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 Nostrum 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.
- **Motivation to use this modelling technique in the Milky Way:** Bovy et al. 2012 .... [TO DO]

- **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...)

- **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 using DF's similar to Binney:**

  - 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, Piffl 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]

2. Dynamical Modelling

2.1. Model

2.1.1. Actions

[TO DO]

2.1.2. Distribution function

[TO DO] [To Do here: Also mention how the density is calculated.]

2.1.3. Potential models

[TO DO] Mention different ways to calculate actions in different potentials.

2.1.4. Selection function: observed volume and completeness

[TO DO]

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 obvious procedure would be to draw $\boldsymbol{J}_i$ from $\mathrm{qDF}(\boldsymbol{J}_i \mid p_{\mathrm{DF}})$ and $\boldsymbol{\theta}_i$ between 0 and 2π, transforming this to $(\mathrm{x}_i, \mathrm{v}_i)$ in the given potential and rejecting all stars that are outside the observed volume. 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 $(\boldsymbol{x}_i, \boldsymbol{v}_i) \longrightarrow (\boldsymbol{J}_i, \boldsymbol{\theta}_i)$. The observed volume is also much smaller than the whole galaxy and the fraction of rejected stars would be enormous. We propose a fast and simple two-step method for drawing mock data from an action distribution function in a given observed volume. [TO DO]

2.2.1. Preparation: Tracer density

[TO DO: This section should be shifted to the distribution function section. ???]

One crucial point in creating the mock data, as well as in the analysis, is to calculate the spatial tracer density $\rho_{\mathrm{DF}}(\boldsymbol{x} \mid p_\Phi, p_{\mathrm{DF}})$ for a given distribution function with parameters p_{DF} in a potential with parameters p_Φ. We do this by integrating the axisymmetric distribution function over the velocity at $N_{\mathrm{spatial}} \times N_{\mathrm{spatial}}$ regular grid points in the (R, z) plane, using a Gauss-Legendre quadrature of order N_{velocity} in each of the three velocity components. The velocity distribution according to the qDF at a given (R_i, z_i) looks approximately Gaussian for v_R and v_z. The v_T distribution peaks somewhere around $v_{\mathrm{circ}}(R_\odot)$ and $v_T > 0$ (cf.

§?????). We approximate the integration over the velocity therefore as

$$\rho_{\mathrm{DF}}(R, |z| \mid p_\Phi, p_{\mathrm{DF}}) = \int_{-\infty}^{\infty} \mathrm{qDF}(J[R, z, \boldsymbol{v} \mid p_\Phi] \mid p_{\mathrm{DF}})\, \mathrm{d}^3\boldsymbol{v} \tag{1}$$

$$\approx \int_{-N_{\mathrm{sigma}}\sigma_R(R|p_{\mathrm{DF}})}^{N_{\mathrm{sigma}}\sigma_R(R|p_{\mathrm{DF}})} \int_{-N_{\mathrm{sigma}}\sigma_z(R|p_{\mathrm{DF}})}^{N_{\mathrm{sigma}}\sigma_z(R|p_{\mathrm{DF}})} \int_0^{1.5v_{\mathrm{circ}}(R_\odot)} \mathrm{qDF}(J[R, z, \boldsymbol{v} \mid p_\Phi] \mid p_{\mathrm{DF}})\, \mathrm{d}v_T\, \mathrm{d}v_z\, \mathrm{d}v_R, \tag{2}$$

where $\sigma_R(R \mid p_{\mathrm{DF}})$ and $\sigma_z(R \mid p_{\mathrm{DF}})$ are the star's radial and vertical velocity dispersion according to the qDF and given by eq. (???) and (???) and N_{sigma} has to be large enough. The size of the $N_{\mathrm{spatial}} \times N_{\mathrm{spatial}}$ grid in (R, z) is chosen cover the extent of the observed volume for $z > 0$. We interpolate then over this grid to be able to evaluate the density at each (R, z) within the observed volume. The total number of grid points and therefore actions that need to be calculated are $N^2_{\mathrm{spatial}} \cdot N^3_{\mathrm{velocity}}$. Fig. ??? shows the importance of choosing N_{spatial}, N_{velocity} and N_{spatial} sufficiently large in order to get the density with an acceptable numerical accuracy. For the creation of the mock data we use $N_{\mathrm{spatial}} = 20$, $N_{\mathrm{velocity}} = 40$ and $N_{\mathrm{sigma}} = 5$.

2.2.2. Step 1: Drawing positions from the selection function

To get $\boldsymbol{x}_i$ for our mock data stars, we first sample random positions (R_i, z_i, ϕ_i) uniformly from the observed volume. Making use of the geometric shapes of our simple observed volumes (spheres, cylinders, wedges...) we can do this efficiently with inverse transform Monte Carlo sampling. In a second step we then apply a rejection Monte Carlo method to these positions using the pre-calculated, interpolated density grid $\rho_{\mathrm{DF}}(R, |z| \mid p_\Phi, p_{\mathrm{DF}})$. 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 on the set of positions we got from the last step. This procedure can be repeated until the required number of positions are reached. The sample then follows the required distribution

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

2.2.3. 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) found in step 1 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. The envelope Gaussian peaks at $(v_R = 0, v_T = max(\mathrm{qDF}(R_i, z_i, 0, vT, 0 \mid p_{Phi}, p_{\mathrm{DF}})), v_z = 0)$ and has a standard deviation of $(2\sigma_R(R_i), 2\sigma_R(R_i), 2\sigma_z(R_i))$. We then pick one of the sampled velocities $\boldsymbol{v}_i$ at and for each (R_i, z_i) and arrive at our desired mock data sample

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

[TO DO: mention fig. 1. ???]

2.2.4. Step 3: Introducing measurement errors

[TO DO]

[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)

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. §2.1.2) 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

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

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

$\mathrm{sf}(\boldsymbol{x}, \boldsymbol{v})$, which is in general a function of the position only, $\mathrm{sf}(\boldsymbol{x})$. 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 statistics. The set of qDF fit parameters is therefore

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

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. §2.1.3).
The total set of model parameters to fit is then

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

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 $\mathrm{qDF}(\boldsymbol{J}_i \mid M) := \mathrm{qDF}(\boldsymbol{J}[\boldsymbol{x}_i, \boldsymbol{v}_i \mid p_\Phi] \mid p_{\mathrm{DF}})$.

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

[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 $\tilde{\boldsymbol{v}} = (\mu_\alpha, \mu_\delta, v_{\mathrm{los}})$ 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. §???).

and it's orbit to be drawn from the qDF with $p_{\rm DF}$, i.e.

$$\mathscr{L}(M \mid D_j) = \prod_i^{N_j} P(\boldsymbol{x}_i, \boldsymbol{v}_i \mid M), \tag{3}$$

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

$$\begin{aligned} P(\boldsymbol{x}_i, \boldsymbol{v}_i \mid M) &= \frac{1}{(r_o v_o)^3} \cdot \frac{\mathrm{qDF}(\boldsymbol{J}_i \mid M) \cdot \mathrm{sf}(\boldsymbol{x}_i)}{\int \mathrm{d}^3x\, \mathrm{d}^3v\ \mathrm{qDF}(\boldsymbol{J} \mid M) \cdot \mathrm{sf}(\boldsymbol{x})} \\ &\propto \frac{1}{(r_o v_o)^3} \cdot \frac{\mathrm{qDF}(\boldsymbol{J}_i \mid M)}{\int \mathrm{d}^3x\ \rho_{\rm DF}(R, |z| \mid M) \cdot \mathrm{sf}(\boldsymbol{x})}. \end{aligned} \tag{4}$$

In the second step we used eq. (1). The factor $\prod_i \mathrm{sf}(\boldsymbol{x}_i)$ is independent of the model parameters, so we use simply eq. (4) 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 $v_{\rm circ}(R = 1) = 1$. $v_{\rm circ}(R_\odot = 8.\mathrm{kpc}) \sim 230\mathrm{km\ s^{-1}}$ 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. (4) 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. (4) 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 §2.2.1. 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. (4) 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.

$$\begin{aligned} \log \mathscr{L}(M \mid D_j) &= \sum_i^{N_j} \log \mathrm{qDF}(\boldsymbol{J_i} \mid M) \\ &\quad -N_j \log(\text{true normalisation}) - N_j \log(1 + \text{rel. error}), \qquad (5) \\ \text{with} &\quad N_j \log(1 + \text{rel. error}) \lesssim 1. \end{aligned}$$

[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. 2 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.]

