rate of the two *MAP*s is fixed to 50%/50%. In Example 3 (Example 4) in the third (fourth) column the "hot" ("cool") *MAP* is shown in dark red (dark blue) and mixed with a qDF whose parameters describe a colder (warmer) population. The 'hotness' of these second *MAP* is varied and approaches the "hot" ("cool") *MAP* 's qDF parameters as the histograms get redder (bluer). The color coding is the same as in fig. 12. [TO DO: Try to include square with color gradient in legend.] [TO DO: make larger distance between the left and right panels.] [TO DO: Write "mixture of 2 qDFs" in legend] [TO DO: Rename example 1 & 2 to example 1a/1b and example 3 & 4 to example 2a/2b]

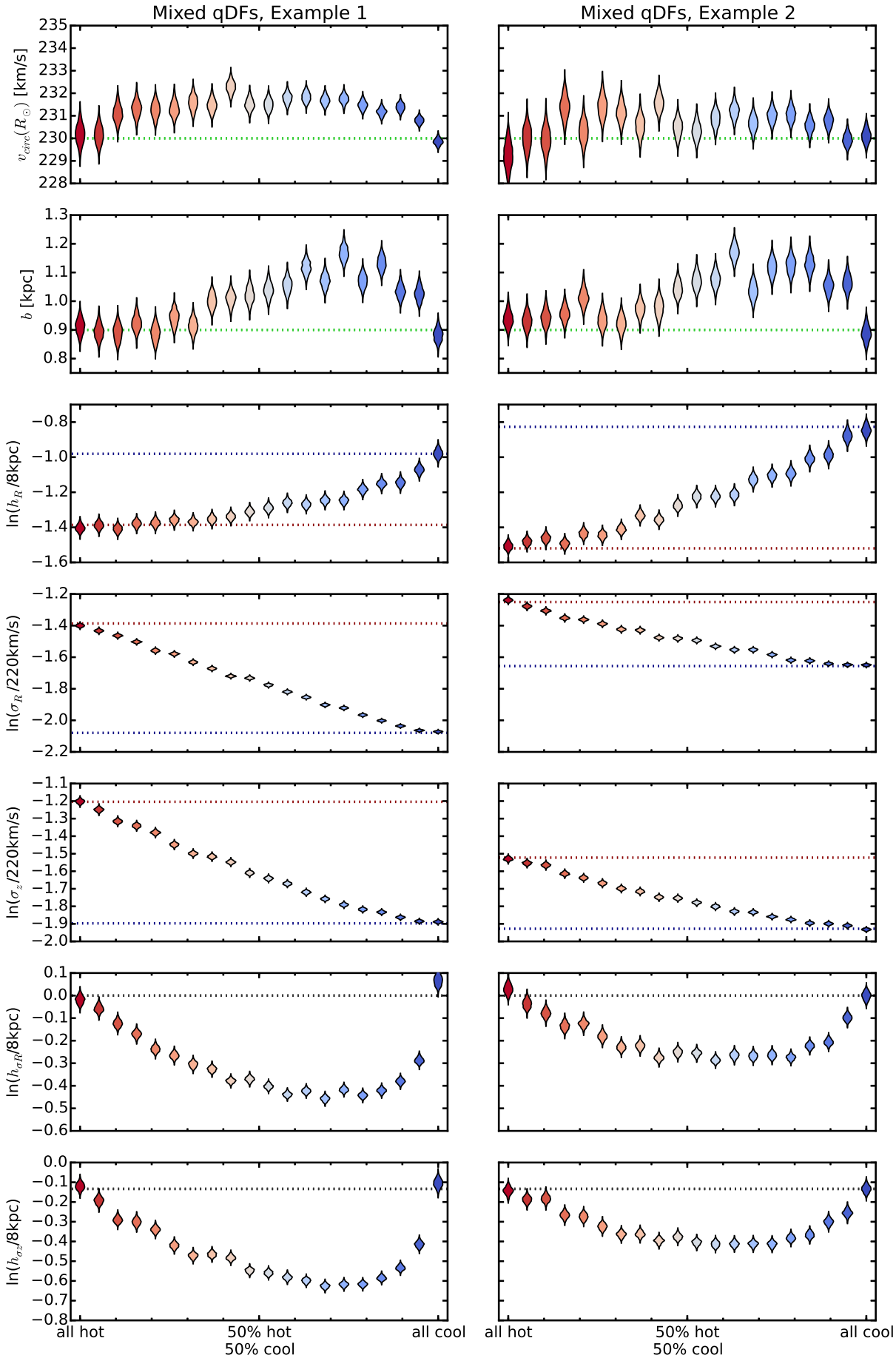

Fig. 11.— (Caption on next page.)

