Monthly Notices of the ROYAL ASTRONOMICAL SOCIETY

MNRAS **511,** 1609–1621 (2022) https://doi.org/10.1093/mnras/stac153
Advance Access publication 2022 January 28

Charting galactic accelerations – II. How to 'learn' accelerations in the solar neighbourhood

A. P. Naik [1]★ J. An [2] C. Burrage[1] and N. W. Evans[3]

[1] *School of Physics & Astronomy, University of Nottingham, University Park, Nottingham NG7 2RD, UK*
[2] *Center for Theoretical Astronomy, Korea Astronomy & Space Science Institute, 776 Daedeok-daero, Yuseong-gu, Daejeon 34055, South Korea*
[3] *Institute of Astronomy, University of Cambridge, Madingley Road, Cambridge CB3 0HA, UK*

Accepted 2022 January 12. Received 2021 December 20; in original form 2021 October 14

ABSTRACT

Gravitational acceleration fields can be deduced from the collisionless Boltzmann equation, once the distribution function is known. This can be constructed via the method of normalizing flows from data sets of the positions and velocities of stars. Here, we consider application of this technique to the solar neighbourhood. We construct mock data from a linear superposition of multiple 'quasi-isothermal' distribution functions, representing stellar populations in the equilibrium Milky Way disc. We show that given a mock data set comprising a million stars within 1 kpc of the Sun, the underlying acceleration field can be measured with excellent, sub-per cent level accuracy, even in the face of realistic errors and missing line-of-sight velocities. The effects of disequilibrium can lead to bias in the inferred acceleration field. This can be diagnosed by the presence of a phase space spiral, which can be extracted simply and cleanly from the learned distribution function. We carry out a comparison with two other popular methods of finding the local acceleration field (Jeans analysis and 1D distribution function fitting). We show our method most accurately measures accelerations from a given mock data set, particularly in the presence of disequilibria.

Key words: methods: data analysis – Galaxy: fundamental parameters – Galaxy: kinematics and dynamics.

1 INTRODUCTION

Given a map of the gravitational acceleration field within a kiloparsec of the Sun, we could learn a wealth of information about the current state of our Galaxy, the distribution of matter (both dark and luminous) and the nature of gravity. For example, if the acceleration due to the luminous component is known, then we can calculate the density distribution of dark matter, uncovering any substructures and measuring the ambient dark matter density in the Solar system. This latter number is of great importance in particle physics, as it is a key parameter in the interpretation of results of dark matter direct detection experiments (Read 2014; de Salas & Widmark 2021). Alternatively, we can use the direction of the acceleration vectors to constrain alternative theories of gravity such as Modified Newtonian Dynamics (Milgrom 1983) or similar. In these theories, there is no dark matter, so the total acceleration is necessarily co-linear with the acceleration due to the baryons, even if the modifications to gravity alter its magnitude (Loebman et al. 2014).

Unfortunately, direct acceleration measurements are challenging. Even so, promising steps have been recently taken in this direction employing measurements of pulsar orbital decay, which give the relative acceleration of a few pulsar systems with respect to the Solar system barycentre. These can then be converted into absolute accelerations using a measurement of the Solar system acceleration, thus providing a small number of direct samples of the Galaxy's acceleration field (Bovy 2020; Chakrabarti et al. 2021b). In future, greater statistical power will be afforded by complementary acceleration measurements, such as those derived from binary eclipse timing (Chakrabarti et al. 2021a) and those from future high-precision radial velocity spectrographs (Quercellini, Amendola & Balbi 2008; Ravi et al. 2019; Silverwood & Easther 2019; Chakrabarti et al. 2020).

In the meantime, we must instead adopt an alternative approach: inferring accelerations (or equivalently the gravitational potential) statistically from the positions and velocities of the stars. If discrete stellar encounters are neglected, then the stellar distribution function (DF), i.e. the probability distribution of the stars in six-dimensional $(\boldsymbol{x}, \boldsymbol{v})$ phase space, can be related to gravitational accelerations via the collisionless Boltzmann equation (CBE),

$$\frac{\partial f}{\partial t} + \boldsymbol{v} \cdot \nabla_{\boldsymbol{x}} f - \nabla_{\boldsymbol{v}} f \cdot \nabla_{\boldsymbol{x}} \Phi = 0, \tag{1}$$

where f is the DF and Φ is the gravitational potential.

It can be difficult to constrain the full DF with a statistically small data set, so it is often preferable to work with the second moments of the DF, i.e. the velocity dispersions, which can be related to the acceleration field via the Jeans equations (e.g. Hagen & Helmi 2018; Sivertsson et al. 2018; Guo et al. 2020; Salomon et al. 2020). While velocity dispersions are comparatively easy to measure from kinematic data, we lose much of the information content of the data compared with techniques working directly with the DF. For this reason, many studies have instead adopted the latter approach, typically in one dimension (e.g. Schutz et al. 2018; Buch, Leung & Fan 2019; Widmark 2019; Widmark & Monari 2019; Li & Widrow 2021; Widmark, de Salas & Monari 2021a). Note that either treatment necessitates the assumption of dynamical equilibrium, so that the

★ E-mail: aneesh.naik@nottingham.ac.uk

time-derivative term in equation (1) can be neglected. The review articles by Read (2014) and de Salas & Widmark (2021) give good overviews of how these methods work in practice.

Despite the statistical advantages of the latter class of technique, there are some limitations. Typical DF models are constructed under a series of assumptions, such as separability, isothermality, and various spatial symmetries. In the era of 'big data', it is worth examining whether an alternative approach can do better justice to the full richness of contemporary data sets and their statistical power. In An et al. (2021, hereafter Paper I), we outlined just such a methodology. Inspired by an idea first proposed by Green & Ting (2020), we described a technique in which a non-parametric DF can be constructed directly from the data using modern deep learning techniques. Such an approach is highly flexible; the resulting learned DF is untrammelled by the limitations of an analytical model, and is instead free to capture the full richness of the training data. After learning a DF in this manner, Paper I shows how to convert the DF into an acceleration map, via an exact inversion of the CBE under the assumption of equilibrium.

In Paper I, we provided a basic demonstration of our technique with a mock data set representing a simple, spherical distribution of stars. In this article, we provide demonstration of the same technique in much more complex context: mock data on stellar kinematics in the solar neighbourhood. The reason to confine ourselves to mock data for the moment is to gain insights into the biases and limitations of the technique before we apply it to real data in a companion paper.

A major obstacle is that the assumption of dynamical equilibrium is not necessarily a good one. Various non-equilibrium structures have been observed in kinematics of the Milky Way (MW) disc stars, such as warping of the disc (Schönrich & Dehnen 2018), north–south asymmetries (Salomon et al. 2020), and the well-known phase spiral (Antoja et al. 2018). Incorrectly assuming that a stellar population is in equilibrium will lead to bias in resulting dynamical inferences (Banik, Widrow & Dodelson 2017, Paper I). To quantify this effect, we additionally examine the application of our methodology to a mock data set resembling a perturbed Galactic disc.

This article is structured as follows. In the following section (Section 2), we recapitulate the methodology described in Paper I – namely, our algorithm to recover acceleration fields from 6D kinematics. After that, in Section 3, we describe the mock data sets we use to test this method. We generate stellar positions and velocities from realistic models of the MW disc. We calculate radial and vertical accelerations within $\sim$1 kpc of the Sun from a mock data set in Section 4. Then, Section 5 compares the accuracy of our measured accelerations against those produced using other techniques, specifically the Jeans analysis method of Salomon et al. (2020) and the 1D DF fitting method of Widmark et al. (2021a). Finally, Section 6 provides a discussion and concluding remarks.

2 METHODS

The two steps in our method are as follows (see Paper I):

(i) Given a stellar kinematic data set, we use a probability estimation technique to 'learn' the underlying DF.

(ii) From the learned DF, we then calculate the gravitational acceleration field using an inversion of the (time-independent) CBE.

2.1 Learning the DF

Given data sampled from some unknown distribution, the problem of trying to derive the underlying probability distribution is known as 'probability estimation'. In our case, the data are the positions $\boldsymbol{x}$ and velocities $\boldsymbol{v}$ of stars, and the probability we wish to estimate is the stellar DF $f(\boldsymbol{x}, \boldsymbol{v})$, i.e. the probability density function of stars in phase space. One way to do this is to write down some parametric model for the probability density then compare the model's predictions with the data until the parameters are optimized. In essence, this is the technique employed by the majority of studies of stellar dynamics, whether they work directly with the DF or with its moments.

We instead adopt a different methodology: we estimate a *non-parametric* DF directly from the data. This data-driven approach has the distinct advantage of being untrammelled by the limitations and underlying assumptions of an explicit model. While non-parametric probability estimation techniques have long existed (e.g. kernel density estimation), recent years have seen a surge of interest in machine learning techniques, which in turn has led to a proliferation of powerful probability estimation algorithms. We employ one such novel algorithm: 'normalizing flows' (Rezende & Mohamed 2015).

The idea behind normalizing flows is simple: we can generate a complex probability distribution by repeatedly transforming a simple one, such as a Gaussian. The input Gaussian can be said to 'flow' through the series of transformations, and after each transformation a (multiplicative) normalizing factor is applied to the new probability distribution to ensure that it is properly normalized, hence the name 'normalizing flows'.

To see this, consider a continuous random variable z, with probability density function $p_z(z)$. We can define a new variable $x \equiv f(z)$, with the only requirement being that the function f is bijective[1] (and thus invertible). The probability distribution for x is then

$$p_x(x) = p_z\left(f^{-1}(x)\right) \left|\det\left(\frac{\partial f^{-1}(x)}{\partial x}\right)\right|. \tag{2}$$

The determinant on the right-hand side is the normalizing factor.

