not done a full exploration of all combinations of p_M and volumina.

3.3. The Significance of Incorrect assumptions on the (In-)Completeness of the Data Set

The selection function of a survey could be 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 §2.3). 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 around the sun and an incompleteness function that drops linearly with distance r from the sun (see Test 5, Example 1, in Table 3 and Figure 8).

This could be understood as a model for t~~he 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 Figure 9). 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 §A.1 we also show that the robustness is somewhat less striking but still ~~given~~ for small misjudgments of the incompleteness in vertical direction, parallel to the disk plane (Figures 19 and 20). This could model the effect of wrong corrections for dust obscuration in the plane. We also investigate in Appendix §A.1 if indeed most of the information is stored in the rotation curve. For this we use the same mock data sets as analysed in Figures 9 and 20, 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 Figure 21).

3.4. ~~Dealing with~~ Measurement Errors and their Effect on the Parameter Recovery

Convergence of the error integral. In §2.5 we introduced how we convolve the model probability with the measurement errors. In the absence of distance errors, the accuracy of

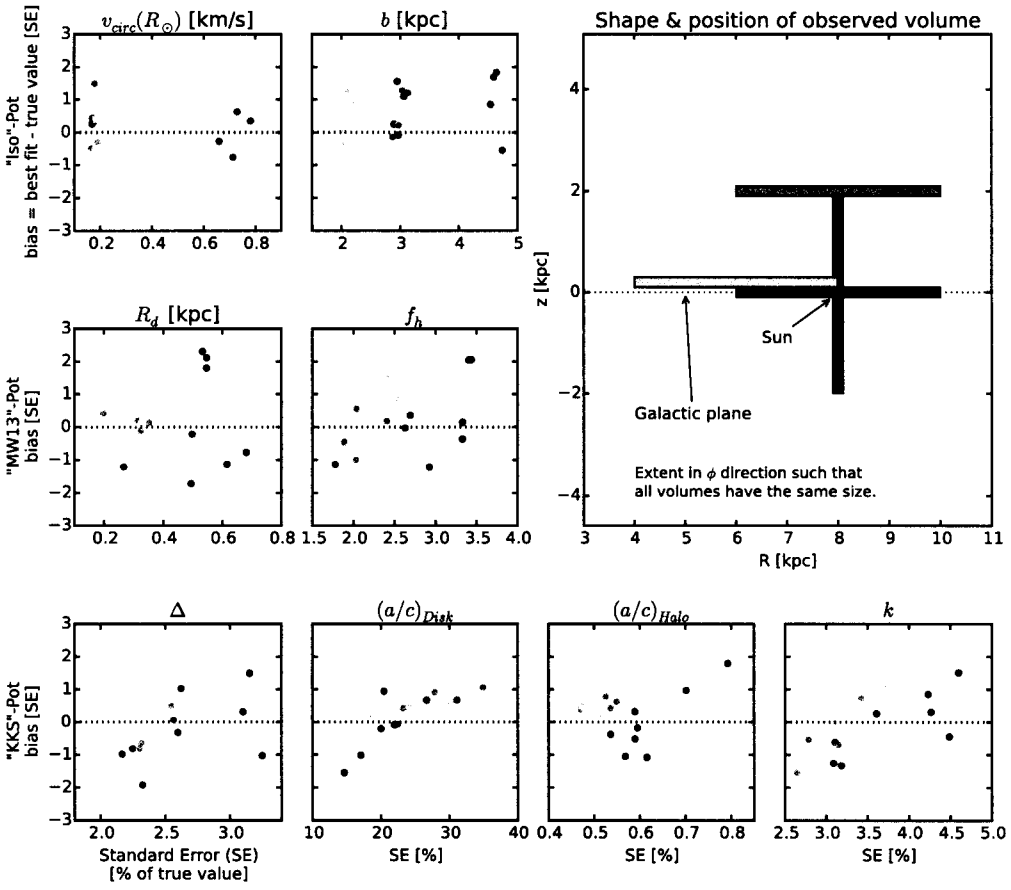


Fig. 7.— Bias vs. standard error in recovering the potential parameters for mock data stars drawn four different test observation volumes within the Galaxy (illustrated in the upper right panel) and three different potentials ("Iso-Pot", "MW13-Pot" and "KKS-Pot" from Table 1, top to bottom row). Standard error and offset were determined as in Figure 6. Per volume and potential we analyse four different mock data realisations; all model parameters are shown as Test 4 in Table 3. The colour-coding represents the different wedge-shaped observation volumes. The angular extent of each wedge-shaped observation volume was adapted such that all have the volume of 4.5 kpc^3, even though their extent in (R, z) is different. Overall there is no clear trend, that an observation volume around the sun, above the disk or at smaller Galactocentric radii should give remarkably better constraints on the potential the other volumes. [TO DO: MW-Pot and KKS-Pot analyses suffer from too low accuracy in action calculation (with StaeckelGrid). Used the StaeckelGrid for BOTH mock data and analysis, but the mock data distribution would actually not look exactly like the desired qDF distribution, i.e. this plot basically is created with a messed up DF. Don't know how higher accuracy would change the plot.]

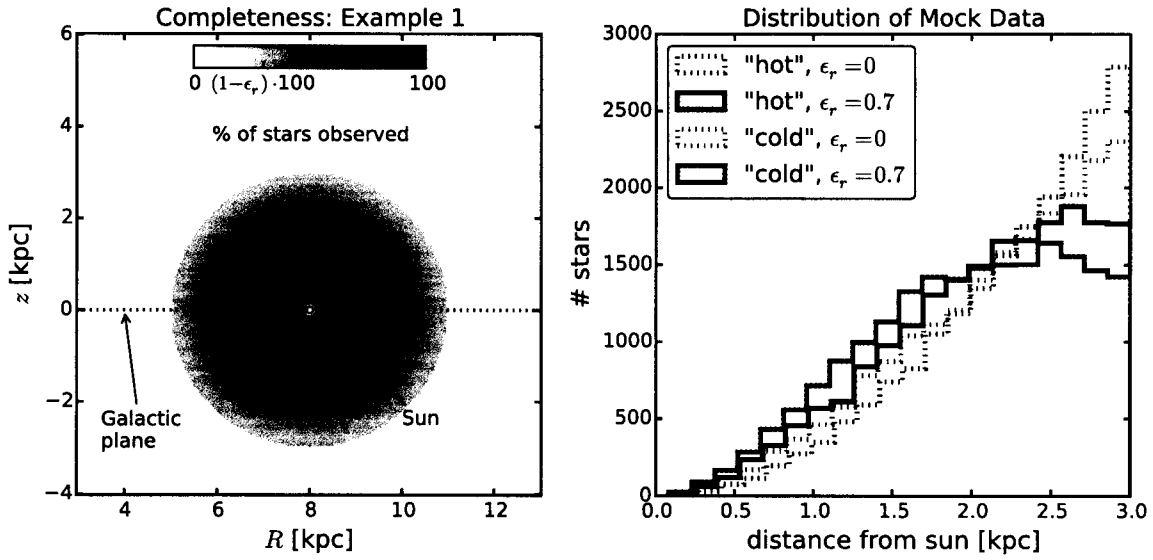
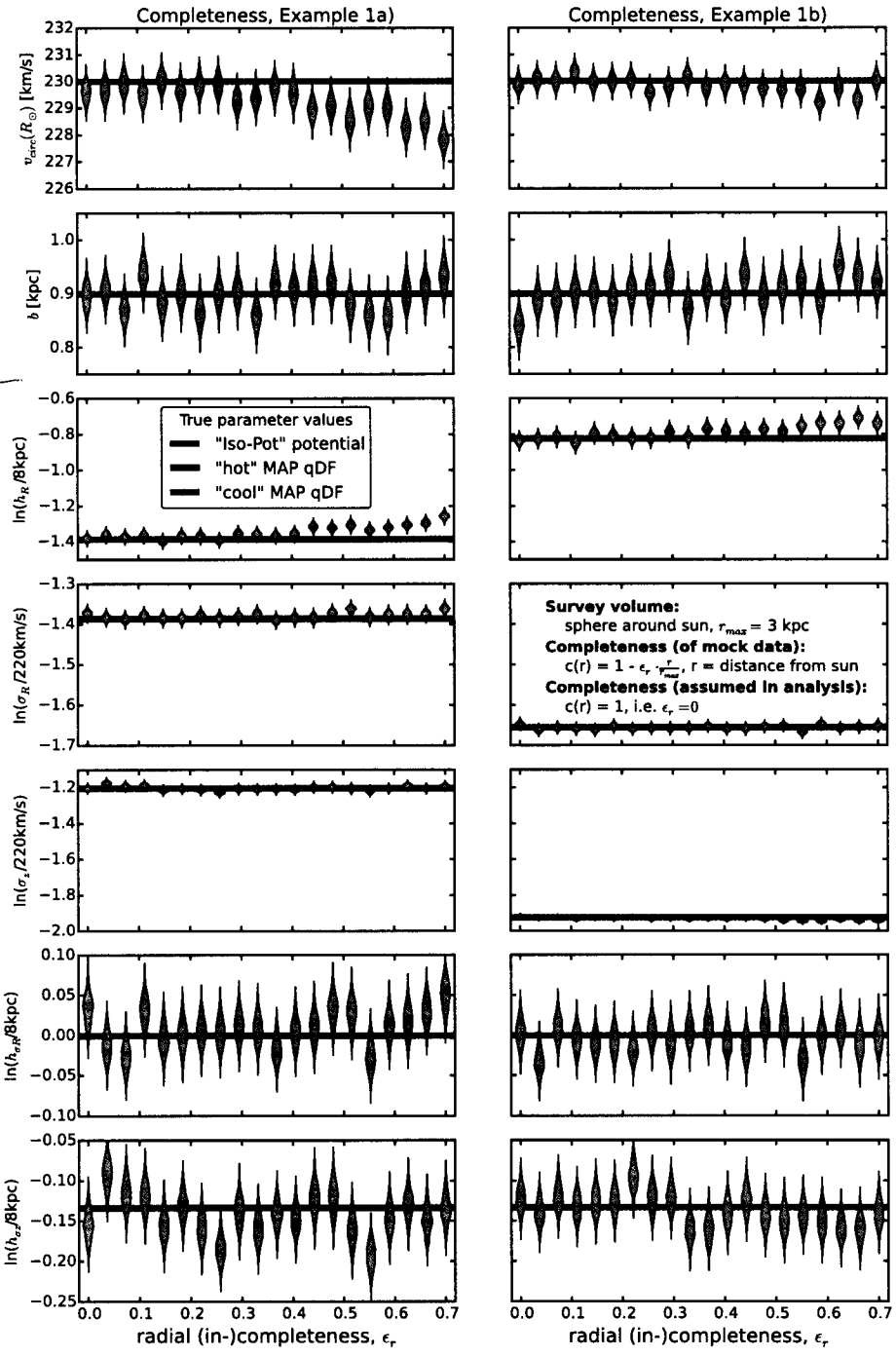


