






The Sloan Digital Sky Survey peculiar velocity catalogue

Cullan Howlett ¹★, Khaled Said ¹, John R. Lucey ², Matthew Colless ³, Fei Qin,⁴
Yan Lai,¹ R. Brent Tully⁵ and Tamara M. Davis ¹

¹*School of Mathematics and Physics, The University of Queensland, Brisbane, QLD 4072, Australia*

²*Centre for Extragalactic Astronomy, Durham University, Durham DH1 3LE, United Kingdom*

³*Research School of Astronomy and Astrophysics, Australian National University, Canberra, ACT 2611, Australia*

⁴*Korea Astronomy and Space Science Institute, Yuseong-gu, Daedeok-daero 776, Daejeon 34055, Korea*

⁵*Institute for Astronomy, 2680 Woodlawn Drive, Honolulu, HI 96822, USA*

Accepted 2022 June 11. Received 2022 June 6; in original form 2021 December 19

ABSTRACT

We present a new catalogue of distances and peculiar velocities (PVs) of 34 059 early-type galaxies derived from fundamental plane (FP) measurements using data from the Sloan Digital Sky Survey (SDSS). This 7016 deg² homogeneous sample comprises the largest set of PVs produced to date and extends the reach of PV surveys up to a redshift limit of $z = 0.1$. Our SDSS-based FP distance measurements have a mean uncertainty of 23 per cent. Alongside the data, we produce an ensemble of 2048 mock galaxy catalogues that reproduce the data selection function, and are used to validate our fitting pipelines and check for systematic errors. We uncover a significant trend between group richness and mean surface brightness within the sample, which may hint at an environmental dependence within the FP or the presence of unresolved systematics, and can result in biased PVs. This is removed by using multiple FP fits as function of group richness, a procedure made tractable through a new analytic derivation for the integral of a three-dimensional (3D) Gaussian over non-trivial limits. Our catalogue is calibrated to the zero-point of the CosmicFlows-III sample with an uncertainty of 0.004 dex (not including cosmic variance or the error within CosmicFlows-III itself), which is validated using independent cross-checks with the predicted zero-point from the 2M++ reconstruction of our local velocity field. Finally, as an example of what is possible with our new catalogue, we obtain preliminary bulk flow measurements up to a depth of 135 h⁻¹Mpc. We find a slightly larger-than-expected bulk flow at high redshift, although this could be caused by the presence of the Shapley supercluster, which lies outside the SDSS PV footprint.

Key words: catalogues – galaxies: distances and redshifts – galaxies: elliptical and lenticular, cD – galaxies: fundamental parameters – galaxies: statistics – cosmology: observations.

1 INTRODUCTION

Galaxies are receding from us due to the expansion of the Universe. The observed relation between galaxy recession velocity and co-moving distance is called the Hubble–Lemaître law. The variation of any galaxy’s observed velocity from its recession is called galaxy peculiar velocity (PV). The main cause of PVs are the gravitational attraction of the growing large-scale structures (LSSs). Hence, robust and accurate measurements of local PVs are essential for inferring the Hubble–Lemaître law and additionally allow for cosmography and precise cosmological studies of gravity in the local Universe.

The PV of a galaxy can be derived if one can independently measure both its distance and redshift. Several distance indicators have been developed that enable the mapping of PVs in the local Universe. Well-known examples include Cepheid variable stars (Leavitt & Pickering 1912), the tip of the red giant branch (Lee, Freedman & Madore 1993), Type Ia supernovae (Phillips 1993), surface brightness fluctuations (Tonry & Schneider 1988), the Tully–Fisher relation (TF) (Tully & Fisher 1977), the Fundamental Plane

(FP) (Djorgovski & Davis 1987; Dressler et al. 1987), and gravitational waves (Holz & Hughes 2005).

Each distance indicator has its own limitations; TF and FP galaxies are relatively abundant and easy to measure, and so far are the only indicators that have been used to derive distances for thousands of galaxies. However, this comes at the cost of large intrinsic scatter in their empirical relationships and so large distance uncertainties.

The current largest individual PV samples are the Cosmicflows-IV Tully–Fisher catalogue (CF4-TF; Kourkchi et al. 2020) containing ~9800 objects, and the FP-based 6-degree Field Galaxy Survey PV sample (6dFGSv; Springob et al. 2014), containing ~8800 objects. In addition to these individual catalogues, the Cosmicflows project (Tully et al. 2008, 2013; Tully, Courtois & Sorce 2016) aims to provide a single comprehensive collection of distance measurements from all of the aforementioned distance indicators. The most recently released iteration, Cosmicflows-III (Tully et al. 2016), contains almost 18 000 galaxies with distance measurements, with progress towards enlarging this substantially (as evident by the recent release of the CF4-TF subsample referenced above). Most previous efforts, including the above cases, have focused on the $z < 0.05$ universe, as this nearby regime is where the FP and TF methods, with large uncertainties that increase with distance, are most useful.

* E-mail: c.howlett@uq.edu.au

PV catalogues have formed the backbone for many science applications over the years. In the 1990s, they were primarily used to constrain Ω_m and linear galaxy bias b (Willick et al. 1997; Sigad et al. 1998). However, as discussed comprehensively in Davis, Nusser & Willick (1996), there were inconsistencies between the velocity fields measured by PV surveys and predicted by redshift surveys; arising from a combination of sparseness in the redshift surveys at high redshift, angular incompleteness in the PV surveys, and potential systematics in the estimation and calibration of the PVs. More recently, the advent of large surveys of nearby galaxies, such as the 2MASS Redshift Survey (Huchra et al. 2012), Sloan Digital Sky Survey (SDSS) (York et al. 2000), 6dF Galaxy Survey (Jones et al. 2004), and the Arecibo Legacy Fast ALFA Survey (Giovanelli et al. 2005), has enabled the creation of large, homogeneous redshift and PV samples. These samples have demonstrated better consistency (e.g. Davis et al. 2011).

Studies using PVs have hence seen a resurgence, including in the areas of cosmography (Springob et al. 2014; Tully et al. 2014; Graziani et al. 2019), measurements of the bulk flow and low-order velocity moments (Watkins, Feldman & Hudson 2009; Feldman, Watkins & Hudson 2010; Nusser & Davis 2011; Ma & Scott 2013; Scrimgeour et al. 2016; Qin et al. 2018, 2019a; Qin 2021; Qin et al. 2021); testing Λ CDM and General Relativity via the velocity correlation function or power spectrum (Johnson et al. 2014; Adams & Blake 2017; Howlett et al. 2017; Huterer et al. 2017; Howlett 2019; Qin, Howlett & Staveley-Smith 2019b; Adams & Blake 2020); and fitting cosmological parameters and the external tidal field using reconstructions of the velocity field from galaxy redshifts (Carrick et al. (Carrick et al. 2015; Said et al. 2020; Boruah, Hudson & Lavaux 2020b; Lilow & Nusser 2021; Stahl et al. 2021). Furthermore, in the era of the Hubble tension (e.g. Verde, Treu & Riess 2019), the importance of robust and accurate PV measurements for correcting low-redshift distance measurements from Type Ia supernovae and gravitational waves has come to the forefront (Scolnic et al. 2014; Guidorzi et al. 2017; Howlett & Davis 2020; Boruah, Hudson & Lavaux 2020a).

In this paper, we capitalize on these previous efforts, particularly 6dFGSv and its predecessors EFAR (Wegner et al. 1996), SMAC (Hudson et al. 1999), and ENEAR (da Costa et al. 2000), to provide distances and PVs for more than 30 000 early-type galaxies using the FP relation and data from SDSS. This catalogue is $\sim 3 \times$ larger than either the 6dFGSv or the CF4-TF catalogue, and also larger than the full ensemble of distances in Cosmicflows-III. The size of this catalogue is in large part due to our inclusion of galaxies up to $z = 0.1$, which extends the reach of our new measurements beyond those typically produced with the FP or TF relationships and into a region of the Universe that will likely be of increasing interest over the coming years. Compared to previous measurements, our new catalogue is limited to a relatively small sky area ($\sim 7000 \text{ deg}^2$), but has a substantially higher number density than other catalogues in the same redshift regime.

As a by-product of this work (specifically, of testing our pipeline for converting the SDSS measurements to PVs), we also provide a suite of 2048 highly realistic and well-calibrated simulations of the SDSS PV catalogue. In combination with the data, these add significant value for future uses of this work, including for cosmological measurements, characterizing the local velocity field; and in understanding potential sources of statistical and systematic errors. The PV catalogue, input data, simulations, and associated data products are all publicly available (see Data Availability).

This paper is organized as follows: We describe the data in Section 2. In Section 3 we present the mock catalogues. Fitting the FP parameters is presented in Section 4, while fitting the distances is

Table 1. Selection criteria applied to SDSS data to create the SDSS PV catalogue. Each row summarizes a different selection criterion, references the section of the text where it is described, and gives the number of remaining galaxies in the sample *after* this selection has been applied.

Selection	Ref.	# remaining
GALAXY with ZWARN = 0	Section 2.1(i)	403 789
Magnitude in range $10.0 \leq m_r \leq 17.0$	Section 2.1(ii)	287 974
Redshift range $0.0033 \leq z \leq 0.1$	Section 2.1(iii)	242 419
de Vaucouleurs profile	Section 2.1(iv)	124 050
Concentration index $r_{90}/r_{50} > 2.5$	Section 2.1(v)	109 614
Axial ratio $b/a > 0.3$	Section 2.1(vi)	102 747
Within the contiguous NGC area	Section 2.1(vii)	87 002
$H\alpha$ EW $< 1 \text{ \AA}$	Section 2.1(viii)	45 716
$g - r$ colour cut	Section 2.1(ix)	43 226
Velocity dispersion cut	Section 2.1(x)	42 170
No spirals or visual inspection rejects	Section 2.4	34 562
No FP outliers	Section 4.2	34 059

presented in Section 5. In Section 6, we provide an example use of the data and simulations by measuring the bulk flow of the catalogue. We conclude in Section 7. Finally, the Appendices provide information on a useful straight-line fitting package we have created for data with errors on x and y , followed by some mathematical results simplifying the application of a 3D Gaussian model for the FP. Unless otherwise stated, in this paper, we assume a flat Λ CDM cosmological model with $\Omega_m = 0.31$ and $H_0 = 100 \text{ h km s}^{-1} \text{ Mpc}^{-1}$. All uses of ‘log’ should be taken to mean logarithms taken to the base 10.

2 SDSS DATA

2.1 Primary selection criteria

The SDSS PV catalogue presented here is based on FP data presented in Said et al. (2020), which was in turn extracted from imaging and spectra provided with the SDSS Data Release 14 (DR14; Abolfathi et al. 2018).

As well as the baseline selection imposed on the SDSS data (which is complete for extended sources at r -band Petrosian magnitudes less than 17.7), we apply a number of additional selection criteria. These again align closely with Said et al. (2020), although there are some differences, and so our full set of criteria are described below. These criteria are designed to isolate dispersion-supported, early-type galaxies with no evidence of recent star formation and robustly measured $z < 0.1$ redshifts. We also make use of existing $H\alpha$ measurements and velocity dispersions from the Portsmouth groups DR8 and DR12 catalogues (Thomas et al. 2013).¹ By cross-matching with the Portsmouth DR8 and DR12 catalogues we have stellar velocity dispersions σ , uncorrected for fibre aperture effects. The numbers of galaxies remaining after each successive selection criterion are summarized in Table 1. Our selection criteria are as follows:

- (i) objects spectroscopically classified as GALAXY with redshift warning flag ZWARN = 0 (i.e. no known problems);
- (ii) de Vaucouleurs magnitude in the SDSS r band in the range $10.0 \leq m_r \leq 17.0$;
- (iii) CMB-frame redshift range $0.0033 \leq z \leq 0.1$;
- (iv) likelihood of the surface brightness profile fit with the de Vaucouleurs model is higher than with the exponential model in both i and r bands;
- (v) concentration index r_{90}/r_{50} in i and r bands greater than 2.5;

¹ Available here: https://www.sdss.org/dr12/spectro/galaxy_portsmouth/

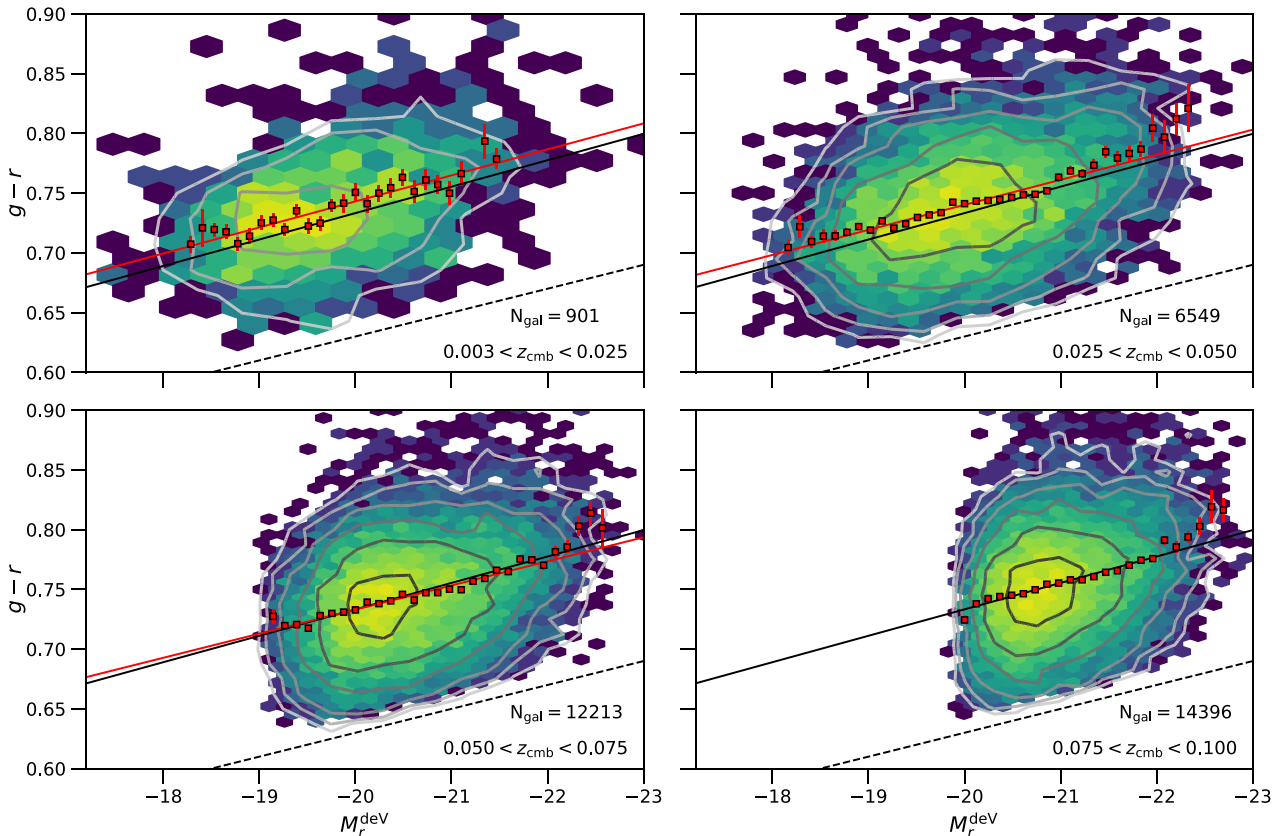


Figure 1. Extinction and k -corrected $g - r$ colour versus r -band absolute magnitude for the SDSS PV catalogue in four different redshift bins. Hexbins and contours show the density of galaxies in the colour–magnitude space, while red points are averages in bins of absolute magnitude. Red lines show the fits to these points in the lowest three redshift bins, while the black line is the fit to the highest redshift bin, replicated in each of the panels. The dashed-line shows the colour-cut we apply to isolate red objects for our catalogue. The typical galaxy colours are extremely uniform across the entire redshift range of our sample, with a comparable slope between the red and black lines in all redshift bins and a maximum difference of ~ 0.01 mag. Also of note is the clear impact of our apparent magnitude limit at higher redshifts, a selection effect that is accounted for when fitting the FP and recovering PVs.

- (vi) axial ratio b/a greater than 0.3 (relatively face-on galaxies) in i and r bands;
- (vii) within the SDSS North Galactic Cap contiguous area;
- (viii) $H\alpha$ measurements from the Portsmouth DR8 or DR12 catalogues with equivalent width $< 1\text{\AA}$;
- (ix) $g - r$ colours in the range $0.63 - 0.02(M_r + 20) \leq g - r \leq 1.03 - 0.02(M_r + 20)$ (inspired by Masters, Springob & Huchra 2008); and
- (x) stellar velocity dispersion greater than 70 km s^{-1} (the SDSS spectral resolution limit) and less than 420 km s^{-1} (which removes objects with spurious measurements).

Table 1 shows that the selection criteria that remove the largest proportions of galaxies are the faint magnitude limit, the redshift range, the requirement that a de Vaucouleurs profile is more likely than an exponential profile, and the lack of detectable $H\alpha$. The magnitude limit ensures that the later photometric properties we select on are robust and trustworthy, while the maximum redshift removes objects that, even if they were on the FP, would have such large distance errors and such a high chance of introducing systematic bias (which is exacerbated as the redshift increases) that we do not feel they are worth including. The cuts on profile likelihood and $H\alpha$ are crucial and required to ensure that the galaxies in our sample are a clean and representative set of early-type galaxies that are *expected* to be on the FP. As we will show in the next section, our selection

criteria are largely, but not perfectly, successful in recovering low- z early-type galaxies, and we implement a further visual classification (see Section 2.4) to remove remaining interloping spirals.

We have critically evaluated all the cuts listed above and ensured they are extremely uniform in terms of the redshift distribution of objects they remove, especially the $H\alpha$ EW cut, the $g - r$ colour cut, and the visual inspection described in Section 2.4. An example (the colour–magnitude diagram) is shown in Fig. 1. This is relevant because, in the course of this work, we found that small changes (i.e. 0.05 mag) in the galaxy properties with redshift can cause large (several hundred km s^{-1}) biases in the average PV of objects at the high redshift end of our sample. Previous works, limited to lower redshift, have rightly assumed or demonstrated that such changes are small enough not to cause a bias in their PVs. However, photometric or spectroscopic systematics that would have previously been negligible are no longer so when we approach $z = 0.1$.

2.2 The FP input parameters

Following our primary selection criteria presented in the previous section, we have a number of SDSS photometric and spectroscopic measurements available for each of our galaxies. These include the r -band scale radius, r^{dev} , magnitude m_r^{dev} , and axial ratio (b/a), all obtained from de Vaucouleurs profile fits to the SDSS photometry.

We also have the de Vaucouleurs g -band magnitude, from which we construct $g - r$ colours and r -band Galactic extinctions A_r (Schlafly & Finkbeiner 2011). From the spectroscopy, we have heliocentric redshifts z_{helio} , from which we compute CMB-frame redshifts z_{CMB} using the angular coordinates of the galaxy and the measured CMB dipole from Planck Collaboration I et al. (2020).

We also further cross-match our input data with the Tempel et al. (2011), Tempel et al. (2014), and Tempel et al. (2017) catalogues to obtain group-averaged CMB-frame redshifts z_{group} and morphologies M , where available.² The Tempel groups are constructed using SDSS DR12 data down to a limiting r -band Petrosian magnitude of 17.77 and the robust Friends-of-Friends algorithm (Turner & Gott 1976). Although this is slightly older than the data that we used to construct the PV sample, it is substantially deeper than our magnitude limit, and the $z \leq 0.1$ data in SDSS is almost identical between DR12 and DR14 – only 418 of our galaxies are not found within the Tempel et al. (2017) group catalogues, for which we assume no group membership.

Given all this input data, as well as uncertainties for these parameters, we are then able to compute measured values and uncertainties for the FP, our distance indicator. The FP relation has the form

$$\log R_e = a \log \sigma_0 + b \log I_e + c \quad (1)$$

where R_e is the effective radius (in $\text{h}^{-1} \text{kpc}$) and is derived from two quantities, the angular effective radius θ_e (in arcsec), which can be measured from our photometry, and the distance, which is the desired quantity. The distance-independent quantities in the FP relation are the central velocity dispersion σ_0 (in km s^{-1}) and the mean surface brightness within the angular effective radius, I_e ; these can be measured from spectroscopy and photometry respectively. The coefficients of the FP are represented as a , b , and c . It is common for the FP to be written in shorthand form as $r = as + bi + c$, where $r = \log R_e$, $s = \log \sigma_0$, and $i = \log I_e$, a convention we also adopt in this work.