Fig. 11.— (Continued.) The dependence of the parameter recovery on degree of pollution and 'hotness' of the stellar population. To model the pollution of a hot stellar population by stars coming from a cool population and vice versa, we mix varying amounts of stars from two very different populations, as indicated on the x-axis. The composite mock data set is then fit with one single qDF. The violines represent the marginalized likelihoods found from the MCMC analysis. The mock data sets are shown in fig. 10, in the same colors as the violins here. All mock data sets come from the same potential ("Iso-Pot") and selection function (sphere with $r_{\rm max} = 2$ kpc). The true potential parameters are indicated by green dotted lines. Example 1 (Example 2) in the left (right) panels mixes the "hot" ("cool") *MAP* with the "cooler" ("hotter") *MAP* in table 2. True parameters of the hotter (colder) of the two populations are shown as red (blue) dotted lines. We find, that a hot population is much less affected by pollution with stars from a cooler population than vice versa. [TO DO: This was done using the current qDF to set the fitting range. Nvelocity=24 and Nsigma=5 is high enough (though not perfect). Maybe redo with fiducial qDF to be consistent with MixDiff test. ???] [TO DO: Rename example 1 & 2 to example 1a/1b and example 3 & 4 to example 2a/2b]

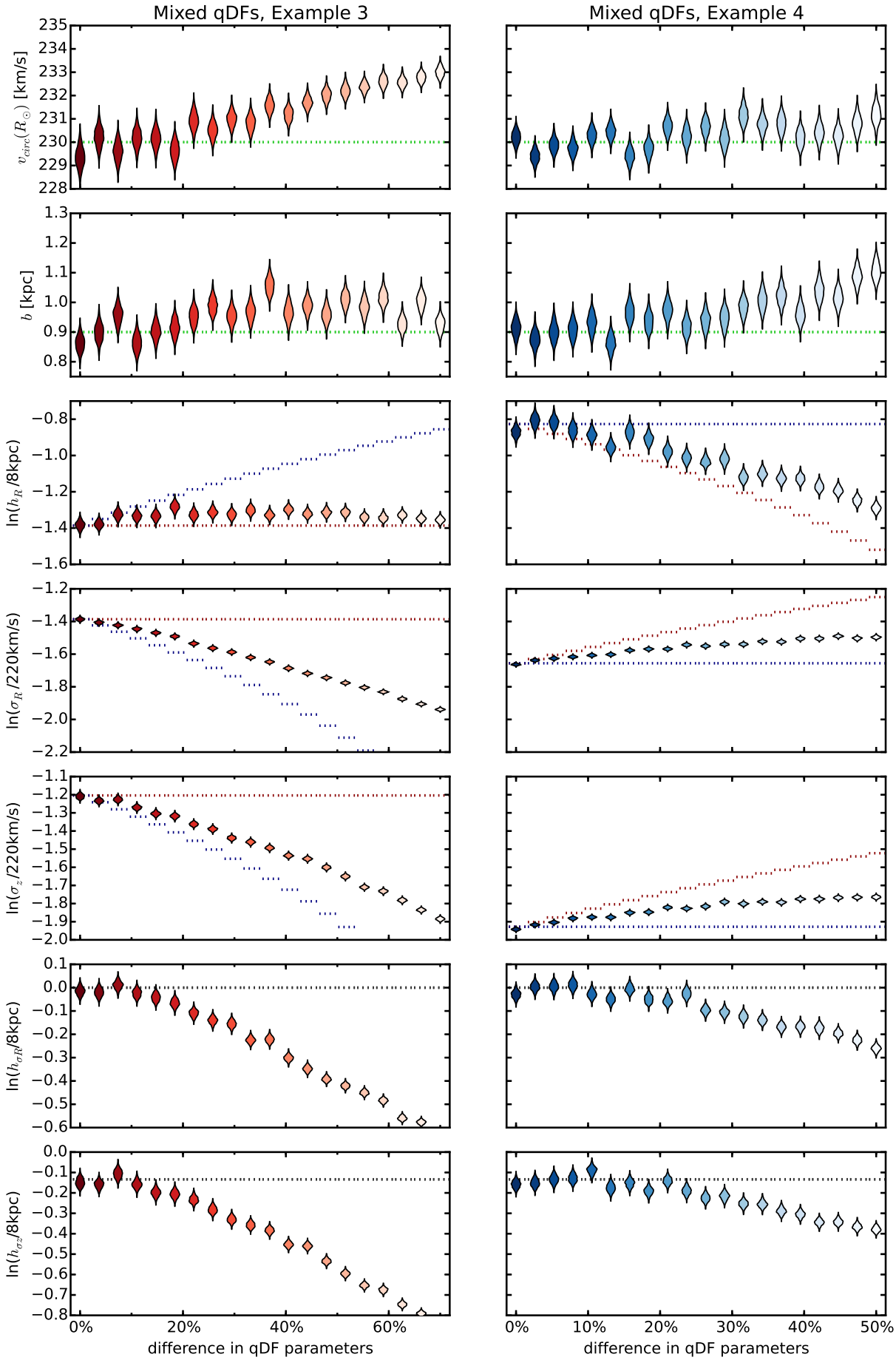

Fig. 12.— (Caption on next page.)

Fig. 12.— (Continued.) The dependence of the parameter recovery on the difference in qDF parameters of the 50%/50% mixture of two stellar populations and their 'hotness'. Each mock data set in Example 3 (Example 4) consists of 20,000 stars, half of them drawn from the "hot" ("cool") qDF in table 2, and the other half drawn from a colder (warmer) population that has $X\%$ smaller (larger) σ_R and σ_z and $X\%$ larger (smaller) h_R. The difference X in these qDF parameters is indicated on the x-axis, and the true parameters of the two qDFs are indicated by the dotted red and blue lines. Each composite mock data set is fitted by a single qDF and the marginalized MCMC likelihoods for the best fit parameters are shown as violines in the third (fourth) column of panels. The mock data was created within the