This can be generalized to a series of bijective transformations, i.e. $x \equiv f(z) = (f_K \circ f_{K-1} \circ \cdots \circ f_2 \circ f_1)(z)$. Now, the probability distribution for x is (here $f^{-1} = f_1^{-1} \circ f_2^{-1} \circ \cdots \circ f_{K-1}^{-1} \circ f_K^{-1}$)

$$p_x(x) = p_z\left(f^{-1}(x)\right) \prod_{i=1}^{K} \left|\det\left(\frac{\partial f_i^{-1}(x)}{\partial x}\right)\right|. \tag{3}$$

By suitably choosing the input distribution p_z and the transformations f_i, we can generate arbitrarily complex probability distributions. In practice, most applications use a simple unit Gaussian for p_z. Furthermore, we typically give each of the transformations f_i the same parametric form, although each transformation can take different parameters. We can then construct a suitable loss function, such that minimizing the loss function (with respect to the transformation parameters) corresponds to generating a probability distribution p_x that best describes the distribution of x.

In practice, we construct an *ensemble* of estimators. Before training, each flow in the ensemble is initialized with different parameters. The high complexity of a typical loss landscape means that each flow is then likely to end its training at a different local minimum, i.e. each flow learns a slightly different probability distribution. The probability p_x (and its gradients) is then given by the mean over all the estimators.

In summary, the user need only specify

(i) the parametric form of an individual transformation,

[1]Technically, we also assume both f and f^{-1} are continuously differentiable; most practical applications are limited to such transformations.

(ii) the loss function,
(iii) the number of 'units' (i.e. transformations),
(iv) the number of flows in the ensemble.

It is worth remarking that the target distribution, the learned DF, is guaranteed to be a well-behaved probability distribution, i.e. it is positive everywhere and has unit normalization. The latter requirement restricts the space of usable transformations to bijective functions, and this space is then restricted further by the desire for computational efficiency. Different normalizing flow techniques differ primarily in their choices of these transformations, as well as the base distributions and flow architectures.

More detailed descriptions of normalizing flows are given in the article by Rezende & Mohamed (2015) first describing the algorithm, and the recent review articles by Kobyzev, Prince & Brubaker (2021) or Papamakarios et al. (2021). Also, Paper I provides a discussion of the advantages of normalizing flows over kernel density estimation.

We do not impose any further physicality requirements on the learned DF. For example, the acceleration vectors calculated from the DF (see Section 2.2) could in principle show negative divergences (i.e. negative mass densities) or non-zero curls (i.e. non-conservative forces). This is advantageous as it allows us to probe deviations from Newtonian gravity or the effects of disequilibrium.

Green & Ting (2020) employed a species of normalizing flow called 'neural spline flows' (Durkan et al. 2019). Applied to our mock data sets, we find that neural spline flows struggle with hard edges of the sample volumes. We instead use 'masked autoregressive flows' (MAFs; Papamakarios, Pavlakou & Murray 2017). For each mock data set, we train an ensemble of 20 estimators, each with 8 transformations along the flow, each transformation being an artificial neural network with one hidden layer of 64 units. We use the implementation of MAFs in the publicly available software package NFLOWS,[2] and train the estimators using the gradient descent algorithm ADAM (Kingma & Ba 2015).

2.2 Calculating accelerations from a known DF

Section 4 of Paper I describes a procedure for calculating gravitational accelerations $\boldsymbol{a}$ given a known DF $f(\boldsymbol{x}, \boldsymbol{v})$. The calculation is based on an inversion of the CBE under the assumption of dynamical equilibrium. Here, we merely state the main result, i.e. the expression giving the acceleration in terms of first derivatives of the DF. At a given point $\boldsymbol{x}$ in configuration space with the cylindrical polar coordinates (R, φ, z), the acceleration vector $\boldsymbol{a} = (a_R, a_\varphi, a_z)$ is given by (cf. equation 24 in Paper I)

$$\boldsymbol{a} = \mathbf{A}^{-1} \sum_{\text{sample}} \boldsymbol{R}, \tag{4}$$

where $\mathbf{A} \equiv [\mathsf{A}_{ij}]$ is the matrix with

$$\mathsf{A}_{ij} = \sum_{\text{sample}} \frac{\partial f}{\partial v_i} \frac{\partial f}{\partial v_j};$$

$$\boldsymbol{R} = \nabla_{\boldsymbol{v}} f \left(\boldsymbol{v} \cdot \nabla_{\boldsymbol{x}} f + \frac{v_\varphi^2}{R} \frac{\partial f}{\partial v_R} - \frac{v_R v_\varphi}{R} \frac{\partial f}{\partial v_\varphi} \right).$$

The gradient operators here are $\nabla_{\boldsymbol{x}} \equiv (\partial/\partial R, R^{-1}\partial/\partial\varphi, \partial/\partial z)$ and $\nabla_{\boldsymbol{v}} \equiv (\partial/\partial v_R, \partial/\partial v_\varphi, \partial/\partial v_z)$.

The sums labelled 'sample' in these expressions are over a number of suitably chosen points in velocity (v_R, v_φ, v_z) space. Formally, we need at least three sample points for the matrix $\mathbf{A}$ to be invertible. This would suffice if using a DF that is exactly correct, but in our context we work with an approximate reconstructed DF that might be more accurate in some regions of phase space than others. It is therefore safer to increase the sample size. We generally find converged results for ∼100 sample points, but to err on the side of caution we choose 1000 sample points every time we calculate an acceleration.

In Paper I, we gave a discussion of how best to sample these velocity points. In particular, we described 'zones of avoidance': regions of velocity space worth avoiding. These are typically areas where the DF approaches zero (e.g. near the escape velocity) or where its first derivatives approach zero (e.g. near $\boldsymbol{v} = \bar{\boldsymbol{v}}$). In these areas, small *absolute* errors in the learned DF and its derivatives are large *relative* errors, leading to large errors in the derived accelerations. With these points in mind, we sample velocities throughout the remainder of this paper as follows: writing $\boldsymbol{v} = \bar{\boldsymbol{v}} + \delta\boldsymbol{v}$ and taking $\bar{\boldsymbol{v}} = (0, 220, 0)\,\mathrm{km\,s^{-1}}$, we sample the magnitude of $\delta\boldsymbol{v}$ uniformly between 10 and 50 $\mathrm{km\,s^{-1}}$, and its orientation isotropically. We have experimented extensively with more sophisticated sampling schemes but these have not yielded any substantial improvement in accuracy.

Having sampled these velocities, it is then straightforward to evaluate equation (4) to give the gravitational acceleration at a given spatial location. As noted in Paper I, a particular benefit here of normalizing flows (and their implementation in NFLOWS) is that the learned DF is everywhere exactly differentiable, irrespective of the complexity of the flow architecture. Using automatic differentiation, we can efficiently calculate the exact partial derivatives of the DF without resorting to potentially noisy finite difference schemes.

3 MOCK DATA

In Paper I, we tested our technique on a sample of stars distributed spherically in a Hernquist profile. Here, we turn to a more complex test case in local stellar kinematics, i.e. the kinematics of stars in a small, heliocentric region of the MW disc. This section describes how the mock data set is constructed.

3.1 Distribution function

One model for the DF of Galactic disc populations is the 'quasi-isothermal' action-based DF (or qDF) first described by Binney (2010, 2012b). Ting et al. (2013) found that the qDF gives a good description of individual 'mono-abundance populations' (MAPs) of stars, i.e. populations of stars with similar [Fe/H] and [α/Fe] abundances. In particular, the qDF gives density profiles that are radially and vertically near-exponential, and velocity dispersion profiles that are radially near-exponential and vertically near-isothermal. Different MAPs will take different parameter values for the qDF (i.e. scale lengths and normalizations), and the overall disc population can then be described as a linear superposition of many qDFs. This idea was then applied to real data by Bovy & Rix (2013), who subdivided some 17 000 G-dwarf stars from SDSS-SEGUE into 43 individual MAPs, modelled each MAP with a qDF, and subsequently derived measurements for the scale length of the MW disc among other important parameters.

For our mock data, we populate the disc with stars drawn from six distinct MAPs. In other words, we construct a DF from the weighted sum:

$$f(\boldsymbol{x}, \boldsymbol{v}) = \sum_{i=1}^{6} w_i \, f_{\mathrm{qDF}}(\boldsymbol{x}, \boldsymbol{v} | \theta_i). \tag{5}$$

[2] NFLOWS: normalizing flows in PYTORCH.

Table 1. Parameters of six MAPs in our disc. The metallicity [Fe/H] and abundances [α/Fe] are not used further, but simply serve to indicate the type of stellar population being emulated by each MAP. The weights w_i give the relative size of each sub-population, cf. equation (5). Finally, h_R and σ_R are two of the five qDF parameters, respectively, representing the radial scale lengths and radial velocity dispersions. The remaining three qDF parameters are either held fixed or depend on the listed parameters; see discussion in the text.

i	[Fe/H]	[α/Fe]	w_i	h_R (kpc)	σ_R (km s^{-1})
1	−0.7	0.2	0.10	2.0	60.0
2	−0.3	0.2	0.15	2.0	52.0
3	−0.3	0.0	0.25	2.6	52.0
4	0.1	0.0	0.25	2.6	44.0
5	0.1	−0.2	0.15	3.2	44.0
6	0.5	−0.2	0.10	3.2	36.0

Here, w_i is the relative weight and θ_i the various qDF parameters for population i. The qDF requires specification of five scale parameters: the radial scale length h_R, the R and z velocity dispersions in the disc-plane σ_R, σ_z, and the radial scale lengths of these dispersions h_{σ_R}, h_{σ_z}. We remark that these numbers are merely scale parameters and not physical, measurable quantities describing the system. For example, a qDF with $\sigma_R = 40$ km s^{-1} will not necessarily generate a stellar population with exactly that radial velocity dispersion.

