An unbiased estimator of peculiar velocity with Gaussian distributed errors for precision cosmology

Richard Watkins* and Hume A. Feldman*

1 Department of Physics, Willamette University, Salem, OR 97301, USA
2 Department of Physics & Astronomy, University of Kansas, Lawrence, KS 66045, USA

Accepted 2015 March 23. Received 2015 February 28; in original form 2014 November 21

ABSTRACT

We introduce a new estimator of the peculiar velocity of a galaxy or group of galaxies from redshift and distance estimates. This estimator results in peculiar velocity estimates which are statistically unbiased and have Gaussian distributed errors, thus complying with the assumptions of analyses that rely on individual peculiar velocities. We apply this estimator to the SFI++ and the Cosmicflows-2 catalogues of galaxy distances and, since peculiar velocity estimates of distant galaxies are error dominated, examine their error distributions. The adoption of the new estimator significantly improves the accuracy and validity of studies of the large-scale peculiar velocity field that assume Gaussian distributed velocity errors and eliminates potential systematic biases, thus helping to bring peculiar velocity analysis into the era of precision cosmology. In addition, our method of examining the distribution of velocity errors should provide a useful check of the statistics of large peculiar velocity catalogues, particularly those that are compiled out of data from multiple sources.

Key words: galaxies: kinematics and dynamics – galaxies: statistics – cosmology: observations – cosmology: theory – distance scale – large-scale structure of Universe.

1 INTRODUCTION

The Doppler effect provides a remarkably accurate method to infer the velocity of a galaxy towards or away from us by measuring the blue or red-shift of its spectral lines, respectively. However, since cosmological expansion also causes a redshift, determination of the peculiar (local) motion \(v\) also requires the measurement of the galaxy’s distance \(r\), so that

\[
v = cz - H_0r,
\]

where \(c\) is the speed of light, \(z\) is the redshift and \(H_0\) is Hubble’s constant. This formula assumes a linear Hubble relation. For more accuracy, particularly at large distances, we can include the effects of cosmic acceleration by replacing \(z\) with \(z_{\text{mod}}\), where

\[
z_{\text{mod}} = z[1 + 0.5(1 - q_o)z - (1/6)(1 - q_o - 3q_o^2 + 1)z^2],
\]

\(q_o\) is the deceleration parameter (see also Davis & Scrimgeour 2014; Springob et al. 2014). In addition, we can achieve additional accuracy by accounting for the fact that redshift is not an additive quantity. Rather than \(cz_{\text{mod}} = H_0r + v\), we instead should write

\[(1 + z_{\text{mod}}) = (1 + H_0 r/c)(1 + v/c),\]

which reduces to the familiar formula at low redshift. Solving for \(v\), we obtain

\[
v = \frac{cz_{\text{mod}} - H_0r}{1 + H_0 r/c} \approx \frac{cz_{\text{mod}} - H_0r}{1 + z_{\text{mod}}},
\]

where in the second expression we replaced \(H_0 r/c\) with \(z_{\text{mod}}\), a good approximation since the difference between them, which is approximately \(v/c\), is always much less than 1. The second expression is easier to work with in practice since it does not introduce new factors of \(r\), a quantity that has large uncertainties.

Whereas redshift can be measured very accurately, distance measurements typically have uncertainties of \(\approx 20\) per cent, so that the uncertainty, \(\delta v\), in a peculiar velocity is approximately \(\delta v \approx 0.20 H_0 r\). Since typical peculiar velocities are thought to be \(\approx 500\) km s\(^{-1}\), we see that for \(H_0 \approx 70\) km s\(^{-1}\) Mpc\(^{-1}\) the uncertainties in peculiar velocities become of the order of their magnitudes for objects at distances \(r > 35\) Mpc, which includes the region that we would like to use peculiar velocities as a tool to probe large-scale structure. Thus individual peculiar velocity measurements have very low signal-to-noise, which makes it is necessary to have a large sample in order to extract meaningful information.