same potential ("Iso-Pot") and selection function (sphere with $r_{\rm max} = 2$ kpc). The true potential parameters are indicated by green dotted lines. The data sets are shown in figure 10, where the histograms have the same colors as the corresponding best fit violines here. By mixing *MAP*s with varying difference in their qDF parameters, we model the effect of bin size in the [Fe/H]-[α/Fe] plane when sorting stars into different *MAP*s : The smaller the bin size, the smaller the difference in qDF parameters of stars in the same bin. We find that the bin sizes should be chosen such that the difference in qDF parameters between neighbouring *MAP*s is less than 20%. [TO DO: Maybe different/same x-axis???] [TO DO: This was done using the current qDF to set the fitting range. Nvelocity=24 and Nsigma=5 is not high enough for the largest differences, i.e. grid search and MCMC converge to different values. Redo with fiducial qDF.] [TO DO: Add in plot a label, that it is a 50%/50% mix of a hot and a cold population.??] [TO DO: Rename example 1 & 2 to example 1a/1b and example 3 & 4 to example 2a/2b] [TO DO: Write in plot, that there is a 50/50 mix of cool and hot])

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

In the long run we would like to incorporate a family of gravitational potential models in *RoadMapping* that is flexible enough to reproduce the essential features of the MW's true mass distribution. Here we want to inspect if we can already give constraints on the true potential, even if our assumed potential is still too rigid - be it because of a low number of free potential parameters, or because our beliefs about the overall shape of the MW's potential are slightly wrong. While our fundamental assumption of axisymmetry springs immediately to mind, being at odds with the obvious existence of a bar and spiral arms in the MW, we will not dive into investigating the implications in the scope of this paper. We rather focus on the case where the mock data was drawn from one axisymmetric potential ("MW14-Pot") and is then analysed using another axisymmetric potential family ("KKS-Pot"), that does *not* incorporate the true potential (compare the second and fourth panel in Fig. ???). The results are shown in Fig. 13.