Fig. 8.— Selection function and mock data distribution for investigating radial incompleteness of the data. All model parameters are summarized as Test 5, Example 1, in Table 3. The survey volume is a sphere around the sun and the percentage of observed stars is decreasing linearly with radius from the sun, as demonstrated in the left panel. How fast this detection/incompleteness rate drops is quantified by the factor ϵ_r. Histograms for four data sets, drawn from two *MAPs* ("hot" in red and "cool" in blue, see Table 2) and with two different ϵ_r, 0 and 0.7, are shown in the right panel for illustration purposes.



Maybe remove all of these & only show potential parameters This figure takes up a lot of space

Fig. 9.— (Caption on next page.)

Fig. 9.— Influence of wrong assumptions about the radial incompleteness of the data on the parameter recovery with *RoadMapping* . Each mock data set was created having different incompleteness parameters ϵ_r (shown on the x-axis and illustrated in Figure 8) and the model parameters are given as Test 5, Example 1, in Table 3. The analysis however didn't know about the incompleteness and assumed that all data sets had constant completeness within the survey volume ($\epsilon_r = 0$). The marginalized likelihoods from the fits are shown as violins. The green lines mark the true potential parameters ("Iso-Pot") and the red and blue lines the true qDF parameters ("hot" *MAP* in red and "cool" *MAP* in blue), which we tried to recover. The *RoadMapping* method seems to be very robust against small to intermediate deviations between the true and the assumed data incompleteness. [TO DO: rename $h_{\sigma R}$ to $h_{\sigma,R}$, σ_R to $\sigma_{R,0}$ and analogous for z]

the parameter recovery is limited by an insufficient MC sampling of the convolution integral in Equation (15). Test 6.1 in Table 3 and Figure 10 investigate how many MC samples are needed, given the size of the velocity error, for the integral to be accurate within certain limits:
For each $\delta\mu \in [2, 3, 4, 5]$mas yr^{-1} we set up four mock data sets and evaluated the likelihood for different $N_{\rm error}$. We used $N_{\rm conv} := 800$ and 1200 MC samples to calculate the numerically converged likelihood for proper motion errors $\delta\mu \leq 3$mas yr^{-1} and $\delta\mu > 3$mas yr^{-1}, respectively (see left panels in Figure 11). We determined the mean bias

$$\mathrm{BIAS}(N_{\rm error}, \delta\mu) \equiv \frac{1}{4}\sum_{j=1}^{4} \left[\langle p_i \rangle (N_{\rm error}, \delta\mu)\right]_j - \left[\langle p_i \rangle (N_{\rm conv}, \delta\mu)\right]_j ,$$

where $[\langle p_i \rangle (N_{\rm error}, \delta\mu)]_j$ is the best estimate for the i-th model parameter $p_i \in p_M$ from the analysis of the j-th mock data realisation with $\delta\mu$ using $N_{\rm error}$ MC samples. From this we then generated the curves $N_{{\rm error},i}(\delta v_{\rm max}, \mathrm{BIAS})$ in Figure 10 by linear interplolation, that show how many MC samples are needed for parameter p_i given a velocity error and a systematic bias in units of the standard error (SE) of the estimate. The proper motion error $\delta\mu$ translates to a velocity error according to

$$\delta v_{\rm max}[\mathrm{km\ s^{-1}}] \equiv 4.74047 \cdot r_{\rm max}[\mathrm{kpc}] \cdot \delta\mu[\mathrm{mas\ yr^{-1}}], \qquad (16)$$

where $r_{\rm max}$ is the maximum distance of stars. We find in Figure 10 the relation

$$N_{{\rm error},i}(\delta v_{\rm max}, \mathrm{BIAS}) \propto (\delta v_{\rm max})^2 .$$

Figure 10 also demonstrates that different model parameters do not have the same sensitivity to the numerical inaccuracies introduced by insufficient sampling. [TO DO: we haven't tested yet, if this plot depends on hotness of population and / or number of stars.] But if it takes forever to actually do the calculations, I guess, we just leave it like this.]

Testing the error convolved likelihood approximation. In absence of distance (modulus) errors our approximation for the likelihood, which is the model probability convolved with the measurement errors in Equation (15), is equal to the true likelihood. In case there *are* distance modulus errors, this likelihood links the range of possible velocities (specified by the measurement errors in line-of-sight velocity, proper motion and distance modulus) to a fixed but slightly wrong position, as we ignore the distance error in the position. As the link between position and velocity provides the information about the potential, this will lead to systematic biases in the parameter recovery the larger the distance error becomes. In Test 6.2 in Table 3 and Figure 11 we investigate the capabilities of Equation (15) with and without distance modulus errors.

The left column of panels in Figure 11 shows how well the approximation works in the absence of distance errors. There seemed to be no biases in the parameter recovery, independent of the size of the proper motion error. (We note that there could be a tiny bias $\ll 1$ SE in the $\ln(\sigma_{z,0}/200 \text{ km s}^{-1})$ qDF parameter, most likely due to insufficient numerical accuarcy, but all the other model parameters, also those not shown in the figure, are very well behaved.) Overall the standard errors on the recovered parameters are quite small (a few percent at most for 10,000 stars), which demonstrates that, if we perfectly knew the measurement errors, we still could get very precise constraints on the potential. The constraints also get tighter the smaller the proper motion error becomes. We found that for $\delta\mu = 1$ mas yr^{-1} the precision of the recovered parameters reduce by $\sim$ half compared to $\delta\mu = 5$ mas yr^{-1}.

The right column of panels in Figure 11 demonstrates the failure of our adopted likelihood approximation in the case of large distance modulus errors. The larger the $\delta(m-M)$, the wronger the recovered parameters become: The systematic biases can get many SEs large. We find however that in case of $\delta(m-M) \leq 0.2$ mag (if also $\delta\mu \leq 2$ mas yr^{-1} and a maximum distance of $r_{\max} = 3$ kpc, see Test 6.2 in Table 3) the parameters can still be recovered within 2 SEs. For most model parameters (except $\ln(\sigma_{z,0}/200 \text{ km s}^{-1})$, as shown in the figure, and $\ln(h_R/8 \text{ kpc})$) even $\delta(m-M) \leq 0.3$ mag still gives biases smaller than 2 SEs. This corresponds to a relative distance error of $\sim 10\%$. This encourages us that for smaller distance modulus errors we really could use our likelihood approximation in Equation (15), which is computationally cheaper than a proper treatment, also on real data sets.