Two major approaches have been used to analyse peculiar velocity catalogues (for reviews, see Jacoby et al. 1992; Strauss & Willick 1995). The first forgoes calculating individual peculiar velocities altogether, and instead uses distance and redshift information to estimate parameters of a model of the peculiar velocity field (for
Gaussian estimator of peculiar velocity

Our goal is to obtain an estimate, \(v_\epsilon\), of the peculiar velocity of a galaxy or group from the galaxy’s redshift \(cz\) and an estimate of its distance \(r_\epsilon\). Given equation (1), the most straightforward estimator is

\[
v_\epsilon = cz - H_0 r_\epsilon,
\]

and this is typically the estimator used in peculiar velocity analyses. However, from a statistical point of view, this estimator has several undesirable qualities. (For a general discussion of the statistics of estimators, see Lupton 1993.) First, distance indicators give distance moduli or log-distances with Gaussian distributed errors. Exponentiating skews the error distribution, resulting in distance errors that are not Gaussian distributed. Secondly, this estimator is biased in a statistical sense: the average of an ensemble of velocity estimates with different errors is not the true value, i.e. \(\langle v_\epsilon \rangle \neq v\). This is the result of the skewness of the distribution of distance errors, which gives rise to \(\langle r_\epsilon \rangle \neq r\). These undesirable features can lead to biases in our analyses and in general invalidate our statistical assumptions about the errors in peculiar velocities. They suggest that we should be investigating other estimators that might be better behaved statistically.

Instead we propose calculating peculiar velocities using the estimator

\[
v_\epsilon = cz \log(cz/H_0 r_\epsilon).
\]

While this estimator may look unfamiliar, it has the statistical properties that we desire in an estimator. First, since it uses the log-distance (or equivalently, the distance modulus), it has Gaussian distributed errors. It is easy to see that the uncertainty in the peculiar velocity, \(\delta v_\epsilon\), is given by \(\delta v_\epsilon = cz \delta l_\epsilon\), where \(\delta l_\epsilon\) is the uncertainty in the log-distance. Secondly, we can use \(\langle \log(r_\epsilon) \rangle = \log(r)\) to show that this estimator is unbiased, as long as the true \(v \ll cz\), which is a good assumption for distant galaxies,

\[
\langle v_\epsilon \rangle = -cz (\log(H_0 r_\epsilon)) - \log(cz)
\]

\[
= -cz (\log(H_0 r)) - \log(cz)
\]

\[
= -cz (\log(cz - v) - \log(cz))
\]

\[
= -cz (\log(1 - v/cz))
\]

\[
\approx v,
\]

where we have used equation (1) to replace \(H_0 r\) with \(cz \sim v\), and we have assumed that the uncertainties in the redshift \(cz\) are negligible.
From equation (3), we see that a more accurate estimator at large redshift is given by
\[ v_e = \frac{cz_{\text{mod}}}{1 + z_{\text{mod}}} \log(c z_{\text{mod}} / H_0 r_i), \]  
(7)
with uncertainty \( \delta v_i = c z_{\text{mod}} 8 k / (1 + z_{\text{mod}}). \) We stress that the assumption that we are making is that the actual velocity of the galaxy or group \((v)\) is small compared to the redshift, not the estimated velocity \((v_{\text{e}}).\) While estimates of peculiar velocities can be a few \( \times \) 10^3 km s^{-1}, it is thought that most actual peculiar velocities are at most a few \( \times \) 10^2 km s^{-1}. Our assumption should hold quite well for galaxies at distances \( \gtrsim 20 \text{ Mpc}. \)

3 STATISTICS OF PECULIAR VELOCITY SURVEYS

The defining characteristic of large-scale peculiar velocity surveys is that they have low signal-to-noise ratio. However, if our goal is to determine the distribution of the noise, then we can turn this to our advantage. In particular, if we consider objects with peculiar velocity errors \( \sigma \) such that \( \sigma \gg \sigma_i, \) where \( \sigma_i \) is the spread in actual peculiar velocities, then we can be assured that the objects’ measured peculiar velocities are dominated by noise, with negligible contributions from actual motions.

Here we will examine the error distribution in two large peculiar velocity surveys. The SFI++ (Masters et al. 2006; Springob et al. 2007) is a sample of 4052 spiral galaxies with TF distances. The Cosmicflows2 (hereafter CF2; Tully et al. 2013) galaxy catalogue is a compendium of distances to 8135 galaxies measured with various methods, including TF, Fundamental Plane, SNIa, surface brightness fluctuations and tip of the red giant branch (TRGB). While the CF2 contains the SFI++ as one of its largest components, in compiling the CF2 a reanalysis of the literature distances was done to ensure consistency between data sets. For both samples, we use the more accurate expressions given by equation (3) for the traditional estimator and equation (7) for the new estimator, following Tully et al. (2013) in assuming the standard cosmological model with \( \Omega_m = 0.27 \) and \( \Omega_\Lambda = 0.73, \) so that \( q_0 = 0.5(\Omega_m - 2\Omega_\Lambda) = -0.595. \)

Another difference in the catalogues is that the SFI++ catalogue provides distances in km s^{-1} so as to be scaled relative to the Hubble constant. In contrast, the CF2 sample attempts to determine an absolute scale, and so reports distances in Mpc. Thus to calculate peculiar velocities from the distances in the CF2, we must assume a value for the Hubble constant. The nominal value given by the authors of the CF2 in Tully et al. (2013) is 74.4 km s^{-1} Mpc^{-1}.