The set of reference potential parameters of the "KKS-Pot" in Table ??? were found by adjusting the 2-component Kuzmin-Kutuzov Stäckel potential by Batsleer & Dejonghe (1994) such that it looks like the "MW14-Pot" from Bovy (2015): the radial and vertical force in $R \in [4, 12]$ kpc, $|z| \in [0, 4]$ kpc, and the rotation curve in $R \in [0, 16]$ kpc (blue??? lines in Fig. 13). This could be understood as optimum, i.e. a fitting result from *RoadMapping* will most likely not be better than a fit directly to the potential. Even though the analysis results from *RoadMapping* shown in Fig. 13 (yellow??? lines) fit the overall density shape less than the optimum (blue???), we only used tracers within the survey volume (marked in red???). And within the survey volume we actually capture the radial and vertical gravitational force very well - and it is the forces to which the stars' orbit are sensitive to. We also get the density structure of the disk inside the survey volume right, as well as the slope of the rotation curve at the sun. We used the "hot" *MAP* from Table ???, which has a short tracer scale length, i.e. probes the inner regions better than the outer regions. This could explain why the halo shape in the outer regions of the survey volume is less well recovered than the disk in the inner regions.

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

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

The potential model used by Bovy & Rix (2013) had only two free parameters (disk scale lentgh and halo contribution to $v_{\rm circ}(R_\odot)$. To circumvent the obvious disadvantage of this being at all not flexible enough, they fitted the potential separately for each *MAP* and recovered the mass distribution for each *MAP* only at that radius for which it was best constrained - assuming that *MAP*s of different scale length would probe different regions of the Galaxy best. Based on our results in Fig. 13 this seems to be indeed a sensible approach [TO DO: Check that this is indeed the case].

Our choice of fitting a superposition of two Stäckel potentials to the mock data was motivated by the work of Batsleer & Dejonghe (1994) and **?**, who aimed to create MW-like Stäckel potentials from a superposition of several Kuzmin-Kutuzov Stäckel potentials. The big advantage of this approach is the exact and fast action calculation in such a potential, which allows to explore a bigger potential parameter space in the same computation time. Our results in Fig. 13 are also very encouraging that already two components alone can give relatively good constraints. Using more components could allow us also to model the bulge or to include more flexibility in modelling the disk structure.