In choosing our qDF parameters, we start by picking pairs of elemental abundances [Fe/H] and [α/Fe] – one pair per MAP – to characterize the six MAPs. These abundances are not used further in our investigation, but serve to give an indication of the type of stellar population being emulated by each MAP, and thus inform the choice of qDF parameters. We manually choose abundances to roughly match those of the Hypatia catalogue of stellar abundances in the local neighbourhood (Hinkel et al. 2014). Note that Bovy & Rix (2013) confined their analysis to stars at large heights above the disc plane, and so found a comparatively greater proportion of α-rich, old thick disc stars than is found in the solar neighbourhood by Hypatia. The elemental abundances we adopt for each MAP are given in Table 1. In matching qDF parameters to these abundances, we then emulate the trends observed by Bovy & Rix (2013): Fe-poor, α-old populations have short radial scale lengths h_R and large velocity dispersions σ_R, while Fe-rich, α-young populations are opposite on both counts. Alongside the abundances, Table 1 gives the assigned h_R, σ_R values, as well as the relative weights w_i of the six MAPs. The remaining three qDF parameters are fixed following Bovy & Rix (2013): $\sigma_z = \sigma_R/\sqrt{3}$ and $h_{\sigma_R} = h_{\sigma_z} = 8$ kpc.

We have thus arrived at a DF that can be used to sample a mock stellar population containing a mix of thick and thin disc sub-populations.

3.2 Milky Way model

The qDF is a function of orbital actions rather than phase space coordinates, and the conversion from one coordinate system to the other requires the specification of a potential.[3] In this way, the underlying gravitational potential is encoded in the stellar kinematics.

The various components and parameters of our adopted MW models are listed in Table 2. It is identical to the `MWPotential2014` model of Bovy (2015): a three-component model, comprising a power-law bulge with exponential cut-off, an NFW dark matter halo, and a Miyamoto–Nagai disc.

[3]For a given potential, we calculate actions utilizing the Stäckel approximation, adopting a focal length of 3.6 kpc (Binney 2012a).

Table 2. Parameters and component normalizations for our MW models. The adopted models for the three components (bulge, halo, and disc), and thus the meanings of the quoted parameters, are discussed further in the text. The component normalizations f given in the lower part of the table are defined such that component x contributes fraction f_x to the MW circular velocity at 8 kpc.

Parameter	Value
Bulge power-law exponent	−1.8
Bulge cut-off radius (kpc)	1.9
Halo scale radius (kpc)	16
Disc scale length (kpc)	3
Disc scale height (pc)	280
Component normalization	
f_b	0.05
f_h	0.35
f_d	0.6

3.3 Sampling

We usually sample 10^6 stars within an annulus between $R = 7$ and 9 kpc, with no restrictions on vertical height z. Note that we assume axisymmetry, and so neglect the azimuthal coordinate and sample 5D data, $(R, z, v_R, v_\varphi, v_z)$. Regarding the size of the region, we generally find that for accurate results, survey regions of size $\gtrsim 300$ pc around the Sun are needed. If smaller regions are used, the flows have difficulty accurately estimating the spatial gradients of the DF, leading (via equation 4) to inaccurate estimates for the acceleration.

The assumption of axisymmetry is not strictly necessary, but leads to a substantial improvement in accuracy. This is due to more than just the reduction in dimensionality: in an axisymmetric or near-axisymmetric system, $\partial f/\partial\varphi$ should be zero or close to zero. However, v_φ is larger than the other velocity components, due to the Galactic rotation. So, small errors in the estimation of $\partial f/\partial\varphi$ are disproportionately amplified in the $\boldsymbol{v}\cdot\boldsymbol{\nabla}_{\boldsymbol{x}} f$ term appearing in equation (4). Assuming axisymmetry (i.e. fixing $\partial f/\partial\varphi = 0$) eliminates this effect.

We sample the data directly from the DF (equation 5) using a Markov chain Monte Carlo (MCMC) technique. For this, we use the affine-invariant ensemble sampler implemented in the software package EMCEE (Foreman-Mackey et al. 2013). To evaluate the DF in this procedure, we use the qDF implementation and various potential models aboard the software package GALPY (Bovy 2015).

4 RESULTS

In Section 4.1, we calculate solar neighbourhood accelerations in our MW model. Then, in Section 4.2, we consider the effects of disequilibria in the MW disc by perturbing this data set.

4.1 The local acceleration field

Training an ensemble of normalizing flows on the mock data set generated from our MW model, we arrive at a learned DF describing the local population of disc stars. Fig. 1 depicts this learned DF, alongside the true DF (equation 5) and residuals. Three phase planes are depicted in Fig. 1: R–z, v_R–v_φ, and z–v_z. In each case, a 2D slice

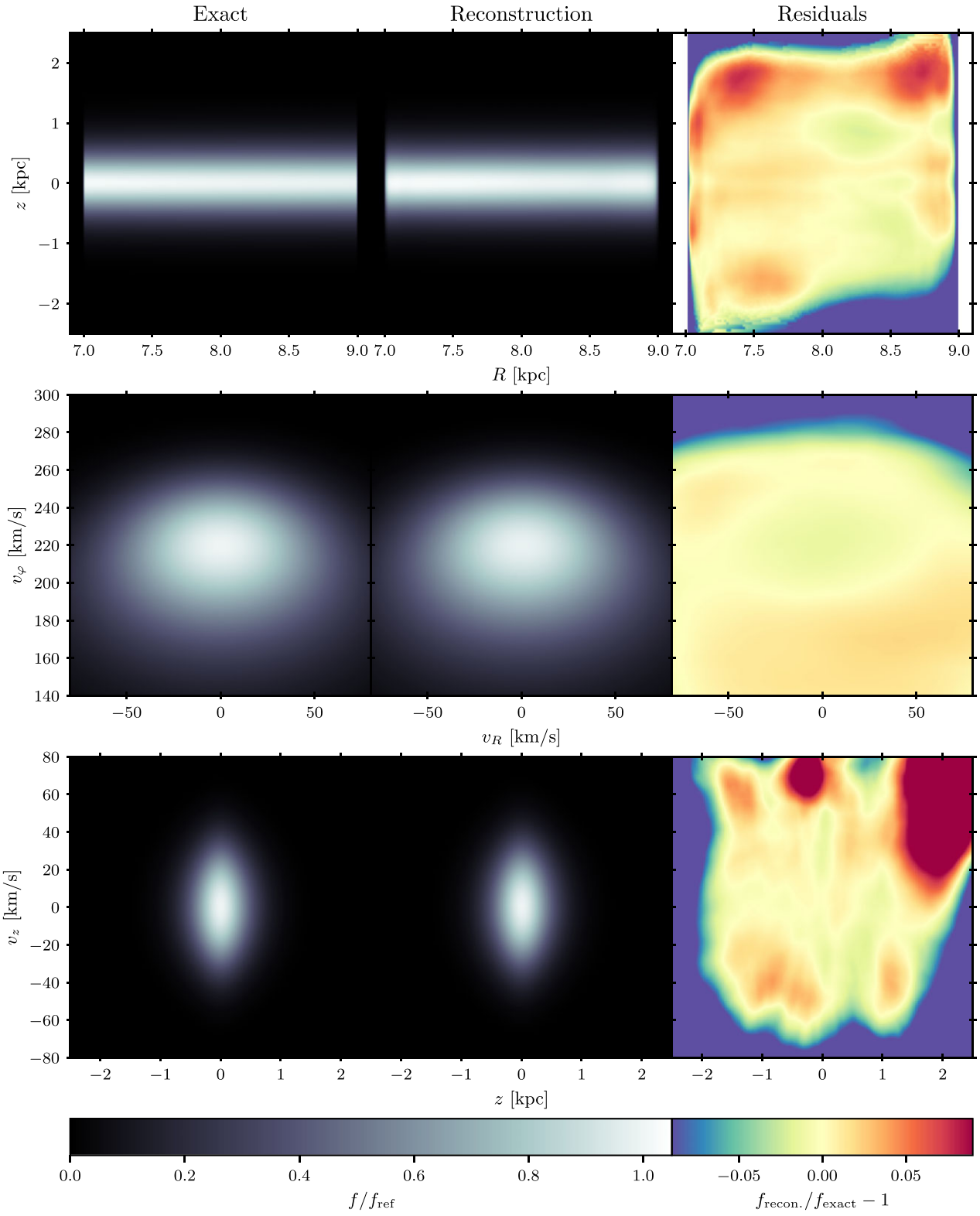

Figure 1. Three slices through phase space: R–z (top), v_R–v_φ (middle), and z–v_z (bottom). In each case, the left-hand panel shows the true, underlying DF (equation 5), the middle panel shows the flow-learned DF, and the right-hand panel shows fractional residuals between the two. In each panel, the DF is evaluated by varying two coordinates while holding the remaining coordinates at fixed values. The values of the DF are shown in units of $f_{\rm ref}$: the DF evaluated at the solar position ($R = 8$ kpc, $z = 10$ pc) and the local standard of rest. This figure demonstrates that our technique is capable of reproducing the underlying stellar DF with excellent accuracy within the well-populated regions of phase space.

through phase space is shown, i.e. two coordinates are varied while the other three are held constant at $R = 8$ kpc, $v_\varphi = 220$ km s^{-1}, $z = v_R = v_z = 0$.

Inspecting the residuals (the right-hand panels of Fig. 1), we find excellent per cent-level agreement between the true DF and the learned DF in the well-populated region of phase space, i.e. within $|v - \bar{v}| \lesssim 60$ km s^{-1} and $|z| \lesssim 2$ kpc. The errors only start to grow large at greater velocities or greater heights above or below the mid-plane, where the estimators have very few data points with which to train.

Another remarkable feature of Fig. 1, which might escape notice at first glance, is in the R–z plane (top row). Here, there are hard edges in the exact DF at $R = 7$ and 9 kpc. These represent the edges of our sample region (Section 3.3). When feeding the data to the normalizing flows during the training procedure, the flows are entirely unaware of these hard edges a priori. None the less, these sharp edges are detected and reproduced excellently in the model DF, albeit with increased residuals immediately inside the edges. This is a demonstration of the flexibility of normalizing flows, which can deal with sharp transitions in the data.