Underestimation of the proper motion error. We found that in case we perfectly knew the measurement errors (and the distance error is negligible), the convolution of the model probability with the measurement errors gives precise and accurate constraints on the model parameters - even if the error itself is quite large. Now we investigate what would happen if the quoted measurement errors, e.g. the proper motion errors, were actually smaller than the true errors. Figure 12 shows the case for two different stellar populations

and an error underestimation of 10% and 50%.
Overall the parameter recovery gets worse the larger the proper motion error and the stronger the underestimation. The relation between the bias due to error misjudgment and the size of the proper motion error seems to be linear.
For the recovery of the isochrone potential scale length b the hotness of the population does not matter (see lower left panel in Figure 12). The circular velocity $v_{\rm circ}(R_\odot)$ is, as always, better measured by cooler than by hotter populations (see upper left panel in Figure 12).
We find that the recovery of the qDF parameters on the other hand is more strongly affected by the misjudgment of the velocity error for *cooler* stellar popluations. The measured velocity dispersion is the convolution of the intrinsic dispersion with the measurement errors. If the proper motion error is underestimated, the deconvolved velocity dispersion is larger than the intrinsic velocity dispersion and the relative difference is bigger for a cooler population (see upper right panel for $\sigma_{z,0}$ in Figure 12). The intrinsic velocity dispersion is also cooler at larger radii than at smaller radii, therefore the deconvolved dispersion is overestimated more strongly at large R and the velocity dispersion scale length will be overestimated as well (see lower left panel for $h_{\sigma,z}$ in Figure 12). We get analogous results for the qDF parameters $\sigma_{R,0}$ and $h_{\sigma,R}$. The recovery of the tracer density scale length h_R is not affected by the misjudgment of velocity errors.
The most important and encouraging result from Figure 12 is, that for an underestimation of 10% the bias is still $\lesssim$ 2 sigma [TO DO: can I say sigma???] - even for proper motion errors of almost 3 mas/yr.

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, *MAP*s do not follow a qDF. Or, even if they do follow a qDF, the finite abundance errors effectively mix different *MAP*s . We investigate both these issues by creating mock data sets (Figure 13) 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. 14 and 15. Following the observational evidence, *MAP*s with cooler qDFs also have longer tracer scale lengths. In the first set of test, we choose qDFs of widely different temperatures and vary their relative fraction (dubbed "Examples 1a/b" in Figure 14 and Test 7 in Table 3); in the second set of tests ("Examples 2a/b" in Figure 15 and Test 7in Table 3), we always mix mock data points from two different qDFs in equal proportion, but vary by how much the qDF's temperatures differ.

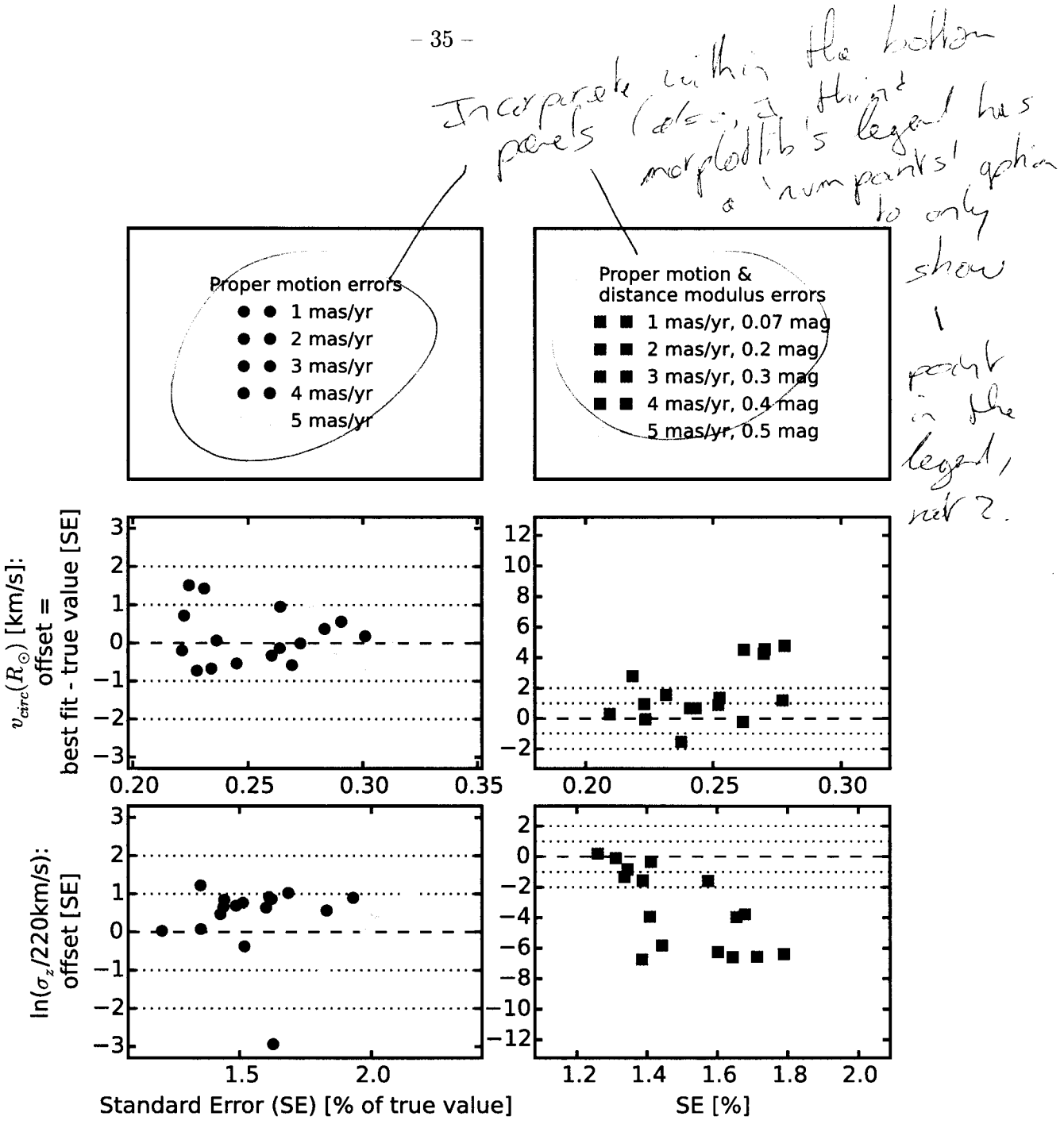


Fig. 11.— (Caption on next page.)

Fig. 11.— Parameter recovery using the approximation for the measurement error convolved likelihood in Equation (15). All model parameters used to create the mock data sets analyzed for this figure are given as Test 6.2 in Table 3. The mock data sets in the left panels have only errors in line-of-sight velocity and proper motions, while the data sets in the right panels also have distance modulus errors, as indicated in the legends in the first row. The size of the error is color coded. The other panels plot the offset of the recovered model parameter to the true parameter vs. the relative standard error for two of the seven model parameters, the potential parameter $v_{\rm circ}(R_\odot)$ and qDF parameter $\sigma_{z,0}$. For data sets with proper motion error errors $\delta(m-M) \leq 3$mas yr^{-1} Equation (15) was evaluated with $N_{\rm error} = 800$, for $\delta(m-M) > 3$mas yr^{-1} we used $N_{\rm error} = 1200$. In the absence of distance errors Equation (15) gives unbiased results, for $\delta(m-M) \geq 3$mas yr^{-1} (which corresponds in this test to $\delta v_{\rm max} \lesssim 43$, see Equation (16)) however biases of several sigma [TO DO: can I say sigma???] are introduced as Equation (15) is only an approximation for the true likelihood in this case. [TO DO: rename σ_z to $\sigma_{z,0}$]

The first set of tests mimics a DF that has wider wings or a sharper core in velocity space than a qDF (Figure 13). The second test could be understood as 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 much cooler qDF with a longer scale length, as opposed to the same pollution of stars drawn from a hotter qDF with a shorter scale length.