To start, we convert the de Vaucouleurs scale radius to an angular effective radius in arcseconds θ_e , using

$$\theta_e = r^{\text{deV}} \sqrt{b/a}. \quad (2)$$

From this, we compute the distance-dependent quantity of the FP, the physical effective radius in units of $\text{h}^{-1} \text{kpc}$

$$r_z = \log(\theta_e) + \log(d(z_{\text{group}})) - \log(1 + z_{\text{helio}}) + \log(\pi/648), \quad (3)$$

where $d(z_{\text{group}})$ is the comoving distance in $\text{h}^{-1} \text{Mpc}$ to the group redshift under our assumed fiducial cosmology. We have specifically included the subscript in r_z to highlight that this is the physical effective radius inferred using the observed redshifts rather than the true comoving distance to the galaxy. The last term accounts for the conversion from arcseconds to $\text{h}^{-1} \text{kpc}$, and we have been careful to distinguish between the use of group redshifts for the comoving distance calculation, and heliocentric redshifts for the conversion from comoving to angular diameter distance (Calcino & Davis 2017). In using the Tempel et al. (2017) group redshifts for equation (3), we are neglecting intra-group PVs, which helps reduce non-linearities in the final PV measurements.

The second, distance-independent, FP parameter, the effective surface brightness in units of $L_{\odot} \text{pc}^{-2}$, is computed from the same angular effective radius, along with the apparent magnitude and a

²In doing so, we recompute the group-averaged redshifts ourselves to avoid the additive redshift approximation used in Tempel et al. (2017).

similar unit conversion,

$$i = 0.4(M_{\odot}^r - m_r^{\text{deV}} - 0.85z_{\text{group}} + k_r + A_r) - \log(2\pi\theta_e^2) + 4 \log(1 + z_{\text{helio}}) + 2 \log(64800/\pi). \quad (4)$$

The additional $4 \log(1 + z_{\text{helio}})$ factor accounts for surface brightness dimming, while $0.85z_{\text{group}}$ is an evolution correction following Bernardi et al. (2003a). k_r is the k -correction computed using the heliocentric redshift and $g - r$ colour following Chilingarian, Melchior & Zolotukhin (2010), while $M_{\odot}^r = 4.65$ is the absolute magnitude of the Sun in the SDSS r band (Willmer 2018).

The last of our three parameters, the velocity dispersion in km s^{-1} , is simply derived as $s \equiv \log(\sigma_0)$, where σ_0 is the central velocity dispersion. This is obtained from the measured velocity dispersion using the aperture correction of Jorgensen, Franx & Kjaergaard (1995) such that

$$s = \log(\sigma) - 0.04(\log(\theta_e) - \log(8\theta_{\text{ap}})), \quad (5)$$

where θ_{ap} is the fibre radius used to obtain the galaxy spectrum. $\theta_{\text{ap}} = 1.5$ arcsec for objects with SDSS plate number < 3510 , whereas $\theta_{\text{ap}} = 1.0$ arcsec for objects observed on later plates, as this demarcates the transition from the older 640-fibre SDSS spectrograph to the newer 1000-fibre BOSS spectrograph.³

Equations (3)–(5) form the basis of our input catalogue and are used to fit the FP and, in turn, obtain PVs. An overview of how this is done is covered in the next sub-section, with detailed descriptions given in Sections 4 and 5.

2.3 From FP parameters to PVs

The measured parameters r_z , s , and i form the input for the FP. Because the observed group redshift has been used to obtain r_z , it does not necessarily represent the true intrinsic size of the galaxy, which we denote as r_t – the PV of the group in which the galaxy resides creates a difference between r_z and r_t equal to the log-distance ratio, $r_z - r_t = \eta \equiv \log(d(z_{\text{group}})/d(\bar{z}))$. This follows from a comparison of how the physical size in equation (3) would be computed from the angular size if one were to use either the (known) group or (unknown, but desired) cosmological redshift (Springob et al. 2014).⁴

Hence, the PV of an object is derived from its offset from the best-fitting FP in the r direction – the FP is fit using the ensemble of measured r_z , s , and i and for each galaxy can be used to predict r_t and then η . Using η is convenient because if the FP is treated as Gaussian, and the uncertainties in each of the input parameters are Gaussian, then η is also Gaussian distributed (modulo some small skewness introduced due to selection functions and other complexities in the

³As detailed here: https://www.sdss.org/dr12/spectro/spectro_basics/.

⁴This equation differs slightly from that in Springob et al. (2014), which includes an additional factor $\log(1 + \bar{z}) - \log(1 + z_{\text{group}})$. However, this factor arises due to a common misconception in the conversion from angular diameter distance (which is proportional to the difference in effective radii) to comoving distance. Although r is a physical size, as pointed out in Calcino & Davis (2017), the conversion from comoving distance to luminosity or angular diameter distance always depends on *heliocentric* redshift, even when estimating the comoving distance from a distance indicator. As a result, the additional factor included by Springob et al. (2014) actually vanishes. Fortunately, the effect of incorrectly including this term only produces a relative error in each galaxy's log-distance ratio and PV v_p that is approximately equal to two times its redshift: $\Delta v_p/v_p \approx \ln(10)z/(1+z)$ (see Appendix D). For the 6dFGSv, this means that the resulting errors in the measured PVs or log-distance ratios given by Springob et al. (2014) are generally < 10 per cent of the typical statistical uncertainty.

fitting process). This is the main quantity provided in the SDSS PV catalogue.

From η we can compute the distance modulus,

$$\mu = 5\log(d(z_{\text{CMB}})) - 5\eta + 25, \quad (6)$$

where $d(z_{\text{CMB}})$ is taken to be in units of Mpc. The PV can be computed fully by first numerically inverting the redshift-distance relation to convert from log-distance ratio to cosmological redshift \bar{z} , then using the equation for propagating redshifts to obtain the PV (Davis & Scrimgeour 2014). The downside of this procedure is that the distribution for the PV is no longer Gaussian and so must be expressed as a non-Gaussian PDF (taking into account the Jacobian of the transformation; Johnson et al. 2014), approximated (i.e. Scrimgeour et al. 2016), or corrected for in some other manner (e.g. Hoffman et al. 2021; Qin 2021).

An alternative, which is used in this work whenever a PV is presented, is to use the approximate conversion (Watkins & Feldman 2015)

$$v_p \approx \frac{cz_{\text{mod}}}{1 + z_{\text{mod}}} \ln(10)\eta, \quad (7)$$

and

$$z_{\text{mod}} = z_{\text{CMB}} [1 + 1/2(1 - q_0)z_{\text{CMB}} - 1/6 \times (j_0 - q_0 - 3q_0^2 + 1)z_{\text{CMB}}^2], \quad (8)$$

where c is the speed of light, and q_0 and j_0 are the present day deceleration and jerk parameters, which for our fiducial cosmology have the values $q_0 = -0.535$ and $j_0 = 1$. This estimator retains the same distribution as the original log-distance ratio. It is derived from a Taylor expansion of $\ln(1 - v_p/cz_{\text{CMB}})$ and so is accurate as long as the true PV (not necessarily the measured value) satisfies $v_p \ll cz_{\text{CMB}}$. Given the low-redshift cut applied to the SDSS PV data ($z_{\text{CMB}} > 0.0033$), we expect this approximation to work well – Howlett et al. (2017) demonstrated that this estimator holds even for the much lower redshift 2MASS Tully–Fisher catalogue.

2.4 Contamination by spirals and other interlopers

Before fitting the FP, there are a few caveats and additional steps that must be explored. Our selection of 42 170 galaxies in Section 2.1 is designed to select photometrically clean red E/S0 galaxies with no H α emission. However, this is not sufficient to isolate a clean sample for the FP analysis. Additional steps are needed to remove unsuitable objects that can act as outliers from the FP, resulting in spurious velocity measurements for these objects and a potential bias in the fit to the FP itself.

In order to identify and remove these interlopers, we used the morphological classifications from GalaxyZoo (Willett et al. 2013) Tempel et al. (2011), and Tempel et al. (2014), and we visually inspected all galaxies on 1×1 arcmin colour cutouts extracted from the Pan-STARRS1 (Chambers et al. 2016) and Legacy Survey images (Dey et al. 2019). The deeper images of the Legacy Survey, particularly the model residual images, allow considerably better discrimination than was possible with previous survey images. We identified the following unsuitable objects:

- (i) spiral galaxies that had not been previously recognized in the shallower images;
- (ii) galaxies where the measurement of the FP photometric parameters is likely to be unreliable due to overlapping sources (either stars or other galaxies);
- (iii) galaxies with strong asymmetries;

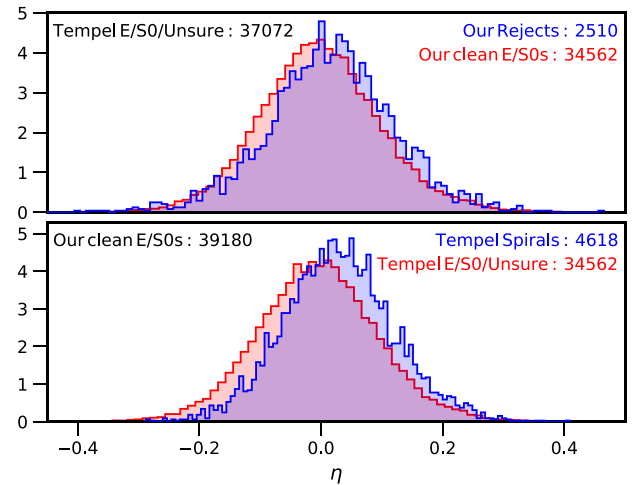


Figure 2. Normalized histograms of the best-fitting log-distance ratios for the SDSS PV sample using different E/S0 classifications. The top panel shows measurements for all galaxies classified by Tempel et al. (2014) as early-type/unsure, and further subdivided into those we keep (red) or reject (blue) after our visual inspection. The bottom panel shows the opposite; galaxies classified by us as photometrically clean early-types split into their Tempel et al. (2014) classification. In both cases there is a remaining subset of interlopers with log-distance ratios offset from the global mean, requiring us to adopt an either/or classification to remove. Note that the red histograms in both panels, despite containing the same numbers of objects, are not quite the same because the FP and log-distance ratios are fit to the full sample of galaxies (red + blue histograms), which differ between the two panels.

(iv) galaxies with strong central dust features, which are likely to bias the velocity dispersion measurements.

While all visual inspection work is subjective, our aim was to identify galaxies that are clearly not suitable to be included in our FP analysis.

Considering first the binary classifications provided by Tempel et al. (2014), we remove 5098 objects, confidently identified as spirals ($M = 1$), fit the FP, and derive log-distance ratios (following the procedures detailed in Sections 4 and 5). We then compare histograms of the log-distance ratios for the remaining objects based on whether they are flagged as E/S0s or rejected by our visual identification. As shown in the top panel of Fig. 2, we find that the remaining 2510 galaxies visually rejected by us pass the Tempel et al. (2014) criteria and exhibit log-distance ratios strongly offset from our clean E/S0 classification. These would hence need to be removed as well.

We then play the reverse game, removing the 2990 galaxies that we have visually rejected before fitting the FP and deriving log-distance ratios. Unfortunately, as shown in the bottom panel of Fig. 2, we then find a remaining 4618 objects that we classify as clean E/S0s but Tempel et al. (2014) classify as spirals that also exhibit biased log-distance ratios. From these results, it is clear that both our classification and the Tempel et al. (2014) classification are picking up different subsets of interlopers, both of which bias our sample. This can be verified by looking at the FP parameters themselves (in Table 2, see also Fig. 7); our rejects are situated primarily at the large radius, low surface brightness end of the FP, whilst Tempel et al. (2014) spirals are brighter and more compact. As such, we adopt a hybrid classification, removing a combined total of 7608 galaxies from our sample if either we or Tempel et al. (2014) have classified the galaxy as a spiral/reject. It is worth noting that we did test using cuts in the GalaxyZoo probability itself for cleaning the sample, but ultimately found that the combination of binary classifications

Table 2. FP parameters for the SDSS PV sample and subsets. Columns $M = 1$ and $J = 0$ are subsets of spiral galaxies identified by Tempel et al. (2014) and us, respectively.

Parameter	Fiducial	$M = 1$	$J = 0$
N_{gal}	34 059	5033	2 931
a	1.274 ± 0.027	1.153	1.296
b	-0.841 ± 0.009	-0.738	-0.821
\bar{r}	0.161 ± 0.016	0.027	0.250
\bar{s}	2.174 ± 0.008	2.154	2.198
\bar{i}	2.688 ± 0.003	2.893	2.638
σ_1	0.0537 ± 0.0006	0.0472	0.0517
σ_2	0.335 ± 0.004	0.315	0.367
σ_3	0.219 ± 0.005	0.191	0.179

removed the contamination from spirals whilst retaining a larger total number of objects. The number of objects remaining after removing spirals/rejects is 34 562.

2.5 Angular Mask

The complete SDSS PV catalogue contains galaxies spread over the contiguous area of the SDSS Northern Galactic Cap. However, the data comes from a combination of various SDSS releases up to and including Data Release 14 (DR14). In order to produce random and mock galaxy catalogues that reproduce the angular distribution of the data, we require an angular mask that describes which regions of sky we expect to be present or missing from the catalogue due to how the data was collected. However, the SDSS PV data is a complex combination of data from the SDSS DR8, plus additional low redshift objects from later data releases. To the best of our knowledge, there are currently no publicly available angular masks describing the distribution of DR8 galaxies across the SDSS footprint that take into account holes arising from centreposts, bright stars, or the tiling geometry of the survey, let alone when newer data up to DR14 is included. All of these are relevant for samples where the clustering may be measured.

As such, rather than tackling the difficult task of tracking back the imaging, targeting, tiling, and veto masks for the SDSS PV catalogue, we instead opt for the simpler solution of identifying the SDSS PV galaxies that belong to a pre-existing, well-defined, angular mask and using only those galaxies for scenarios where the angular distribution may be important (for example in measuring the clustering of the galaxy density field). For this purpose, we use modified versions of the MANGLE (Hamilton & Tegmark 2004; Swanson et al. 2008) polygon files provided alongside the NYU Value-Added Galaxy Catalogue Data Release 7 (DR7; Blanton et al. 2005).⁵ We then identify and flag any galaxies that are outside this mask, and exclude them from clustering measurements and any computation requiring knowledge of the precise sky coverage of the data (such as computing the number density per unit volume). It should be noted that all galaxies are still retained in the SDSS PV catalogue and have measured PVs; those that are within the mask are merely flagged in the data file as IN_MASK. Of the 34 562 early-type galaxies retained in the sample so far, 944 are outside the DR7 footprint; a small fraction that we expect to have little bearing on further analysis of the data. The total number of SDSS PV catalogue galaxies in the mask is thus 33 618 and the total area covered by the angular mask is 7016 deg^2 ; the sky coverage of the SDSS PV

catalogue relative to existing PV data from 6dFGSv and the CF4-TF sample is shown in Galactic coordinates in Fig. 3.

2.6 Random catalogue

In order to measure the clustering of the SDSS PV sample, both in this and in future work, we require a random unclustered sample of data points that can be used to compute the relative overdensity of the galaxies. Given the angular mask, we first use MANGLE to produce a random sample of points that matches the angular distribution of the data. The redshift distribution of the data was then fitted with a smoothed spline to capture the general shape of the survey selection function without incorporating real LSS along the line-of-sight. We then used this smoothed spline to sample redshifts for the random catalogues, and to down-sample the simulated galaxy catalogues presented in Section 3. The number of SDSS PV galaxies as a function of redshift is shown in Fig. 4 alongside the 6dFGSv and CF4-TF data. We can see that the three samples are highly complementary, with the CF4-TF sample peaking at lower redshift than the 6dFGSv or SDSS PV samples. The two FP samples contain a similar number of objects at $z < 0.055$ (although not the same number density; 6dFGSv is spread over more than twice the area of the SDSS PV), but the SDSS data extends to higher redshift, up to $z < 0.1$, where the bulk of our new PV measurements lie. This figure emphasizes that our new data push into a previously unexplored redshift regime, although it is certainly important to point out that the improvements in terms of constraining power from SDSS compared to previous data sets are more modest than the number of galaxies would suggest, as the points at higher redshifts have proportionately larger errors on their PVs (see Sections 5 and 6).

3 MOCK GALAXY SURVEYS

Mock galaxy catalogues (mocks) that reproduce the clustering and selection functions of the PV data are essential for validating the methodology used to extract PVs from a distance indicator in the presence of survey selection effects. They are also a necessity for the interpretation of cosmological results from the data. In this section, we describe the production of 2048 mocks for the SDSS PV data that are designed to reproduce all relevant aspects of the data while encapsulating the effects of cosmic variance on the PV field.

3.1 Characterising the SDSS data

To begin, we first derive a series of simple fitting functions for the properties of the SDSS data that will enable us to add realistic uncertainties and observed quantities to the simulations.

3.1.1 Measurement uncertainties

We start by looking at the errors in the measured FP parameters. We denote the errors on r_z , s , and i as e_r , e_s , and e_i , respectively. They are derived from the errors on r^{deV} , b/a , m_r^{deV} , and σ provided with the publicly available SDSS data by error propagation through equations (2)–(5).

First, as the fainter galaxies in our sample have larger uncertainties in their apparent magnitude, we find [via equation (4)] a strong correlation of the surface brightness uncertainties (e_i) with apparent magnitude. The distribution of e_i is found to be close to lognormal, so we instead quantify the relationship between de Vaucouleurs r -band magnitude and the logarithm of the uncertainty in the surface

⁵ Available here: <http://sdss.physics.nyu.edu/vagc/>

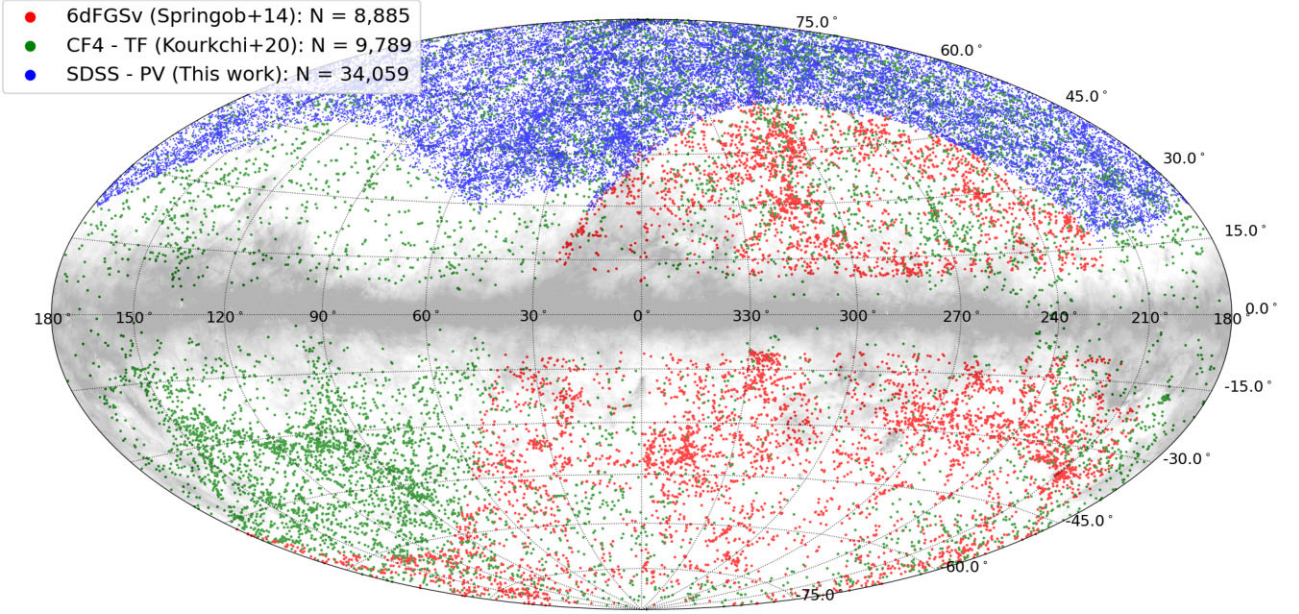


Figure 3. The distribution of the SDSS PV data (blue) in Galactic coordinates compared to data from the 6° Field Galaxy Survey (red; Springob et al. 2014) and Cosmicflows-4 TF (green; Kourkchi et al. 2020). The shaded region shows areas of high galactic extinction around the plane of the Milky Way, grey-scaled by $\log[E(B - V)]$.

brightness, $\log(e_i)$, by binning the data in magnitude bins and fitting the mean and standard deviation in each bin. e_i , being the measurement error, can only be positive and hence be treated in this manner. So, although the form of this quantity may appear strange ($\log(e_i)$ being the log of the error in a quantity that is already logarithmic), we use this purely as it is more Gaussian distributed than the uncertainty itself, and hence allows us to randomly assign errors to the mocks more easily (fitting the e_i as a function of apparent magnitude instead and then drawing from a lognormal distribution would be mathematically equivalent).