2.3.6. Marginalization over coordinates

[TO DO]

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

2.3.7. Measurement Errors

[TO DO]

2.4. Fitting Procedure

We search the $(p_\Phi, p_{\rm DF})$ parameter space for the maximum of the likelihood in eq. (3). 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.

[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 $h_R^{\rm in} = h_R$.

- *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 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. (4), $N^2_{\rm spatial} \times N^3_{\rm velocity}$ 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. (2). 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..]

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. 3 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. Figure4 then illustrates that the *pdf* width, σ_p indeed scales as $1/N_{sample}$. Fig.5 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. 3 and 4: 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. 5: 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.

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. 6. 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.6 the illustrate the ability of *RoadMapping* to constrain model parameters (in this case two p_Φ parameters). The two top right panels of Fig.6 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 5 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. 6:

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. 5

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 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 and 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 either with distance from the sun (left panels in fig. 7) or with distance from the Galactic plane (right panels in fig. 7). We demonstrate that the potential recovery with *RoadMapping* is very robust against somewhat wrong assumptions about the (in-)completeness of the data (see the tests for the radial incompleteness function in fig. 8

Stuff that needs to be further examined:

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

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. What if our assumed distribution function differs from the stars' DF?

Collection of possible tests and plots

- *Test 1:* mix hot and cold populations, 5 free qdf parameters in analysis!, use code that estimates the best velocity integration ranges. h_sigma_r & h_sigma_z are the same for both populations, sigma_r and sigma_z have the same ratio, but are 50% different for the two populations. h_R is also 50% different. Vary the fraction of pollution. Idea behind this: What if the stellar distribution has a different shape, e.g. added "wings", or had a different tracer density decrease with R. Would be however great, if we could show how the mixture of qdf's quanlitatively changes the shape of the df. Any ideas? *Plot 1:* Violin plot: x-axis - fraction of pollution. y-axis: b-parameter and one or two qdf parameters.

- *Test 2:* same as Test 1, but this time vary the degree of difference and make it 50% pollution. Idea behind this: What happens, if we have errors in the abundances and mix different MAPs? For this it would be could to compare how much the qdf parameters of neighbouring MAPs differ and how big the difference between MAPs can be, such that it still can reproduce the potential.
  Plot 2: Violin plot: x-axis - difference in qdf parameters. y-axis: b-parameter and one or two qdf parameters.

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

Collection of possible tests and plots *Test 1:* Try to recover a Miyamoto-Nagai disk + power-law halo potential by fitting a 2-component Stäckel potential.
Plot 1:

- (R,z)-plane: color coding: difference between true potential's F_R and best fit potential F_R

- (R,z)-plane: color coding: difference between true potential's F_z and best fit potential F_z
  Any idea how to account for the error bars on the best fit potential?

4. Conclusion

[TO DO]

5. 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

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

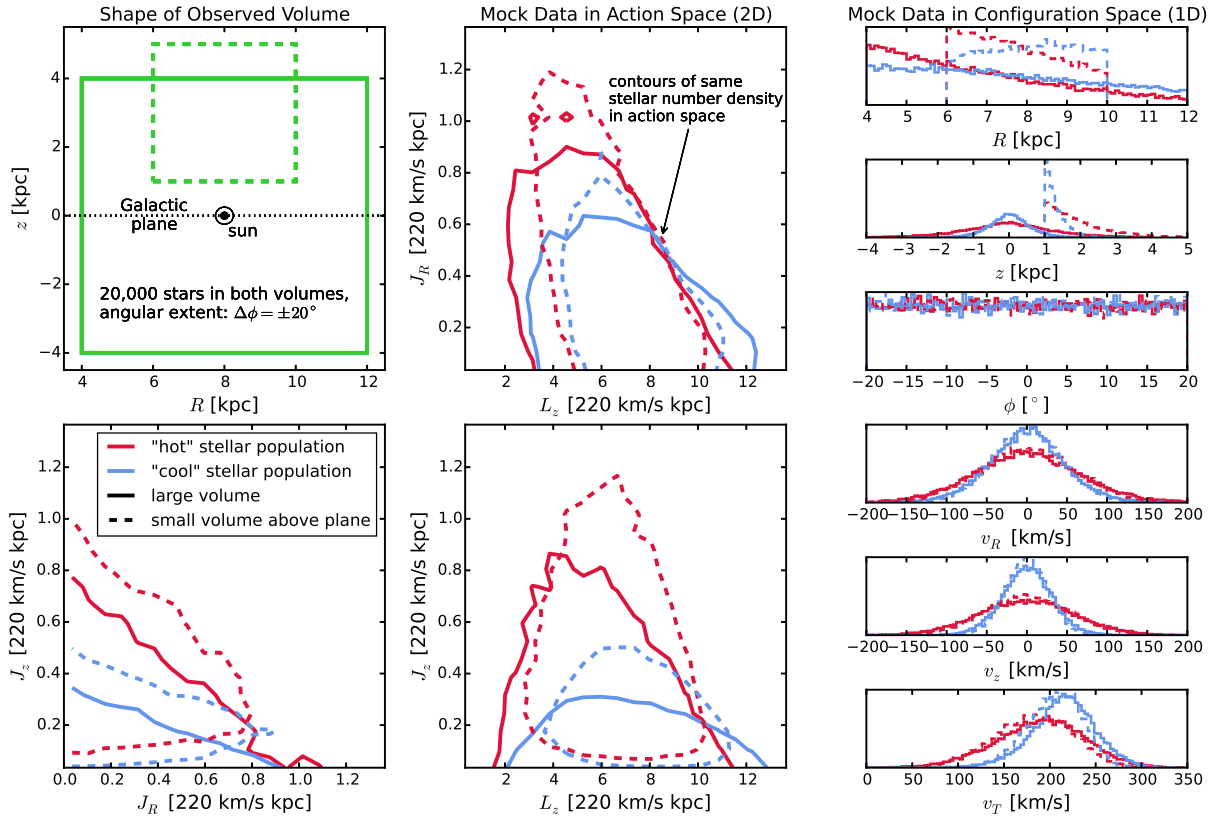

Fig. 1.— Distribution of mock data in action space (2D iso-density contours in the two central and the lower left panel) and configuration space (1D histograms in right panels), depending shape and position of observation volume (cf. lower left panel) and 'hottness' of the stellar population. The mock data was created in a 2-component KK-Staeckel-potential (cf. ???) with parameters $p_\Phi = \{v_{\rm circ}, \Delta, (a/c)_{\rm disk}, (a/c)_{halo}, k)\} = \{230 \text{ km s}^{-1}, 0.3, 20., 1.07, 0.28\}$ [TO DO: Does Δ have units????] (which is an approximate fit to the MilkyWay2014 potential in Galpy). We use two stellar populations, a 'hot' one with $p_{DF,hot} = \{h_R, \sigma_R, \sigma_z, h_{\sigma_R}, h_{\sigma_z}\} = \{2 \text{ kpc}, 55 \text{ km s}^{-1}, 66 \text{ km s}^{-1}, 8 \text{ kpc}, 7 \text{ kpc }\}$ (red lines) and a 'cool' population with $p_{DF,cool} = \{h_R, \sigma_R, \sigma_z, h_{\sigma_R}, h_{\sigma_z}\} = \{3.5 \text{ kpc}, 42 \text{ km s}^{-1}, 32 \text{ km s}^{-1}, 8 \text{ kpc}, 7 \text{ kpc }\}$ (blue lines). In the upper left panel we demonstrate the shape of the two different observation volumes within which we were creating each a 'hot' and a 'cool' mock data set: a large volume centered on the Galactic plane (solid lines in all plots) and a smaller one above the plane (dashed lines in all plots). Both volumes have an angular extent of $\Delta\phi \pm 20°$. Each of the four mock data sets compared in this plot has 20,000 stars in it. The stars of the 'cool' population have in general lower radial and vertical actions, i.e. are on more circular obrits. The different ranges of L_z's in the two volumes reflect $L_z \sim R v_{\rm circ}$ 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 Galacto-centric 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. [TO DO: How many percent do the contours enclose????]