We now have a DF model that we can input to the machinery of Section 2.2 to calculate gravitational accelerations. The result of doing so is shown in Fig. 2, which plots derived vertical and radial accelerations alongside the actual accelerations in our MW model. As in Fig. 1, the fractional residuals in the well-populated regions are at the sub-per cent level. One exception is the region near $z = 0$ where the residuals artificially grow large as a result of dividing by small numbers; by eye, it is clear that the agreement remains good in this region. On the other hand, the radial acceleration residuals do truly grow large in the regions immediately near the edges at 7 and 9 kpc, as a result of the DF derivatives being poorly estimated in these regions. As suggested by Fig. 1, similar issues also arise at large heights above and below the mid-plane.

Edge effects aside, Fig. 2 encapsulates the key result of this paper: given 10^6 stars in equilibrium in an annulus between $R = 7$ and 9 kpc, we can calculate the underlying gravitational acceleration field with excellent accuracy.

Before forecasting such accuracy for application to the *Gaia* data, it is worth ensuring that this accuracy persists in the presence of realistic errors. We perform this test by adding errors to the parallax, line-of-sight (los) velocity, and proper motions of each mock star, neglecting errors in the sky positions which we assume to be subdominant. We universally assign Gaussian errors of $\sigma_\varpi = 25$ μas, $\sigma_\ell = 1$ km s^{-1}, and $\sigma_\wp = 25$ μas yr^{-1} to each star's parallax (σ_ϖ), los velocity (σ_ℓ), and proper motion ($\sigma_\wp$), respectively, assuming zero covariance. In the real *Gaia* data, these uncertainties correlate with the apparent G-band magnitudes: brighter stars have more precise astrometry. Our chosen errors correspond to stars with $G \lesssim 14.5$ in *Gaia* EDR3 (Gaia Collaboration 2021; see also the $\sigma_\wp$ fitting function of Dong-Páez, Vasiliev & Evans 2022). The subset of stars with $G < 14.5$ is the most kinematically robust sample, and at least for *Gaia* DR2, Schönrich, McMillan & Eyer (2019) recommend restricting kinematic analyses to this subset to avoid serious systematic errors. Our assigned errors therefore closely resemble the typical errors in the kind of subset of *Gaia* data to which our method is likely to be applied in future.

We propagate these errors following the method suggested in Paper I: when training an ensemble of flows, each flow is provided with a different data set, representing a different realization of the error distribution. In practice, each flow takes in the original error-free data set, transforms the data coordinates to the heliocentric spherical frame, inverts distances to parallaxes, shifts parallaxes, los velocities, and proper motions by random amounts as generated from Gaussian distributions of widths $\sigma_\varpi, \sigma_\ell, \sigma_\wp$, respectively, transforms the data back to the original Galactocentric cylindrical frame, then finally commences the training procedure as normal. Subsequently, the differences in the DFs learned by different flows quantify not only the variability intrinsic to the technique (see Section 2.1), but also the statistical uncertainty in the training data.

Fig. 3 plots vertical and radial accelerations measured after propagating uncertainties in this way (orange points with error bars). Each point represents the median measured value across the flow ensemble, while the accompanying error bars give the 16th and 84th percentile values. Reassuringly, the accuracy remains excellent, with sub-per cent level residuals everywhere except near the radial edges and large $|z|$ as before.

Another obstacle facing the application of our technique to real data is that many *Gaia* stars do not have accompanying los velocity data. Of the EDR3 stars with $G < 14.5$, around 30 per cent have measured los velocities. The full third data release (scheduled 2022) will fill in many gaps and we can boost the proportion even further by cross-matching the *Gaia* stars with those from independent radial velocity surveys, but it remains inevitable that a significant proportion of the data set will lack this sixth dimension.

There are a number of ways to circumvent this issue. Arguably the simplest is to assume the los velocity selection has minimal kinematic bias, so that all of the stars can be used to learn the stellar density $\nu(\boldsymbol{x})$, and the subset of stars with available los velocities can be used to learn the (position-dependent) velocity distribution $p(\boldsymbol{v} \mid \boldsymbol{x})$. Normalizing flows can be employed in both cases, and the full DF is then given by the product of the two probability distributions. Fig. 3 shows the results of such an approach, plotting accelerations (blue points) obtained from the same mock data as that used for Figs 1 and 2, but now with a randomly chosen 50 per cent sample of stars taken as having missing los velocity measurement. In other words, the full data set is used to learn $\nu(R, z)$, but only half of the data set is used to learn $p(\boldsymbol{v} \mid R, z)$. The residuals remain generally small, indicating that the issue of missing los velocities is not insurmountable.

There are, however, some issues arising: the residuals are somewhat noisier and grow larger (~ 10 per cent) at large z. Both of these facts result from half of the data being discarded when learning $p(\boldsymbol{v} \mid R, z)$. In particular, the *spatial* gradients of $p(\boldsymbol{v} \mid R, z)$ are less well estimated as a consequence. Given that the discarded stars do have two dimensions of velocity information (i.e. their proper motions), we might attain better results by instead retaining these stars and estimating their missing los velocities. A possible technique to do so has been suggested by the work of Dropulic et al. (2021), which demonstrated that artificial neural networks can be successful in recreating the missing los velocities of *Gaia* stars.

4.2 Disequilibria

There is a growing body of evidence for non-equilibrium structure in the stellar kinematics of the MW disc. Whereas the first step of our methodology (learning the DF) assumes only axisymmetry, the second step (converting to an acceleration field) requires the assumption of dynamical equilibrium so that the time-derivative term can be neglected in the CBE.

In Paper I, we showed that the incorrect assumption of equilibrium leads to a bias in the derived accelerations that is linear in $\partial f/\partial t$. Similarly, Banik et al. (2017) estimated that, under plausible perturbations, the bias induced by incorrect assumption of equilibrium in measurements of vertical accelerations is at the 10 per cent level or so. However, their assumed methodology was different from that

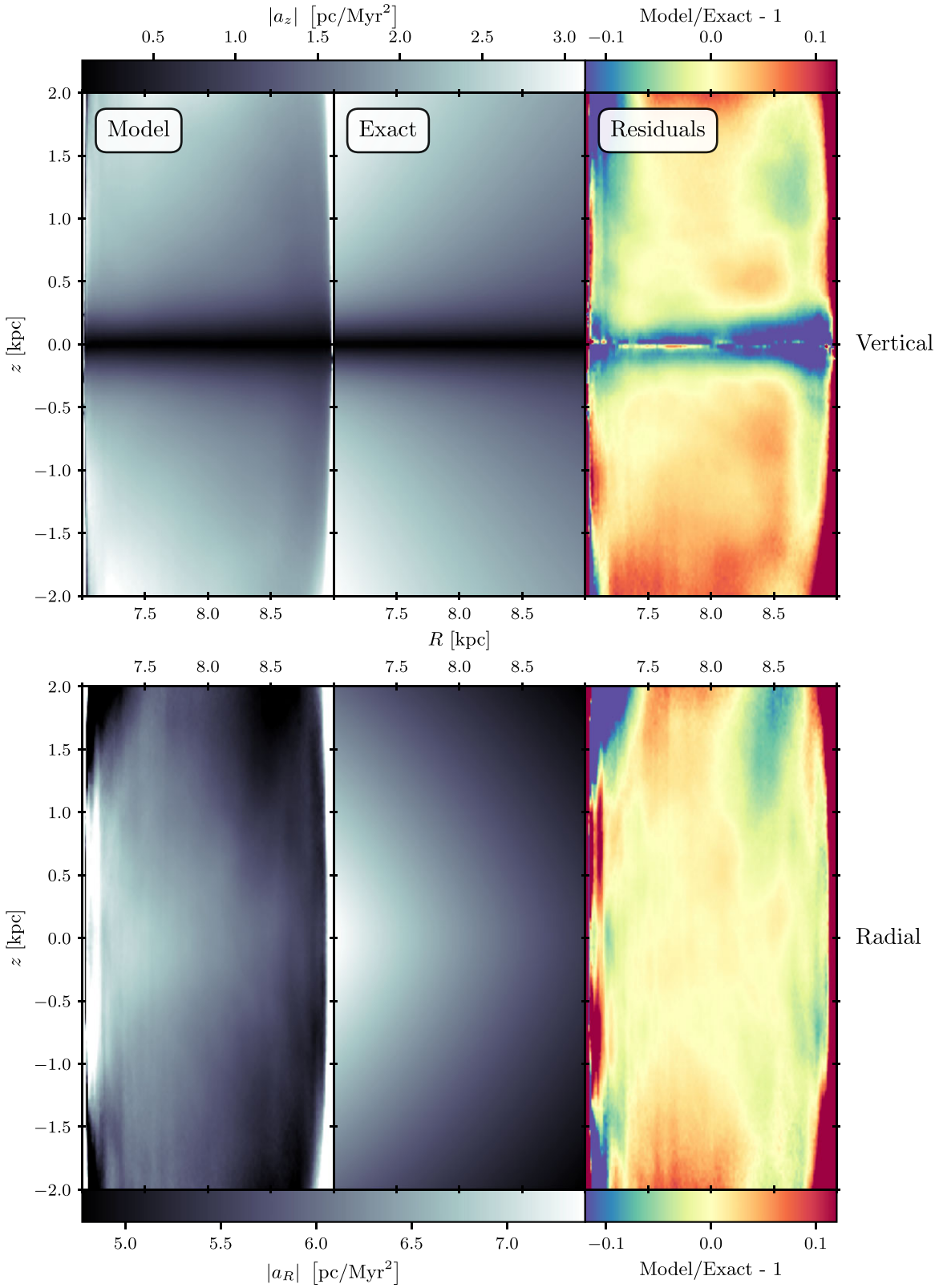

Figure 2. Maps of vertical (top) and radial accelerations (bottom). In each case, the left-hand panel plots our reconstructed accelerations, the middle panel the exact accelerations of the model, and the right-hand panel the fractional residuals. This figure demonstrates that, assuming dynamical equilibrium, our technique is capable of calculating accelerations in the solar neighbourhood with excellent accuracy.