We find that the mean is well represented by a piece-wise function that is quadratic for faint galaxies and asymptotes to a constant for galaxies brighter than some limiting magnitude. The Gaussian scatter, denoted $\sigma_{\log(e_i)}$ is fitted well by a straight line. Our best-fitting relationship is given by

$$\langle \log(e_i) \rangle = \begin{cases} -2.35 & m_r^{\text{deV}} < 12.77 \\ 0.02(m_r^{\text{deV}})^2 - 0.43m_r^{\text{deV}} - 0.13 & \text{otherwise} \end{cases} \quad (9)$$

$$\sigma_{\log(e_i)} = 0.0034m_r^{\text{deV}} + 0.0052. \quad (10)$$

The data, binned mean and scatter, best-fitting relationship, and an example set of random values generated using the above fitting formulae are shown in the left-hand panel of Fig. 5.

By construction, Said et al. (2020) set the uncertainty on the effective radius r to be equal to half the uncertainty on the surface brightness. They also set the correlation coefficient between r and i to be -1.0 , i.e. perfectly anticorrelated given they are produced using the same data, but have the opposite dependence on angular effective radius. We adopt the same procedure here, so $e_r = 0.5e_i$.

The only remaining measurement uncertainty to quantify is the error on the velocity dispersion e_s . When compared to the properties of the input data, we find that this is correlated strongly with spectral S/N – unsurprisingly, velocity dispersions can be measured more accurately in spectra with a high S/N ratio. This in turn introduces a correlation between velocity dispersion and apparent magnitude because brighter objects can reach a higher spectral S/N

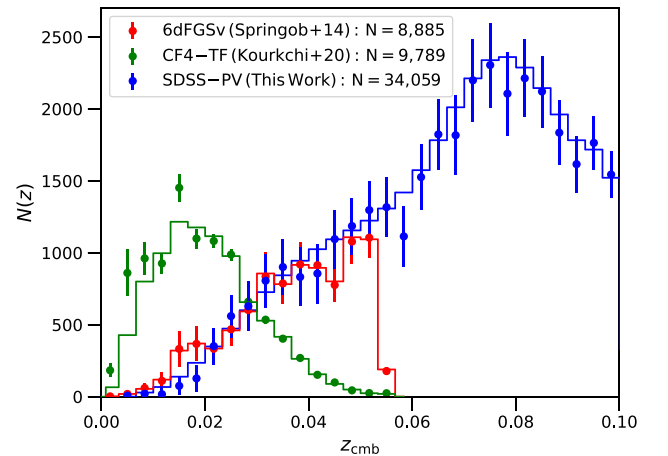


Figure 4. The redshift distribution of the SDSS PV sample compared to the 6-degree Field Galaxy Survey (red; Springob et al. 2014) and Cosmicflows-4 TF (green; Kourkchi et al. 2020). Points show the number of galaxies per bin of width 1000 km s^{-1} , error bars encapsulate the cosmic variance from ensembles of simulated mock surveys (see Section 3 for our SDSS PV mocks and Qin et al. 2019b and Qin et al. 2021 for 6dFGSv and CF4-TF mocks respectively), and lines are the mean distributions from these mocks.

in fixed observing time. However, for the purposes of generating the simulations, we do not have spectral S/N ratios from which we could draw velocity dispersion uncertainties. Furthermore, we opt *not* to use apparent magnitude for this, as to do so would introduce correlations between the effective radius/surface brightness and velocity dispersion that are observational rather than intrinsic in nature. In fitting the FP in Section 4, we treat the correlation between photometric and spectroscopic observational errors as zero, and using apparent magnitude to generate mock velocity dispersion errors would run counter to this.

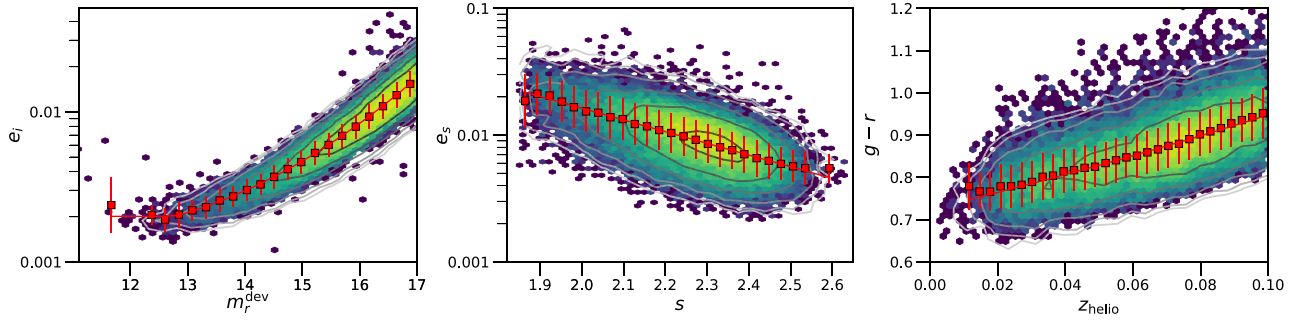


Figure 5. Characterization of the measurement errors in surface brightness and velocity dispersion (e_i and e_s , respectively) and $g - r$ colours in the SDSS PV sample. In each panel, the real data are shown as a hex-binned distribution colour-coded by the number of galaxies in each bin. The contours show a random distribution of points drawn using the x -axis values of the real data (de Vaucouleurs r -band magnitude, velocity dispersion, and heliocentric redshift, respectively, from left-hand to right-hand panel) and the fitting formulae provided in Section 3.1. The points in each panel show the mean and scatter to which these formulae were fit, and the red lines show the fit itself.

We instead make use of a third correlation found between the error on the velocity dispersion and the velocity dispersion itself, in the same way as was done in Scrimgeour et al. (2016) and Qin et al. (2018). This correlation arises in combination with those previously identified – larger velocity dispersions can be measured more precisely both because it is easier to fit the absorption features, but also because the galaxies are, by virtue of the FP, brighter and have higher spectral signal-to-noise ratio. This choice has the benefit of relying only on a spectral property, and one which we produce as part of our simulations.

To produce a fit, we follow the same method as before. We again verify that e_s is close to log-normally distributed, and so we work with $\log(e_s)$ and find a good fit using the relation

$$\langle \log(e_s) \rangle = -0.766s - 0.121 \quad (11)$$

$$\sigma_{\log(e_s)} = -0.049s + 0.236. \quad (12)$$

The data, fitted relationship, and an example distribution drawn from this fit are shown in the middle panel of Fig. 5.

3.1.2 Colours and k -corrections

The final aspect of the data we need to characterize are the k -corrections. We aim to reproduce these in our simulations so that we can be sure that these are not impacting our recovered PVs for galaxies close to the edge of the magnitude limit, where a small change in the k -correction can potentially scatter a galaxy in or out of our sample.

k -corrections in the SDSS PV data are obtained using the bivariate fits from Chilingarian et al. (2010) as a function of $g - r$ colour and heliocentric redshift. To replicate these in simulations, we need to assign $g - r$ colours to our mock galaxies based on some other properties that we already have to hand. Looking at the correlations between various aspects of the data, we find that the colour is strongly correlated with redshift. Such a trend can be inferred from Fig. 1 due to the way intrinsically fainter, bluer galaxies fall out of the sample as we go to higher redshifts. We also found a weaker correlation with velocity dispersion, but that was difficult to model well due to large scatter, and not clearly causative because the velocity dispersion itself also increases with redshift. For simplicity in producing the mocks, we therefore just fit the colour as a function of heliocentric redshift, using a lognormal distribution as above. We note that this has the benefit of simply reducing the k -corrections used in the mocks to a

univariate redshift-dependent function. Our best-fitting relationship is given by

$$\langle \log(g - r) \rangle = 1.167z_{\text{helio}} - 0.140 \quad (13)$$

$$\sigma_{\log(g-r)} = -0.042z_{\text{helio}} + 0.028. \quad (14)$$

The right-hand panel of Fig. 5 shows the colours of the true SDSS PV data as a function of the colour modelled with this fit. The data are scattered normally about the expected one-to-one line and is well reproduced by a random set of points drawn with the scatter given above. Given we can use these formulae to compute colours, we can then combine these with the same redshifts to calculate a k -correction.

Overall, we have been able to effectively reproduce the most important aspects of the SDSS PV data using a small number of fitting formulae. In the next section, these fitting formulas will be used to create a large ensemble of simulations that fully reproduce the data.

3.2 Producing the simulations

Our method for producing the simulations is presented in Qin et al. (2019b). First, an ensemble of 256 dark matter simulations is produced using the approximate N-body code L-PICOLA. Each simulation consists of 2560^3 dark matter particles in a box of edge length $1800 h^{-1}$ Mpc evolved to $z = 0$. Dark matter haloes are then identified in the simulation using the 3D friends-of-friends algorithm (Davis et al. 1985) with a minimum of 20 dark matter particles. This corresponds to a minimum halo mass of $\sim 5 \times 10^{11} h^{-1} M_{\odot}$. From our simulations, we have catalogues of halo positions, velocities, and masses.

Galaxies are placed within these haloes using a variant of sub-halo abundance matching (Conroy, Wechsler & Kravtsov 2006). However, the approximate nature of the L-PICOLA simulations means that only haloes, not sub-haloes, can be reliably identified from the dark matter field using the friends-of-friends algorithm. Instead, we artificially add in sub-haloes by using a power law with two free parameters (A and α) to describe the number of sub-haloes, N_{sub} , as a function of the mass ratio between parent haloes and sub-haloes, f_M ,

$$N_{\text{sub}}(f_M) = A f_M^{-\alpha}. \quad (15)$$

Integrating this function between some minimum value of f_M and 1 (where we set the minimum using a sub-halo mass equivalent to 20

dark matter particles) gives us the *expected* number of sub-haloes in each parent halo.

However, to account for the observed scatter in the sub-halo to halo mass distribution (Giocoli, Tormen & van den Bosch 2008; Elahi et al. 2018), in practice we draw an actual number of sub-halo masses using a Poisson distribution with mean N_{sub} . The sub-haloes are then placed within their parent haloes using an orbital radius and velocity drawn from the NFW profile (Navarro, Frenk & White 1997), using the algorithms/equations in Cole & Lacey (1996), Robotham & Howlett (2018), and mass–concentration relation from Prada et al. (2012).

Finally, galaxies are drawn from the FP distribution with the best-fitting SDSS parameters given in Table 2. Given our set of halo and sub-halo positions, velocities and masses, these are assigned to the haloes and sub-haloes based on rank-ordering the masses and the value of $2r + i$ (which acts as the proxy for luminosity). We add scatter to this rank-ordering based on a third free parameter, $\sigma_{\log L}$, between the log of the masses in units of M_{\odot} , and the log of the luminosities in units of L_{\odot} . Galaxies are given the position and velocity of their host halo/sub-halo.

3.2.1 Incorporating the selection function

After FP parameters have been generated, we apply the selection effects and incorporate measurement uncertainties into the mocks. We place 8 separate observers in each of our 256 simulations spaced maximally far apart ($\sim 600h^{-1}$ Mpc) and for each one:

- (i) Use MANGLE and the angular mask from Section 2.5 to down-sample the mock catalogues to match the footprint of the data.
- (ii) Perturb the effective radius r for each galaxy based on its PV (via the log-distance ratio).
- (iii) Generate $g - r$ colours for the mock galaxies based on the fit to the heliocentric redshift and velocity dispersion s identified in Section 3.1.
- (iv) Generate r band k -corrections for each galaxy using the $g - r$ colour, heliocentric redshift and fitting formulae of Chilingarian et al. (2010).
- (v) Compute the r -band Milky Way extinction for each galaxy based on its location on the sky using the implementation of the Schlafly & Finkbeiner (2011) dust maps in the Python DUSTMAPS package (Green 2018).⁶
- (vi) Combine the FP parameters with the k -correction, Galactic extinction, and luminosity distance to each galaxy to compute the r -band apparent magnitude by inverting equations (3) and (4).
- (vii) Use the magnitude and FP parameters to produce correlated uncertainties on r , s , and i based on the fitting formulae in Section 3.1 and then use these to generate *observed* FP measurements centred on the true values.
- (viii) Apply cuts to the simulated observed redshifts, velocity dispersions and magnitudes, matching those in the SDSS data: $0.0033 < z < 0.1$, $s > \log(70)$ and $m_r^{\text{dev}} < 17.0$.
- (ix) Subsample the observed redshifts of the mock galaxies using the smooth spline fit to the data from Section 2.6. After populating the simulation with galaxies and applying the other selection effects, the number density of galaxies in the catalogues is larger than in the data by an approximately constant factor – the trend as a function of redshift matches the data well due to the inclusion of the magnitude and velocity dispersion cuts, but our model is quite simplistic, is

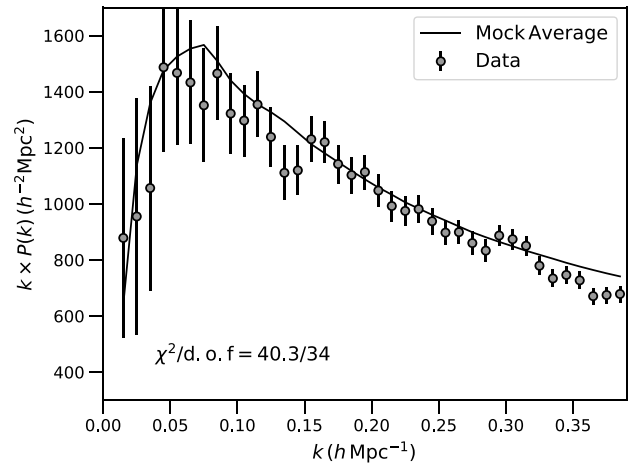


Figure 6. The spherically averaged density power spectrum of the SDSS PV sample (points) compared to the mean of the mocks (line). The error bars come from the variance in the mocks, which captures both cosmic variance and shot noise. The chi-squared value is calculated from the difference between the data and the mock-mean, and shows that the mocks reproduce the clustering in the data well.

fit only to the monopole of the clustering (see Section 3.2.2), and does not account for things like the redshift success rate. All of these effects cause the number density obtained from the previous steps to be larger than in the data, which we remedy by randomly subsampling. Doing this randomly ensures the clustering properties of the mocks remain unchanged by the subsampling.

3.2.2 Tuning the mocks

Given that the relationships between galaxy properties and uncertainties have been characterized in Section 3.1, our selection function is well-defined, and parent halo concentrations are computed based on fits in the literature to high-resolution N-body simulations, our entire procedure contains three remaining free parameters, A , α , and $\sigma_{\log L}$. These are the slope and normalization of the sub-halo mass ratio distribution and the scatter between halo/sub-halo mass and galaxy luminosity.

These parameters are tuned by fitting the monopole of the SDSS PV galaxy density power spectrum. We optimize by brute-force; iteratively populating the halo catalogues, applying the selection function, then computing the galaxy power spectrum and computing the χ^2 relative to the SDSS PV power spectrum. As the covariance matrix for this comparison itself has to be computed from mocks, we repeat this entire optimisation process iteratively; using a first guess for the free parameters to generate the first set of mocks and covariance matrix, which is then subsequently used to tune the next generation of mocks. We perform the entire procedure over four generations, after which the best-fitting parameters do not change considerably between successive generations. In the end, we find best-fitting values of $A = 1.850$ and $\alpha = 1.175$ for the power law amplitude and slope, respectively, and $\sigma_{\log L} = 0.138$ for the scatter. These are comparable to the values found for 6dFGSv (Qin et al. 2019b), but with a lower slope and higher amplitude. This is consistent with our finding that the clustering amplitude (which is mainly set by α) is lower for SDSS than 6dFGSv, but the number density (set mainly by the overall normalization, A) is higher. After iterating, the clustering of the mocks reproduces the data well, as shown in Fig. 6. The χ^2 difference between the power spectrum of

⁶<https://dustmaps.readthedocs.io/en/latest/>

the data and the average of the mocks, computed using the covariance matrix estimated from the mocks, is 40.3 for 34 degrees of freedom.

In the following section, we explain how we fit the FP from the data and the mocks, and demonstrate that the mocks satisfactorily reproduce the expected distribution of the SDSS data.

4 FITTING THE FP

We calculate log-distance ratios and PVs from the SDSS PV FP data using the Maximum Likelihood Gaussian method introduced by Saglia et al. (2001) and Colless et al. (2001) for analysis of the EFAR sample, and as also used for 6dFGSv data (Magoulas et al. 2012; Springob et al. 2014). This is a three-step process, where first the FP itself is fit to the data assuming no PVs, then the offset from the best-fitting plane is used to infer the PV of the individual galaxies. Finally, the zero-point of the FP is calibrated, which is akin to correcting for the fact that the first step of the process assumed an average bulk motion of zero across the *entire* SDSS PV sample, which is unlikely to be true in reality. We note that the procedure can actually all be carried out in a single Bayesian hierarchical model, as done in Dam (2020) and Said et al. (2020). However, this is computationally expensive and so works best when the goal is to evaluate the posterior for a smaller number of derived parameters (such as the growth rate of structure or bulk flow) rather than to produce a catalogue of individual velocities for each galaxy. None the less, this is a clear place for future work to improve on.

For the first two stages, we assume the FP is described by a censored 3D Gaussian, so that the probability of observing N galaxies with effective radii, velocity dispersions and surface brightnesses $\mathbf{x}_n = \{r_n, s_n, i_n\}$ can be written

$$\mathcal{L} = \prod_{n=1}^N \left(\frac{1}{(2\pi)^{3/2} |\mathbf{C}_n|^{1/2} f_n} \times \exp \left[-\frac{1}{2} (\mathbf{x}_n - \bar{\mathbf{x}}) \mathbf{C}_n^{-1} (\mathbf{x}_n - \bar{\mathbf{x}})^T \right] \right)^{1/S_n}, \quad (16)$$

where $\bar{\mathbf{x}} = \{\bar{r}, \bar{s}, \bar{i}\}$ is the mean of the FP and f_n normalizes the likelihood of the observed galaxy n to 1 over the observed parameter space; $f_n < 1$ unless the data is uncensored and contains no selection effects. S_n is an inverse weighting to account for galaxies that are missing from our sample due to the selection function and is based on the commonly used $1/V_{\max}$ weighting (Schmidt 1968). The full likelihood would properly require calculation of the integral of the PDF times the selection probability for each galaxy and FP model. This is computationally expensive, and one can instead use a $1/S_n$ weighting to approximate the full likelihood. As discussed in Eadie, Drijard & James (1971), this results in the same maximum likelihood FP values, but underestimates the variance in the parameters. This is fine for the purposes of our calculation, as we are only interested in the best-fitting values for the FP, and estimate the uncertainties using the variance across our ensemble of mock catalogues, which also incorporate the effects of cosmic variance.

Finally, the covariance matrix \mathbf{C}_n describes the scatter in the FP, which consists of both intrinsic scatter Σ and measurement uncertainty \mathbf{E}_n . Assuming both of these individual components are Gaussian and the measurements are unbiased (so that the inclusion of measurement noise does not bias the measurements away from the mean $\bar{\mathbf{x}}$), we write this as

$$\mathbf{C}_n = \Sigma + \mathbf{E}_n = \begin{pmatrix} \sigma_r^2 & \sigma_{rs} & \sigma_{ri} \\ \sigma_{rs} & \sigma_s^2 & \sigma_{si} \\ \sigma_{ri} & \sigma_{si} & \sigma_i^2 \end{pmatrix} + \begin{pmatrix} \epsilon_r^2 + \epsilon_r^2 & 0 & -e_r e_i \\ 0 & e_s^2 & 0 \\ -e_r e_i & 0 & e_i^2 \end{pmatrix}. \quad (17)$$

The components of the error matrix \mathbf{E}_n are obtained directly from the n^{th} galaxy's measurement uncertainties. We also assume a perfect anticorrelation between r and i (cf. Section 3.1 and Said et al. 2020), no correlation between the photometric and spectroscopically obtained measurements, and add a minimal contribution to the uncertainty on r , $\epsilon_r = \log(1 + 300/cz_n)$, to account for non-linear velocities that may bias the fit when the uncertainty on the effective radius for a given galaxy is small.

Following Saglia et al. (2001) and Magoulas et al. (2012), the scatter matrix Σ is decomposed into orthogonal unit eigenvectors $\hat{\mathbf{v}}_1, \hat{\mathbf{v}}_2, \hat{\mathbf{v}}_3$. These can be defined in terms of the FP parameters from the relationship in equation (1). From these definitions, the Jacobian can then be used to write the terms in equation (17) as functions of the FP parameters a and b and the intrinsic scatter along each orthogonal direction, σ_1, σ_2 , and σ_3 . The conversions between the two coordinate systems are given in Appendix B. In our fit, we follow Saglia et al. (2001), Colless et al. (2001), and Magoulas et al. (2012), and assume *a priori* that the longest axis of the 3D Gaussian lies exactly in the r - i plane.