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

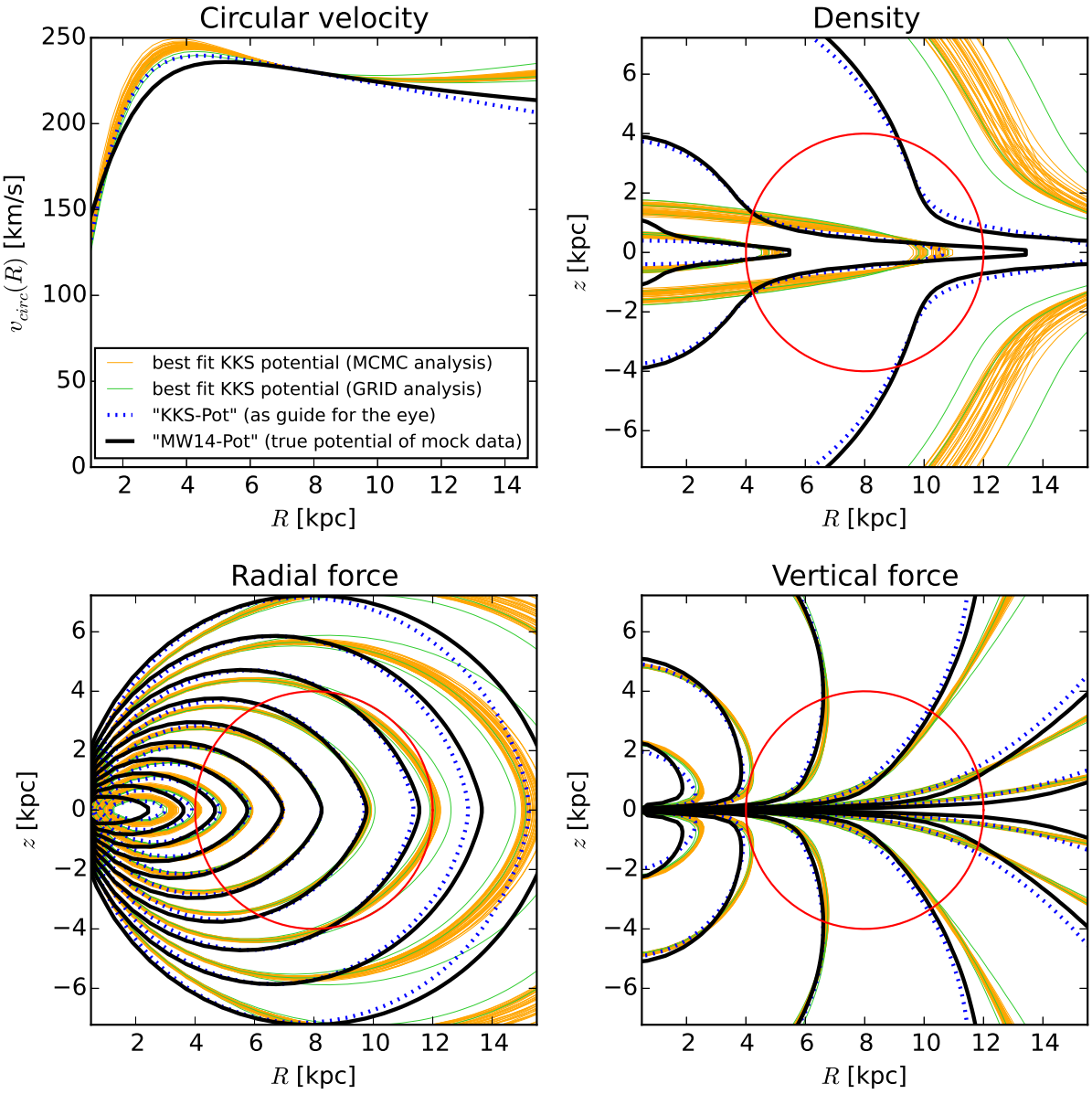

Fig. 13.— Caption [TO DO] [TO DO: redo analyses with fiducial qDF for integration range] [TO DO: include selection function in legend] [TO DO: Correct typo in figure name.]

A. Appendix

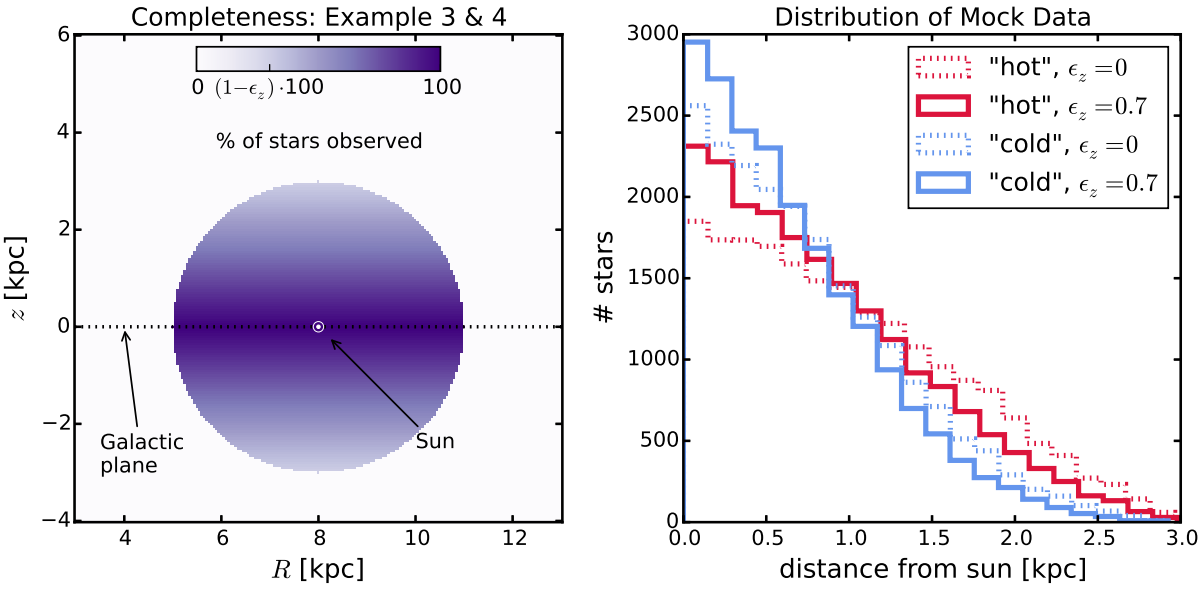

Fig. 14.— [TO DO: Rewrite Caption] Selection function and mock data distribution for investigating radial (Example 1 & 2, left) and vertical (Example 3 & 4, right) incompleteness of the data. The survey volume is a sphere around the sun with $r_{\rm max} = 3$ kpc. In Example 1 & 2 (Example 3 & 4) the percentage of observed stars is decreasing linearly with radius from the sun (height above the Galactic plane), as demonstrated in the first row of panels. How fast this detection rate drops is quantized by the factor ϵ_r (ϵ_z) in eq. (**??**) (eq. (**??**)). Different mock data sets have different ϵ_r (ϵ_z). Histograms for four data sets, each with 20,000 and drawn from two *MAP*s ("hot" in red and "cool" in blue, see table 2) and with two different ϵ_r (ϵ_z), 0 and 0.7, are shown in the lower two panels for illustration purposes.[TO DO: Re-do, if new analyses are in violin plot.]