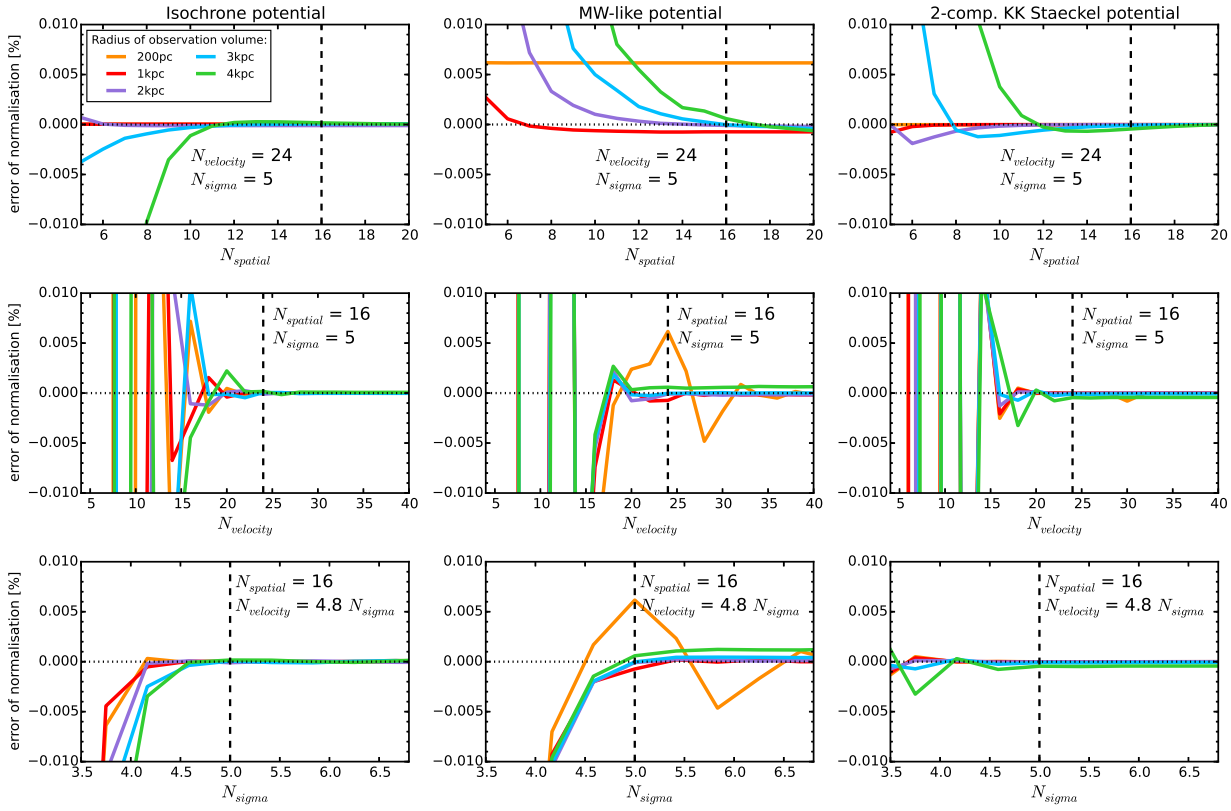

Fig. 2.— Relative error of the likelihood normalization in eq. (4) and (5) depending on the accuracy of the density calculation in §2.2.1. The different colors represent calculations for different radii of the spherical observation volume around the sun, as indicated in the legend. $N_{\rm spatial}$ 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. (2). 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 $p_{\rm DF} = \{h_R, \sigma_R, \sigma_z, h_{\sigma_R}, h_{\sigma_z}\} = \{2 \text{ kpc}, 55 \text{ km s}^{-1}, 66 \text{ km s}^{-1}, 8 \text{ kpc}, 7 \text{ kpc }\}$. We show the results for three different potentials, an isochrone potential with parameters $p_\Phi = \{v_{\rm circ}, b\} = \{230 \text{ km s}^{-1}, 0.9 \text{ kpc }\}$, a MW-like potential (cf. ???) with parameters $p_\Phi = \{v_{\rm circ}, R_d, z_h, f_h, \frac{\mathrm{d} \ln v_c}{\mathrm{d} \ln R}]\} = \{230 \text{ km s}^{-1}, 2.5 \text{ kpc}, 400 \text{ pc }, 0.8, 0\}$ and a 2-component KK-Staeckel potential with parameters $p_\Phi = \{v_{\rm circ}, \Delta, (a/c)_{\rm disk}, (a/c)_{halo}, k)\} = \{230 \text{ km s}^{-1}, 0.3, 20., 1.07, 0.28\}$. (Caption continues on next page.)

Fig. 2.— (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 $N_{spatial}$.???]