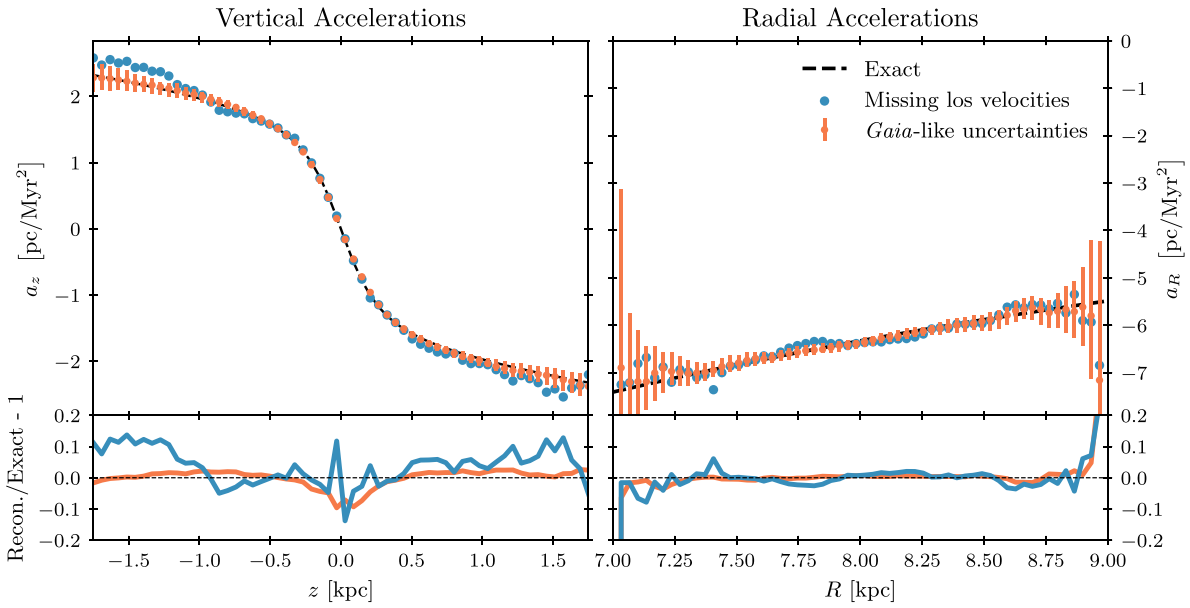

Figure 3. Accelerations measured using a mock data set with *Gaia*-like uncertainties in parallax, los velocity, and proper motions (orange points with error bars) and a mock data set with 50 per cent missing los velocities (blue points). The vertical accelerations (left) are taken at $R = 8$ kpc and the radial accelerations (right) are taken at $z = 0$. In the case of the mock data with uncertainties, these are propagated by providing a different realization of the error distribution to each flow in the ensemble. The points give the median values across the ensemble, while the error bars show the 16th and 84th percentile values. The black dashed line plots accelerations in the underlying MW model. The smaller panels below show fractional residuals between the measured and true values. Even with realistic errors or missing los velocities, our method recovers the underlying acceleration field with sub-per cent accuracy in well-populated regions.

of the present work, and so the applicability of this estimate is not entirely clear.

Here, we quantify the disequilibrium bias by applying our methodology to a mock data set representing a perturbed stellar population. To generate this perturbed data set, we start by employing the method described in Section 3 to sample an equilibrium data set comprising 2.5×10^7 stars between $R = 1$ and 16 kpc, under our fiducial MW model. Note that this population size gives roughly the desired number of stars (10^6) in our region of interest, 7 kpc $< R <$ 9 kpc.

Next, we apply a 'kick' to these equilibrium stars, mimicking the procedure of Li & Widrow (2021): we randomly choose 10 per cent of the stars and boost their vertical velocities by $\delta v_z = +20$ km s^{-1}. Such a kick can be understood as being roughly resemblant to the impact of the Sagittarius dwarf passing through the Galactic disc (e.g. Laporte et al. 2019; Bland-Hawthorn & Tepper-García 2021): under the impulse approximation, $\delta v \approx 2GM/bv$, where M, v, and b are, respectively, the mass, speed, and impact parameter of the perturber. Adopting plausible values of $M = 10^{10}$ M$_\odot$, $v = 300$ km s^{-1}, and $b = 15$ kpc, the resulting kick is $\delta v \approx 20$ km s^{-1}.

After applying this perturbation, we evolve the stars' orbits under the (unperturbed) MW potential for 500 Myr, saving snapshots of this evolution at $t = 0$, 200, and 500 Myr after the initial perturbation. At each snapshot, we isolate the stars between $R = 7$ and 9 kpc and feed them through the pipeline of Section 2 to measure accelerations. Fig. 4 shows the resulting accelerations at these times.

Immediately after the perturbation, accelerations are everywhere overestimated by 10 per cent or so: a similar level of bias to that predicted by Banik et al. (2017). Note that as in Fig. 2, we are still disregarding the residuals immediately around $z = 0$. After 200 Myr, the magnitude of the bias has decreased to ~ 5 per cent, and is confined to larger heights, $|z| \gtrsim 0.5$ kpc. The stars confined to lower heights appear to have equilibriated more quickly, as expected given their shorter dynamical times.

Finally, after 500 Myr, the perturbation appears to have decayed beyond our sensitivity: the residuals are everywhere comparable to the equilibrium case (cf. Fig. 2). There is a feature in the residuals at $z \approx -1.5$ kpc, but it is unclear whether this is due to lingering effects of the perturbation at large heights or the smaller sampling densities there.

Our finding that the stars have largely equilibriated after 500 Myr is at odds with Li & Widrow (2021), who still see a significant bias 500 Myr after an identical perturbation. There are a number of possible causes for this discrepancy. First, a denser Galactic disc has a shorter dynamical time and thus faster equilibriation. However, the difference in the two models does not appear to be great enough: the density in our model is only around 35 per cent larger, meaning the dynamical time is only around 15 per cent shorter (taking $t_{\rm dyn} \propto \rho^{-1/2}$). Another possibility is the dimensionality. Under our treatment, we evolve the stellar orbits in three-dimensional configuration space after the initial perturbation, whereas Li & Widrow (2021) use a one-dimensional approximation. This is tantamount to assuming integrability, as all Hamiltonians with one degree of freedom are exactly integrable. A bundle of trajectories in phase space spreads linearly with time in an integrable system, and so mixing times are longer. A final possibility is the difference in DF models. Whereas we learn a non-parametric DF, Li & Widrow (2021) fit the stellar kinematics with an analytical DF. It is possible that after 500 Myr, $\partial f/\partial t$ is sufficiently small that equation (4) can be employed with minimal resulting bias provided the correct DF is used, but the analytical DF used by Li & Widrow (2021) is not (yet) a good fit for the stars, which still retain some memories of the perturbation in their distribution. In other words, the bias Li & Widrow (2021) find at 500 Myr might not be directly induced by disequilibrium, but indirectly, via the misapplication of their analytical DF model.

The estimated biases and time-scales shown here can only be used as a rough guide. The real perturbation in the MW disc due

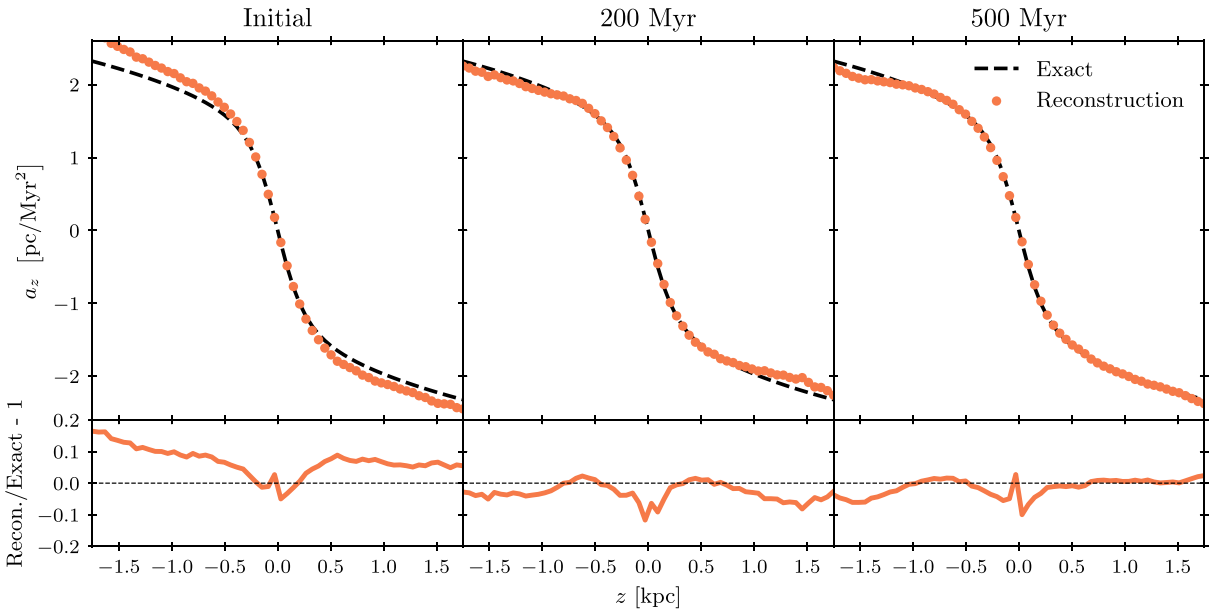

Figure 4. Vertical accelerations derived by applying our methodology to a stellar population following a perturbation. The three main panels present different snapshots in time: immediately following the perturbation (left), 200 Myr later (middle), and 500 Myr later (right). In each case, the coloured points plot the derived accelerations, and the black dashed line plots the true accelerations under our adopted MW model. The three smaller panels below show fractional residuals. This figure illustrates the bias induced by incorrectly assuming equilibrium, and how the bias varies with height and decays with time.

to the Sagittarius dwarf could well be larger, and thus induce a longer-lasting bias in the measured accelerations. Moreover, there are other potential sources of vertical perturbation beyond the Sagittarius dwarf, such as stellar bar buckling (Khoperskov et al. 2019). Beyond these vertical perturbations, there are also in-plane perturbations to consider. For example, moving groups (i.e. coherent kinematic structures in the local v_x–v_y space) are either dynamical footprints of the Galactic spiral arms and bar (e.g. Antoja et al. 2008; Michtchenko et al. 2018), or dissolving open clusters and associations (e.g. Oh & Evans 2020; Gagné et al. 2021). These could provide additional contributions to the systematic bias in our estimation of the Galactic acceleration field, but a full accounting is beyond the scope of the present work. However, such effects can be mitigated in practice by masking the stars known to belong to these substructures.