The best-fitting FP hence consists of a total of 8 free parameters: $a, b, \bar{r}, \bar{s}, \bar{i}, \sigma_1, \sigma_2$, and σ_3 . We fit these by maximising the log of the likelihood function in equation (16). The maximisation is done using SCIPY's implementation of the differential evolution optimisation algorithm (Storn & Price 1997), which provides robust fits in large, multidimensional parameter spaces without requiring gradients. The fitting algorithm can be made extremely fast (taking less than a minute to find the best fit on a single core) by utilising analytic computation wherever possible. This includes writing the determinant and inverse of the covariance matrix for each galaxy as a function of the matrix components rather than numerically inverting (trivial given this is only a 3×3 matrix), as this enables the calculation over all galaxies to be fully vectorized. The only remaining obstacle is the inclusion of selection effects, which as we will show in the next section can also be accounted for 'analytically' and vectorized using elementary functions.

4.1 Selection effects

The SDSS PV sample has several selection effects that need to be accounted for both when fitting the FP and deriving log-distance ratios. These are:

- (i) lower and upper redshifts limit of $z_{\min} = 0.0033$ and $z_{\max} = 0.1$, respectively;
- (ii) a lower limit on velocity dispersions arising from the instrumental resolution of the SDSS spectrograph $s_{\min} = \log(70)$; and
- (iii) magnitudes limited to the range $10.0 \leq m_r^{\text{deV}} \leq 17.0$.

The likelihood function in equation (16) can account for these selection functions via the normalization f_n and the $1/S_n$ weights. However, which of these are required and how they are computed depends on what exactly is being fit and so changes whether we are fitting the FP or using the fitted FP to extract PVs for each galaxy. In this section we will focus on the former; modelling of the selection function when fitting log-distance ratios will be tackled in Section 5.

4.1.1 Magnitude and redshift limits

Firstly, for a survey with both upper and lower redshift limits and a magnitude limit, like ours, we need to account for the fact that our observed sample is not a complete representation of the underlying FP from which the galaxies are drawn. At certain distances, galaxies

will fall below the magnitude limit of the survey, which cuts a slice through the FP as a function of r and i . To account for this, we upweight the galaxies that we *have* observed using the $1/S_n$ factor. This is computed based on the fraction of the enclosed survey volume in which each galaxy with apparent magnitude $m_{r,n}^{\text{dev}}$ could be observed,

$$S_n = \begin{cases} 1 & z_{\text{lim}} \geq z_{\text{max},n} \\ \frac{d_L^3(z_{\text{lim},n}) - d_L^3(z_{\text{min}})}{d_L^3(z_{\text{max}}) - d_L^3(z_{\text{min}})} & z_{\text{min}} < z_{\text{lim},n} < z_{\text{max}} \\ 0 & z_{\text{lim},n} \leq z_{\text{min}}, \end{cases} \quad (18)$$

where $d_L(z)$ is the luminosity distance to the redshift z and $d_L(z_{\text{lim},n})$ is the limiting distance for each galaxy. The latter is computed, given our r -band magnitude limit, using $d_L(z_{\text{lim},n}) = d_L(z_{\text{group},n}) \times 10^{(17 - m_{r,n}^{\text{dev}})/5}$, where $z_{\text{group},n}$ is the group-averaged redshift to galaxy n . The limiting redshift is then obtained from the limiting distance by inverting the redshift-distance relation. The form of the S_n calculation is such that for galaxies bright enough that the limiting redshift is greater than the maximum redshift, we are complete and all possible galaxies at this magnitude have been included in our sample (modulo the s_{min} cut that will be discussed shortly); hence the weight is 1. For galaxies with limiting redshift below z_{max} , we have only observed, on average, S_n of these galaxies. We account for the impact of the missing galaxies on the FP fit by upweighting, taking the likelihood of the n th galaxy to the power $1/S_n$.

4.1.2 Velocity dispersion limits

As the S_n weighting accounts for both the redshift and magnitude limits of our survey, only the velocity dispersion remains to be dealt with. This is accounted for using the f_n normalization to rescale the PDF of each galaxy based on the volume of the FP parameter space that is not observed. Note that a similar approach could also have been utilized for the magnitude limits (and this is what is done when fitting the log-distance ratios). However, while this is the more principled approach (in a Bayesian sense), it is also far harder to implement than the S_n weighting as it requires knowledge of the magnitudes *and* log-distance ratios of galaxies that by definition have not been observed. This would require simulations for each set of FP parameters enclosed within a Bayesian hierarchical model. Such an approach is not necessary for modelling only the s cut, as the impact of this cut on each galaxy does not depend on the galaxy's cosmological distance. It is also not required for fitting the log-distance ratio for the galaxies that *have* successfully made it into our sample once the FP parameters are fixed.

Given the simplification that we only need to consider the limit imposed by s_{min} , the computation of f_n can be written as a 3D integral over the Gaussian in equation (16). Such an integral can be reduced to a single complementary error function, as demonstrated in Appendix C2, which depends primarily on the value of s_{min} , and only weakly on the properties of each individual galaxy through the error matrix. This dependence is simple to understand; the portion of the FP missing from the observations is related to how far the velocity dispersion cut is from the mean of the sample given the scatter. A sample with a mean far higher than the cut and small scatter will be effectively complete, whereas a sample with cut close to or higher than the mean will be heavily censored. For the SDSS PV sample, the velocity dispersion cut is considerably lower than the sample mean, and the covariance is dominated by the intrinsic scatter in the FP, which is the same for all galaxies, rather than the measurement errors. Consequently, f_n is similar, and close to unity, for most galaxies; 99 per cent of the SDSS PV sample have $f_n \geq 0.998$.

4.2 Outlier Rejection

Using the above methodology, we fit the FP iteratively in order to test the impact of outliers on the fitting procedure. At each iteration, we fit the FP, compute the χ^2 difference between each data point and the best-fitting model, and then compute the p -value for each galaxy given the total log-likelihood of the fit and number of galaxies. Outliers are identified as galaxies with $p < 0.01$ and then excluded from the fit for the next iteration. The reduced χ^2 is originally found to be ~ 0.9 , similar to that found in Magoulas et al. (2012), which indicates that the best-fitting intrinsic scatter is being overestimated to accommodate the outliers, but this quickly converges to 1.005 over several iterations as outliers are removed. We find that the procedure converges after 5 iterations, with 503 outlier galaxies rejected from our sample (these are removed entirely from the PV catalogue), leaving us with a total of 34,059.

4.3 Results

The final best-fitting FP parameters for the SDSS data are given in Table 2. In the same table, we also provide FP fits for galaxies identified as spirals by Tempel et al. (2014) ($M = 1$) and/or rejected by us ($J = 0$). In each case, we adopt our iterative outlier rejection method. These fits demonstrate that the rejected galaxies clearly populate distinct regions of the FP parameter space – as identified in Section 2.4, the Tempel et al. (2014) spirals are typically more compact, being smaller with higher surface brightness, which is reflected in the drastic difference in \bar{r} and \bar{i} compared to our fiducial sample. The objects we reject are the opposite, being substantially larger in extent than our fiducial sample, and with smaller average surface brightness (although in such a way that they seem to shift along the plane in Fig. 7). Although comprising only ~ 20 per cent of the underlying sample, the rejected galaxies are enough to lead to a systematic shift in the FP fit when they are included compared to our fiducial case. This would lead to biased log-distance ratios for the rejects, *and* for the clean early-types if they were left in.

We also apply the same fitting procedure to all of our mock galaxy catalogues. This allows us to test how well our simulations match the distribution of the data, and also estimate uncertainties on our fiducial sample using the standard deviations between mock realisations. The distribution of mock FP parameters is shown in Fig. 7. We see that the mocks match the fit from the data extremely well.

Overall, the key assumption in the above fitting methodology is that the data is well described by a 3D Gaussian. Although we do not plot them here, looking at the individual distributions of r and i provided with the publicly released data file, one can see that this is clearly true. There is some skewness in s , with a slightly elongated tail of low velocity dispersion galaxies. None the less, the small number of outliers removed during our fitting procedure, chi-squared per degree of freedom very close to one and log-likelihood that is well-reproduced by mocks that *are*, by construction, 3D Gaussian distributed, demonstrates that this skewness is not unduly affecting our fits and that the model works well for the SDSS PV data. We do caution, however, that this may not be the case for future, larger data sets.

5 FITTING THE LOG-DISTANCE RATIOS

Given the relationship between log-distance ratio and the difference in effective sizes described in Section 2.3, one can then use the probability distribution of the FP given in equation (16) to fit the log-distance ratio of each galaxy by using the modified set of variables

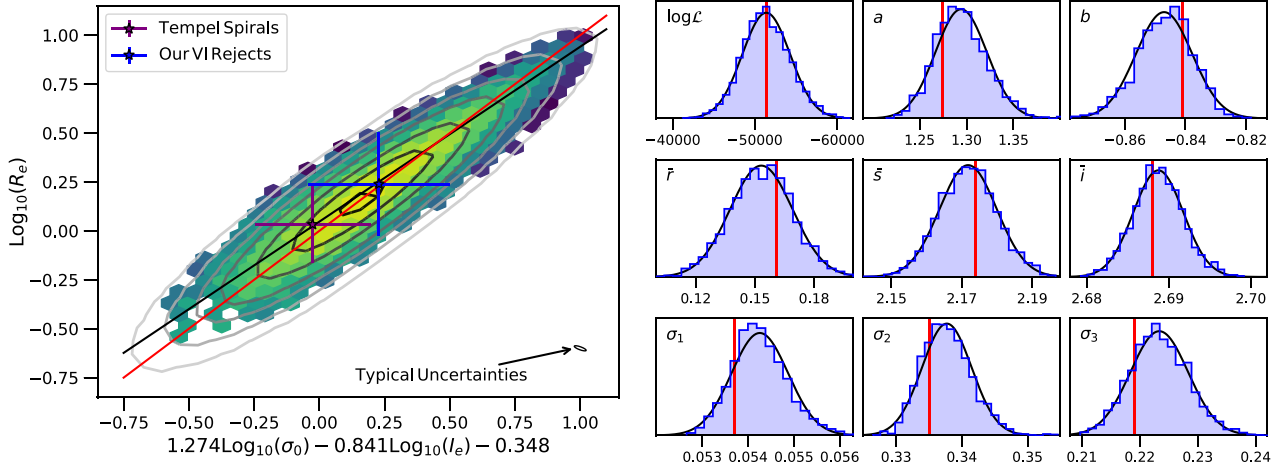


Figure 7. A comparison of the data and mock FP and best-fitting parameters. Left-hand panel: The measured effective radii against the predicted effective radii based on the best-fitting FP parameters for the SDSS data, weighted by the $1/S_n$ factor (equation 18). Coloured bins show the sum of the $1/S_n$ weighted data points, contours show the average distribution of the mocks. The impact of the $1/S_n$ weights is to up-weight galaxies with small effective radius that are generally faint enough they could have been missed at higher redshifts. The red line is the one-to-one line, which alongside the χ^2 of our 3D Gaussian fit (102 704 for 102 169 degrees of freedom), demonstrates excellent agreement between the predicted and measured effective radii. The black line is the fit without including the $1/S_n$ weighting, which clearly demonstrates that the weighting is accounting for faint galaxies missing from our sample (which would have resided in the lower left corner of the FP). The blue and purple points show the mean and variance of the samples rejected by us, or because Tempel et al. (2014) classify them as spirals, respectively. Finally, the small ellipse in the lower right shows the average (correlated) uncertainties for the data. Right-hand panel: The distribution of FP parameters measured from the mocks (blue histograms), overlaid with a Gaussian centred on the mean of mocks and with variance given as the uncertainty on the data in Table 2. The top left-hand sub-panel is the log-likelihood for the best-fitting FP. The vertical red-lines are the best-fitting parameters for the data, which are consistent with the distribution of mock realisations.

$\mathbf{x}_n = \{r_n - \eta_n, s_n, i_n\}$ and fixing $\bar{\mathbf{x}}$ and \mathbf{C}_n based on the best-fitting FP parameters and observational uncertainties for each galaxy.

To ‘fit’ the log-distance ratio, we generate 1001 uniformly distributed candidate values for the log-distance ratio of each galaxy in the range $[-1.5, 1.5]$ and compute the log-likelihood for each. We then combine this with a flat prior on the log-distance ratio and normalize to obtain a finely tabulated posterior PDF for each galaxy, $P(\eta_n | r_n, s_n, i_n, \bar{\mathbf{x}}, \mathbf{C}_n)$. Summary statistics are then produced by assuming the posterior PDF of each galaxy can be represented by a skew-normal distribution (O’Hagan & Leonard 1976; Azzalini 1985) with location (ξ_n), scale (ω_n), and shape (α_n) parameters

$$P(\eta_n | r_n, s_n, i_n, \bar{\mathbf{x}}, \mathbf{C}_n) = \frac{1}{\sqrt{2\pi}\omega_n} \exp\left[-\frac{(\eta_n - \xi_n)^2}{2\omega_n^2}\right] \times \left(1 + \operatorname{erf}\left[\alpha_n \frac{\eta_n - \xi_n}{\sqrt{2}\omega_n}\right]\right), \quad (19)$$

where $\operatorname{erf}(z)$ is the error function. We find that this distribution provides an excellent representation of the SDSS galaxy posteriors, and captures the small skew in the log-distance ratio PDFs that arises from the f_n correction for the selection function described below. Example PDFs and corresponding skew-normal distributions are shown in Fig. 8, where we purposely plot the objects with the largest/smallest skewness and mean, and the largest uncertainty. A similar distribution was used in Springob et al. (2014). The location, scale, and shape parameters for each galaxy can be estimated from the mean ($\langle\eta_n\rangle$), standard deviation (σ_{η_n}), and skewness (γ_n) of the tabulated posteriors using the following relations,

$$\xi_n = \langle\eta_n\rangle - \omega_n \delta_n \sqrt{\frac{2}{\pi}}, \quad \omega_n = \sigma_{\eta_n} \sqrt{\frac{\pi}{\pi - 2\delta_n^2}},$$

$$\alpha_n = \frac{\delta_n}{\sqrt{1 - \delta_n^2}}, \quad |\delta_n| = \sqrt{\frac{\pi|\gamma_n|^{2/3}}{2|\gamma_n|^{2/3} + (\sqrt{2}(4 - \pi))^{2/3}}}. \quad (20)$$

When $\alpha_n = 0$, equation (19) reduces to a normal distribution with the mean and standard deviation used above. Though the α_n values shown in Fig. 8 are consistently non-zero, the skewness of the SDSS galaxies is small enough that it will likely have a negligible effect on subsequent analyses of our catalogue, but we recommend confirming this where possible.

When evaluating the log-likelihood as a function of η there are two terms that complicate matters: a multiplication by $1/S_n$ and the addition of $-\ln(f_n)$. Unlike the case when fitting the FP, we can now ignore the $1/S_n$ weight because it is fixed for a given galaxy; however, we cannot ignore the f_n term.

5.1 f_n correction

f_n describes the normalization of each galaxy’s PDF and mathematically encapsulates the ‘selection bias’ in our sample. Hence, it must be accurately computed for each galaxy and each possible distance to that galaxy. The effect of including f_n is to up-weight larger distances for each galaxy to account for (1) the sampling of the FP being less complete at large distances where fainter galaxies drop below our magnitude limit and (2) observed galaxies being more likely at larger distances than smaller distances because there is simply more volume at larger distances.

Mathematically, f_n is the integral of the Gaussian likelihood, but with limits imposed by our selection function. These are a lower limit for velocity dispersions, $s > \log(70)$, which doesn’t depend on distance and so actually normalizes out of our fitting anyway, and a limited range in magnitude, $10.0 \leq m_r^{\text{dev}} \leq 17.0$. The latter cuts through the FP in a way that can be written in terms of the minimum and maximum effective sizes given a value for the surface brightness. This can be seen by rearranging the relationship between apparent magnitude, effective radius, and surface brightness in equations (3)

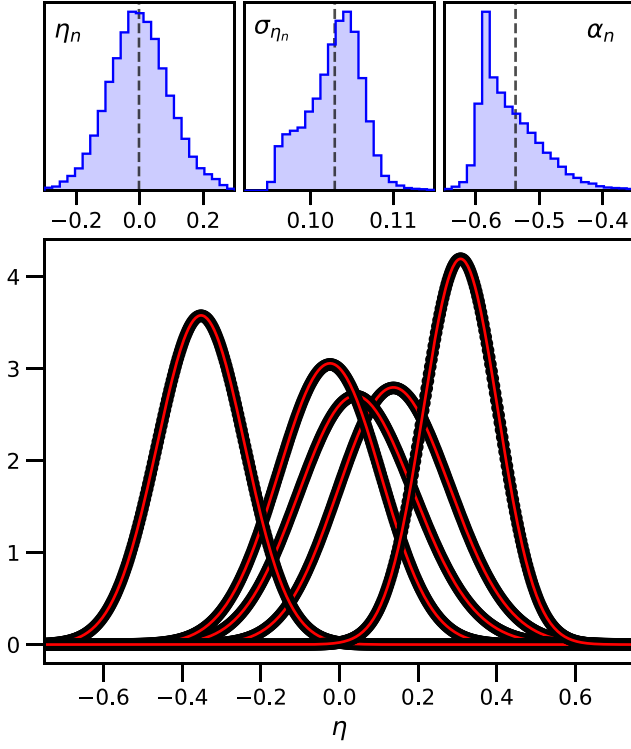


Figure 8. Top panel: The distribution of means, standard deviations, and skewness parameters for the SDSS PV data. Bottom panel: Example normalized probability distributions for the log-distance ratio of galaxies in the SDSS PV sample. We purposely plot the galaxies with the largest/smallest mean and skewness, and the largest uncertainty. In all cases, the skew-normal provides an excellent fit (red line), and the distributions are close to Gaussian. Note that these distributions include the f_n correction described in Section 5.1 and are plotted with $2.5 \times$ fewer points than what is actually used in our fitting.

and (4), and substituting in the magnitude limits. Hence,

$$f_n = \frac{1}{(2\pi)^{3/2} |\mathbf{C}_n|^{1/2}} \int_{-\infty}^{\infty} \int_{r_{\min}^{-i/2}}^{r_{\max}^{-i/2}} \int_{s_{\min}}^{\infty} ds dr di \times \exp \left\{ -\frac{1}{2} (\mathbf{x}_n - \bar{\mathbf{x}}) \mathbf{C}_n^{-1} (\mathbf{x}_n - \bar{\mathbf{x}})^T \right\} \quad (21)$$

where

$$r_{\min} = (10 + M_{\odot}^r + 5 \log(1 + z_{\text{helio}}) - 0.85 z_{\text{group}} - 2.5 \log(2\pi) + k_r + A_r + 5 \log(d(\bar{z})) - 17)/5, \quad (22)$$

$$r_{\max} = (10 + M_{\odot}^r + 5 \log(1 + z_{\text{helio}}) - 0.85 z_{\text{group}} - 2.5 \log(2\pi) + k_r + A_r + 5 \log(d(\bar{z})) - 10)/5, \quad (23)$$

and both the cosmological redshift \bar{z} and the comoving distance to that redshift, $d(\bar{z})$, are computed based on each galaxy's observed redshift and the candidate log-distance ratios. Hence there are actually 1001 values of f_n for each galaxy, varying as the proposed distance to the galaxy increases. The individuality of the f_n for each galaxy in the SDSS sample is also apparent in the fact that the k -correction, extinction, and covariance matrix that enter into the above equations vary from galaxy-to-galaxy.