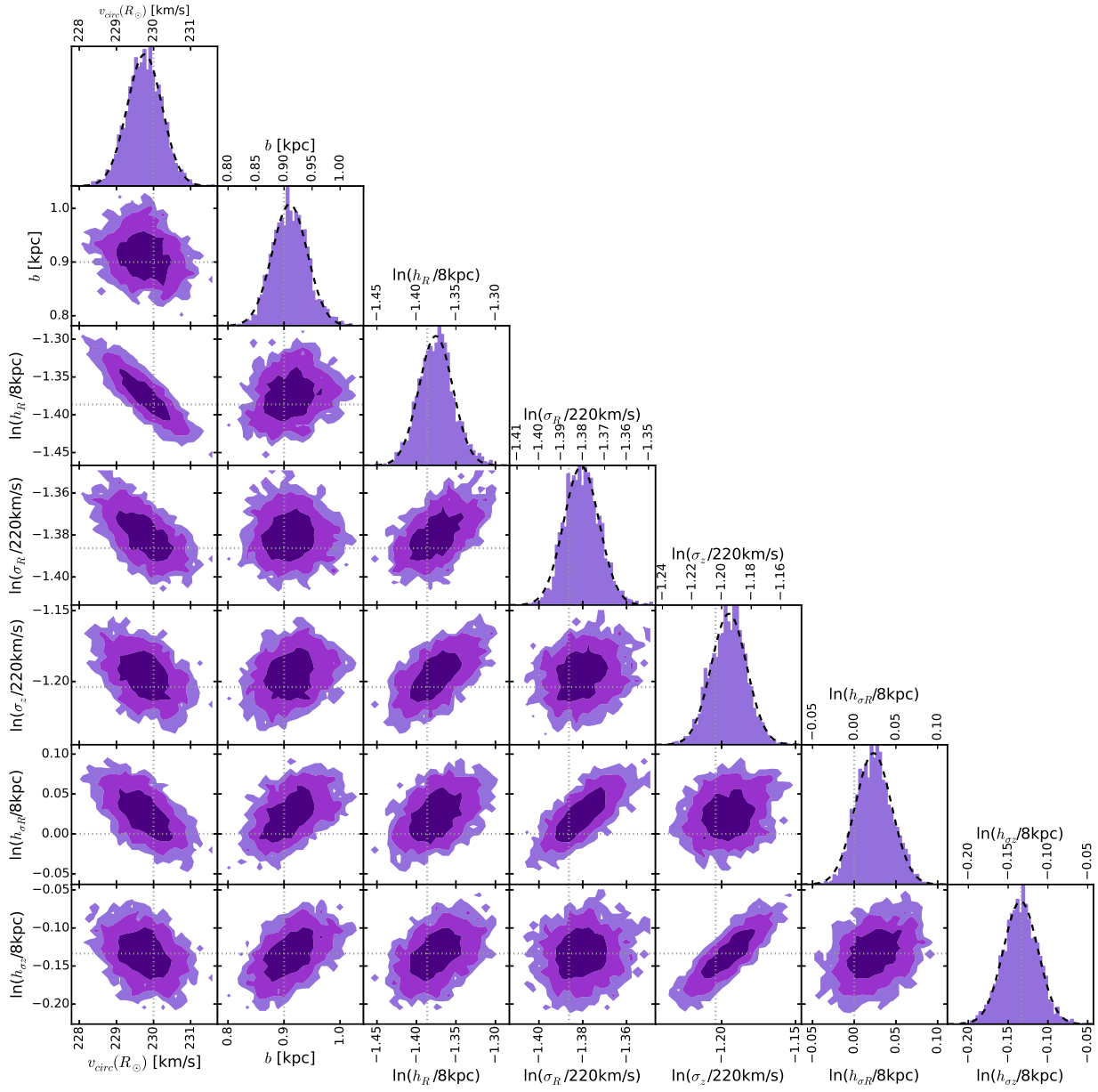

Fig. 3.— The likelihood in eq. (???) in the parameter space $\{p_\Phi, \ln(p_{\rm DF})\}$ for one example mock data set. This mock data set has 20,000 stars and was created in a isochrone potential with $p_\Phi = \{v_{\rm circ}, b\} = \{230 \text{ km s}^{-1}, 0.9 \text{ kpc }\}$, observed within a spherical volume around the sun of radius $r = 2$ kpc , and represents a rather hot stellar population with DF parameters $p_{\rm DF} = \{h_R, \sigma_R, \sigma_z, h_{\sigma_R}, h_{\sigma_z}\} = \{2 \text{ kpc}, 55 \text{ km s}^{-1}, 66 \text{ km s}^{-1}, 8 \text{ kpc}, 7 \text{ kpc }\}$. The true parameters are marked by dotted lines. The dark, medium and bright purple 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 multi-variate 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." ???

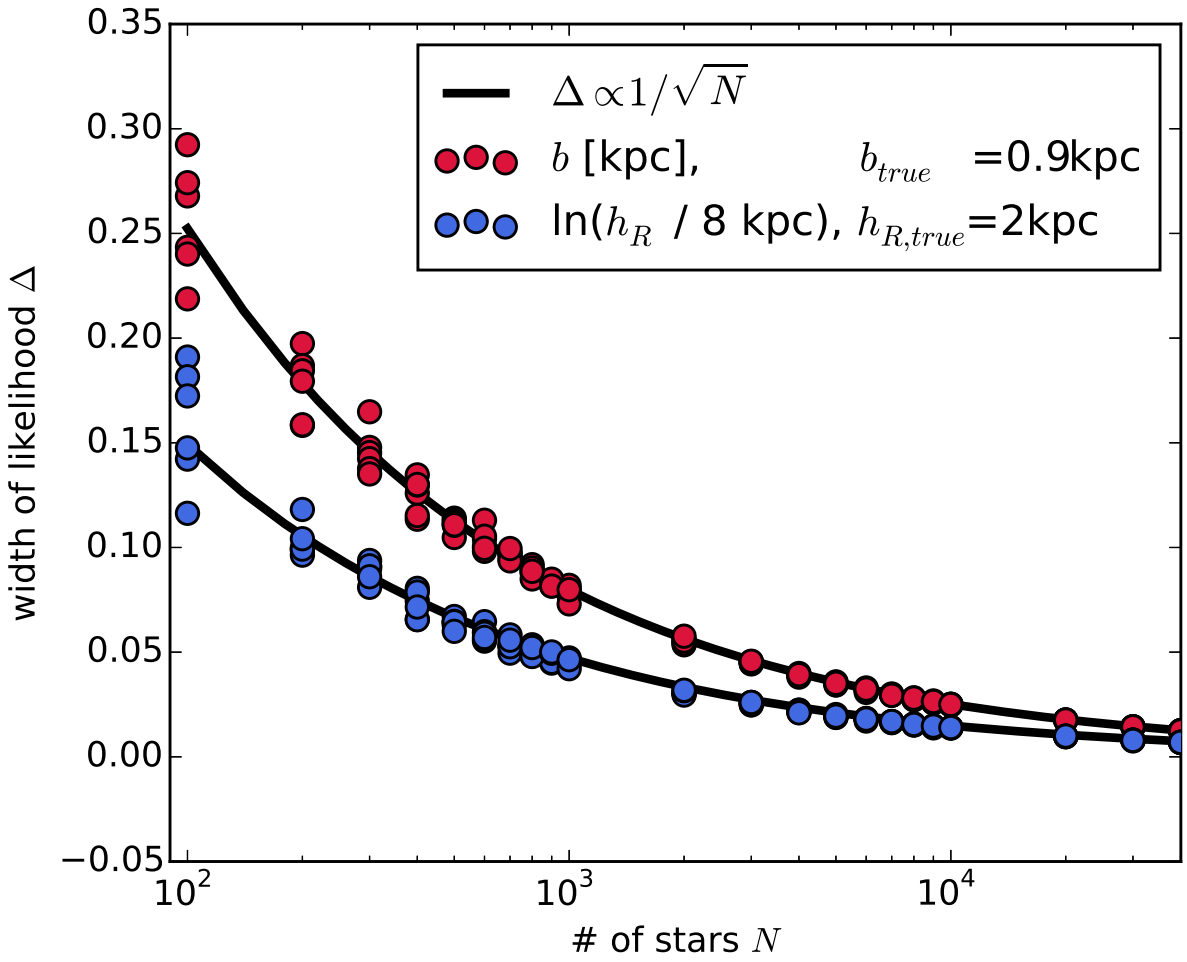

Fig. 4.— The width of the likelihood for two fit parameters found from analyses of 132 mock data sets vs. the number of stars in each data set. All mock data sets were created in an isochrone potential with $p_\Phi = \{v_{\rm circ}, b\} = \{230 \text{ km s}^{-1}, 0.9 \text{ kpc }\}$, for a qDF with $p_{\rm DF} = \{h_R, \sigma_R, \sigma_z, h_{\sigma_R}, h_{\sigma_z}\} = \{2 \text{ kpc}, 55 \text{ km s}^{-1}, 66 \text{ km s}^{-1}, 8 \text{ kpc}, 7 \text{ kpc }\}$ and within a spherical observation volume around the sun of radius $r = 3$ kpc . The data sets have different sample sizes and contain between 100 and 40,000 stars, as indicated on the x-axis. The likelihood was evaluated on a grid in the parameters $\{b, \ln(h_R/8\text{kpc}), \ln(\sigma_R/230\text{km s}^{-1}), \ln(h_{\sigma_R}/8\text{kpc})\}$, while all other parameters were assumed to be known and kept at their true values. For each fit parameter the likelihood was then marginalized by summing over the grid and then a Gauss curve was fitted to the marginalized likelihood. The standard deviation of these best fit Gaussians Δ is shown on the y-axis for b in kpc (red dots) and for $\ln(h_R/8\text{kpc})$ 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).???]