Given this uncertainty, it is worth asking whether our framework provides any way to directly detect the presence of disequilibria. Li & Widrow (2021) achieved this by comparing their best-fitting model of the DF directly with their perturbed data binned in z–v_z space, and found that a clear 'phase spiral' emerged in the residuals. As a star progresses along its orbit, it exhibits oscillatory motion in the z–v_z plane. In particular, defining the 'vertical energy' $E_z = v_z^2/2 + \Phi_z(z)$, where Φ_z is the vertical part of the galactic potential, stars moves on clockwise 'circles' of constant E_z. In any potential except a harmonic ($\Phi_z \propto z^2$) potential, the orbital period in this plane is not constant with respect to E_z; there is differential rotation. An initial overdensity in z–v_z space is thus stretched, after the passage of time, into a phase spiral. Eventually, the spiral is stretched and wound to the point where it is no longer detectable, and the population is 'phase mixed'. The detection of a phase spiral in a stellar population constitutes clear evidence that the population is not fully phase mixed, i.e. not in equilibrium.

Inspired by Li & Widrow (2021), we search for a phase spiral in the DF trained on the perturbed data. Unlike in their case, a phase spiral will not emerge in our residuals, because any phase spiral encoded in the data will be similarly encoded in the learned DF. We instead consider a symmetrized DF $\bar{f}$ constructed from the learned DF f via

$$\bar{f}(z, v_z, R, v_R, v_\varphi) \equiv \frac{f(z, v_z, R, v_R, v_\varphi) + f(\text{-}z, \text{-}v_z, R, v_R, v_\varphi)}{2}. \quad (6)$$

In other words, we take average of f and its 180° rotation in the z–v_z plane, holding R, v_R, v_φ fixed. We then calculating residuals by comparing f with $\bar{f}$, and so extract any asymmetric component of f.

Fig. 5 plots $f/\bar{f} - 1$ after 200 Myr, fixing $R = 8$ kpc, $v_\varphi = 220$ km s^{-1}, and $v_R = 0$. A remarkably clear phase spiral emerges. The spiral is sharply defined: the residuals swing from ~ 10 per cent to ~ -10 per cent very rapidly between the overdense and underdense regions, although the innermost part of the spiral is a bit fainter. We find that the spiral emerges much more cleanly and clearly in these DF slices rather than in projections (i.e. integrating f over v_φ, v_R), where the residuals are ~ 1 per cent. This is because the orientation and winding of the spiral vary as a function of v_φ and v_R, so that integrating over velocity space serves to partially wash out the signal.

This detection of the phase spiral is an encouraging result: it suggests that a phase spiral would be easy to detect in real data, enabling a straightforward diagnosis of disequilibrium. Moreover, a phase spiral has various uses beyond the simple diagnosis of disequilibrium. In particular, the exact shape of the spiral encodes a wealth of information. For example, Li & Widrow (2021) fit the phase spiral shape (both their mock spiral and the real Gaia DR2 spiral) to find the time elapsed since the spiral-inducing perturbation. Meanwhile, Widmark, Laporte & de Salas (2021b) and Widmark et al. (2021c) use the spiral shape to derive the vertical potential in the Galactic disc, and from there the tightest constraints to date on a thin dark disc.

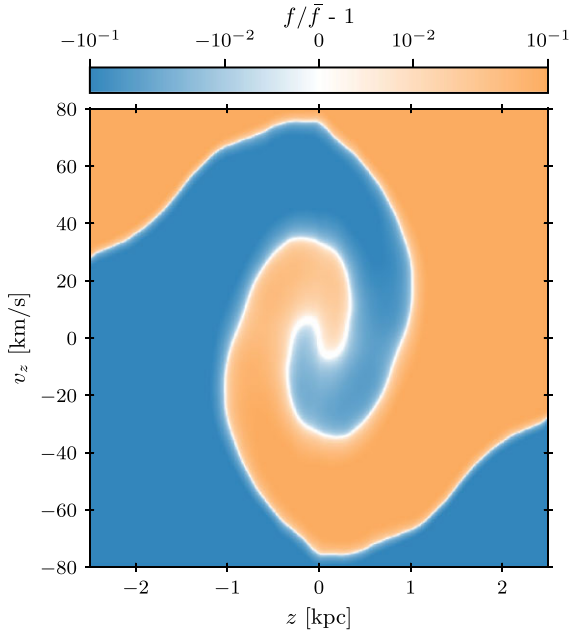

Figure 5. The asymmetric component of the DF trained on the non-equilibrium stellar population, 200 Myr after the perturbation. The asymmetric component is extracted by dividing by a symmetrized DF $\bar{f}$ (definition: equation 6). A clear phase spiral emerges when plotting the learned DF in this manner.

5 COMPARISON WITH OTHER METHODS

Here, we test our technique by a direct performance comparison with other, competing methods. Two such techniques are described in Section 1: Jeans analysis (e.g. Salomon et al. 2020) and the 1D DF-fitting approach (e.g. Widmark et al. 2021a). We outline the two methods here, and give more detailed descriptions of our implementations of them in Appendices A and B, respectively.

In the Jeans analysis of Salomon et al. (2020), stars are binned into radial and vertical bins, and radial and vertical velocity dispersions are computed in each bin, along with the stellar density. Parametrized functional forms are assumed for the spatial variation of the density and dispersions, and these functions are fit to the values obtained from the bins. Given the functional forms and the best-fitting parameters, the vertical Jeans equation is solved to give the vertical acceleration. Along with the assumed functional forms, another key assumption concerns the tilt of the local velocity ellipsoid (i.e. the covariance between radial and vertical motions), which is assumed to be spherically aligned. This has been shown empirically to be a generally good assumption, except perhaps very close to the disc plane (Everall et al. 2019).

By contrast, the DF-fitting approach of Widmark et al. (2021a) does not bin the data, but directly fits the positions and motions of individual stars. Here, the key assumptions are that the DF is separable, i.e. $f(\boldsymbol{x}, \boldsymbol{v}) = f_\perp(z, v_z) f_\parallel(x, y, v_x, v_y)$, and that the 'vertical energy' $E_z \equiv v_z^2/2 + \Phi(z)$ is an integral of motion. These assumptions, taken together with an assumed parametrization for the potential $\Phi(z)$, give an analytical expression for $f_\perp$ which can be directly fit to the data. The vertical accelerations are then given by first derivative of $\Phi(z)$, taking the best-fitting parameters.

Fig. 6 shows the results of this test for two mock data sets in particular: the unperturbed data set studied in Section 4.1, and the perturbed data set studied in Section 4.2, 500 Myr after the initial perturbation. In the unperturbed case (left-hand panel), the vertical accelerations calculated with our method reproduce the underlying model with excellent sub-per cent level accuracy, as already demonstrated in Section 4.1. Here, the DF-fitting approach of Widmark et al. (2021a) performs nearly as well, giving residuals of $\lesssim$5 per cent everywhere. As the data are binned and only the second moments are considered rather than the full shape in a Jeans analysis, much information content is lost. It is then perhaps predictable that the Jeans approach performs the least well of the three. Turning to the perturbed case (right-hand panel), our measured accelerations are of comparable accuracy to the unperturbed case (cf. Fig. 4, right-hand panel). The accuracy of other two methods, however, appears to worsen. The change is most marked in the DF-fitting approach, where the residuals approximately double to the ~ 10 per cent level. This finding lends support to our speculation of Section 4.2: while our technique assumes equilibrium and is thus susceptible to disequilibrium-induced bias, it less susceptible than other techniques.

6 CONCLUSIONS

A novel procedure for calculating gravitational accelerations from stellar kinematical data was introduced in An et al. (2021). In this article, we test the methodology in the context of the neighbourhood of the Sun, with a view towards an upcoming paper in which we apply our technique to the *Gaia* data.

The procedure is split into two stages. First, we 'learn' the phase space DF of the data by training normalizing flows (Rezende & Mohamed 2015; Kobyzev et al. 2021; Papamakarios et al. 2021). In so doing, we construct a data-driven, non-parametric DF, without recourse to any assumptions about the underlying kinematics of the stars, e.g. we do not assume the stellar populations are isothermal or reduce the problem to one dimension. One assumption we do make is axisymmetry, i.e. we learn a five-dimensional DF in $(R, z, v_R, v_\varphi, v_z)$, but this assumption can be easily relaxed if desired. In the second step of the procedure, we convert the learned DF to gravitational accelerations, via an inversion of the CBE. At this stage, we assume that the stars are in dynamical equilibrium.

To test our method, we apply it to a mock data set resembling a population of MW disc stars in equilibrium. The stars are drawn from multiple 'MAPs', and so represent a mix of stars mirroring the mix of sub-populations in the real Galactic disc. Following Bovy & Rix (2013) and Ting et al. (2013), we construct the mock data set by assuming that each MAP can be individually described by a 'quasi-isothermal' DF (Binney 2010, 2012b), tracing the underlying bulge + halo + disc MW model of Bovy (2015). We sample 10^6 stars between 7 and 9 kpc in Galactocentric radius.