When considering the case of a 50/50 mix of contributions from different qDFs in Example 2, there is a systematic, but only small, error in recovering the potential parameters, monotonically increasing with the qDF parameter difference; in particular for fractional differences in the qDF parameters of $\lesssim 20\%$ the systematics are insignificant even for samples sizes of 20,000, as used in the mock data.

Overall, a cooler DF seems to always give tighter constraints on the circular velocity at the sun $v_{\rm circ}(R_\odot)$, because the rotation curve can be constrained easier if more stars are on near-circular orbits. But the recovered $v_{\rm circ}(R_\odot)$ does not necessarily have to be right. The hotter DFs give less tight constraints and are therefore more forgiving.

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

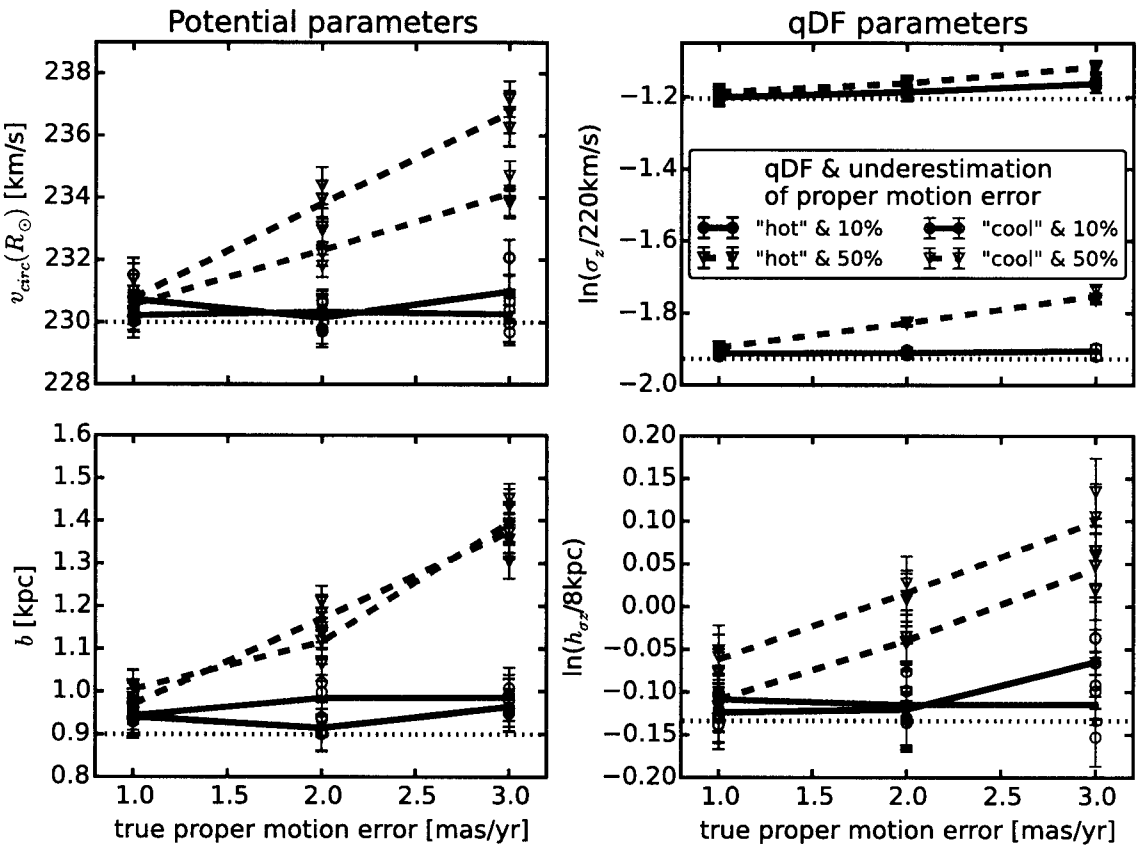


Fig. 12.— Effect of ~~an~~ systematic underestimation of proper motion errors in the recovery of the model parameters. The true model parameters used to create the mock data are summarized as Test 6.3 in Table 3, four of them are given on the y-axes and the true values are indicated as black dashed lines. The velocities of the mock data were perturbed according to Gaussian errors in the α and δ proper motions as indicated on the x-axis. The circles and triangles are the best fit parameters of several mock data set assuming the proper motion ~~error~~, with which the model probability was convolved, was underestimated in the analysis by 10% or 50%, respectively. The error bars correspond to 1 sigma [TO DO: Can I say sigma????] confidence. The lines connect the mean of each two data realisations and are just ~~guides to the eyes.~~ [TO DO: rename $h_{\sigma z}$ to $h_{\sigma,z}$, σ_z to $\sigma_{z,0}$]

small radii, while stars from the cooler qDF (with its longer tracer scale length) dominate at large radii. The recovered tracer scale lengths, h_R vary smoothly between the input values of the two qDFs that entered the mix of mock data, with again the impact of contamination by a hotter qDF (with its shorter scale length in this case) being more important.

3.6. The Implications of Assuming a Potential Model which Differs from the Real Potential

We inspect if we can give constraints on the true potential, if our beliefs about the overall parametric form of the MW's potential are slightly wrong. We ignore deviations from axisymmetry and focus on a test 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 Figure 1). In the analysis we assume the circular velocity at the sun to be fixed and known and only fit the parametric potential form. The results are shown in Figure 16.

The reference potential parameters of the "KKS-Pot" in Table 1 were found by adjusting the 2-component Kuzmin-Kutuzov Stäckel potential by Batsleer & Dejonghe (1994) such that it generates radial and vertical force profiles similar to the "MW14-Pot" from Bovy (2015) (dotted gray lines in Figure 16). The analysis results from *RoadMapping* shown in Figure 16, red for a "hot" mock data *MAP* and blue for a "cool" *MAP* , give an comparable good or even better agreement with the true potential than the (by-eye) fit directly to the potential: especially the force contours, to which the orbits are sensitive, and the rotation curve are very tightly constrained and reproduce the true potential even outside of the observed volume of the mock tracers. This demonstrates that *RoadMapping* provides an optimal best fit potential within the capabilities of the parametric potential model.
The density contours are less tightly constrained than the forces, but we still capture the essentials: The "hot" *MAP* from Table 2 constrains the halo; especially at smaller radii it is equally good or better than the "cool" *MAP* . The "cool" *MAP* gives tighter constraints on the halo in the outer region and recovers the disk better than the "hot" *MAP* . This is in concordance with expectations as the "cool" *MAP* has a longer tracer scale length and is more confined to the disk than the "hot" *MAP* and therefore also probes the Galaxy in these regions better.
Overall the best fit disk is less dens in the midplane than the true disk.

Figure 17 compares the true qDF parameters with the best fit parameters. While tracer scale length and radial velocity dispersion profile are very well recovered, we misjudge the radial profile of the vertical velocity dispersion: $\sigma_{0,z}$ [TO DO: consistent] and $h_{\sigma,z}$ are both