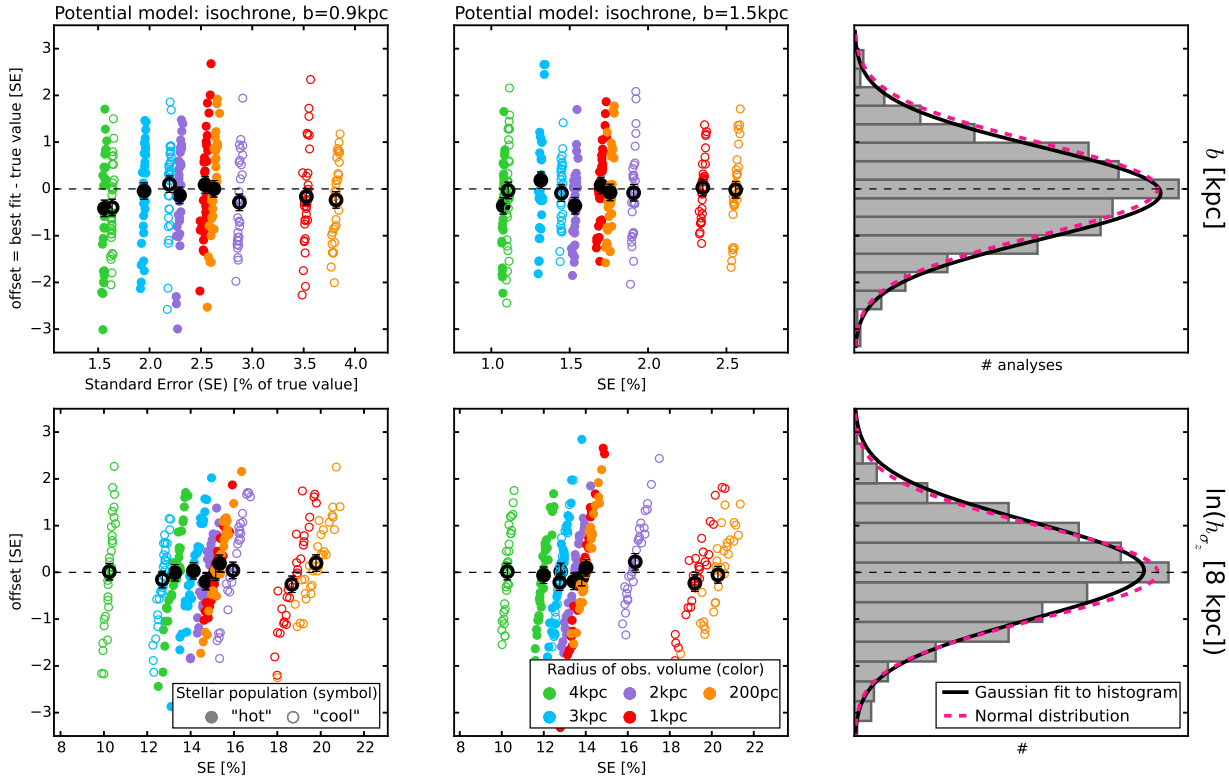

Fig. 5.— (Un-)bias of the parameter estimate: According to the central limit theorem the likelihood will follow a Gaussian distribution for a large number of stars. From this follows that also for a large number of data sets the corresponding best fit values for the model parameters have to follow a 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. 5.— (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????????]

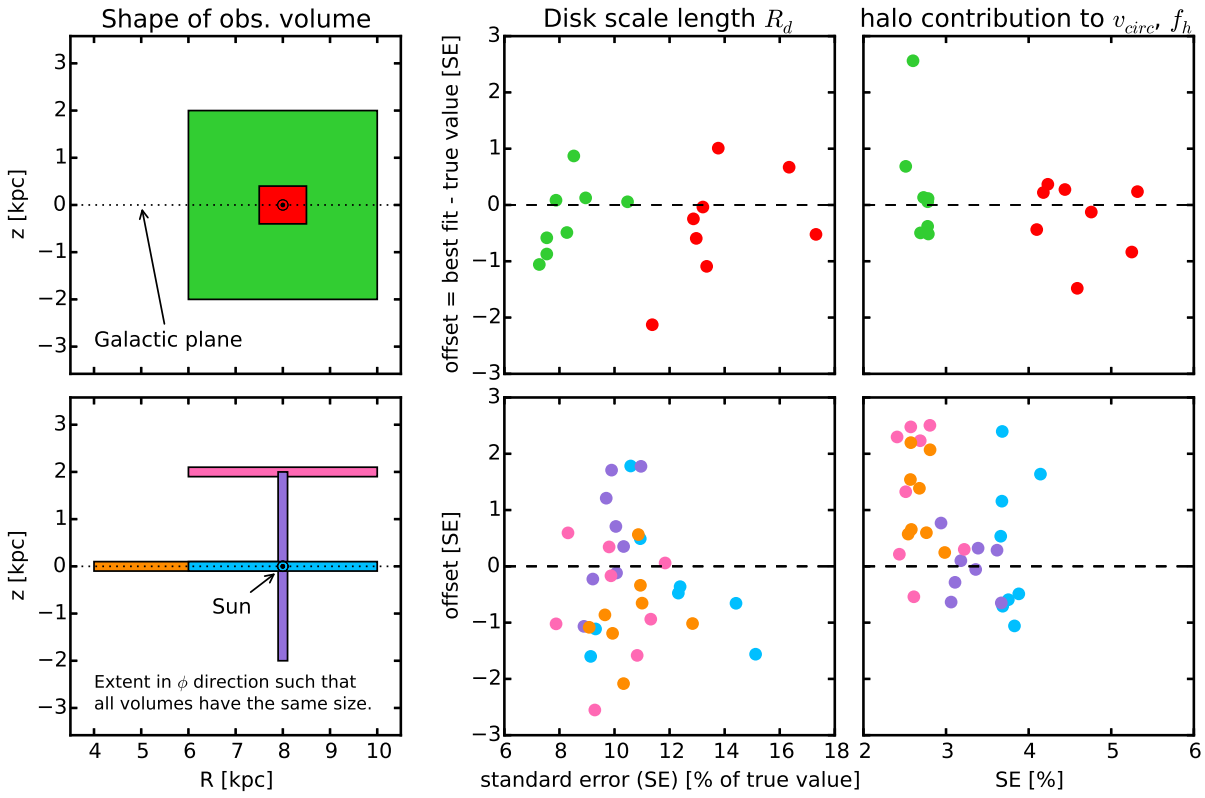

Fig. 6.— We demonstrate that for a given size of the observation volume the shape and position of the volume does not matter much as long as we have both large radial and/or vertical coverage. The left column shows the position of our test observation volumes within the Galaxy with respect to the Galactic 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 $p_\Phi = \{v_{\rm circ}, R_d, z_h, f_h, \frac{{\rm d}\ln v_c}{{\rm d}\ln R}]\} = \{230 \text{ km s}^{-1}, 2.5 \text{ kpc}, 400 \text{ pc }, 0.8, 0\}$ and a 'hot' stellar population with $p_{\rm DF} = \{h_R, \sigma_R, \sigma_z, h_{\sigma_R}, h_{\sigma_z}\} = \{2 \text{ kpc}, 55 \text{ km s}^{-1}, 66 \text{ km s}^{-1}, 8 \text{ kpc}, 7 \text{ kpc }\}$. We evaluate the likelihood on a grid in the fit parameter $\{R_d, f_h, \ln(h_R/8\text{kpc}), \ln(\sigma_R/230\text{km s}^{-1}), \ln(h_{\sigma_R}/8\text{kpc})\}$. All other parameters are kept at their true values in the modelling. Standard error and offset were determined as in fig. 5. 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.