Given this mock data set, we apply our outlined technique, i.e. train normalizing flows to learn the DF, then convert the DF to a map of accelerations. We find an excellent sub-per cent level match between the measured radial and vertical accelerations and the underlying acceleration field in the adopted MW model. As we will apply our method to data from the *Gaia* satellite (Gaia Collaboration 2021), we check that this excellent accuracy persists even when realistic errors are added to the data and propagated to the measured accelerations, and when a substantial proportion of los velocities are unavailable. This is the key result of the paper: given the observed positions and motions of a million bright ($G \lesssim 14.5$) stars within a kiloparsec of the Sun, we can robustly determine the underlying gravitational accelerations.

A potential source of systematic bias in our technique is disequilibrium. In converting a learned DF to accelerations, we assume dynamical equilibrium. We test this by employing our technique on a mock data set following a perturbation emulating the passage of

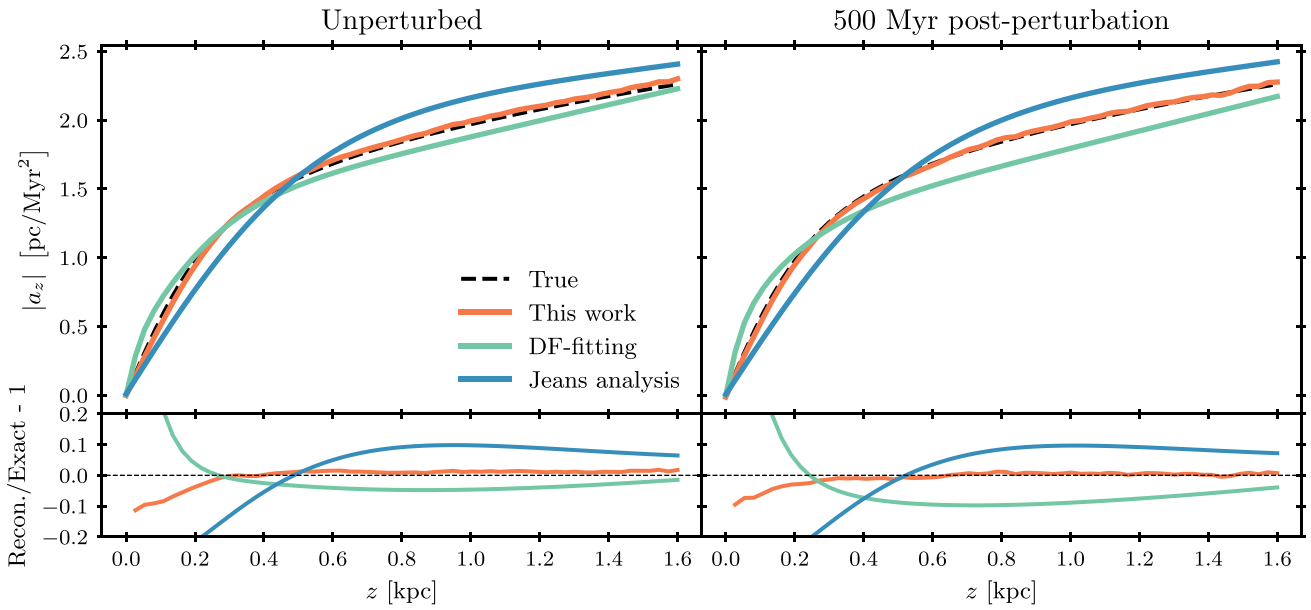

Figure 6. Vertical accelerations at the solar radius ($R = 8$ kpc) inferred from mock data using different techniques. The two panels correspond to two different mock data sets: the unperturbed data set of Section 4.1 (left), and the perturbed data set of Section 4.2, 500 Myr following the initial perturbation (right). In each case, the black dashed line gives the true accelerations under the assumed MW model, and the coloured lines plot the accelerations inferred via the different techniques as labelled in the legend. By bypassing various limiting assumptions made by other techniques, our method is capable of more accurately measuring accelerations.

the Sagittarius dwarf through the outer disc. Immediately following the perturbation, we find that accelerations are overestimated by ~ 10 per cent. This bias decays over time, until it is no longer detectable in the residuals (at least within $|z| \lesssim 1$ kpc) by 500 Myr. Additionally, we find that a remarkably clear and distinct phase spiral can be extracted from the DF trained on the perturbed data set. Not only can such a phase spiral be used to diagnose disequilibrium, but its shape can also reveal insights into the MW potential and perturbation history (e.g. Li & Widrow 2021; Widmark et al. 2021b, c)

We compare the performance of our method to that of two other widely used methods: solution of the Jeans equations and fitting the vertical (one-dimensional) DF to parametrized models. Using the same mock data set as an input, our method measures accelerations the most accurately. This is particularly true in the aftermath of a perturbation, suggesting that our technique is less susceptible to disequilibrium-induced bias than competing techniques.

In summary, we provide a new algorithm to accurately determine the local acceleration field from stellar kinematical data by non-parametrically reconstructing the stellar DF. We argue that it is the most robust technique yet devised for this purpose. Its strength derives largely from the fact that the DF is constructed directly from the data, thereby bypassing the limiting assumptions and model-sensitivity of our existing methods. In the *Gaia* era, such data-driven techniques have the potential to reveal new insights into fundamental physics and the makeup of our Galactic neighbourhood.

ACKNOWLEDGEMENTS

We thank the anonymous referee for their constructive comments and George Papamakarios for useful discussions. APN and CB were supported by a Research Leadership Award from the Leverhulme Trust. We are grateful for access to the University of Nottingham's Augusta HPC service.

DATA AVAILABILITY

The source code and plotting scripts used in this paper have all been made publicly available at https://github.com/aneeshnaik/LocalFlows, and the mock data sets are available at https://doi.org/10.5281/zenodo.5781350.

REFERENCES

An J., Naik A. P., Evans N. W., Burrage C., 2021, MNRAS, 506, 5721 (Paper I)
Antoja T., Figueras F., Fernández D., Torra J., 2008, A&A, 490, 135
Antoja T. et al., 2018, Nature, 561, 360
Banik N., Widrow L. M., Dodelson S., 2017, MNRAS, 464, 3775
Binney J., 2010, MNRAS, 401, 2318
Binney J., 2012a, MNRAS, 426, 1324
Binney J., 2012b, MNRAS, 426, 1328
Binney J., Tremaine S., 2008, Galactic Dynamics, 2nd edn. Princeton Univ. Press, Princeton, NJ
Bland-Hawthorn J., Tepper-García T., 2021, MNRAS, 504, 3168
Bovy J., 2015, ApJS, 216, 29
Bovy J., 2020, preprint (arXiv:2012.02169)
Bovy J., Rix H.-W., 2013, ApJ, 779, 115
Buch J., Leung J. S. C., Fan J., 2019, J. Cosmol. Astropart. Phys., 2019, 026
Chakrabarti S. et al., 2020, ApJ, 902, L28
Chakrabarti S., Stevens D. J., Wright J., Rafikov R. R., Chang P., Beatty T., Huber D., 2021a, preprint (arXiv:2112.08231)
Chakrabarti S., Chang P., Lam M. T., Vigeland S. J., Quillen A. C., 2021b, ApJ, 907, L26
de Salas P. F., Widmark A., 2021, Rep. Prog. Phys., 84, 104901
Dong-Páez C. A., Vasiliev E., Evans N. W., 2022, MNRAS, 510, 230
Dropulic A., Ostdiek B., Chang L. J., Liu H., Cohen T., Lisanti M., 2021, ApJ, 915, L14
Durkan C., Bekasov A., Murray I., Papamakarios G., 2019, in Wallach H., Larochelle H., Beygelzimer A., d'Alché-Buc F., Fox E., Garnett R., eds, Advances in Neural Information Processing Systems, Vol. 32,

33rd Conference on Neural Information Systems (NeurIPS 2019). Curran Associates, Inc., Vancouver, Canada, p. 7511
Everall A., Evans N. W., Belokurov V., Schönrich R., 2019, MNRAS, 489, 910
Foreman-Mackey D., Hogg D. W., Lang D., Goodman J., 2013, PASP, 125, 306
Gagné J., Faherty J. K., Moranta L., Popinchalk M., 2021, ApJ, 915, L29
Gaia Collaboration, 2021, A&A, 649, A1
Green G. M., Ting Y.-S., 2020, in Machine Learning & the Physical Sciences, Workshop at the 34th Conference on Neural Information Systems (NeurIPS2020 ML4PS). p. 12
Guo R., Liu C., Mao S., Xue X.-X., Long R. J., Zhang L., 2020, MNRAS, 495, 4828
Hagen J. H. J., Helmi A., 2018, A&A, 615, A99
Hinkel N. R., Timmes F. X., Young P. A., Pagano M. D., Turnbull M. C., 2014, AJ, 148, 54
Khoperskov S., Di Matteo P., Gerhard O., Katz D., Haywood M., Combes F., Berczik P., Gomez A., 2019, A&A, 622, L6
Kingma D. P., Ba J., 2015, in Bengio Y., LeCun Y., eds, 3rd International Conference on Learning Representations. Conference Track Proceedings (ICLR 2015). Available at:https://arxiv.org/abs/1412.6980
Kobyzev I., Prince S. J. D., Brubaker M. A., 2021, IEEE Trans. Pattern Anal. Mach. Intell., 43, 3964
Laporte C. F. P., Minchev I., Johnston K. V., Gómez F. A., 2019, MNRAS, 485, 3134
Li H., Widrow L. M., 2021, MNRAS, 503, 1586
Loebman S. R. et al., 2014, ApJ, 794, 151
Michtchenko T. A., Lépine J. R. D., Barros D. A., Vieira R. S. S., 2018, A&A, 615, A10
Milgrom M., 1983, ApJ, 270, 365
Oh S., Evans N. W., 2020, MNRAS, 498, 1920
Papamakarios G., Pavlakou T., Murray I., 2017, in Guyon I., Luxburg U. V., Bengio S., Wallach H., Fergus R., Vishwanathan S., Garnett R., eds, Advances in Neural Information Processing Systems, Vol. 30, 31st Conference on Neural Information Systems (NIPS 2017). Curran Associates, Inc., Long Beach, CA, p. 2338
Papamakarios G., Nalisnick E., Rezende D. J., Mohamed S., Lakshminarayanan B., 2021, J. Mach. Learn. Res., 22, 57
Quercellini C., Amendola L., Balbi A., 2008, MNRAS, 391, 1308
Ravi A., Langellier N., Phillips D. F., Buschmann M., Safdi B. R., Walsworth R. L., 2019, Phys. Rev. Lett., 123, 091101
Read J. I., 2014, J. Phys. G: Nucl. Part. Phys., 41, 063101
Rezende D. J., Mohamed S., 2015, in Bach F., Blei D., eds, Proceedings of Machine Learning Research, Vol. 37, Proceedings of the 32nd International Conference on Machine Learning (PMLR). p. 1530
Salomon J.-B., Bienaymé O., Reylé C., Robin A. C., Famaey B., 2020, A&A, 643, A75
Schönrich R., Dehnen W., 2018, MNRAS, 478, 3809
Schönrich R., McMillan P., Eyer L., 2019, MNRAS, 487, 3568
Schutz K., Lin T., Safdi B. R., Wu C.-L., 2018, Phys. Rev. Lett., 121, 081101
Silverwood H., Easther R., 2019, PASA, 36, e038
Sivertsson S., Silverwood H., Read J. I., Bertone G., Steger P., 2018, MNRAS, 478, 1677
Ting Y.-S., Rix H.-W., Bovy J., van de Ven G., 2013, MNRAS, 434, 652
Widmark A., 2019, A&A, 623, A30
Widmark A., Monari G., 2019, MNRAS, 482, 262
Widmark A., de Salas P. F., Monari G., 2021a, A&A, 646, A67
Widmark A., Laporte C., de Salas P. F., 2021b, A&A, 650, A124
Widmark A., Laporte C. F. P., de Salas P. F., Monari G., 2021c, A&A, 653, A86