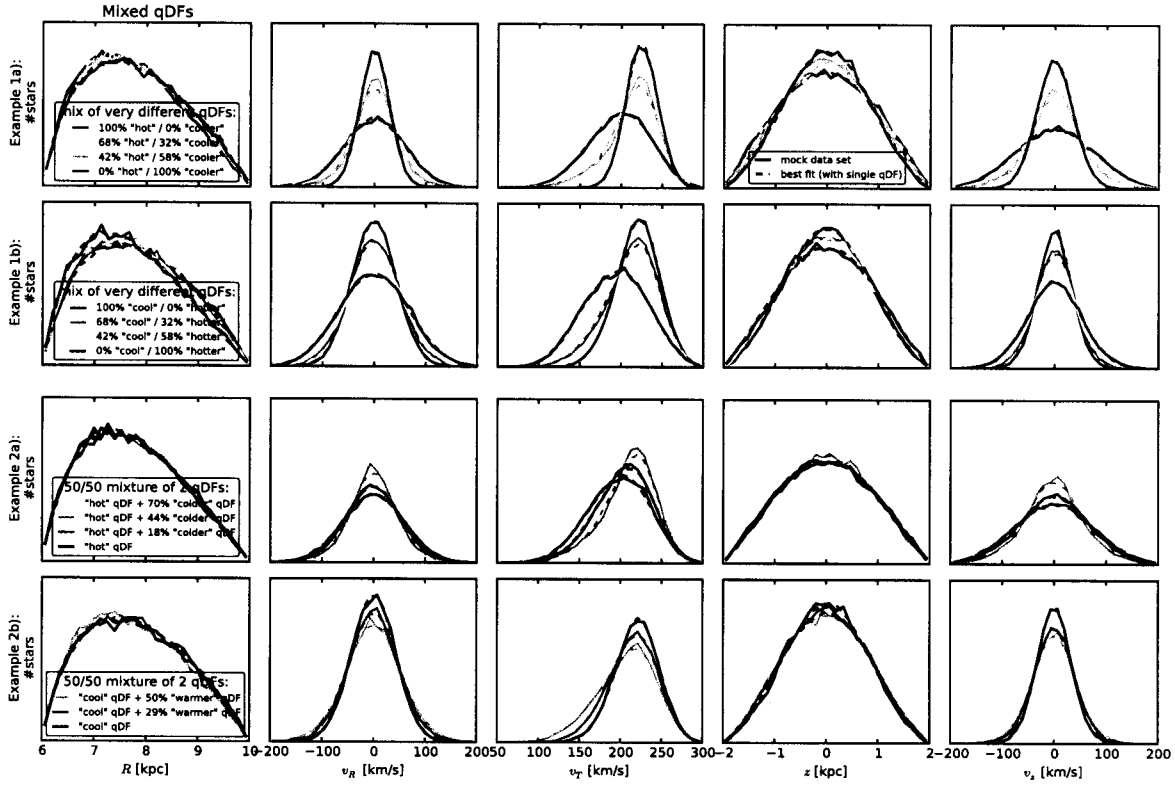


Fig. 13.— Distribution of mock data, created by mixing stars drawn from two different qDFs (solid lines), and the distribution predicted by the best fit of a single qDF and potential to the data (dashed lines). The model parameters to create the mock data (solid lines) are given in Table 3 as Test 7, and the qDF parameters referenced in the figure's legend are given in Table 2. The corresponding single qDF best-fit curves (dashed lines) were created by drawing mock data from the best fit parameters found in Figures 14 and 15. *Example 1:* Distribution of mock data drawn from a superposition of two very different (but fixed) qDFs at varying mixing rates. *Example 2:* Mock data distribution of two *MAP*s that were mixed at a fixed rate of 50%/50%, but the difference of the qDF parameters of one *MAP* was varied with respect to the qDF parameters of the other *MAP* by X% (see Table 2). The data sets are color coded in the same way as the corresponding analyses in Figures 14 and 15. This figure demonstrates how mixing two qDFs can be used as a test case for changing the shape of the DF to not follow a pure qDF anymore, e.g. by adding wings or slightly changing the radial density profile. When comparing the mock data and best fit distribution, we see that especially for the most extreme deviations it becomes obvious that a single qDF is a bad assumption for the stars' "true" DF.

I feel like just showing one of these examples might be clearer, because they essentially demonstrate the same thing

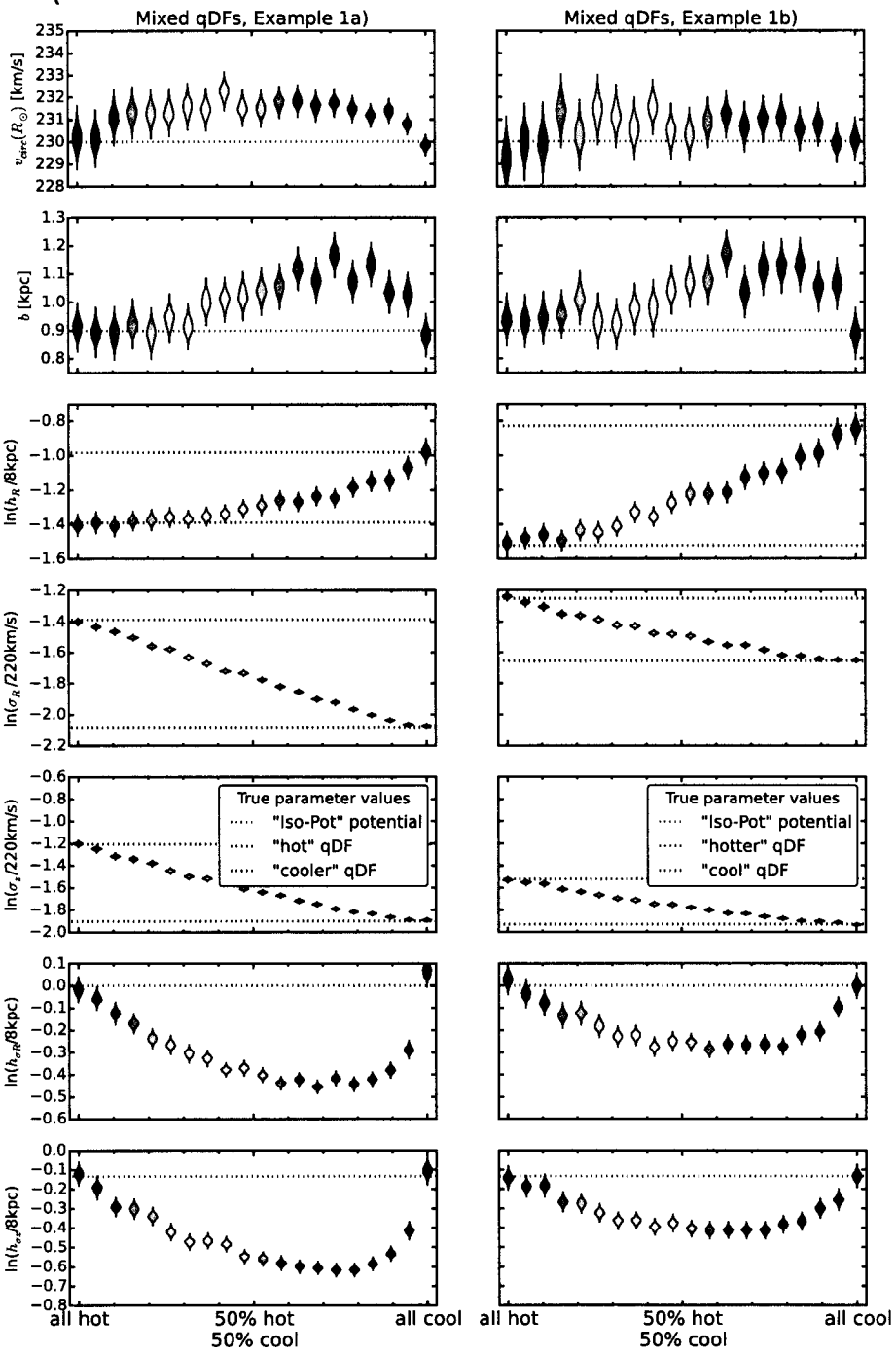


Fig. 14.— (Caption on next page.)

Fig. 14.— The dependence of the parameter recovery on degree of pollution and temperature of the stellar population. To model the pollution of a hot stellar population by stars coming from a cool population and vice versa, we mix varying amounts of stars from two very different populations, as indicated on the X-axis. The composite mock data set is then fit with one single qDF. The violins represent the marginalized likelihoods found from the MCMC analysis. *Example 1a* (*Example 1b*) in the left (right) panels mixes the "hot" ("cool") *MAP* with the "cooler" ("hotter") *MAP* in Table 2. All model parameters used to create the mock data are given in Test 7, *Example 1a) & b)* in Table 3. Some mock data sets are shown in Figure 13, first two rows, in the same colors as the violins here. We find that a hot population is much less affected by pollution with stars from a cooler population than vice versa. [TO DO: rename $h_{\sigma R}$ to $h_{\sigma,R}$, σ_R to $\sigma_{R,0}$ and analogous for z]

underestimated, which leads to a steeper profile and a lower dispersion around the sun. This is a direct result of the surface density underestimation in the midplane, the corresponding lower vertical forces around the sun (see also Figure 16) and therefore lower vertical actions [TO DO: I have honestly no idea, if this is a proper explanation. In configuration space both models, original mock data set and best fit, have exactly the same radial dispersion and velocity profile.]. Figure 18 demonstrates that even though the misjudgment of the potential lead to biases in the qDF parameters, the model is still a very good fit to the data.

4. Discussion and Summary

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

We present *RoadMapping*, an improved implementation of the dynamical modelling machinery by Bovy & Rix (2013), to recover the potential and orbit distribution function of stellar *MAP*s within the Galactic disk. In this work we investigated the capabilities, strengths and weaknesses of *RoadMapping* by testing its robustness against the breakdown of some of its assumptions - for well defined, isolated test cases using mock data. Overall the method works very well and reliable, also if there are small deviations of the model assumptions from the real world galaxy.