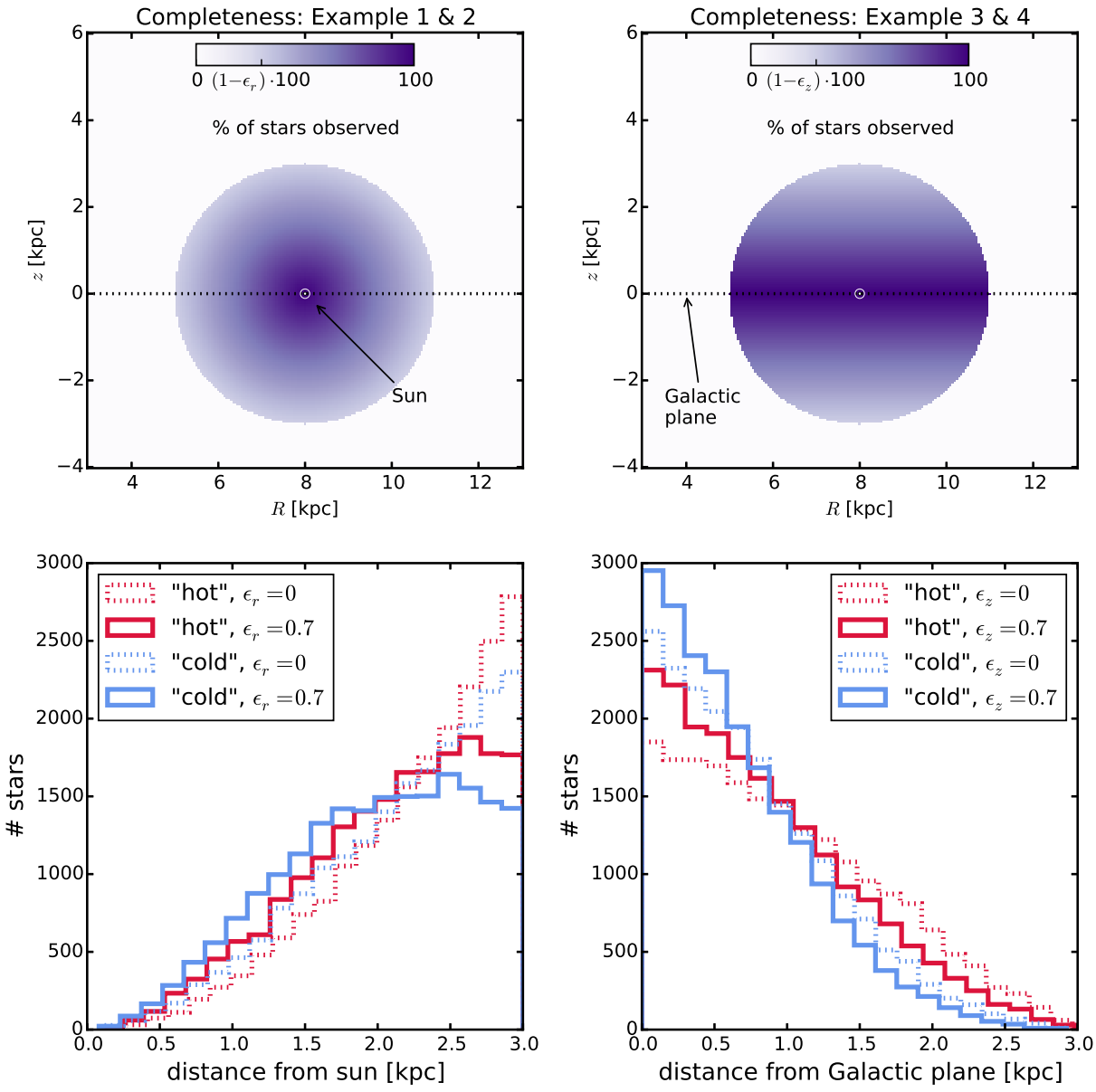

Fig. 7.— Caption [TO DO] ([TO DO] Re-do, if new analyses are in violin plot.)

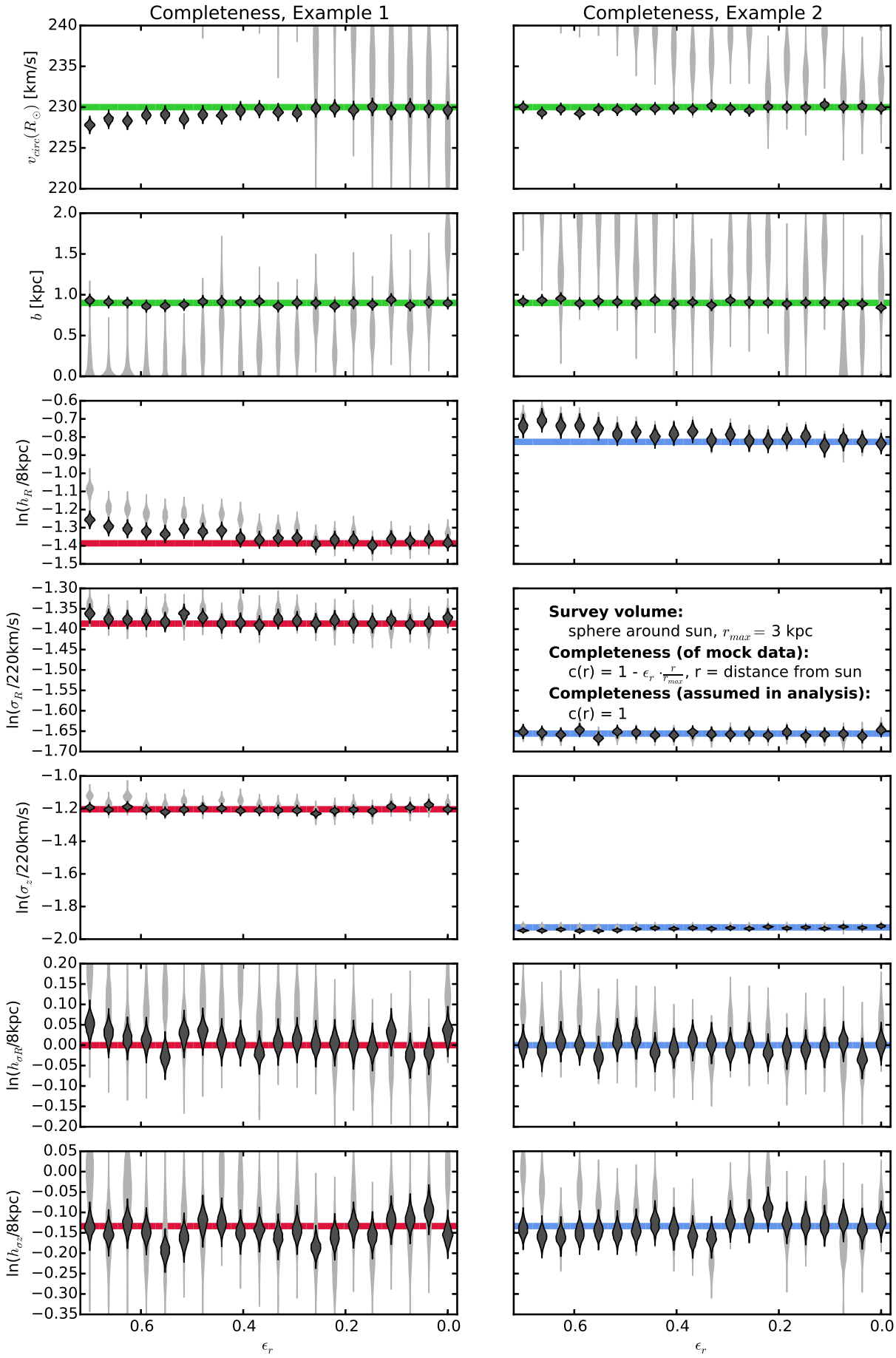

Fig. 8.— Caption [TO DO] (This was done using the current qDF to set the fitting range.