In Fig. 1, we show histograms for the values of the peculiar velocity divided by their uncertainty, \( v_i / \sigma_i, \) calculated using both the traditional estimator and the new estimator for galaxies with \( \sigma_i > 1000 \text{ km s}^{-1} \) in the CF2 survey. If our statistical assumptions are correct, and if actual motions make only a small contribution, then the values \( v_i / \sigma_i \) should be unit Gaussian variates, and the histograms should match the Gaussian of unit standard deviation shown in the figure. We see that the histogram using the new estimator is a good match to the unit Gaussian, but that the traditional estimator results in a skewed distribution.

As seen in the figure, the exponentiation of the Gaussian distributed log-distances results in a distribution of errors that is skewed in a complicated way, with the peak shifted towards negative velocities but with a shortened tail on the negative side and an elongated tail on the positive side. This effect is more clearly seen in Fig. 2 where we plot the histogram using a logarithmic scale. This skewness cannot be corrected for by simply shifting the centre of the distribution. Nor can the skewness be corrected by adjusting only negative velocities, as is proposed by Tully et al. (2013) to correct for what they call ‘error bias’.

In Fig. 3, we show \( v_i / \sigma_i \) histograms for galaxies with higher \( \sigma_i > 1000 \text{ km s}^{-1}. \) As in Fig. 2, we plot the histograms on a logarithmic scale to accentuate the tails of the distribution. The catalogue provides both Malmquist-corrected and -uncorrected distances. Malmquist bias correction methods account for the skewness of the distribution of distance errors, and Lynden-Bell et al. (1988) showed that velocities calculated with Malmquist-corrected distances should be approximately Gaussian distributed. It does not make sense to use Malmquist-corrected distances with the new estimator, since, as we discuss below, different corrections apply for distance moduli or log-distances than for distances. We
thus show histograms for the new estimator using uncorrected values and the traditional estimator using corrected values. First, we see that the histograms for the new estimator and the traditional estimator using Malmquist-corrected distances do in fact have the same peak. However, the histograms are not centred on zero.

One possible explanation for the skewing of the error histograms from zero in Fig. 2 is a coherent outflow in the volume occupied by the survey. While random velocities, or even bulk motions, should have little effect on the histograms since their affect would average out over different directions, a coherent outflow would be expected to shift the peak of the histograms towards positive velocities, exactly the effect we see in the figure. We can test this explanation by considering a simple model where the outflow adds a constant peculiar velocity to each galaxy in the survey. In Fig. 4, we show the same histograms as in Fig. 2 except that we have subtracted 400 km s$^{-1}$ from the peculiar velocity of each galaxy. We see that subtracting a relatively modest outflow has resulted in an error distribution that is Gaussian and centred on zero. The existence of a coherent outflow would support recent arguments that we live in a low-density region, dubbed the ‘local hole’ (Whitbourn & Shanks 2014).

The disagreement between the CF2 and SFI++ catalogues regarding the existence of a coherent outflow is a consequence of the addition of new data and the reanalysis of literature distances that was done when the CF2 catalogue was assembled. Tully et al. (2013) compared distance estimates for galaxies and groups that appeared in more than one component sample and used this comparison to rescale and reanalyze distances to achieve statistical consistency between all the components of the CF2 sample. This reanalysis was anchored by the zero-point distance provided by Cepheid and TRGB distances. They found that the resulting CF2 catalogue was consistent with a Hubble constant $H_0 = 74.4$ km s$^{-1}$ Mpc$^{-1}$ that did not vary with redshift. It is worth noting that although this relatively low redshift ($cz < 0.1$) measurement of $H_0$ is in tension with microwave background results, it agrees well with a recent measurement of $H_0$ using SN1a at much higher redshift (Neill et al. 2014).

It is possible that the rescaling of the SFI++ could have inadvertently ‘erased’ a real coherent outflow. However, this suggests another explanation of the skewness in the error distribution of the SFI++ survey, a systematic error in the scaling of distances. In Fig. 5, we show the same histograms as in Fig. 3, but with all distances increased by 5 per cent. This scaling is equivalent to changing the zero-point of the SFI++ or increasing its Hubble constant. Again, we see that there is good agreement between the new estimator histogram and the unit Gaussian centred on zero. This roughly corresponds to the rescaling of the SFI++ in the CF2. A direct comparison of the distances given in the SFI++ and the CF2 distances for the same galaxies in the CF2, using $H_0 = 74.4$ km s$^{-1}$ Mpc$^{-1}$, shows that CF2 distances are about 6.8 per cent larger on average.