The large number of galaxies in an FP sample means that evaluating the above integral becomes computationally demanding if one were to use numerical integration. To circumvent this, Springob et al. (2014) used simulations before and after the selection functions

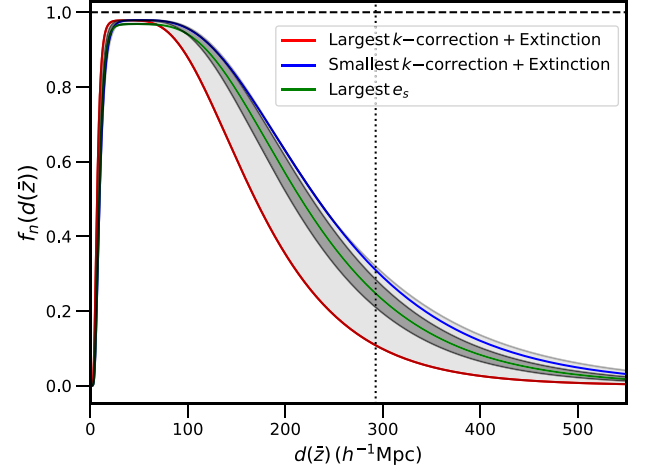


Figure 9. The f_n normalization for each galaxy in the SDSS sample as a function of proposed distance. The vertical dotted line shows the comoving distance to the maximum redshift ($z = 0.1$) of our sample assuming our fiducial cosmology. The light grey area shows the region in which *all* galaxies reside, while the dark grey region contains 95 per cent of the galaxies. In addition, curves for three individual galaxies are shown to highlight the typical shape of the curve and also the sources of variation in the relationship between galaxies. Galaxies with smaller k -corrections and extinctions have f_n relations typically shifted to greater distances compared to those with larger values; if they both had the same effective size and surface brightness, the one with the smaller sum of k -correction and extinction would be observable to a larger distance. In addition, galaxies with larger velocity dispersion uncertainties (e_s) are more likely to have scattered into the sample from below the velocity dispersion cut.

were applied as a form of Monte Carlo integration and assumed the same correction for each galaxy. Besides this assumption, the disadvantage of this method is that the f_n calculation remains quite inaccurate at large distances, where, by construction, the number of mock galaxies with which we can compute f_n quickly goes to zero, even when large samples are generated. In Appendix C, we show that, although complex, the above integral can actually be reduced to a sum of elementary functions that are fast to evaluate using standard computational libraries (such as SCIPY in Python) and give an exact solution. As a result, we are able to compute f_n for every galaxy in the SDSS sample at each of their 1001 candidate distances in less than a minute.

The resulting function is shown in Fig. 9. The characteristic shape of f_n as a function of distance is a peak lying in the range 10–100 h^{-1} Mpc with a value close to unity (but not exactly, due to the s_{\min} cut) that drops towards zero at very small distances and at larger distances due to the bright and faint magnitude limits respectively. For about 95 per cent of galaxies, the curves are very similar, which validates the assumption made by Springob et al. (2014); however, there are some outlying galaxies with more extreme distributions as a function of distance. Broadly speaking, the combination of k -correction and extinction that enters into the conversion from apparent magnitude to effective radius and surface brightness varies the distance scale of the f_n relationship; two galaxies with the same effective radius and surface brightness but different colours or at different proximity to the Galactic plane may not be observable to the same distance. The other trend seen in Fig. 9 is related to the galaxy's velocity dispersion uncertainty. Although all galaxies are assumed to be drawn from the same best-fitting FP and are subject to the same s_{\min} cut, galaxies with larger velocity dispersion uncertainties are

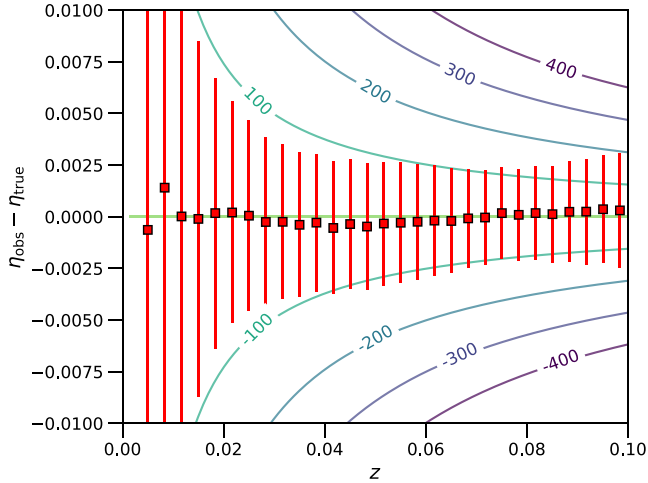


Figure 10. Bias in the measured log-distance ratios from SDSS PV mocks, binned as a function of redshift. Each data point corresponds to the error-weighted mean, while the error bar represents the standard deviation of the mocks in each bin (*not* the standard error of the mean, which would be $\sqrt{2048}$ times smaller). Lines of constant PV are also shown. Our pipeline produces measured log-distance ratios that are unbiased across the entire redshift range of our sample and in excellent agreement with the truth values from the simulations.

more likely to have scattered into the sample erroneously, and so, at its peak, the f_n value is slightly further from unity.

5.2 Tests on mock catalogues and residual bias corrections

Given the method for estimating the log-distance ratio in the presence of selection effects described above, we turn to validating how well we recover the true log-distance ratios in our mocks. We test the extent of residual biases needing to be corrected in the data. This also allows us to assign realistic measured values and errors to the mock data, which will be useful for computing (for instance) the uncertainty on cosmological parameters measured from the SDSS PV sample.

5.2.1 Trends as a function of redshift

We first examine the difference between the true and measured log-distance ratios (η_{true} and η , respectively) in the mocks as a function of redshift. A bias as a function of redshift translates into a spurious inflow/outflow, and even a small offset in log-distance ratio can lead to a large bias in PV at the high redshift end of our data. Fig. 10 shows that our methodology for first fitting the FP, then extracting log-distance ratios, when applied to all our mocks, produces results that are unbiased. There is excellent agreement between the measured and true log-distance ratios averaged over the 2048 simulations, and there is no evidence that our pipeline is introducing spurious flows.

5.2.2 Trends as a function of magnitude

Further investigation into the mock catalogues and data reveals a trend between the absolute magnitude and the recovered log-distance ratios. In both the mock catalogues and the data, intrinsically bright (faint) galaxies have log-distance ratios that are systematically higher (lower) than the mean. This translates into a similar trend as function of apparent magnitude, and is shown in Fig. 11.

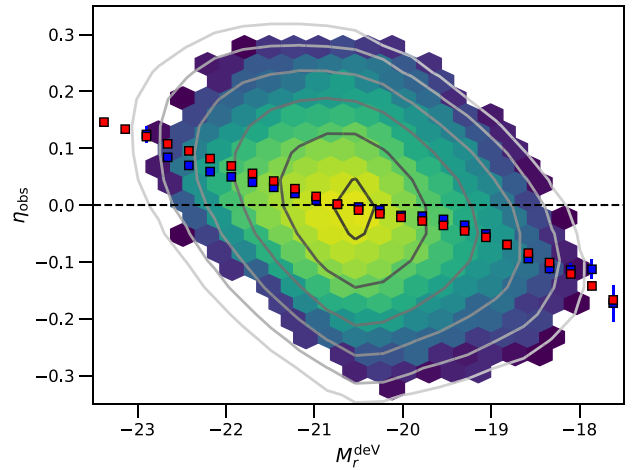


Figure 11. A plot of the trend between absolute r -band magnitude and log-distance ratio seen in our mocks and data. Hexbins show real galaxies, with brighter colours indicating a higher density; contours show mock galaxies, with darker contours indicating a higher density. Red (blue) points are the average log-distance ratio of the mocks (data) in bins of absolute magnitude. The log-distance ratios clearly show a trend with absolute magnitude, but this does not lead to any biases.

Although at first glance, this is a concern, we have verified that this trend is an expected result of identifying the best-fitting log-distance ratio from the offset between r and the 3D Gaussian FP. This can be understood by considering that, from equations (3) and (4), the absolute magnitude of a galaxy $M_r^{\text{dev}} \propto r + 0.5i$. However, the maximum likelihood log-distance ratio is given by the offset from the FP in the r -direction, which from equation (1) (ignoring the f_n term and other complexities) means $\eta \propto r + 0.841i$ for the SDSS PV sample (which has $b = -0.841$; see Table 2). Putting these together clearly shows that we expect $\eta \propto M_r^{\text{dev}}$, albeit with some residual dependence on the surface brightness and velocity dispersion.

This can be seen graphically in Fig. 12, where we plot the SDSS FP data as in Fig. 7, but with each bin colour-coded by the average absolute magnitude. Although subtle, there is a preference for intrinsically brighter/fainter galaxies to be situated above/below the plane, which then results in the brighter/fainter galaxies having positive/negative log-distance ratios, exactly as discussed mathematically above, and seen in Fig. 11. This trend might be diminished if one were to use an alternate method of fitting the FP and extracting log-distance ratios. However, Saglia et al. (2001) explored a number of these alternatives and found that the Maximum likelihood 3D Gaussian we use here gives the most unbiased PVs, despite the apparent trend with absolute magnitude, as it more accurately accounts for the range of sizes, velocity dispersions and surface brightnesses seen in a typical FP sample as well as simultaneous (and potentially correlated) errors in all three parameters.

We did investigate whether this trend could lead to biases in subsequent uses of the SDSS PV catalogue and should be corrected. We concluded that this trend does not result in a bias because: (i) we have already demonstrated that the mocks are unbiased as a function of redshift (Fig. 10), even though there is a trend with absolute magnitude; (ii) our sample is almost volume limited up to even large distances, at $z = 0.05$, corresponding to a comoving distance of $150h^{-1}$ Mpc, the limiting absolute magnitude is only $M_r^{\text{dev}} \leq -18.9$; (iii) our tests of the bulk flow and growth rate measurements obtained from the mocks (presented here in Section 6 and in Lai et al., in prep., respectively) show no significant biases in the recovered

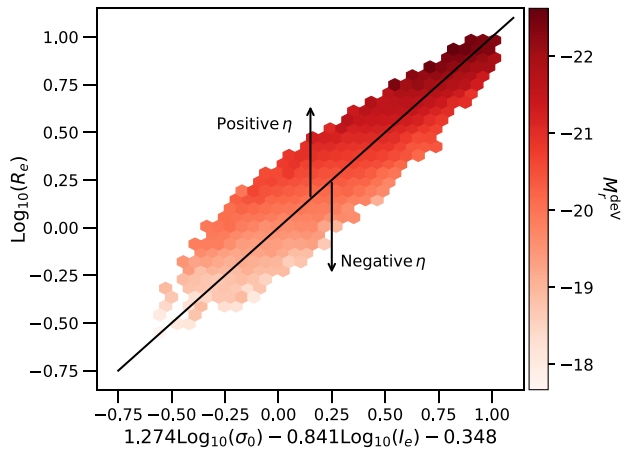


Figure 12. The measured effective radii against the predicted effective radii based on the best-fitting FP parameters for the SDSS data, as per Fig. 7, but with bins coloured according to the mean absolute magnitude. The black line shows the one-to-one line, and the vertical arrows show how the recovered log-distance ratio is related to this projection of the FP. One can see that there is a correlation between absolute magnitude and position above/below the FP (rather than just along it) that inevitably creates the observed trend between recovered log-distance ratio and absolute magnitude.

measurements; and (iv) as discussed in the next section, we do find a bias associated with group richness, that could conceivably have been due to groups containing galaxies that do not span the full range of absolute magnitudes. However, we implemented a correction forcing the log-distance ratio in the data and mocks to be flat as a function of absolute magnitude, and found it had *no* effect on the bias with group richness (which is therefore corrected differently). Hence, we do not ‘correct’ for the trend between log-distance ratio and absolute magnitude seen here.

It is important to note that this could cause bias in subsequent analyses if the data were later cut to a brighter magnitude limit and averaged over because the FP and f_n correction for all galaxies in the SDSS sample have been fit/computed assuming a magnitude limit of 17.0. If the data is cut to a brighter limit, then the correct procedure would be to refit the FP and log-distance ratios using the updated magnitude limit.

5.3 Trends as a function of group richness

In the previous sections, we have demonstrated that we are able to recover log-distance ratios that are unbiased as a function of redshift, and the observed trend with absolute magnitude is as expected and does not need correction. However, we do observe one source of systematic bias in the SDSS sample that does require correction: a correlation between the recovered log-distance ratio and the number of galaxies in the same group. We verified that this is uncorrelated with the trend with absolute magnitude (i.e. due to the fact that groups with more members may contain different distributions of bright or faint galaxies), since forcing the average observed log-distance ratio to be constant as a function of absolute magnitude did not remove the bias with group richness.

The bias is demonstrated in Fig. 13, where we plot the average log-distance ratio over groups of different sizes, where for size we use the number of galaxies (not all of which are in the SDSS PV sample) belonging to the same Tempel et al. (2017) group. We see a clear trend of decreasing log-distance ratio with increasing group richness. This is a problem because larger groups will have more

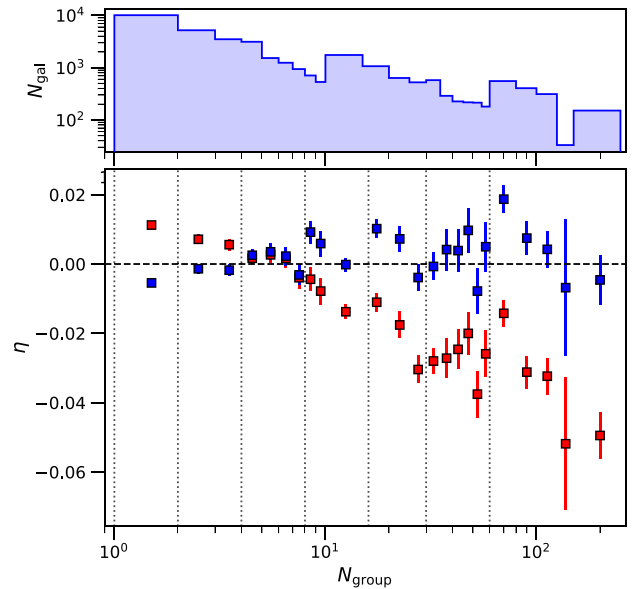


Figure 13. Top panel: Number of SDSS PV galaxies as a function of group richness. Bottom panel: Average log-distance ratio in the SDSS PV catalogue as a function of Tempel et al. (2017) group richness. Red points show the log-distance ratios recovered using our default pipeline where all data is fit using a single FP, and that single set of FP parameters is used to recover log-distance ratios for the full sample. Blue points show the log-distance ratios when separate FPs are fit and used for different subsamples based on group richness as discussed in Section 5.3 and denoted by the vertical dotted lines.

measured PVs, and it is common to average over the PVs within each group, exacerbating this bias.

The existence and origin of environmental dependencies in the FP, either in terms of local properties such as the distance from the galaxy to the cluster centre, or global properties such as cluster richness or radius, is a long-standing question. Previous studies using large samples of galaxies such as those of Bernardi et al. (2003b); D’Onofrio et al. (2008), La Barbera et al. (2010), Magoulas et al. (2012), and Hou & Wang (2015) find correlations between the FP offset/residuals with local surface density, but less evidence of correlations with group richness – although large differences are seen between galaxies in groups (of any size) and in the field. One possibility is that the observed change in the FP with cluster richness is the result of a more elementary correlation between the FP and the stellar age of a galaxy, with richer clusters containing more evolved stellar populations (d’Eugenio et al. 2021).

Another likely possibility is that the trend arises due to data systematics. There are a number of other trends in the data correlated with group richness; including redshift, apparent magnitude, and angular size. However, a substantial fraction of this is to be expected – for a magnitude limited sample, richer groups contain, on average, fainter galaxies, which are found at lower redshifts with larger angular sizes on the sky. This effect should be partially compensated for by our accounting of the selection function when fitting the FP and log-distance ratios, and we see no bias in our log-distance ratios with redshift (which would clearly translate into a bias on group richness if this also varies with redshift). The bias remains if we remove all use of group redshifts in our FP data and fitting, which indicates it is not due to misidentification of, or misassociation with, groups.

It is hence difficult to disentangle what proportion of the observed trends are due to our sample being magnitude limited, data systematics or environmental effects, and we do not address this further in

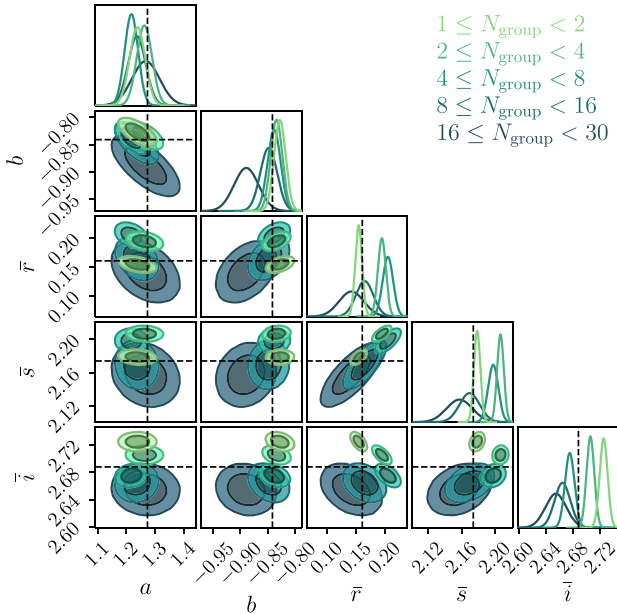


Figure 14. Distributions of FP parameters for subsamples of the SDSS PV catalogue split as a function of group richness. Contours and histograms are derived from 200 simulations centred on the best-fitting parameters from the data, each with the same number of simulated data points as the real data subsamples. These are used to demonstrate the expected spread in FP parameters for each subsample. The black-dashed lines show the FP parameters fit to the full SDSS PV sample. There is considerable scatter in the best-fitting parameters with group richness, but the slope b and mean surface brightness \bar{i} exhibit clear trends.

the current work. None the less, we seek a way to ensure this does not lead to biased PVs, which is most easily achieved by fitting the FP to different subsamples of our full data set based on the group richness.

We do this by splitting our sample into roughly logarithmic bins in group richness, where each subsample contains at least 1,500 galaxies and 70 distinct groups. We then run the separate subsamples through our default fitting methodology, recovering both the best-fitting FP parameters and new log-distance ratios for each galaxy.

The FP parameters for the most constraining sub-samples along with errors derived from 200 simulations centred on the fits from the data are shown in Fig. 14. The simulations are generated as for the single FP sample in Section 3, but without populating an underlying N-body simulation (and so do not have any LSS or clustering). We also do not show the constraints for group sizes larger than 30, as the parameter constraints are too weak to deduce anything meaningful. None the less, there are clear systematic variations of the slope b and mean surface brightness \bar{i} , both of which decrease with increasing group richness. Correlations with other parameters are less clear. The trends with b and \bar{i} seem to be detected at high significance, but for other parameters, the field population does not seem statistically different from the different size groups.

The origin of this remains unclear – the trend identified by La Barbera et al. (2010) is with the parameter combination $c = \bar{r} - a\bar{s} - b\bar{i}$ and not reproduced here, while Magoulas et al. (2012) did not see a strong correlation between b and group richness, but did with local galaxy surface density. However, we find that performing separate fits by group richness and then combining the resulting subsamples leads to log-distance ratios that exhibit far less bias with group richness. This is shown in Fig. 13 as the blue points.

As we are able to remove the bias associated with group richness by simply fitting separate FPs to the subsamples of the data, we use this as our empirical correction going forward. The ‘corrected’ log-distance ratios provided with the SDSS PV catalogue are those obtained using these multiple FP fits rather than a single fit to the entire sample.

5.4 Zero-point

A key assumption when fitting the FP is that the net velocity of the sample is zero. This is unlikely to be true in reality. To correct for this assumption, we need to make one final ‘zero-point’ correction to our sample. In Springob et al. (2014), this was done by assuming that a sample drawn from an approximate great circle (in this case, close to the celestial equator) truly does have net velocity equal to zero. However, this makes use of the hemispherical sky-coverage afforded by 6dFGSv, which is not available with the smaller footprint of the SDSS PV sample. Instead, we calculate the zero-point of the SDSS PV sample by cross-matching to overlapping galaxies that also contain distance measurements in the Cosmicflows-III catalogue (CF3; Tully et al. 2016), using the individual redshifts to convert from distance moduli presented in CF3 to log-distance ratio. CF3 itself is calibrated using a distance ladder containing first galaxies hosting Cepheid variable stars (Freedman et al. 2012), Tip of the Red Giant Branch stars (Rizzi et al. 2007) and maser emission (Humphreys et al. 2013); then Type Ia Supernovae (Rest et al. 2014). By linking the SDSS PV sample to CF3 we are hence extending this distance ladder and relying on the calibration of the intermediate rungs to set our zero-point. This also means that our result for the bulk flow in Section 6 will be strongly correlated with the same measurement from CF3. However, we do validate our zero-point using independent data from the 2M++ reconstruction of the local velocity field.

We perform the calibration in two ways, careful to fairly compare log-distance ratios before and after the correction for group richness in Section 5.3. We first look at the 296 individual galaxies in common between CF3 and the SDSS PV sample, of which 8, 6, 2, and 285 have previous distances from Type Ia supernovae, surface brightness fluctuations, the TF relation, and the FP, respectively.⁷ The CF3 FP-based distances were derived from the EFAR, SMAC, ENEAR, and 6dFGSv surveys. The presence of SDSS PV measurements that also have TF distances is worrying, but both of the above cases are lenticular galaxies where it is unclear from the visual inspection whether the FP or TF relation (or indeed both!) is more appropriate. To be conservative, we remove these from the zero-point calibration, but find that doing so changes the zero-point by only 0.1σ . The remaining 294 overlapping galaxies are distributed across the full redshift range of the SDSS PV sample; however, the calibration is dominated by low redshift objects – 90 per cent have $z < 0.05$.