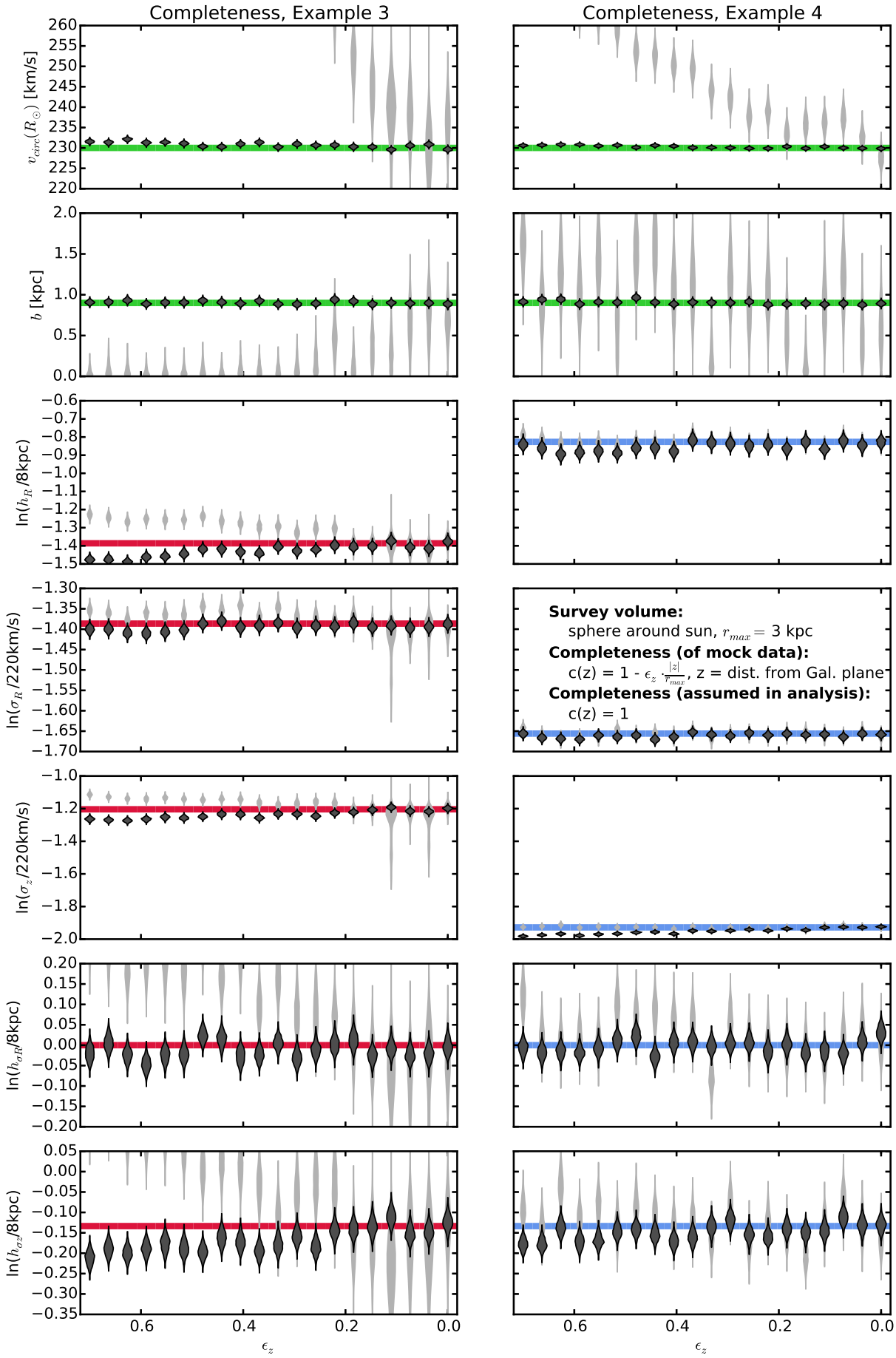

Fig. 9.— Caption [TO DO] (This was done using the current qDF to set the fitting range.

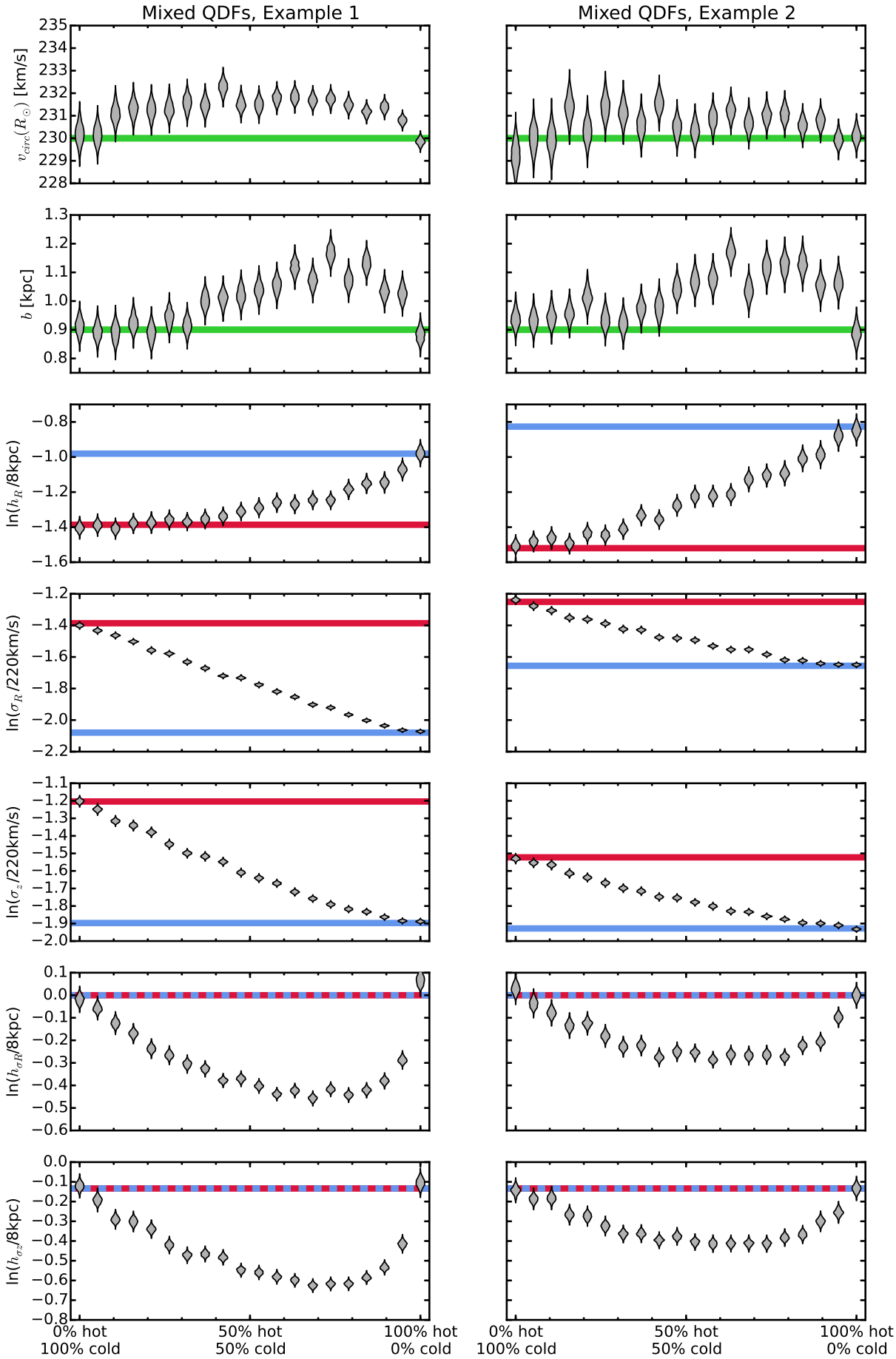

Fig. 10.— (Caption on next page.)