Need to discuss this subtlety because the 'correct' potential is not part of the fitted potential family, we shouldn't expect the qDF parameters that best fit the data, to be the 'input' qDF parameters. From your comment, it seems like you recover the DF correctly, although the parameters are off.

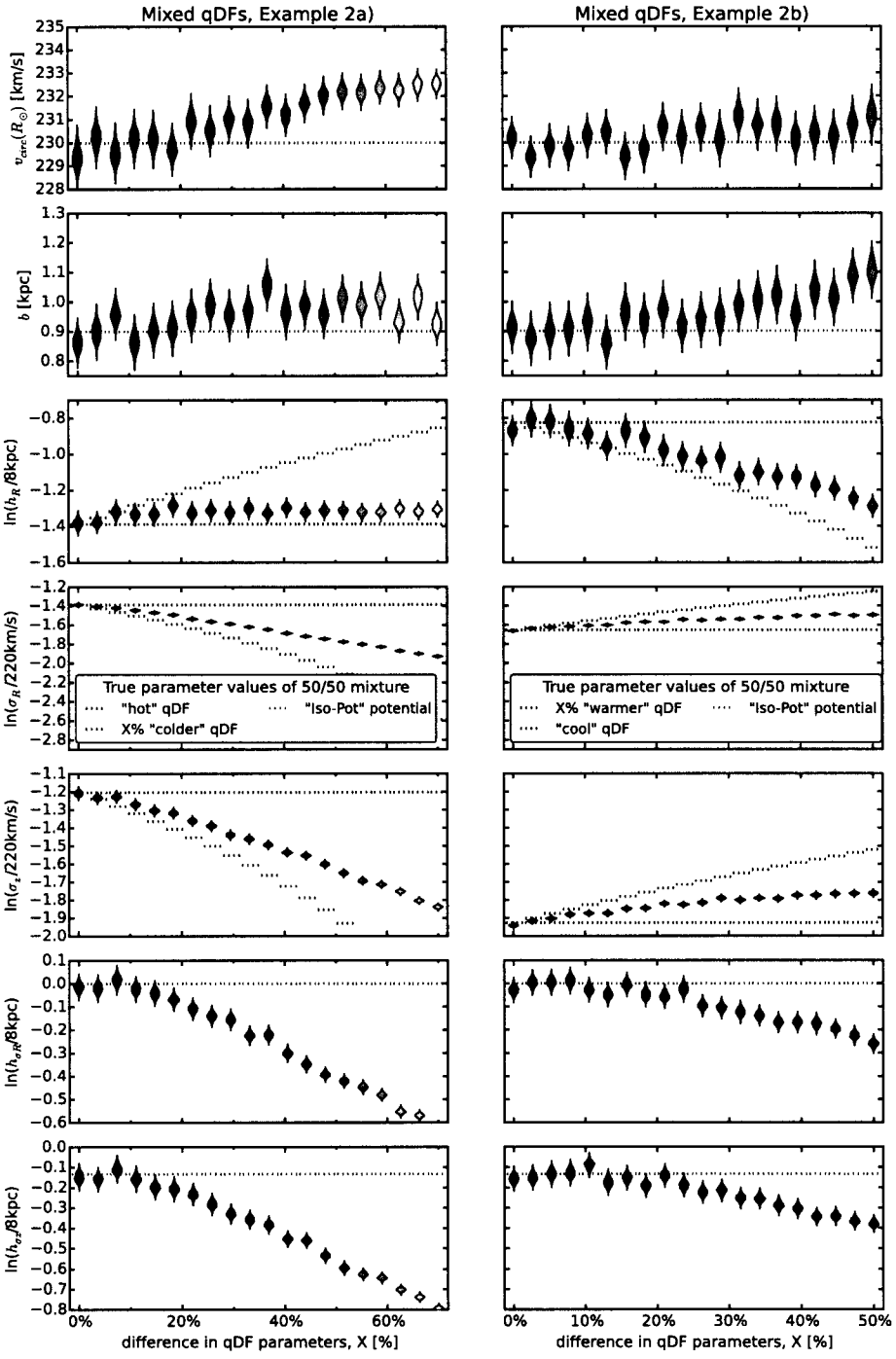


Fig. 15.— (Caption on next page.)

Fig. 15.— The dependence of the parameter recovery on the difference in qDF parameters of a 50%/50% mixture of two stellar populations and their temperature. Half of the star in each mock data set in *Example 2a* (*Example 2b*) was drawn from the "hot" ("cool") qDF in Table 2, and the other half drawn from a "colder" ("warmer") population that has $X\%$ smaller (larger) $\sigma_{R,0}$ and $\sigma_{z,0}$ and $X\%$ larger (smaller) h_R. Each composite mock data set is then fitted by a single qDF and the marginalized MCMC likelihoods for the best fit parameters are shown as violins. The model parameters used for the mock data creation are given as Test 7, *Example 2a) & b)* in Table 3. Some mock data sets are shown in figure 13, last two rows, where the distributions have the same colors as the corresponding best fit violins 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: rename $h_{\sigma R}$ to $h_{\sigma,R}$, σ_R to $\sigma_{R,0}$ and analogous for z]

4.1. Improved Computational Speed for Application to Larger Data Sets

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

Large data sets in the age of Gaia require more, and more accurate, likelihood evaluations for more flexible models. To be able to deal with these increased computational demands and explore larger parameter spaces, we sped up the code by combining a nested grid approach with MCMC and by faster action calculation using the Stäckel (Binney 2012) interpolation grid by Bovy (2015). Especially accurately determining the likelihood normalisation will be of crucial importance for large data sets. The nested-grid approach automatizes the search for the optimal normalisation integration ranges ("fiducial qDF") and start position for the MCMC walkers, which helps the MCMC to converge fast and to reduce biases due to insufficient accuracy. However, application of *RoadMapping* to millions of stars simultaneously with acceptable accuracy will still be a task for supercomputers and calls for even more improvements and speed-up in the fitting machinery.