From the 294 galaxies, we compute the weighted mean difference between CF3 and log-distance ratios *from our single FP fit to the full sample* (i.e. with no correction for group richness) as

$$\langle \eta_{\text{CF3}} - \eta_{\text{SDSS}} \rangle = -0.0028 \pm 0.0080, \quad (24)$$

The second way we compute the zero-point is by using groups that share both a CF3 and SDSS PV measurement. Using the (Tempel et al. 2017) group catalogue, we identify 292 groups that contain at least one CF3 and one SDSS PV measurement. We then first

⁷This adds up to more than 296 galaxies as some galaxies have measurements using multiple tracers. In this case, the individual CF3 measurements are averaged into a single value for the galaxy.

average the measurements within the two catalogues (i.e. if a group contains two CF3 and four SDSS PV measurements, we average the two to obtain a single CF3 consensus value, and the four to obtain a single SDSS PV consensus value). We then compare the difference at the group-averaged level between CF3 and our log-distance ratios obtained from multiple FPs fit to the full sample as a function of group richness (i.e. correcting for the bias in Section 5.3). We find

$$\langle \eta_{\text{CF3}} - \eta_{\text{SDSS}}^{\text{corr}} \rangle = -0.0037 \pm 0.0040. \quad (25)$$

This is fully consistent with the zero-point from individual objects. However, the use of group-averages provides a smaller uncertainty, and so we adopt this as the official zero-point for the SDSS PV sample. As a final check, we predict the velocities for each of the SDSS PV galaxies using the SDSS PV redshifts and reconstructed velocity field of 2M++ (Carrick et al. 2015; processed as in Carr et al. 2021). We then convert these to log-distance ratios and evaluate the zero-point. For the single FP fit and multiple FP fit log-distance ratios, we find

$$\langle \eta_{2\text{M}++} - \eta_{\text{SDSS}} \rangle = -0.0019 \pm 0.0006, \quad (26)$$

$$\langle \eta_{2\text{M}++} - \eta_{\text{SDSS}}^{\text{corr}} \rangle = -0.0013 \pm 0.0005, \quad (27)$$

respectively, which are both also consistent with our other methods of determining the zero-point. We do caution that this last method is not fully model independent – the predicted velocities of Carrick et al. (2015) depend on cosmological parameters – and so this is used only as a cross-check of the empirical zero-point (equation 25) we actually adopt.

A comparison of the log-distance ratios for individual objects in both CF3 and the SDSS PV sample, and distance moduli for groups containing overlapping galaxies from both data sets after applying our zero-point correction is shown in Fig. 15. The best-fitting and 1σ shaded regions are obtained by taking into account the uncertainties in both axes using HYPERFIT (Appendix A). The fit after the correction is consistent with a one-to-one line, with small intrinsic scatter, demonstrating excellent agreement between the two independent sets of measurements. Given the close agreement across a wide range of values, we do not apply a change to the slope in addition to the zero-point offset.

This choice is further justified in Fig. 16, which shows the mean log-distance ratio in bins of redshift. Aside from the presence of LSSs along the line-of-sight, there is no evidence of radial systematics in the data (as would be hoped given the mock validation in Section 5.2.1), and reasonable agreement between the three methods of estimating the log-distance ratio. Note that the reconstruction of Carrick et al. (2015) only extends up to $z = 0.067$, so beyond this the predicted velocity is simply forced to gradually tend to zero and does not include any inhomogeneities that may exist at these higher redshifts.

It is important to note that we do not include the uncertainty on \bar{r} of 0.016 (most of which comes from cosmic variance) from Table 2 in the zero-point error budget in equation (25). As we are comparing multiple measurements for the same objects in this calibration, which are subject to the same cosmic variance, the expected error is hence much smaller than the 0.016 found from our ensemble of mocks. None the less, if one were to compare the log-distance ratios in our catalogue to the prediction from LCDM (wherein cosmic variance must be accounted for), the error in the zero-point that should be considered would be the combination of the zero-point uncertainty from the comparison to CF3 (0.0040) and that due to cosmic variance (0.016). How exactly that is done would depend on the method, as it may be that the cosmic variance contribution is instead incorporated

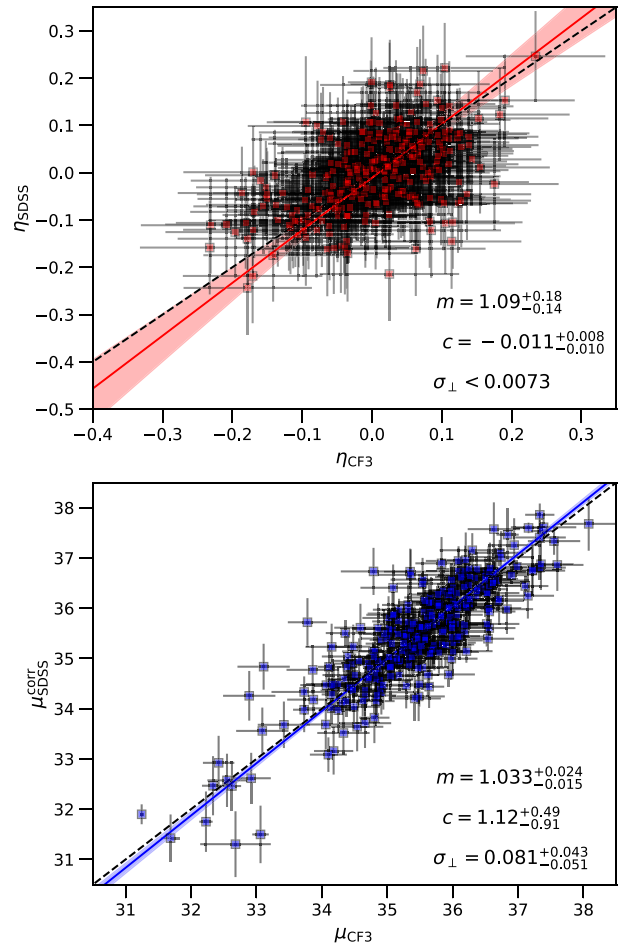


Figure 15. A comparison of log-distance ratios from galaxies in common, and distance moduli from groups in common, between the SDSS PV catalogue and CosmicFlows-III. We use log-distance ratios from single FP fits and multiple FP fits for the upper and lower plots respectively. The line and shaded region show the best-fitting and corresponding 1σ region from a fit to the data after the zero-point correction, while the dashed-line is the expected one-to-one line.

into the theory calculation (as is often done in bulk flow studies, see Section 6), and so should be kept separate to avoid double counting.

We also do not include the error in the zero-point from CF3 (which for $H_0 = 75 \pm 2 \text{ km s}^{-1} \text{ Mpc}^{-1}$ gives an uncertainty on the log-distance ratio of 0.0116). We do so as to enable other choices of zero-point, Hubble Constant, or distance ladder anchor to be made. Any uncertainties in the calibration of the CF3 distances themselves to Cepheids or other local distance anchors should hence be included as an additional error contribution to the 0.004 we quote.

5.5 Summary

In this section, we have provided a thorough explanation for how to extract measurements of the log-distance ratio and PV from a sample of FP galaxies. Our approach includes a new analytic method for accounting for selection bias, which makes it tractable for us to apply the same technique to all our mocks. We have demonstrated that the method produces unbiased log-distance ratios, but that a residual correlation between the group richness and log-distance ratio requires us to fit subsamples with different cluster sizes separately. Our final combined catalogue achieves a mean uncertainty on the log-distance

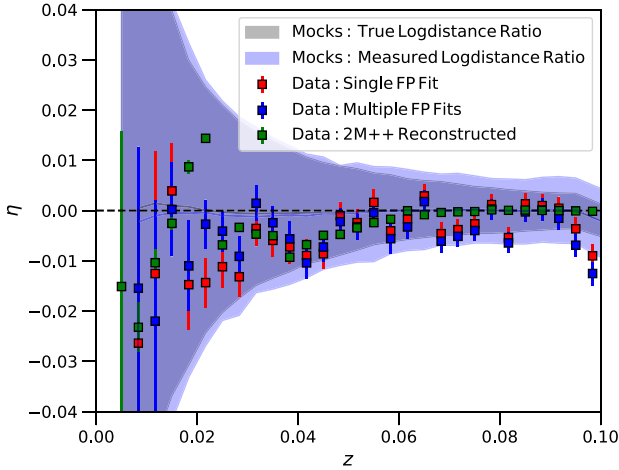


Figure 16. Log-distance ratios in the SDSS PV catalogue (points) and simulations (faint lines/bands), binned as a function of redshift. The grey and blue filled-in bands and corresponding lines show the mean and 95 per cent (2σ) bounds for the simulations using the true and measured log-distance ratios, respectively, which are centred on the horizontal dashed line at $\eta = 0$ as expected. We show log-distance ratio measurements from single and multiple FP fits as a function of group richness, and predicted from the 2M++ reconstruction of Carrick et al. (2015).

ratio of 0.1 dex, which translates to a ~ 23 per cent uncertainty on the distance. This is slightly better than was achieved with 6dFGSv (26 per cent; Springob et al. 2014), and could potentially be improved further with a more detailed understanding of the observed correlation between group richness and the FP parameters. Finally, we tie the zero-point of our sample to the CosmicFlows-III data, demonstrating consistency using both individual objects and objects within the same cluster, and recovering a relative zero-point uncertainty of 0.004 dex (not including cosmic variance or the uncertainty in the CF3 zero-point itself).

6 BULK FLOW

In the last part of this work, we present measurements of the bulk flow from the data and mocks as an example of the analysis that can be performed using our publicly available SDSS PV catalogue and associated simulations. All measurements are performed in Supergalactic Cartesian coordinates.

In Fig. 17, we show measurements of the bulk flow estimated from our 2048 mock catalogues using the Maximum Likelihood method (Kaiser 1988) applied directly to the log-distance ratios as in Qin et al. (2018) and Qin et al. (2019a). The ‘true’ bulk flow is defined simply as the weighted average of the underlying velocities in each direction, where the measurement error is used as the weight to ensure that the two sets of bulk flows are at the same effective depth.

Fig. 17 shows that we recover bulk flow measurements that are, on average, unbiased and well correlated with the true bulk flow in each simulation. However, the error bars do not represent well the scatter between the measured and true values, as can be seen by the very large reduced χ^2 difference between the observed and true values from the mocks in each of the three separate directions included in the figure. A similar result can be seen in Qin et al. (2018, 2021) and other work using the Maximum Likelihood Estimator. Possible reasons for this are a failure of our assumption that the velocity of each galaxy can be represented simply as a bulk flow, without higher order components, and/or that the distribution of each measured velocity can be treated

as an independent Gaussian. However, we leave detailed testing of this hypothesis, and of whether the scatter can be reduced or the error bars made more reasonable, for future work. Instead, when necessary in this work, we simply enlarge the observational errors in each of the three components by a factor equal to $\sqrt{\chi_{\text{red}}^2}$, such that the reduced chi-squared is renormalized to one.⁸

We also find that the uncertainty is smallest and most underestimated in the y -axis. This is interesting because our use of the Supergalactic coordinate system places the y axis almost perfectly along the observer’s line of sight, with x - and z - transverse to this. The SDSS PV survey is a somewhat narrow but long cone (compared to a survey like 6dFGSv) and so the large χ_{red}^2 in this direction indicates that there is likely some systematic in the measurement technique (i.e. neglecting higher-order moments) that becomes more important when we are able to average over a larger volume.

Our results applying the same procedure to the full set of data are shown in Table 3. In Fig. 18, we show the same measurements but cutting the upper redshift limit of the data at seventeen different values of $z_{\text{max}} \in (0.02, 0.10)$ with $\Delta z_{\text{max}} = 0.005$. When making such measurements, we remove any data with log-distance ratios scattered more than 4σ from the mean to ensure objects with outlying PVs do not bias our results. We also correct all the observational uncertainties for the data by the same factor as was found to be necessary to bring the true and measured values for the mocks into statistical agreement. We compute this scaling separately for each value of z_{max} .

In both the table and figure we also provide the theoretical, cosmic-variance expectation of a Λ CDM model with our fiducial/simulation cosmology (given at the end of Section 1) accounting for the actual geometry of the SDSS PV catalogue and the uncertainty on each measurement. This is done using the methods of Feldman et al. (2010) and Ma, Gordon & Feldman (2011), where the theoretical covariance between each PV measurement is computed using

$$\mathcal{W}_{mn} = \frac{\Omega_m^{-1} H_0^2}{2\pi^2} \int f_{mn}(k) P(k) dk, \quad (28)$$

which depends on both the fiducial cosmological model (through the matter power spectrum, $P(k)$, matter density Ω_m and Hubble constant H_0) and the relative location of each pair of galaxies in the SDSS PV sample (through f_{mn} ; which is given in Equation A11 of Ma et al. 2011). For the SDSS PV catalogue \mathcal{W}_{mn} is a $34,059 \times 34,059$ matrix. To reduce this down to the theoretical covariance matrix for the 3 bulk flow components $R_{pq}^{(v)}$, we multiply by the vector of Maximum Likelihood weights for each galaxy n , $R_{pq} = w_{p,n} w_{q,n} \mathcal{W}_{mn}$ (where we adopt the Einstein summation convention), and where

$$w_{p,n} = A_{pj}^{-1} \frac{\hat{x}_{j,n}}{\sigma_{v,n}^2 + \sigma_*^2}; \quad A_{pj} = \frac{\hat{x}_{p,n} \hat{x}_{j,n}}{\sigma_{v,n}^2 + \sigma_*^2}. \quad (29)$$

These weights depend only on the unit vector defining the position of the galaxy with respect to our three bulk flow directions \hat{x} , and the uncertainty on the PV measurement. We convert the errors on the log-distance ratio to those on velocities using the approximation of Watkins & Feldman (2015), $\sigma_{v,n} = \ln(10) c z_{\text{mod}} / (1 + z_{\text{mod}}) \times \sigma_{\eta,n}$,

⁸The reduced χ^2 in Fig. 17 has been computed assuming the three directions are independent. This is not true in practice as they are fitted at the same time from data with only radial PVs. However, we verified that χ_{red}^2 is similar when using the full 3×3 covariance for each mock or just summing the individual χ^2 values for each direction. This is demonstrated further by the fact that after rescaling the uncertainty on each *individual* direction so that the reduced chi-squared for each component is one, the reduced chi-squared accounting for the cross-correlation is also close to one ($\chi_{\text{red}}^2 = 1.08$).

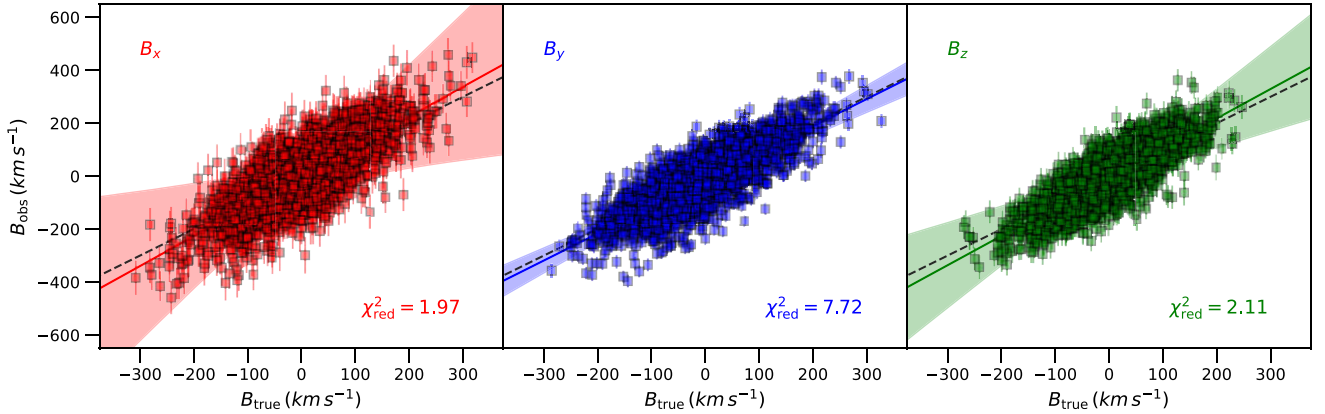


Figure 17. A comparison of the measured and true bulk flow in the Supergalactic x -, y - and z -directions (left-hand, middle, and right-hand columns, respectively) in the SDSS mocks. The solid lines and shaded regions show a linear fit, plus 1σ errors on the fit, where we have assumed the reported errors are accurate. In this case, we recover the expected one-to-one line (dashed black line) to within 1σ . However, the error bars are not fully representative of the scatter seen in the measurements, as can be seen from the reduced chi-squared χ_{red}^2 between the observed and true values from the mocks. This can also be seen in Qin et al. (2018, 2021) and thus requires further study in subsequent work to improve.

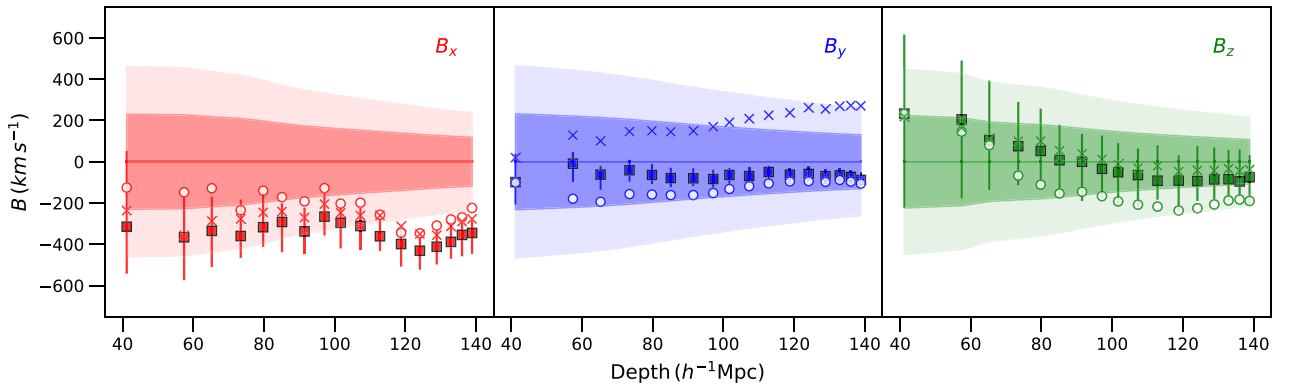


Figure 18. Bulk flow measurements from the SDSS PV catalogue as a function of weighted depth. Each point is computed by cutting the full catalogue to different values of $z_{\text{max}} \in (0.02, 0.10)$ with $\Delta z_{\text{max}} = 0.005$. Circles and squares show measurements using log-distance ratios for the sample with single and multiple FP fits as a function of group richness, respectively. Crosses are measurements for the multiple FP-fitting sample, but where we offset the zero-point by $+0.01$ dex (a 2.5σ shift compared to our zero-point uncertainty), to demonstrate that such errors have only a small effect on the x - and z -directions. Error bars have been corrected for the underestimation seen in the mocks and are of similar size for all three data samples, so, are only plotted for one set for clarity. The shaded regions show the 1 and 2σ expectations for our fiducial Λ CDM model.

and add a small additional contribution to the measurement uncertainties, $\sigma_* = 350 \text{ km s}^{-1}$ to account for non-linearities in the velocity field. In this way, the weights encode the relative contribution of each galaxy in the sample to the overall bulk flow measurement, and so the theoretical prediction also takes into account that galaxies with larger errors will contribute less, and galaxies with radial PVs aligned with, for example, the y -direction will not contribute to the x or z direction bulk flow.

Finally, we compute the chi-squared value for a measured bulk flow \mathbf{B} given our theoretical model using

$$\chi_{BF}^2 = B_p (R_{pq}^{(v)} + \beta_p \beta_q R_{pq}^{(\epsilon)})^{-1} B_q^T \quad (30)$$

where we have included the measurement error for the bulk flow, $R_{pq}^{(\epsilon)}$ scaled by β to account for the underestimation of the errors seen in the mocks (Fig. 17).

Considering both cosmic variance and the measurement error, the probability of obtaining a larger bulk flow in Λ CDM is 20.0 per cent and 10.8 per cent for the single and multiple FP fit samples, respectively – too high to rule out the null hypothesis (that our fiducial

Λ CDM model is correct). However, looking in more detail at the measurements when cutting the SDSS data at different depths, we see that there is a preference for a larger-than-expected bulk flow in the x direction that is persistent when including data above $z = 0.08$ and actually more discrepant with the Λ CDM prediction when the highest redshift data is *not* included. Using only data for $z \leq 0.08$, we find $P(>\chi^2) = 8.1$ per cent and 6.1 per cent for the single and multiple FP fits, respectively, which is closer to the 5 per cent confidence level but still not enough to disfavour Λ CDM with confidence. The measurements using our preferred multiple FP fits as a function of group richness are not quite consistent with those for a single FP fit to the full sample, which is perhaps not surprising given the bias in the ‘single FP’ sample identified in Section 5.3. None the less, we present both to highlight that (regardless of the choice of data) both recover slightly larger than expected bulk flows at large distances, at least in the x -direction.