Fig. 10.— (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. In total there are always 20,000 stars in the data set. Both populations come from same potential, an isochrone potential with $p_\Phi = \{v_{\rm circ}, b\} = \{230 \text{ km s}^{-1}, 0.9 \text{ kpc }\}$ (true parameters are indicated by green lines). The composite data set is then fit them with one single qDF. Example 1 (left) mixes the 'hot' population $p_{\rm DF,hot,1} = \{h_R, \sigma_R, \sigma_z, h_{\sigma_R}, h_{\sigma_z}\} = \{2 \text{ kpc}, 55 \text{ km s}^{-1}, 66 \text{ km s}^{-1}, 8 \text{ kpc}, 7 \text{ kpc }\}$ with the 'cool' population $p_{\rm DF,cool,1} = \{h_R, \sigma_R, \sigma_z, h_{\sigma_R}, h_{\sigma_z}\} = \{2 \text{ kpc} + 50\%, 55 \text{ km s}^{-1} - 50\%, 66 \text{ km s}^{-1} - 50\%, 8 \text{ kpc}, 7 \text{ kpc }\}$. Example 2 (right) mixes the 'cool' population $p_{DF,cool,2} = \{h_R, \sigma_R, \sigma_z, h_{\sigma_R}, h_{\sigma_z}\} = \{3.5 \text{ kpc}, 42 \text{ km s}^{-1}, 32 \text{ km s}^{-1}, 8 \text{ kpc}, 7 \text{ kpc }\}$ with the 'hotter' population $p_{DF,hot,2} = \{h_R, \sigma_R, \sigma_z, h_{\sigma_R}, h_{\sigma_z}\} = \{3.5 \text{ kpc} - 50\%, 42 + 50\% \text{ km s}^{-1}, 32 + 50\% \text{ km s}^{-1}, 8 \text{ kpc}, 7 \text{ kpc }\}$. The parameters were chosen such, that the two parameter sets have the same σ_R/σ_z ratio and 'hotter' populations have shorter tracer scale lengths. The velocity dispersion scale lengths were fixed according to ???Bovy2012???. True parameters of the 'hotter' population are shown as red lines, those of the 'cooler' populations as blue lines. The violines represent the marginalized likelihoods found from the MCMC analysis. [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: Change 'cold' in plot to 'cool'. ???] [TO DO: Write 'all hot' instead of '100% hot'.???]

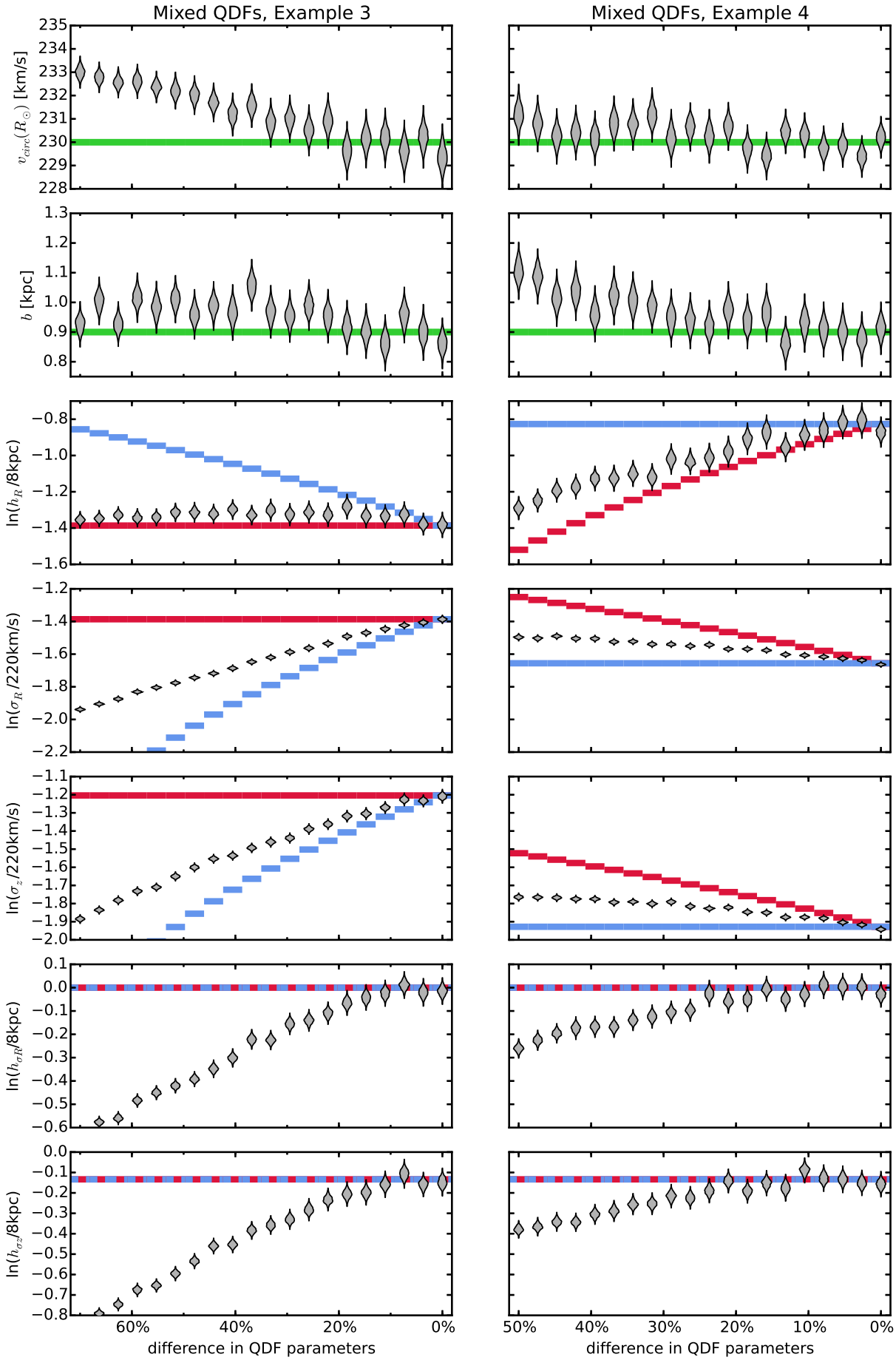

Fig. 11.— (Caption on next page.)

Fig. 11.— (Continued.) The dependence of the parameter recovery on the difference in DF parameters of the mixture of two stellar populations and their 'hotness'. [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] [TO DO: Add in plot a label, that it is a 50%/50% mix of a hot and a cold population.???])

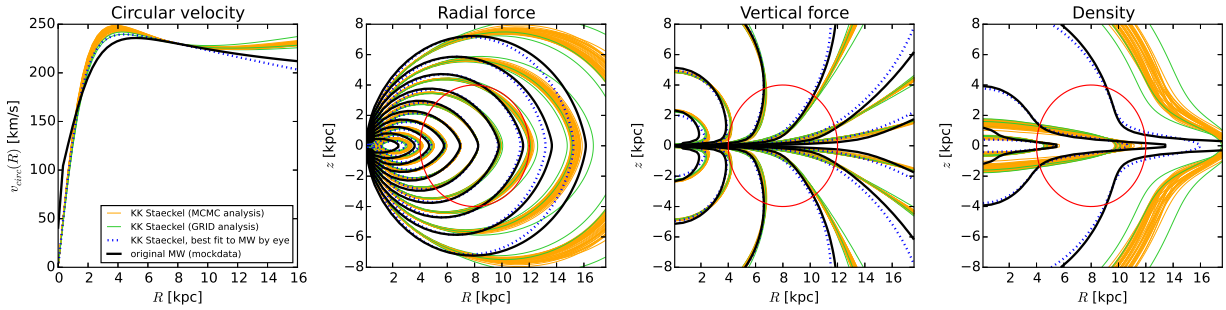

Fig. 12.— [TO DO]