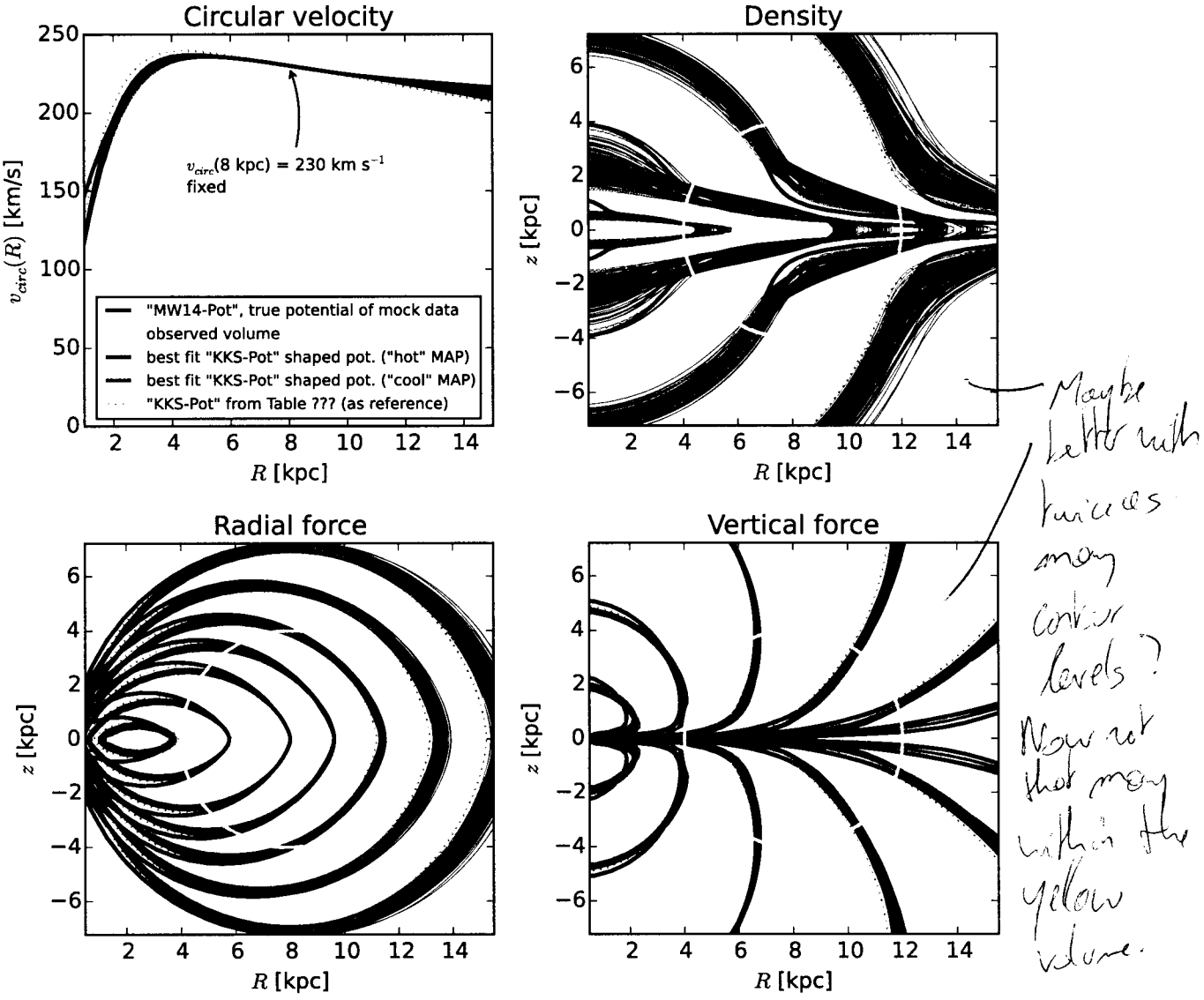


Fig. 16.— Recovery of the gravitational potential if the assumed potential model ("KKS-Pot" with fixed $v_{\rm circ}(R_\odot)$) and the true potential of the (mock) stars ("MW14-Pot" in Table 1) is slightly different. We show the circular velocity curve, as well as contours of equal density, radial and vertical force in the R-z-plane, and compare the true potential with 50 [TO DO: CHECK] sample potentials drawn from the posterior distribution function found with the MCMC for a "hot" (red) and a "cool" *MAP* (blue). All model parameters are given as Test 8 in Table 3. [TO DO: Do more analyses???] [TO DO: fancybox Legend]

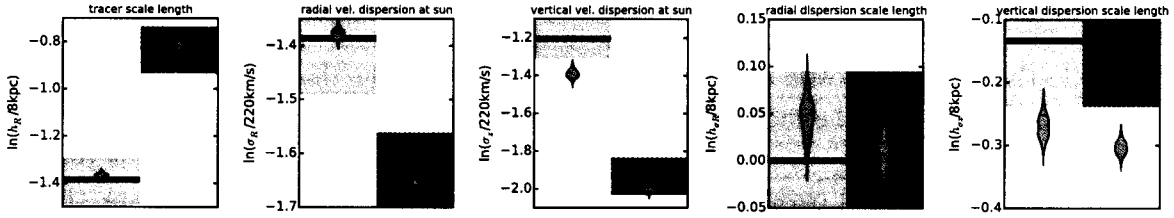


Fig. 17.— Recovery of the qDF parameters for the case where the true and assumed potential deviate from each other (Test 8 in Table 3). The thick red (blue) lines represent the true qDF parameters of the "hot" ("cool") qDF in Table 2 used to create the mock data, surrounded by a 10% error region. The grey violins are the marginalized likelihoods for the qDF parameters found simultaneously with the potential constraints shown in Figure 16. [TO DO: rename $h_{\sigma R}$ to $h_{\sigma,R}$, σ_R to $\sigma_{R,0}$ and analogous for z]

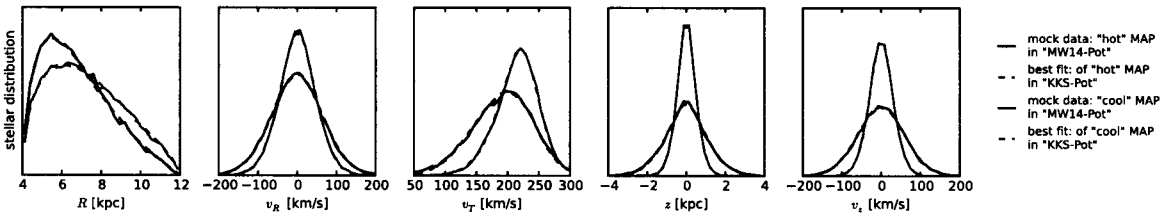


Fig. 18.— Comparison of the distribution of mock data in configuration space created in the "MW14-Pot" potential (solid lines) with a "hot" (red) and "cool" (blue) *MAP* (Test 8 in Table 3), and the best fit distribution using a "KKS-Pot" potential (dashed lines). The best fit potentials are shown in Figure 16 and the corresponding best fit qDF parameters in Figure 17. The best fit

Add something about the fit being good.

4.2. Modelling Sensitivity to Properties and Unaccounted Imperfections of the Data Set

Choice of observation volume. We found that the *position* of the survey volume matters little, in the sense that there are no regions in the Galaxy that contain intrinsically stars on manifestly more diagnostic orbits than others. Closer to the disk and at smaller Galactocentric radii it is only the increased number of stars that will lead to tighter constraints. Concerning the *shape* of the survey volume, a large radial *and* vertical coverage is best. In the axisymmetric case phi coverage doesn't matter. Making a volume cut for stars, that lie around $R_\odot$ but at larger *phi*, could therefore improve the results, if their measurements are very uncertain.

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

Selection function misjudgment. Surprisingly *RoadMapping* seems to be very robust against misjudgments in the selection function of the data. The reason for this robustness could be that missing stars in the data set do not affect the connection between a star's velocity and position, which is given by the potential. A lot of information about the potential profile is stored in the rotation curve, but even when not including measurements of tangential velocities in the analysis, small misjudgments of the incompleteness do not affect the potential recovery.

That we reproduce the qDF equally well, could be due to the symmetry of our assumed incompleteness profiles around the sun. We investigated a decrease in knowledge of the data completeness in distance from the sun and Galactic plane. Our test with the radial incompleteness profile could be understood as a decreasing detection rate due to the lower apparent brightness of stars at larger distances. The test with the planar incompleteness profile could mimic a misjudgment of the dust obscuration within the Galactic plane. Both effects would show the same symmetries as tested in this work.

This result is encouraging for future studies, but nevertheless surprising as it was previous believed that knowing the (spatial) selection function very precisely is of large importance for dynamical modelling (Rix & Bovy 2013).

Measurement errors. Properly convolving the likelihood with measurement errors is computational very expensive. By ignoring positional errors and only including distance errors as part of the velocity error, we can drastically reduce the computational costs. For stars within 3 kpc from the sun this approximation works well for distance errors of $\sim 10\%$ or

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

4.3. Data Deviations from the Modelling Assumptions about the Distribution Function and the Potential

Deviations from the qDF Assumption. Our modelling is founded on the assumption, that we can identify *a priori* sub-components of the Galactic disk that follow a qDF (e.g. by considering *MAPs*). There are two reasons why any chosen sub-sample of star (here a *MAP*) may not be well described by any qDF. Either, because nature is more complex, or because even if perfect *MAPs* would be well described by qDFs, finite abundance errors would mix *MAPs*. We have considered both cases.

In Example 1 in §3.5 we investigated how well we can recover the potential, if this assumption was not perfectly satisfied, i.e., the *MAPs* true DF does not perfectly follow a qDF. We considered two cases: a) a hot DF, that has less stars at small radii and more stars with low velocities than predicted by the qDF (reddish data sets in Figure 13), or b) a cool DF that has broader velocity dispersion wings and less stars at large radii than predicted by the qDF (bluish data sets). We find that case a) would give more reliable results for the potential parameter recovery. If we assumed that the distribution of stars from one *MAP* is caused by radial migration away from the initial location of star formation, it would more likely that the qDF overestimates the true number of stars at smaller radii than underestimating it at larger radii. [TO DO: Is this actually a sensible argument???]
Following this, focusing the analysis especially on hotter *MAPs* could be an advisable way to go in the future, if there is doubt that the stars truly follow the qDF.

Another critical point is the binning of stars into *MAPs* depending on their metallicity and α abundances. Example 2 in §3.5 could be understood as a model scenario for decreasing bin sizes in the metallicity-α plane when sorting stars in different *MAPs*, 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 Figure 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%. Figure 14 and 15 suggest that especially the tracer scale length h_R needs to be recovered to get the potential right. For this parameter however the bin sizes in Figure 6 of Bovy & Rix (2013) might not yet be small enough to ensure no more than 20% of difference in neighbouring h_R. This is especially the case in the low-α ($[\alpha/Fe] \lesssim 0.2$), intermediate-metallicity ($[Fe/H] \sim -0.5$) *MAP*s . The above is valid for 20,000 stars per *MAP* . In case there are less than 20,000 stars in each bin the constraints are less tight and due to Poisson noise one could also allow larger differences in neighbouring *MAP*s while still getting reliable results.