This is interesting because several other studies have reported larger-than-expected bulk flows at similar depth (Pike & Hudson 2005; Feldman & Watkins 2008; Kashlinsky et al. 2008; Feldman

Table 3. Estimates of the bulk flow from the SDSS data. ‘Data’ columns correspond to bulk flows measured from the SDSS PV catalogue using either a single FP fit to the entire sample, or our preferred method of fitting separate FPs to the sample as a function of group richness to remove the bias demonstrated in Section 5.3. In both cases, the uncertainties have been increased by a constant factor to correct for the underestimation of the observational uncertainties seen in the mocks (Fig. 17). ‘ Λ CDM’ columns correspond to the predictions from our fiducial cosmological model for the SDSS survey geometry. We list the three individual components B_i (for which the expected value is always 0), the bulk flow amplitude $|\mathbf{B}|$, and the weighted depth of the measurements d_{MLE} . The last row gives the probability of recovering a χ^2 difference between the data and Λ CDM that is larger than the value we actually find. For the individual components, the expectation is a Gaussian with a zero-mean and standard deviation $\sim 120 \text{ km s}^{-1}$ and for the bulk flow amplitude (which is Maxwell–Boltzmann distributed), we list the ‘most probable’ value and 68 per cent confidence limits.

	Single FP fit		Multiple FP fits	
	Data	Λ CDM	Data	Λ CDM
B_x (km s^{-1})	-224^{+73}_{-97}	± 118	-345^{+66}_{-101}	± 119
B_y (km s^{-1})	-106^{+39}_{-34}	± 131	-88^{+40}_{-29}	± 131
B_z (km s^{-1})	190^{+140}_{-105}	± 107	-75^{+107}_{-132}	± 107
$ \mathbf{B} $ (km s^{-1})	323^{+82}_{-76}	98^{+50}_{-43}	381^{+85}_{-79}	99^{+51}_{-43}
d_{MLE} ($h^{-1} \text{Mpc}$)	139		139	
$P(>\chi^2)$	20.0 per cent		10.8 per cent	

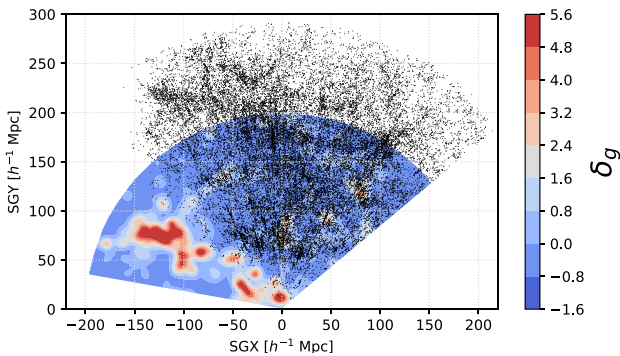


Figure 19. Map of the SDSS PV catalogue (black points) alongside a $-1 h^{-1} \text{Mpc} < \text{SGZ} < 1 h^{-1} \text{Mpc}$ slice of the 2M++ reconstruction of the local density field (Carrick et al. 2015) in Supergalactic coordinates. The large overdensity at $\text{SGX} \approx -125 h^{-1} \text{Mpc}$, $\text{SGY} \approx 75 h^{-1} \text{Mpc}$ is the Shapley supercluster, which provides a possible explanation for our larger-than-expected bulk flow measurement.

et al. 2010; Lavaux et al. 2010). One possible explanation for our result is systematic errors, but we would typically expect the impact of any residual systematics to only become larger when the higher redshift data is included, and, moreover, the choice of coordinates used in this bulk flow analysis places the axis of increasing redshift (which is the one most prone to systematic errors) almost purely in the Cartesian y -direction. It is in this direction that any zero-point calibration errors would be mostly confined – as can be seen in Fig. 18, a change to the zero-point of $+0.01$ dex (a 2.5σ change) causes only a small change in the x - and z -direction bulk flow compared to the y -direction.

As a physical explanation, it is worth noting that the direction of our measured bulk flow aligns well with the position of the Shapley supercluster, as can be seen in Fig. 19 using the reconstructed density field from 2M++ (Carrick et al. 2015). However, it is difficult to say whether the amplitude of our measurement is in agreement with

what would be expected given the gravitational influence from this structure. This hence provides an interesting avenue for further study, either with improved techniques for measuring the bulk flow (such as the Minimum Variance estimator; Watkins et al. 2009), such that we can avoid having to correct the observational uncertainties as done here, or with data at the same or larger depths and over a wider area, which may be possible with upcoming surveys.

7 CONCLUSIONS

In this paper, we present the SDSS PV catalogue, a collection of 34,059 high-quality PV measurements up to $z = 0.1$, subtending an area of 7016 deg^2 . We also provide a detailed analysis of the characteristics of the data, identifying and correcting systematic errors, and find excellent agreement with overlapping measurements from existing surveys. A key finding is the trend between mean surface brightness, slope b , and group richness. Whether this reflects an intrinsic dependence of the FP due to the differing assembly histories in different environments, or is due to unresolved systematics remains unclear. However, for the purposes of the PV catalogue, we account for this by fitting separate FPs to samples of different group richness.

Alongside our data, we make public an ensemble of 2,048 simulated catalogues that almost exactly reproduce the selection function and quality of the data and were run through the same pipeline to enable accurate systematic calibration. As a necessary step towards this, we created improved techniques for fast fitting of the FP and extraction of PVs, which also set the stage for next-generation samples of data that we may expect from upcoming surveys such as DESI (DESI Collaboration et al. 2016) or Southern hemisphere surveys on the 4MOST facility (4HS; PIs Cluver and Taylor). However, we caution that the 3D Gaussian model that we assume here may need to be extended to better incorporate skewness arising from the inclusion of fainter galaxies from these samples, in which case our analytic corrections for the selection functions will also need to be revisited.

In terms of future work, our preliminary tests weakly suggest a bulk flow from the SDSS PV data that is higher than expected from Λ CDM. We have demonstrated that the alignment of this flow implies it is not a result of either our correction for group richness or an inaccurate zero-point calibration, but could be a result of the proximity of our data to the Shapley supercluster. Further work is required to uncover the full origin of this large bulk flow. We hope that the publicly available SDSS PV data and associated data products will provide all the necessary ingredients to do just that, as well as enabling further cosmological and cosmographic analysis of our local Universe.

ACKNOWLEDGEMENTS

CH and KS acknowledge support from the Australian Government through the Australian Research Council’s Laureate Fellowship funding scheme (project FL180100168). JRL acknowledges support from the UK Science and Technology Facilities Council through the Durham Astronomy Consolidated Grants ST/P000541/1 and ST/T000244/1. FQ is supported by the project 우주거대구조를이 용한 암흑우주은구 (‘Understanding Dark Universe Using Large Scale Structure of the Universe’), funded by the Ministry of Science. MC acknowledges support from the Australian Government through the Australian Research Council’s Discovery Projects funding scheme (project DP160102075). This research has made use of NASA’s Astrophysics Data System Bibliographic Services and the astro-ph pre-print archive at <https://arxiv.org/>, the MATPLOTLIB

plotting library (Hunter 2007), and the CHAINCONSUMER and EMCÉE packages (Hinton 2016; Foreman-Mackey et al. 2013). Computations were performed on the OzSTAR national facility at Swinburne University of Technology, which receives funding in part from the Astronomy National Collaborative Research Infrastructure Strategy (NCRIS) allocation provided by the Australian Government, and with the assistance of resources and services from the National Computational Infrastructure (NCI), which is also supported by the Australian Government.

DATA AVAILABILITY

The SDSS PV catalogue and associated data products and simulations are available on Zenodo: <https://zenodo.org/record/6640513>. The catalogue can also be accessed in a modified form with slight additional metadata at the Extragalactic Distance Data base <https://edd.ifa.hawaii.edu/> in the section Summary Distances, in a file called FP: SDSS Distances. Raw SDSS data was obtained from the SDSS Casjobs server. Exact queries used for the SDSS PV data and its supersets in Table 1 will be shared upon reasonable request to the corresponding author, as will all other codes or data.

REFERENCES

- Abolfathi B. et al., 2018, *ApJS*, 235, 42
 Adams C., Blake C., 2017, *MNRAS*, 471, 839
 Adams C., Blake C., 2020, *MNRAS*, 494, 3275
 Azzalini A., 1985, *Scand. J. Stat. Theory Appl.*, 12, 171
 Bernardi M. et al. 2003b, *AJ*, 125, 1866
 Bernardi M. et al., 2003a, *AJ*, 125, 1849
 Blanton M. R. et al., 2005, *AJ*, 129, 2562
 Boruah S. S., Hudson M. J., Lavaux G., 2020a, *MNRAS*, 507, 2697
 Boruah S. S., Hudson M. J., Lavaux G., 2020b, *MNRAS*, 498, 2703
 Calcino J., Davis T., 2017, *J. Cosmology Astropart. Phys.*, 2017, 038
 Carr A., Davis T. M., Scolnic D., Said K., Brout D., Peterson E. R., Kessler R., 2021, preprint ([arXiv:2112.01471](https://arxiv.org/abs/2112.01471))
 Carrick J., Turnbull S. J., Lavaux G., Hudson M. J., 2015, *MNRAS*, 450, 317
 Chambers K. C. et al., 2016, preprint ([arXiv:1612.05560](https://arxiv.org/abs/1612.05560))
 Chilingarian I. V., Melchior A.-L., Zolotukhin I. Y., 2010, *MNRAS*, 405, 1409
 Cole S., Lacey C., 1996, *MNRAS*, 281, 716
 Colless M. et al., 2001, *MNRAS*, 328, 1039
 Conroy C., Wechsler R. H., Kravtsov A. V., 2006, *ApJ*, 647, 201
 d'Eugenio F. et al., 2021, *MNRAS*, 504, 4
 D'Onofrio M. et al., 2008, *ApJ*, 685, 875
 da Costa L. N., Bernardi M., Alonso M. V., Wegner G., Willmer C. N. A., Pellegrini P. S., Rit e C., Maia M. A. G., 2000, *AJ*, 120, 95
 Dam L., 2020, *MNRAS*, 497, 1301
 Davis M., Efstathiou G., Frenk C. S., White S. D. M., 1985, *ApJ*, 292, 371
 Davis M., Nusser A., Masters K. L., Springob C., Huchra J. P., Lemson G., 2011, *MNRAS*, 413, 2906
 Davis M., Nusser A., Willick J. A., 1996, *ApJ*, 473, 22
 Davis T. M., Scrimgeour M. I., 2014, *MNRAS*, 442, 1117
 DESI Collaboration et al., 2016, preprint ([arXiv:1611.00036](https://arxiv.org/abs/1611.00036))
 Dey A. et al., 2019, *AJ*, 157, 168
 Djorgovski S., Davis M., 1987, *ApJ*, 313, 59
 Dressler A., Lynden-Bell D., Burstein D., Davies R. L., Faber S. M., Terlevich R., Wegner G., 1987, *ApJ*, 313, 42
 Eadie W. T., Drijard D., James F. E., 1971, *Statistical methods in experimental physics*. Taylor & Francis, Amsterdam, NL
 Elahi P. J., Welker C., Power C., Lagos C. D. P., Robotham A. S. G., Ca nas R., Poulton R., 2018, *MNRAS*, 475, 5338
 Feldman H. A., Watkins R., 2008, *MNRAS*, 387, 825
 Feldman H. A., Watkins R., Hudson M. J., 2010, *MNRAS*, 407, 2328
 Foreman-Mackey D., Hogg D. W., Lang D., Goodman J., 2013, *PASP*, 125, 306
 Freedman W. L., Madore B. F., Scowcroft V., Burns C., Monson A., Persson S. E., Seibert M., Rigby J., 2012, *ApJ*, 758, 24
 Giocoli C., Tormen G., van den Bosch F. C., 2008, *MNRAS*, 386, 2135
 Giovanelli R. et al., 2005, *AJ*, 130, 2598
 Graziani R., Courtois H. M., Lavaux G., Hoffman Y., Tully R. B., Copin Y., Pomar ede D., 2019, *MNRAS*, 488, 5438
 Green G., 2018, *J. Open Source Softw.*, 3, 695
 Guidorzi C. et al., 2017, *ApJ*, 851, L36
 Hamilton A. J. S., Tegmark M., 2004, *MNRAS*, 349, 115
 Hinton S. R., 2016, *J. Open Source Softw.*, 1, 00045
 Hoffman Y., Nusser A., Valade A., Libeskind N. I., Tully R. B., 2021, *MNRAS*, 505, 3380
 Holz D. E., Hughes S. A., 2005, *ApJ*, 629, 15
 Hou L., Wang Y., 2015, *Res. Astron. Astrophys.*, 15, 651
 Howlett C. et al., 2017, *MNRAS*, 471, 3135
 Howlett C., 2019, *MNRAS*, 487, 5209
 Howlett C., Davis T. M., 2020, *MNRAS*, 492, 3803
 Huchra J. P. et al., 2012, *ApJS*, 199, 26
 Hudson M. J., Smith R. J., Lucey J. R., Schlegel D. J., Davies R. L., 1999, *ApJ*, 512, L79
 Humphreys E. M. L., Reid M. J., Moran J. M., Greenhill L. J., Argon A. L., 2013, *ApJ*, 775, 13
 Hunter J. D., 2007, *Comput. Sci. Eng.*, 9, 90
 Huterer D., Shafer D. L., Scolnic D. M., Schmidt F., 2017, *J. Cosmology Astropart. Phys.*, 2017, 015
 Johnson A. et al., 2014, *MNRAS*, 444, 3926
 Jones D. H. et al., 2004, *MNRAS*, 355, 747
 Jorgensen I., Franx M., Kjaergaard P., 1995, *MNRAS*, 276, 1341
 Kaiser N., 1988, *MNRAS*, 231, 149
 Kashlinsky A., Atrio-Barandela F., Kocevski D., Ebeling H., 2008, *ApJ*, 686, L49
 Kourkchi E. et al., 2020, *ApJ*, 902, 145
 La Barbera F., Lopes P. A. A., de Carvalho R. R., de La Rosa I. G., Berlind A. A., 2010, *MNRAS*, 408, 1361
 Lavaux G., Tully R. B., Mohayaee R., Colombi S., 2010, *ApJ*, 709, 483
 Leavitt H. S., Pickering E. C., 1912, *Harvard College Observatory Circular*, 173, 1
 Lee M. G., Freedman W. L., Madore B. F., 1993, *ApJ*, 417, 553
 Lilow R., Nusser A., 2021, *MNRAS*, 507, 2
 Ma Y.-Z., Gordon C., Feldman H. A., 2011, *Phys. Rev. D*, 83, 103002
 Ma Y.-Z., Scott D., 2013, *MNRAS*, 428, 2017
 Magoulas C. et al., 2012, *MNRAS*, 427, 245
 Masters K. L., Springob C. M., Huchra J. P., 2008, *AJ*, 135, 1738
 Navarro J. F., Frenk C. S., White S. D. M., 1997, *ApJ*, 490, 493
 Nusser A., Davis M., 2011, *ApJ*, 736, 93
 O'Hagan A., Leonard T., 1976, *Biometrika*, 63, 201
 Phillips M. M., 1993, *ApJ*, 413, L105
 Pike R. W., Hudson M. J., 2005, *ApJ*, 635, 11
 Planck Collaboration I et al., 2020, *A&A*, 641, A1
 Prada F., Klypin A. A., Cuesta A. J., Betancort-Rijo J. E., Primack J., 2012, *MNRAS*, 423, 3018
 Qin F., 2021, *Res. Astron. Astrophys.*, 21, 242
 Qin F., Howlett C., Staveley-Smith L., 2019b, *MNRAS*, 487, 5235
 Qin F., Howlett C., Staveley-Smith L., Hong T., 2018, *MNRAS*, 477, 5150
 Qin F., Howlett C., Staveley-Smith L., Hong T., 2019a, *MNRAS*, 482, 1920
 Qin F., Parkinson D., Howlett C., Said K., 2021, *ApJ*, 922, 59
 Rest A. et al., 2014, *ApJ*, 795, 44
 Rizzi L., Tully R. B., Makarov D., Makarova L., Dolphin A. E., Sakai S., Shaya E. J., 2007, *ApJ*, 661, 815
 Robotham A. S. G., Howlett C., 2018, *Res. Notes Am. Astron. Soc.*, 2, 55
 Robotham A. S. G., Obreschkow D., 2015, *PASA*, 32, e033
 Saglia R. P., Colless M., Burstein D., Davies R. L., McMahan R. K., Wegner G., 2001, *MNRAS*, 324, 389
 Said K., Colless M., Magoulas C., Lucey J. R., Hudson M. J., 2020, *MNRAS*, 497, 1275

- Schlafly E. F., Finkbeiner D. P., 2011, *ApJ*, 737, 103
 Schmidt M., 1968, *ApJ*, 151, 393
 Scolnic D. et al., 2014, *ApJ*, 795, 45
 Scrimgeour M. I. et al., 2016, *MNRAS*, 455, 386
 Sigad Y., Eldar A., Dekel A., Strauss M. A., Yahil A., 1998, *ApJ*, 495, 516
 Springob C. M. et al., 2014, *MNRAS*, 445, 2677
 Stahl B. E., de Jaeger T., Boruah S. S., Zheng W., Filippenko A. V., Hudson M. J., 2021, *MNRAS*, 505, 2
 Storn R., Price K., 1997, *J. Glob. Optim.*, 11, 341
 Swanson M. E. C., Tegmark M., Hamilton A. J. S., Hill J. C., 2008, *MNRAS*, 387, 1391
 Tempel E. et al., 2014, *A&A*, 566, A1
 Tempel E., Saar E., Liivamägi L. J., Tamm A., Einasto J., Einasto M., Müller V., 2011, *A&A*, 529, A53
 Tempel E., Tuvikene T., Kipper R., Libeskind N. I., 2017, *A&A*, 602, A100
 Thomas D. et al., 2013, *MNRAS*, 431, 1383
 Tonry J., Schneider D. P., 1988, *AJ*, 96, 807
 Tully R. B. et al., 2013, *AJ*, 146, 86
 Tully R. B., Courtois H. M., Sorce J. G., 2016, *AJ*, 152, 50
 Tully R. B., Courtois H., Hoffman Y., Pomarède D., 2014, *Nature*, 513, 71
 Tully R. B., Fisher J. R., 1977, *A&A*, 54, 661
 Tully R. B., Shaya E. J., Karachentsev I. D., Courtois H. M., Kocevski D. D., Rizzi L., Peel A., 2008, *ApJ*, 676, 184
 Turner E. L., Gott J. R. I., 1976, *ApJS*, 32, 409
 Verde L., Treu T., Riess A. G., 2019, *Nat. Astron.*, 3, 891
 Watkins R., Feldman H. A., 2015, *MNRAS*, 450, 1868
 Watkins R., Feldman H. A., Hudson M. J., 2009, *MNRAS*, 392, 743
 Wegner G., Colless M., Baggle G., Davies R. L., Bertschinger E., Burstein D., McMahan Robert K. J., Saglia R. P., 1996, *ApJS*, 106, 1
 Willett K. W. et al., 2013, *MNRAS*, 435, 2835
 Willick J. A., Strauss M. A., Dekel A., Kolatt T., 1997, *ApJ*, 486, 629
 Willmer C. N. A., 2018, *ApJS*, 236, 47
 York D. G. et al., 2000, *AJ*, 120, 1579

APPENDIX A: HYPERFIT

In the course of this work, we have frequently required a fast, simple method to fit a line or plane to data, allowing for either, or both, intrinsic scatter and (potentially correlated) errors on all the input variables (i.e. in both the ‘x’ and ‘y’ variables for a 2D fit). A general method for such fitting is detailed in Robotham & Obreschkow (2015) and implemented through the associated R-package HYPERFIT. For our purposes, we found it useful to produce a similar package in Python. This package has been fully documented and made ‘pip-installable’. It provides vectorized methods to simply find the best fit given the data or to return a full set of posterior samples for the model given the data. Real, astrophysical, test data is provided with the package, demonstrating that it typically takes only a few seconds to find the best fit or a couple of minutes for a fully converged MCMC run. More details can be found at <https://hyperfit.readthedocs.io/en/latest/>.