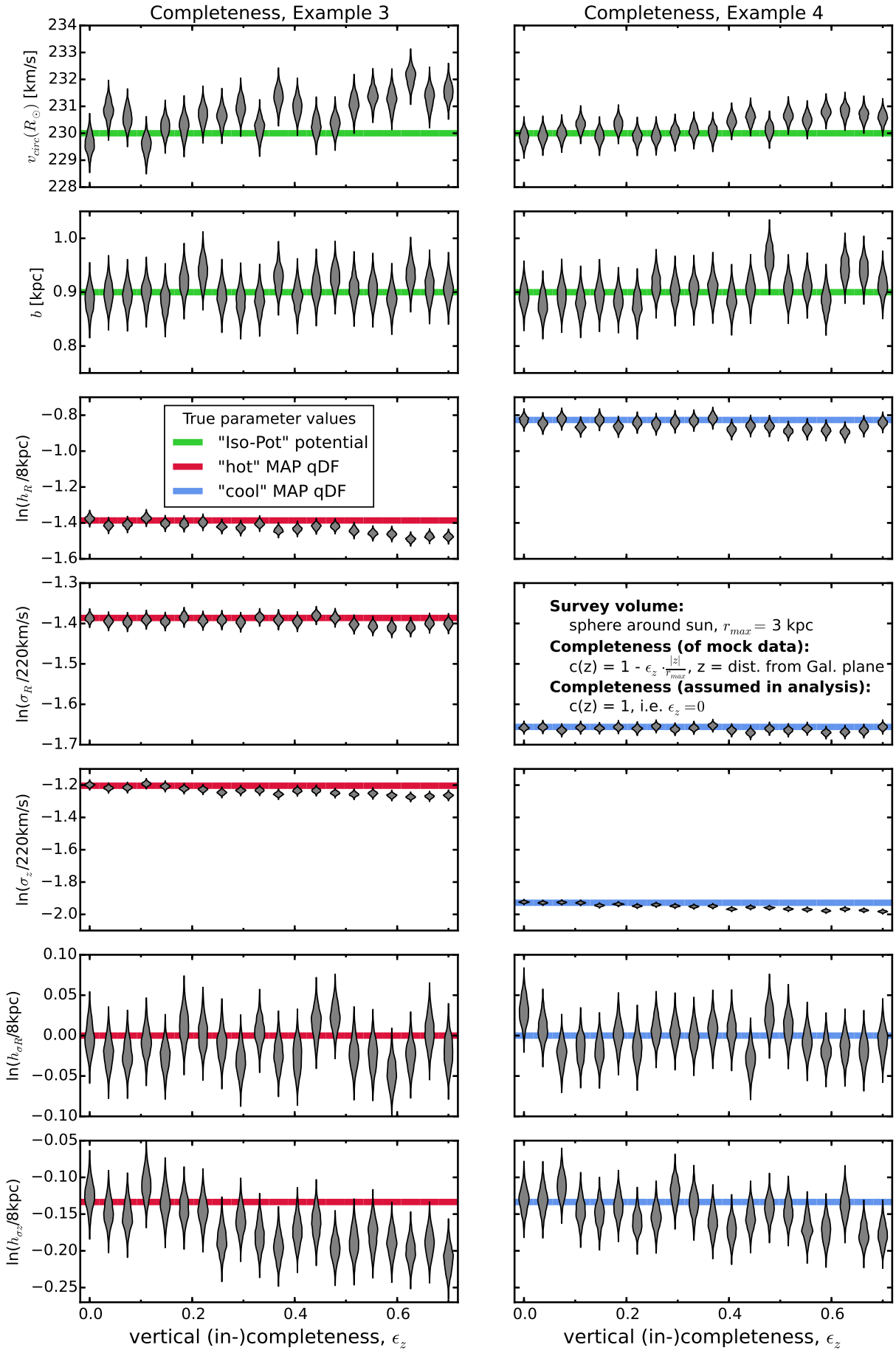

Fig. 15.— Caption [TO DO] (This was done using the current qDF to set the fitting range.