Gravitational Potential beyond the Parameterized Functions Considered. In the long run *RoadMapping* should incorporate a family of gravitational potential models that can reproduce the essential features of the MW's true mass distribution. While our fundamental assumption of the Galaxy's axisymmetry is at odds with the obvious existence of non-axisymmetries in the MW, we will not dive into investigating this implications in this paper. Instead we test how a misjudgment of the parametric potential form affects the recovery by fitting a potential of Stäckel form (Batsleer & Dejonghe 1994) to mock data from a different potential family with halo, bulge and exponential disk. The recovery is quite successful and we get the best fit within the limits of the model. However, even a strongly flattened Stäckel potential component has difficulties to recover the very flattened mass distribution of an exponential disk. This leads to misjudgment of the qDF parameters describing the vertical action distribution, $\sigma_{z,0}$ and $h_{\sigma,z}$. As the qDF parameter $\sigma_{z,0}$ corresponds to the physical vertical velocity dispersion at the sun, a comparison with direct measurements could be a valuable cross-checking reference. [TO DO: This might not be true. For isochrone and Staeckel potential I get this behaviour, but not for the MW14-Pot!!! Might be, because it's not separable??? Check!!!] In case of as many as 20,000 stars we should therefore already be able to distinguish between different potential models.
The advantage of using a Stäckel potential with *RoadMapping* is firstly the exact and fast action calculation via the numerical evaluation of a single integral, and secondly that the potential has a simple analytic form, which greatly speeds up calculations of forces and frequencies (as compared to potentials in which only the density has an easy description like the exponential disk). A superposition of several simple Kuzmin-Kutuzov Stäckel components can successfully produce MW-like rotation curves (see Batsleer & Dejonghe (1994), Famaey & Dejonghe (2003) and Figure 16) and one could think of adding even more components for more flexibility, e.g. a small roundish component for the bulge.

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 Figure 16 this seems to be indeed a sensible approach [TO DO: Check that this is indeed the case - it is not clear to me from the plot. ???].
We suggest that combining the flexibility and computational advantages of a superposition of several Stäckel potential components with probing the potential in different regions with different *MAP*s as done by Bovy & Rix (2013), could be a promising approach to get the best possible constraints on the MW's potential.

4.4. Different Modelling Approaches using Action-based Distribution Functions

We have focussed for the time being on *MAP*s for a number of reasons: First, they seem to permit simple DFs (Bovy et al. 2012b,c,d), i.e. approximately qDFs (Ting et al. 2013). Second, all stars, e.g. those from different *MAP*s, must orbit in the same potential. Therefore each *MAP* will and can yield quite different DF parameters; but each *MAP* will also provide a (statistically) independent estimate of the potential parameters. At the same time – if all is well – those potential parameters, derived from different *MAP*s, should be mutually consistent. In some sense, this approach focusses on constraining the potential, treating the DF parameters as nuisance parameters.

The main drawback is that we have many astophysical reasons that the DF properties (for reasons of galaxy evolution and chemical evolution) are astrophysically linked between different *MAP*s . Ultimately, the goal is to do a fully consistent chemodynamical model that simultaneously fits the potential and DF($\boldsymbol{J}$, [X/H]) simultaneously (where [X/Fe] denotes the full abundance space) with a full likelihood analysis.
This has not yet been attempted.

Since the first application of *RoadMapping* by Bovy & Rix (2013) there have been two similar efforts to constrain the Galactic potential and/or orbit distribution using action-based distribution functions:

Piffl et al. (2014) fitted both potential and a $f(\boldsymbol{J})$ to giant stars from the RAVE survey (Steinmetz et al. 2006) and the vertical stellar number density profiles in the disk by Jurić et al. (2008). They did not include any chemical abundances in the modelling. Instead,

they used a superposition of action-based DFs to describe the overall stellar distribution at once: a superposition of qDFs for cohorts in the thin disk, a single qDF [TO DO: CHECK] for the thick disk stars and an additional DF for the halo stars. Taking proper care of the selection function requires a full likelihood analysis and the calculation of the likelihood normalisation is computational expensive. Piffl et al. (2014) choose to circumvent this problem by directly fitting a) histograms of the three velocity components in eight spatial bins to the velocity distribution predicted by the DF and b) the vertical density profile predicted by the DF to the profiles by Jurić et al. (2008). The vertical force profile of their best fit mass model nicely agrees with the results from Bovy & Rix (2013) for $R > 6.6$ kpc. The disadvantage of their approach is, that by binning the stars spatially, a lot of stellar information is not used.

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

4.5. On the Assumption of Axisymmetry

The key assumption in our modelling, as well as in the approaches by Piffl et al. (2014) and Sanders & Binney (2015) described above, is the overall axisymmetry of Galaxy's potential and DF. This has the convenient advantage, that actions are conserved in axisymmetric potentials and can be calculated straightforward via the "Stäckel Fudge" by Binney (2012) and/or a single integration (in the case of separable potentials). Of course the Galactic disk is in reality not axisymmetric and actions are not conserved: Spiral arms in the disk and the Galactic bar lead to angular momentum exchange and therefore radial migration of stars, i.e. the orbit on which the stars were born on are modified (Minchev et al. 2011; Kawata et al. 2014). Apart from these obvious non-axisymmetries in the Galaxy, the disk itself is not smooth, as there is lot of sub-structure, streams and moving groups within the disk. A famous example is the Arcturus moving group, for which Navarro et al. (2004) found

that the stars might have their origin in a disrupted satellite.
Spiral arms, the stirring of the Galactic bar and the disk sub-structure will undoubtedly affect our modelling with *RoadMapping* . How strong this effect will be and if the modelling can still work, needs to be investigated in detail in future work, e.g. by trying to recover the potential in N-body simulations. But as actions are conserved under adiabatic changes of the potential (Binney & Tremaine (2008)), and the vertical action under radial migration [TO DO: REF], there is some hope, that the modelling could still work.

The ultimate goal would be to theoretically describe those non-axisymmetries and sub-structures in terms of DFs in action/angle-abundance space. We see the current axisymmetric modelling approaches as intermediate step to this goal: Dehnen (1998) described the disk's overall stellar distribution as a smooth background distribution superimposed with sub-structure; the axisymmetric DFs used and to be found by *RoadMapping* and similar approaches could be treated as this smooth background. Firstly, this smooth background DF could help to actually find and identify the disk substructure in action space (see e.g. Sellwood (2010); Klement et al. (2008) for similar approaches to find disk substructure, which then helps to find DF descriptions). Secondly, and because simple superposition of action-based DFs is possible (see e.g. [TO DO: REF (see Payel's hint)]), the substructure DFs could then be directly added to the background DF and incorporated in the modelling. In other words, modelling the Galaxy together with its non-axisymmetries and substructure could be approached as applying perturbations to an axisymmetric equilibrium model. And *RoadMapping* will help finding this equilibrium model.

Such a Galaxy model will be also important in the meantime: Many studies of Galaxy structure and evolution use orbits as tracers and therefore require a reliable fiducial potentials to turn stellar positions and velocities into orbits. And as long as we are as far away from realistic Galaxy models as we are now, the axisymmetric case will need to be our reference.

5. Acknowledgments

[TO DO]

The authors thank Glenn van de Ven for the idea of using Kuzmin-Kutuzov Stäckel potentials in this case study.

A. Appendix

A.1. Influence of wrong assumptions about incompleteness of the data parallel to the Galactic plane

In §3.3 we found a striking robustness of the *RoadMapping* modelling approach against wrong assumptions about the radial incompleteness of the data set. To further test this result, we investigate a different completeness function that drops with distance from the Galactic plane (see Test 5, Example 2, in Table 3 and Figure 19). We get a similar robust behaviour for small deviations, and only slightly less robustness for larger deviations. That an explanation for this robustness could be, that ~~a lot of information~~ about the potential comes from the rotation curve, which is not affected by incompleteness, is demonstrated in Figure 21.

(much of the information

Marginalization over v_T. The likelihood in Equation (11) is marginalized over the coordinate v_T as follows

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

where $P(\boldsymbol{x}, \boldsymbol{v} \mid p_M)$ is the same as in Equation (11) and the numerical integral over v_T is performed as a 24th order Gauss-Legendre quadrature. The additional factor of v_0 is needed to get the units of $P_{(v_T \text{ marg.})}(\boldsymbol{x}_i, v_{R,i}, v_{z,i} \mid p_M)$ right.