In both Figs 4 and 5, we see that although the histogram using the traditional estimator with Malmquist-corrected data is indeed approximately Gaussian, the tails of this distribution are still noticeably skewed. This demonstrates that our new estimator is more effective at correcting for the skewness of peculiar velocity errors than the Malmquist bias corrections used in the SFI++ survey. Since these correction methods are substantially similar to those used in other surveys, it is likely that this is true in general.

4 DISCUSSION

Peculiar velocity analysis methods that work with velocity measurements for individual galaxies, groups or clusters assume that the errors in velocity measurements have a Gaussian distribution. However, the estimator that is traditionally used is known to have a skewed, non-Gaussian error distribution. Malmquist bias corrections include a correction that shifts the peak of the error distribution to zero, but these corrections do not remove the skewness in the tails of the distribution. These tails are particularly important since they represent objects with measurement errors that are larger than expected given their uncertainties. Given that measurement uncertainties typically dominate over the true velocities, these objects have large ratio of estimated peculiar velocity to uncertainty. Since analyses of velocity moments typically weight by uncertainty, these velocities may have a large impact on results, and can potentially introduce biases if the uncertainties are not distributed symmetrically about the central value.

As peculiar velocity catalogues become larger, with a corresponding decrease in the calculated uncertainties in low-order moments such as the bulk flow, it becomes increasingly important to address potential systematic errors arising from non-Gaussian velocity error distributions. We have introduced a simple, easy-to-use peculiar velocity estimator that results in velocities with unbiased, Gaussian errors. We have shown that this estimator works well when applied to the CF2 catalogue and, with some adjustment, the SFI++ catalogue of galaxy distances. This new estimator is an important step in bringing peculiar velocity analyses into the era of precision cosmology.

Our new estimator should not be used to estimate peculiar velocities with Malmquist-corrected distances, since currently implemented Malmquist correction procedures already account for the skewness of the traditional estimator. While we have shown that Malmquist correction does result in approximately Gaussian velocity errors in the SFI++ survey, we have seen that our new
estimator does a better job of producing a distribution with symmetric tails.

Modifying Malmquist correction methods to be used with the new estimator is straightforward. Specifically, in the SFI++ survey, Malmquist bias corrections to distances are implemented by calculating the corrected probability \( p(r_i) \) of a galaxy being at a distance \( r_i \), through the convolution (Springob et al. 2007)

\[
p(r_i) = k_i P_{TF}(r_i) p_{mag}(r_i) p_{hus}(r_i),
\]

where \( k_i \) is a normalization constant, \( P_{TF}(r_i) \) is the probability distribution for the galaxy being at a position \( r_i \) as given by the TF measurement, \( p_{mag}(r_i) \) is the probability of finding a galaxy with its apparent magnitude at a distance \( r_i \) and \( p_{hus}(r_i) \) is the density distribution along the line of sight as given by a redshift survey. To use the new peculiar velocity estimator, we instead calculate the corrected probability \( p(\mu_i) \) of finding a galaxy with a distance modulus \( \mu_i \), as

\[
p(\mu_i) = k_i P_{TF}(\mu_i) p_{mag}(\mu_i) p_{hus}(\mu_i),
\]

where the probabilities \( P_{TF} \), \( p_{mag} \) and \( p_{hus} \) have now been expressed in terms of the distance modulus. Note that in these expressions \( P_{TF}(\mu_i) \) is calculated using the distance modulus, while \( P_{TF}(r_i) \) is not. Note also that the maximum of \( p(r_i) \) will not correspond to the distance modulus that maximizes \( p(\mu_i) \), so that Malmquist-corrected values of \( r_i \) cannot be used in our new estimator; instead, Malmquist-corrected values of \( \mu_i \) must be calculated using \( p(\mu_i) \).