APPENDIX A: JEANS ANALYSIS

In this appendix, we describe the Jeans analysis used for method comparison in Section 5. Except where noted, we follow the procedure of Salomon et al. (2020), who used the method to measure the vertical force and local density of dark matter using red clump stars from *Gaia DR2*.

The three Jeans equations relate stellar velocity dispersions and densities to gravitational accelerations, and can be obtained by integrating the CBE (equation 1) over the three velocity dimensions (e.g. Binney & Tremaine 2008). Under the assumptions of axisymmetry and steady state, the time-independent vertical Jeans equation is

$$\frac{\partial}{\partial z}\left(\nu\sigma_z^2\right) + \frac{1}{R}\frac{\partial}{\partial R}\left(R\nu\sigma_{Rz}^2\right) = -\nu\frac{\partial\Phi}{\partial z}, \tag{A1}$$

where ν is the stellar density, and $\sigma_z^2, \sigma_{Rz}^2$ are the z–z and R–z components of the velocity dispersion tensor. All three of these quantities are themselves functions of R and z. Assuming that the local velocity ellipsoid is spherically aligned, σ_{Rz}^2 is given in turn by

$$\sigma_{Rz}^2 = Rz\frac{\sigma_R^2 - \sigma_z^2}{R^2 - z^2}, \tag{A2}$$

where σ_R^2 is the R–R component of the velocity dispersion tensor.

To proceed, we need to measure $\nu, \sigma_z^2, \sigma_R^2$ and their derivatives in order to solve equations (A1) and (A2) for $a_z \equiv -\partial\Phi/\partial z$. We do this by binning the stars in R and z, calculating the density and velocity dispersions in each bin, and fitting these with functional forms.

Radially, we use only three bins of width 0.6 kpc, centred at $R =$ 7.4, 8, 8.6 kpc. We discard stars beyond 7.1 and 8.9 kpc. Vertically, we use adaptive bin sizes, with smaller bins closer to the disc plane. The bin sizes are chosen so that exactly 400 stars (across all radii) fall into each vertical bin. With $\sim 10^6$ in each mock data set, this gives a few thousand vertical bins. After the radial binning is additionally imposed, the stellar count in each 2D bin varies, but there is still a statistically sufficient number of stars in each bin. Note that we assume mirror symmetry around the disc plane, and so invert z and v_z for each star below the plane and restrict all analysis to positive z.

In each 2D bin, we measure the stellar density ν and the velocity dispersions σ_z^2, σ_R^2, and assign Poisson errors ($\propto 1/\sqrt{N}$) to each measurement. Given these data points, we can fit the functional forms

$$\nu(R, z) = \nu_0\,\mathrm{sech}^2\left(\frac{z}{h}\right)\exp\left(-\frac{R - R_\odot}{d_\nu}\right), \tag{A3}$$

$$\sigma_z^2(R, z) = \sigma_{z,0}^2\exp\left(-\frac{R - R_\odot}{d_{\sigma_z}}\right) + \alpha z, \tag{A4}$$

$$\sigma_R^2(R, z) = \sigma_{R,0}^2\exp\left(-\frac{R - R_\odot}{d_{\sigma_R}}\right) + \beta z. \tag{A5}$$

In all, there are nine free parameters: $\nu_0, h, d_\nu, \sigma_{z,0}^2, d_{\sigma_z}, \alpha, \sigma_{R,0}^2, d_{\sigma_R}$, and β. Note these functional forms differ slightly from those adopted by Salomon et al. (2020); we found that these functions gave better fits to our mock data, and ultimately more accurate acceleration measurements (cf. Fig. 6). This is likely just a reflection of the differences between our mock data set and the red clump sample investigated by Salomon et al. (2020).

To fit equations (A3)–(A5) to the measured densities and dispersions, we use an MCMC technique to find the parameters which maximize the Gaussian likelihood of Sivertsson et al. (2018):

$$\mathcal{L} = \mathcal{L}_\nu\mathcal{L}_{\sigma_z}\mathcal{L}_{\sigma_R}, \tag{A6}$$

where $\mathcal{L}_x$ represents the likelihood in quantity x, given by

$$\mathcal{L}_x = \prod_i \exp\left[-\frac{1}{2}\left(\frac{x_{\mathrm{data},i} - x_{\mathrm{model},i}}{x_{\mathrm{error},i}}\right)^2\right], \tag{A7}$$

where the product is over the data points (i.e. the measured values in each 2D bin).

Given the best-fitting parameters, everything on the left-hand side of equation (A1) can be evaluated to give the vertical acceleration.

APPENDIX B: 1D DF-FITTING

This appendix describes the DF-fitting approach we use for comparison in Section 5. It follows the procedure of Widmark & Monari (2019), Widmark (2019), and Widmark et al. (2021a), who use it to measure the dynamical matter density in the solar neighbourhood using *Gaia* data.

Here, the key assumptions are that the DF is separable: $f(\boldsymbol{x}, \boldsymbol{v}) = f_\perp(z, v_z) f_\parallel(x, y, v_x, v_y)$, the vertical energy $E_z \equiv v_z^2/2 + \Phi(z)$ is an integral of motion, and the stellar population comprises three kinematically distinct sub-populations with different velocity dispersions. Under these assumptions, the vertical DF $f_\perp$ can be written as

$$f_\perp(z, v_z) = \sum_{j=1}^{3} \frac{c_j}{(2\pi\sigma_j^2)^{1/2}} \exp\left(-\frac{v_z^2 + 2\Phi(z)}{2\sigma_j^2}\right), \tag{B1}$$

where c_j and σ_j are the relative weights and velocity dispersions of sub-population j. Note that $\sum c_j = 1$ and $c_j \geq 0$.

We assume the underlying matter density takes the parametrized form

$$\rho(z) = \rho_1 \,\mathrm{sech}^2\left(\frac{z}{h_1}\right) + \rho_2 \,\mathrm{sech}^2\left(\frac{z}{h_2}\right) + \rho_3 \,\mathrm{sech}^2\left(\frac{z}{h_3}\right) + \rho_4, \tag{B2}$$

where the heights h_1, h_2, and h_3 are fixed, respectively, at 40, 100, and 300 pc. These heights are slightly different from those used by Widmark et al. (2021a), and we find they give slightly better fits to our mock data. This density model corresponds to a potential

$$\Phi(z) = 4\pi G \sum_{i=1}^{3} \rho_i h_i^2 \ln\cosh\left(\frac{z}{h_i}\right) + 2\pi G \rho_4 z^2. \tag{B3}$$

Equations (B1) and (B3) together specify an analytical DF model that can be fit directly to the data. There are nine free parameters in the model: four density normalizations ρ_i, three velocity dispersions σ_j, and two population weights c_j (not three: the third is fixed by $\sum c_j = 1$).

To obtain the best-fitting parameters, we maximize the likelihood

$$\mathcal{L} = \prod_i \frac{S(z_i) f_\perp(z_i, v_{z,i})}{\iint S(z) f_\perp(z, v_z)\,\mathrm{d}z\,\mathrm{d}v_z}, \tag{B4}$$

where the product is over individual stars and $S(z)$ is the spatial selection function. We keep only stars with $|z| < 2$ kpc, so $S(z) = 1$ if $|z| < 2$ kpc and $S(z) = 0$ otherwise. Integrating the denominator over v_z, this expression simplifies to

$$\mathcal{L} = \prod_i \frac{S(z_i) f_\perp(z_i, v_{z,i})}{\int_{z_{\min}}^{z_{\max}} \mathrm{d}z \sum_j c_j \exp\left(-\frac{\Phi(z)}{\sigma_j^2}\right)}. \tag{B5}$$

We use an MCMC technique to find the parameters which maximize the this likelihood. We find the best results are obtained if the radial range of the data is restricted to [7.8, 8.2] kpc. Once the best-fitting parameters have been found, the vertical accelerations are given by the first derivative of equation (B3),

$$a_z(z) \equiv -\frac{\partial\Phi}{\partial z} = -4\pi G \sum_{i=1}^{3} \rho_i h_i \tanh\left(\frac{z}{h_i}\right) + 4\pi G \rho_4 z. \tag{B6}$$

This paper has been typeset from a TEX/LATEX file prepared by the author.