APPENDIX B: TRANSFORMATION BETWEEN FP EIGENVECTORS AND PARAMETERS

In the most general 3D Gaussian method (Saglia et al. 2001; Colless et al. 2001; Magoulas et al. 2012), the FP is defined by three

orthonormal unit eigenvectors,

$$\begin{aligned}\hat{v}_1 &= \frac{\hat{r} - a\hat{s} - b\hat{i}}{|\mathbf{v}_1|} \\ \hat{v}_2 &= \frac{b\hat{r} - bk\hat{s} + (1 - ka)\hat{i}}{|\mathbf{v}_2|} \\ \hat{v}_3 &= \frac{(ka^2 - a + kb^2)\hat{r} + (ka - 1 - b^2)\hat{s} + (kb + ab)\hat{i}}{|\mathbf{v}_1||\mathbf{v}_2|}\end{aligned}\quad (\text{B1})$$

where

$$\begin{aligned}|\mathbf{v}_1| &= \sqrt{1 + a^2 + b^2} \\ |\mathbf{v}_2| &= \sqrt{1 + b^2 + k^2(a^2 + b^2 - 2a/k)}.\end{aligned}\quad (\text{B2})$$

Using these expressions and the Jacobian, we can write the scatter matrix components for the FP parameter space shown in equation (17) in terms of $\sigma_1, \sigma_2, \sigma_3$ (the scatter in each of the orthonormal coordinates) as

$$\sigma_r^2 = \frac{\sigma_1^2}{|\mathbf{v}_1|^2} + \frac{b^2\sigma_2^2}{|\mathbf{v}_2|^2} + \frac{(ka^2 - a + kb^2)^2\sigma_3^2}{|\mathbf{v}_1|^2|\mathbf{v}_2|^2}\quad (\text{B3})$$

$$\sigma_s^2 = \frac{a^2\sigma_1^2}{|\mathbf{v}_1|^2} + \frac{k^2b^2\sigma_2^2}{|\mathbf{v}_2|^2} + \frac{(ka - 1 - b^2)^2\sigma_3^2}{|\mathbf{v}_1|^2|\mathbf{v}_2|^2}\quad (\text{B4})$$

$$\sigma_i^2 = \frac{b^2\sigma_1^2}{|\mathbf{v}_1|^2} + \frac{(1 - ka)^2\sigma_2^2}{|\mathbf{v}_2|^2} + \frac{(kb + ab)^2\sigma_3^2}{|\mathbf{v}_1|^2|\mathbf{v}_2|^2}\quad (\text{B5})$$

$$\sigma_{rs} = -\frac{a\sigma_1^2}{|\mathbf{v}_1|^2} - \frac{kb^2\sigma_2^2}{|\mathbf{v}_2|^2} + \frac{(ka - a + kb^2)(ka - 1 - b^2)\sigma_3^2}{|\mathbf{v}_1|^2|\mathbf{v}_2|^2}\quad (\text{B6})$$

$$\sigma_{ri} = -\frac{b\sigma_1^2}{|\mathbf{v}_1|^2} + \frac{b(1 - ka)\sigma_2^2}{|\mathbf{v}_2|^2} + \frac{(ka - a + kb^2)(kb + ab)\sigma_3^2}{|\mathbf{v}_1|^2|\mathbf{v}_2|^2}\quad (\text{B7})$$

$$\sigma_{si} = \frac{ab\sigma_1^2}{|\mathbf{v}_1|^2} - \frac{kb(1 - ka)\sigma_2^2}{|\mathbf{v}_2|^2} + \frac{(ka - 1 - b^2)(kb + ab)\sigma_3^2}{|\mathbf{v}_1|^2|\mathbf{v}_2|^2}.\quad (\text{B8})$$

In practice, previous works (Saglia et al. 2001; Colless et al. 2001; Magoulas et al. 2012) all found that the second eigenvector has a very weak dependence on s , so that the longest axis of the 3D Gaussian is confined to the r - i plane. If one assumes *a priori* that this is true, then $k = 0$ in the above expressions. We make the same assumption in our fits.

APPENDIX C: DERIVATION OF ANALYTIC \mathbf{f}_n

In this appendix, we derive the expression for the integral over the 3D Gaussian of the FP in terms of elementary functions. Although the exact derivation here is somewhat specific to the typical selection function imposed on FP measurements, it can be adapted to other scenarios requiring the integral over a 3D Gaussian function.⁹

C1 General case

The integral we seek to solve has the form

$$\begin{aligned}f_n &= \frac{1}{(2\pi)^{3/2}|\mathbf{C}|^{1/2}} \int_{-\infty}^{\infty} \int_{r_{\min} - i/2}^{r_{\max} - i/2} \int_{s_{\min}}^{\infty} ds dr di \\ &\times \exp\left\{-\frac{1}{2}(\mathbf{x} - \bar{\mathbf{x}})\mathbf{C}^{-1}(\mathbf{x} - \bar{\mathbf{x}})^T\right\},\end{aligned}\quad (\text{C1})$$

⁹This derivation is heavily indebted to anonymous user PRZEMO’s derivation of the ‘Multivariate Gaussian integral over positive reals’ on MATHEMATICS STACK EXCHANGE.

where $\mathbf{x} = (r, s, i)$, $\bar{\mathbf{x}} = (\bar{r}, \bar{s}, \bar{i})$ and

$$\mathbf{C}^{-1} = \begin{pmatrix} \Psi_{rr} & \Psi_{rs} & \Psi_{ri} \\ \Psi_{rs} & \Psi_{ss} & \Psi_{si} \\ \Psi_{ri} & \Psi_{si} & \Psi_{ii} \end{pmatrix}. \quad (\text{C2})$$

For our FP scenario, the values of $r_{\max(\min)}$ and the covariance matrix \mathbf{C} change for each galaxy n ; $r_{\max(\min)}$ also vary as a function of distance. The first thing to note is that we are free to re-centre the integral about the mean values by performing a simple change of base,

$$f_n = \frac{1}{(2\pi)^{3/2} |\mathbf{C}|^{1/2}} \int_{-\infty}^{\infty} \int_{r_{\min} - \bar{r} - \bar{i}/2 - i/2}^{r_{\max} - \bar{r} - \bar{i}/2 - i/2} \int_{s_{\min} - \bar{s}}^{\infty} ds dr di \exp\left\{-\frac{1}{2} \mathbf{x} \mathbf{C}^{-1} \mathbf{x}^T\right\}. \quad (\text{C3})$$

Focussing on the integrand, we can expand the exponent as

$$\mathbf{x} \mathbf{C}^{-1} \mathbf{x}^T = \Psi_{ss} \left(s + \frac{\Psi_{si} i + \Psi_{rs} r}{\Psi_{ss}} \right)^2 + \left(\Psi_{ii} - \frac{\Psi_{si}^2}{\Psi_{ss}} \right) i^2 + \left(\Psi_{rr} - \frac{\Psi_{rs}^2}{\Psi_{ss}} \right) r^2 + 2 \left(\Psi_{ri} - \frac{\Psi_{rs} \Psi_{si}}{\Psi_{ss}} \right) ri. \quad (\text{C4})$$

Substituting this into equation (C3), we can solve the integral over s ,

$$f_n = \frac{1}{4\pi \sqrt{\Psi_{ss} |\mathbf{C}|}} \int_{-\infty}^{\infty} \int_{r_{\min} - \bar{r} - \bar{i}/2 - i/2}^{r_{\max} - \bar{r} - \bar{i}/2 - i/2} dr di \exp\left\{-\frac{1}{2} \times \left[\left(\Psi_{ii} - \frac{\Psi_{si}^2}{\Psi_{ss}} \right) i^2 + \left(\Psi_{rr} - \frac{\Psi_{rs}^2}{\Psi_{ss}} \right) r^2 + 2 \left(\Psi_{ri} - \frac{\Psi_{rs} \Psi_{si}}{\Psi_{ss}} \right) ri \right] \right\} \times \text{erfc}\left(\frac{\Psi_{si} i + \Psi_{rs} r + \Psi_{ss} (s_{\min} - \bar{s})}{\sqrt{2\Psi_{ss}}} \right), \quad (\text{C5})$$

where $\text{erfc}(u) = 1 - \text{erf}(u)$ is the complementary error function. We now make the substitution $u = (\Psi_{si} i + \Psi_{rs} r + \Psi_{ss} (s_{\min} - \bar{s})) / \sqrt{2\Psi_{ss}}$, so that

$$f_n = \frac{-\exp\left\{-\frac{1}{2} (s_{\min} - \bar{s})^2 \Psi_{ss} \left(\frac{\Psi_{ss} \Psi_{ii}}{\Psi_{si}^2} - 1 \right)\right\}}{\Psi_{si} \sqrt{8\pi^2 |\mathbf{C}|}} \int_{-\infty}^{\infty} \int_{\ell_{\min}}^{\ell_{\max}} dr du \times \text{erfc}(u) \exp\left\{-\left(\frac{\Psi_{ss} \Psi_{ii}}{\Psi_{si}^2} - 1 \right) \times \left(u^2 - \sqrt{2\Psi_{ss}} (s_{\min} - \bar{s}) u \right)\right\} \times \exp\left\{-\frac{1}{2} \left[\frac{2\sqrt{2\Psi_{ss}}}{\Psi_{si}} \left(\Psi_{ri} - \frac{\Psi_{rs} \Psi_{si}}{\Psi_{ss}} \right) \times \left(u - \sqrt{2\Psi_{ss}} (s_{\min} - \bar{s}) \right) r + \frac{\Psi_{rs}}{\Psi_{si}} \left(\frac{\Psi_{rs} \Psi_{ii}}{\Psi_{si}} + \frac{\Psi_{rr} \Psi_{si}}{\Psi_{rs}} - 2\Psi_{ri} \right) r^2 \right]\right\}, \quad (\text{C6})$$

where

$$\ell_{\min(\max)} = \frac{2\Psi_{si} (r_{\min(\max)} - \bar{r} - \bar{i}/2) - \sqrt{2\Psi_{ss}} u + \Psi_{ss} (s_{\min} - \bar{s})}{2\Psi_{si} - \Psi_{rs}}. \quad (\text{C7})$$

At first glance it, may seem that our choice of substitution is a poor one and the resulting expression is untenable. However, what we have actually done is isolate the parts of the integral that depend on r in a single exponential. In doing so, we have arrived at an expression of

the form $\int_{\ell_{\min}}^{\ell_{\max}} \exp[-1/2(Ar + Br^2)] dr$, which can be expressed as a difference of error functions times an exponential. After performing this integral, substituting $\ell_{\min(\max)}$, and an exhaustive amount of algebra,

$$f_n = \frac{1}{4\sqrt{\pi\delta} |\mathbf{C}|} \int_{-\infty}^{\infty} du \text{erfc}(u) \exp\left\{-\frac{1}{\delta |\mathbf{C}|} f^2(u)\right\} \times \left[\text{erf}\left\{\frac{Gf(u)}{\sqrt{\delta}} + R_{\min}\right\} - \text{erf}\left\{\frac{Gf(u)}{\sqrt{\delta}} + R_{\max}\right\} \right], \quad (\text{C8})$$

where

$$f(u) = u - \sqrt{\frac{\Psi_{ss}}{2}} (s_{\min} - \bar{s}), \quad (\text{C9})$$

$$R_{\min(\max)} = \frac{\sqrt{2\delta} (r_{\min(\max)} - \bar{r} - \bar{i}/2)}{2\Psi_{si} - \Psi_{rs}}, \quad (\text{C10})$$

$$\delta = \Psi_{rr} \Psi_{si}^2 + \Psi_{ii} \Psi_{rs}^2 - 2\Psi_{si} \Psi_{rs} \Psi_{ri}, \quad (\text{C11})$$

$$G = \sqrt{\Psi_{ss}} \frac{\Psi_{ri} (2\Psi_{si} + \Psi_{rs}) - \Psi_{rr} \Psi_{si} - 2\Psi_{ii} \Psi_{rs}}{2\Psi_{si} - \Psi_{rs}}. \quad (\text{C12})$$

We have managed to reduce the original 3D Gaussian integral to a single integration. Although this is complex to write, it is significantly faster to compute. However, we can go further still. To make clear how, we rewrite this expression using the substitution $x = \sqrt{\frac{1}{\delta |\mathbf{C}|}} f(u)$, so

$$f_n = \frac{1}{4\sqrt{\pi}} \int_{-\infty}^{\infty} dx \exp\{-x^2\} \text{erfc}\left\{\sqrt{\delta |\mathbf{C}|} x + \sqrt{\frac{\Psi_{ss}}{2}} (s_{\min} - \bar{s})\right\} \times \left[\text{erf}\left\{G\sqrt{|\mathbf{C}|} x + R_{\min}\right\} - \text{erf}\left\{G\sqrt{|\mathbf{C}|} x + R_{\max}\right\} \right]. \quad (\text{C13})$$

Using the relationship between the standard and complementary error functions, this last integral can be written as the addition/subtraction of four separate integrals of the form $\int_{-\infty}^{\infty} \exp[-x^2] \text{erfc}[Ax + B]$ or $\int_{-\infty}^{\infty} \exp[-x^2] \text{erfc}[Ax + B] \text{erfc}[Cx + D]$. Both of these can be solved by differentiating under the integral sign using the Leibniz integration rule, sometimes called the ‘Feynman integration trick’.¹⁰ The following identities for these integrals are

$$\int_{-\infty}^{\infty} \exp[-x^2] \text{erfc}[Ax + B] = \sqrt{\pi} \text{erfc}\left[\frac{B}{\sqrt{1 + A^2}}\right] \quad (\text{C14})$$

and

$$\int_{-\infty}^{\infty} \exp[-x^2] \text{erfc}[Ax + B] \text{erfc}[Cx + D] = \sqrt{\pi} \left[1 + 4T(F_1, F_2) + 4T(F_3, F_4) + \frac{2}{\pi} \tan^{-1}(F_5) - \frac{2}{\pi} \tan^{-1}(F_2) - \frac{2}{\pi} \tan^{-1}(F_4) + \text{erf}\left(\frac{F_1}{\sqrt{2}}\right) + \text{erf}\left(\frac{F_3}{\sqrt{2}}\right) \right], \quad (\text{C15})$$

where $T(h, x)$ is Owen’s T function and

$$F_1 = \frac{-\sqrt{2}B}{\sqrt{1 + A^2}}, F_2 = \frac{ABC - (1 + A^2)D}{B\sqrt{1 + A^2 + C^2}}, F_3 = \frac{-\sqrt{2}D}{\sqrt{1 + C^2}}, \quad (\text{C16})$$

¹⁰Again, with significant help from anonymous user PRZEMO’s solution to an ‘Integral of product of exponential function and two complementary error functions’ on MATHEMATICS STACK EXCHANGE.

$$F_4 = \frac{ACD - (1 + C^2)B}{D\sqrt{1 + A^2 + C^2}}, \quad F_5 = \frac{AC}{\sqrt{1 + A^2 + C^2}}. \quad (\text{C17})$$

Utilizing these identities and comparing them to the expression in equation (C13), we arrive at our final result, a sum of elementary functions,

$$f_n = \frac{1}{4} \operatorname{erf}\left(\frac{G_0^{\max}}{\sqrt{2}}\right) - \frac{1}{4} \operatorname{erf}\left(\frac{G_0^{\min}}{\sqrt{2}}\right) + \sum_{j=0}^1 \left[T(G_j^{\max}, H_j^{\max}) - T(G_j^{\min}, H_j^{\min}) + \frac{1}{2\pi} \tan^{-1}(H_j^{\min}) - \frac{1}{2\pi} \tan^{-1}(H_j^{\max}) \right], \quad (\text{C18})$$

where

$$G_0^{\min(\max)} = -\sqrt{\frac{2}{1 + G^2|\mathbf{C}|}} R_{\min(\max)}, \quad (\text{C19})$$

$$G_1^{\min(\max)} = -\sqrt{\frac{\Psi_{ss}}{1 + \delta|\mathbf{C}|}} (s_{\min} - \bar{s}), \quad (\text{C20})$$

$$H_0^{\min(\max)} = \frac{G|\mathbf{C}|\sqrt{\delta} - \sqrt{\frac{\Psi_{ss}}{2} \frac{(s_{\min} - \bar{s})}{R_{\min(\max)}}} (1 + G^2|\mathbf{C}|)}{\sqrt{1 + G^2|\mathbf{C}| + \delta|\mathbf{C}|}}, \quad (\text{C21})$$

$$H_1^{\min(\max)} = \frac{G|\mathbf{C}|\sqrt{\delta} - \sqrt{\frac{2}{\Psi_{ss}} \frac{R_{\min(\max)}}{(s_{\min} - \bar{s})}} (1 + \delta|\mathbf{C}|)}{\sqrt{1 + G^2|\mathbf{C}| + \delta|\mathbf{C}|}}, \quad (\text{C22})$$

and δ , $R_{\min(\max)}$, and G are as defined in equations (C10)–(C12).

This result is a little complex but can still be coded up relatively easily. However, the benefit over numerically evaluating the 3D integral is substantial: it is both exact and can be evaluated much, much faster. This is especially true when the determinant and inverse of the covariance matrix are also computed in terms of the individual elements ‘manually’ rather than relying on numerical matrix operations, which is trivial for a 3×3 matrix. As a comparison, numerically computing the determinant and inverse of the covariance matrix and using these as input to SCIPY.TPLQUAD takes ~ 1 s to evaluate numerically for a single galaxy at one distance to reasonable precision. The analytic form above can be easily vectorized for many galaxies or values of $r_{\min(\max)}$ and then requires on average only $\sim 1\mu\text{s}$ per evaluation, a factor of $\sim 1000\,000$ times faster.

C2 Derivation without magnitude limits

When fitting the FP parameters, rather than the individual distances to each galaxy, we do not need to account for the magnitude limits of the data in the f_n term as explained in Section 4. In this case, the f_n term becomes

$$f_n = \frac{1}{(2\pi)^{3/2} |\mathbf{C}|^{1/2}} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \int_{s_{\min}}^{\infty} ds dr di \times \exp\left\{-\frac{1}{2}(\mathbf{x} - \bar{\mathbf{x}})\mathbf{C}^{-1}(\mathbf{x} - \bar{\mathbf{x}})^T\right\}, \quad (\text{C23})$$

which can also be evaluated in terms of elementary functions. Following the steps in the previous derivation, we see this integral

can be solved by substituting $R_{\min} = -\infty$ and $R_{\max} = \infty$ into equation (C13), leaving

$$f_n = \frac{1}{2\sqrt{\pi}} \int_{-\infty}^{\infty} dx \exp\{-x^2\} \operatorname{erfc}\left\{\sqrt{\delta|\mathbf{C}|}x + \sqrt{\frac{\Psi_{ss}}{2}}(s_{\min} - \bar{s})\right\}. \quad (\text{C24})$$

We can again use the identity in equation (C14) to write this as

$$f_n = \frac{1}{2} \operatorname{erfc}\left(-\frac{G_1^{\min}}{\sqrt{2}}\right). \quad (\text{C25})$$

This is equivalent to the 1D integral in Appendix A of Magoulas et al. (2012), but again is much faster to compute.

APPENDIX D: PV ERROR FROM INCORRECT DISTANCE RATIO

The correct relation between the log-distance ratio and physical size (effective radius) is

$$\eta = \log R_e(z) - \log R_e(\bar{z}), \quad (\text{D1})$$

where z is the observed redshift, and \bar{z} is the cosmological redshift corresponding to the true comoving distance. These redshifts are related by

$$(1 + z) = (1 + \bar{z})(1 + z_p), \quad (\text{D2})$$

where z_p is the redshift corresponding to the PV, $v_p = cz_p$. The incorrect relation used by Springob et al. (2014) is

$$\begin{aligned} \eta' &= \log R_e(z) - \log R_e(\bar{z}) + \log[(1 + z)/(1 + \bar{z})] \quad (\text{D3})(\text{D4})(\text{D5}) \\ &= \log R_e(z) - \log R_e(\bar{z}) + \log(1 + z_p) \\ &= \eta + \log(1 + z_p). \end{aligned}$$

To proceed, we use the velocity estimator of Watkins & Feldman (2015). This allows us to write

$$\frac{v_p}{c} \approx \frac{z}{1 + z} \ln(10)\eta, \quad (\text{D6})$$

which is accurate as long as $v_p \ll cz$. Substituting this into equation (D5) and performing a Taylor expansion around $v_p = 0$, we find

$$\eta' \approx \eta \left(1 - \frac{z}{1 + z} \ln(10)\right). \quad (\text{D7})$$

Hence, the relative error in the log-distance [and the PV, according to equation (D6)] is given by

$$\frac{\eta v_p}{v_p} \equiv \frac{v'_p - v_p}{v_p} \approx \frac{z}{1 + z} \ln(10). \quad (\text{D8})$$

Thus, the relative error in the peculiar redshift (or PV) from using the incorrect relation is approximately equal to two times the observed redshift.

This paper has been typeset from a $\text{\TeX}/\text{\LaTeX}$ file prepared by the author.