As a specific example, consider the simple case of a uniform density of galaxies, where, for mathematical simplicity, we will work with the equivalent log-distance \( l_i = \log(r_i) \) instead of the distance modulus \( \mu_i \). In this case there is a bias, sometimes called the homogeneous Malmquist bias, whereby galaxies are more likely to have scattered from larger than smaller radius due the increasing distance. This comes into our calculations through the fact that in this case \( p_{hus}(r_i)dr_i \propto r^2dr_i \), so that \( p_{hus}(l_i)d\mu_i \propto e^{\mu_i}d\mu_i \). Assuming that \( p_{mag}(l_i) \) is constant and that \( P_{TF}(l_i) \) is a Gaussian distribution centred on the value \( l_0 \) with uncertainty \( \Delta l \), we have

\[
p(l_i) \propto \exp\left(-\frac{(l_i - l_0)^2}{2\Delta l^2}\right)e^{\mu_i}.
\]

Thus we see that \( p(l_i) \) remains Gaussian, with the effect of the Malmquist bias correction being to shift the peak of the distribution outwards by \( 3\Delta l \). The size of this shift matches the result of a similar calculation given in Lynden-Bell et al. (1988). More generally, this calculation suggests that as long as the product \( p_{mag}(r_i)p_{hus}(r_i) \) can be approximated by a power law in \( r_i \) in the region around the galaxy’s location, the effect will be to shift the peak of \( p(\mu_i) \) relative to \( P_{TF}(\mu_i) \), while maintaining a Gaussian distribution. Since \( p_{mag} \) and \( p_{hus} \) are typically slowly varying compared to \( P_{TF} \), it is thus reasonable to expect that Malmquist-corrected \( \mu_i \) will still have Gaussian errors. We will investigate this issue in more detail in future work.

Large-scale motion analyses require estimates of both the radial peculiar velocities and positions of a set of galaxies. While we must use both redshift and a distance estimate to calculate peculiar velocity, either of these quantities can be used to estimate position. Analyses that use distance estimates to estimate position are said to be done in ‘real space’, while those that use redshift are said to be done in ‘redshift space’. While it may seem more intuitive to use a distance estimate to estimate position, it is important to remember that redshift is often a more accurate estimate of distance, particularly in situations where distance units of \( h^{-1} \) Mpc or km s\(^{-1} \) are used, so that uncertainty in the value of the Hubble constant does not enter the calculation. Due to the small uncertainty in redshift, the scatter of the redshift about the distance (as measured in km s\(^{-1} \)) is caused almost entirely by peculiar velocities, which are thought to be of the order of 500 km s\(^{-1} \) at most. This is to be contrasted with distance estimates, which for many distance indicators have uncertainties of the order of 20 per cent. Thus redshift begins to be a more accurate measure of a galaxy’s position around distances of 2500 km s\(^{-1} \), or 25h\(^{-1} \) Mpc, and for distances of the order of 100 h\(^{-1} \) Mpc, redshift can be a factor of 4 more accurate than distance estimates on average. We note that the advantage of redshift over distance estimates can be somewhat smaller for clusters and groups of galaxies, where distance uncertainties can be reduced by \( \sqrt{N} \), where \( N \) is the number of galaxies in the cluster with measured distances, and for SNIa, where distance uncertainties are closer to 5 per cent. Because position uncertainties are much smaller in redshift space, particularly for objects at large distances, Malmquist bias effects that are caused by position uncertainties are much less important and can be neglected. However, in redshift space one must account for other forms of Malmquist bias that affect the determination of distance relation parameters, e.g. the slope and zero-point of the TF relation. For a more detailed discussion of Malmquist bias in real and redshift space, see Strauss & Willick (1995).

The new estimator we introduced here will prove particularly useful for peculiar velocity analyses that are done in redshift space with data that has not been Malmquist corrected. In this case, peculiar velocities calculated with the traditional estimator have an error distribution that is biased in addition to being skewed. For example, the estimator we have introduced alleviates the problem of ‘error bias’ noted in Tully et al. (2013). We have shown that peculiar velocities calculated from CF2 distances using the new estimator have a symmetric, Gaussian error distribution and do not require any further correction.

We have also presented a method to check large catalogues of peculiar velocities to confirm that they have the expected distribution of errors. We stress that skewness or non-Gaussianity in velocity error distributions can lead to results which do not accurately reflect the large-scale flows we are trying to study. This method should provide a useful tool for compiling large peculiar velocity catalogues, particularly when combining data from different sources.

Finally, we have seen that the distribution of errors in the SFI++ survey is not centred on zero. This can be explained by an approximately 400 km s\(^{-1} \) coherent outflow in the survey volume or by a systematic error in the scaling of distances of about 5 per cent. Which of these explanations is correct is an interesting question that should be pursued in further research.

ACKNOWLEDGEMENTS

We would like to thank Brent Tully and Mike Hudson for useful comments.

REFERENCES