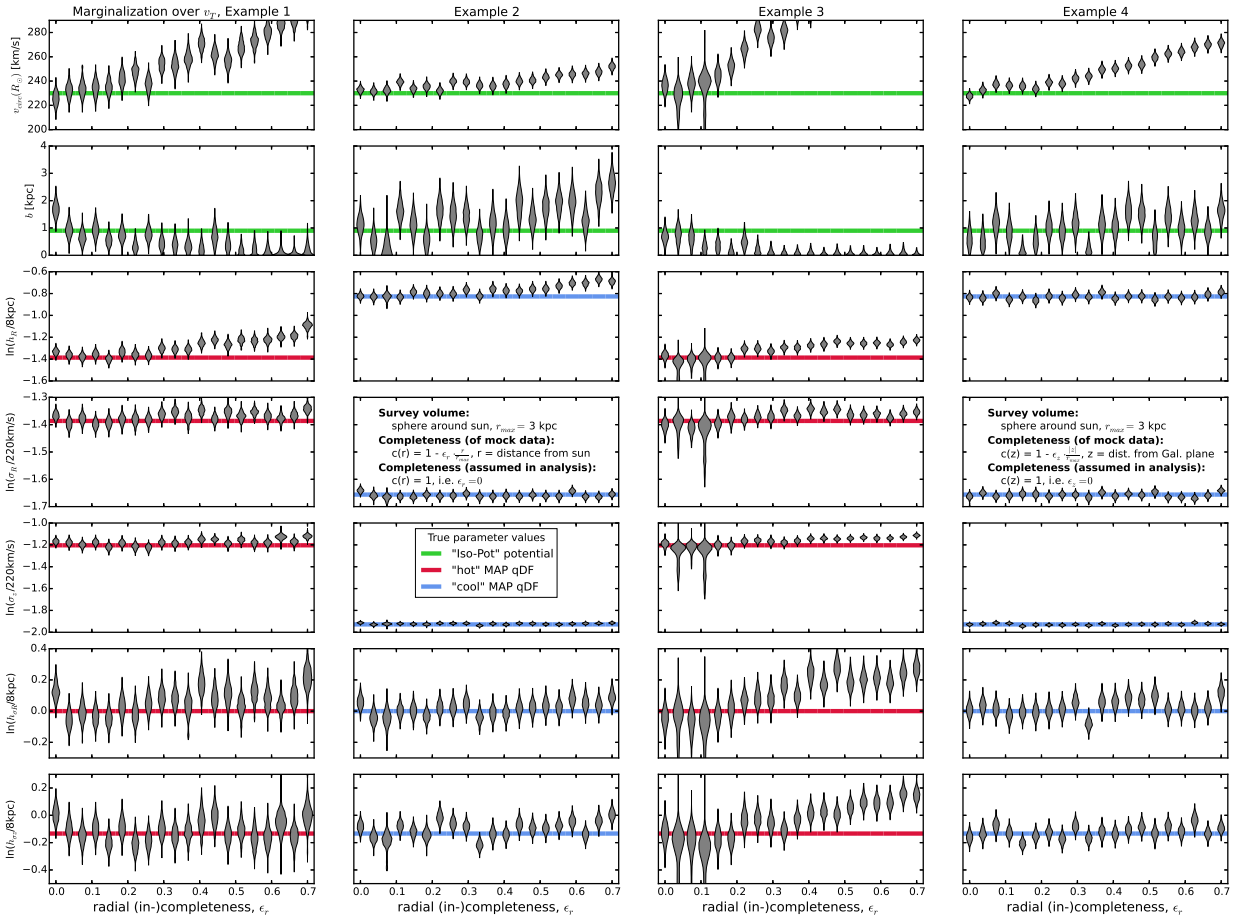

Fig. 16.— Caption [TO DO] ([TO DO: Redo all analyses for which MCMC did not converge to expected peak, and for which b ¡0 was not excluded. ???])

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

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

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

REFERENCES

[TO DO]

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

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

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

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

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

Bovy, J., & Rix, H.-W. 2003, ApJ, 779, 115

Piffl, T., Binney, J., & McMillan, P. J. et al., 2014, MNRAS, 455, 3133

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

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

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

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

[TO DO] Bovy (2015) Galpy paper