Incomplete Data and Measurement Error in the Galactic Mass Estimation Problem

Gwendolyn Eadie* William Harris† Aaron Springford‡

Abstract
It is widely accepted that the Milky Way Galaxy resides within a massive dark matter halo. The mass and cumulative mass profile of this halo (and the Galaxy as a whole) are two of the most fundamental properties of the Galaxy. Estimating these properties, however, is not a trivial problem. We rely on the kinematic information of satellites which orbit the Galaxy, such as globular clusters and dwarf galaxies, and this data is incomplete and subject to measurement uncertainty. In particular, the complete 3D velocity and position vectors of objects are sometimes unavailable, and there are selection biases due to the distribution of objects around the Galaxy and our measurement position. On the other hand, the instrumental uncertainties of telescopes that collect this data is fairly well understood. Thus, we would like to incorporate these uncertainties into our estimate of the Milky Way's mass. The Bayesian paradigm offers a way to deal with both the missing kinematic data and measurement errors using a hierarchical model.

Key Words: Bayesian, astrostatistics, astronomy, astrophysics, incomplete data, Galaxy mass profiles, dark matter, hierarchical

1. Introduction

The universe is filled with galaxies of various sizes and morphological types, including disk-shaped spirals, giant ellipticals, and dwarfs, as well as irregular galaxies with no discernible patterns. Despite the visible differences between galaxies, however, it is strongly believed that all galaxies have one thing in common: they each reside in their own massive, invisible, dark matter (DM) halo. Although DM has yet to be detected directly (its composition is still unknown), there is mounting evidence to support the idea of this missing, unseen mass both in galaxy clusters and galaxies.

Some of the first evidence for DM dates back to work by Zwicky (1933, 1937), who looked at the dynamics of galaxies in the Coma Cluster. By applying the virial theorem, Zwicky found the mass of the cluster to be much larger than what was implied by the luminous matter contained therein. Another famous paper by Rubin et al. (1980) showed that the orbital velocities of stars in spiral galaxies are too large if the only gravitational forces acting on them are from visible matter. Gravitational lensing around galaxy clusters also suggests the presence of DM; distant background galaxies are seen as distorted, lensed arcs around galaxy clusters. Another indication of DM in galaxy clusters comes from analysing X-ray images of the intracluster gas. Finally, cosmological simulations of the universe require DM for regular (baryonic) matter to properly collapse and coalesce into the galaxies we see today.

Barring that general relativity is incorrect, it seems that DM is necessary for describing not only the overall evolution of the universe, but also the dynamical behaviour of objects we observe today. Thus, estimating the mass and cumulative mass profiles of the DM halos of galaxies is incredibly important if we are to understand the nature of DM and its role in galaxy formation.

*Physics & Astronomy Department, McMaster University, Hamilton, ON L8S 4M1, Canada
†Physics & Astronomy Department, McMaster University, Hamilton, ON L8S 4M1, Canada
‡Mathematics & Statistics Department, Queen's University, Kingston, ON K7L 3N6, Canada

Estimating the Milky Way's mass profile is a great place to start; the Galaxy is a typical spiral with many natural satellites (such as dwarf galaxies and globular clusters) orbiting under the influence of gravity. Using the positions and velocities of these satellites, and assuming a model for the gravitational potential, one can infer the mass of the DM halo responsible for the tracers' orbits. This basic idea has been utilized in different ways under both frequentist and Bayesian methods for many years (for example, Bahcall and Tremaine 1981; Little and Tremaine 1987; Wilkinson et al. 2003; Battaglia et al. 2005; An and Evans 2011; Boylan-Kolchin et al. 2013). However, one drawback associated with using satellites as tracers of the DM halo remains: the 3-dimensional velocities of satellites are critical to the analysis, but they are not always available.

The 3-dimensional velocity vector of a satellite is split up into two components: 1) the 1-dimensional line-of-sight velocity, which is measured along the line-of-sight between the Earth and the satellite, and 2) the 2-dimensional proper motion $\boldsymbol{\mu}$, which is measured on the plane of the sky. While measurements of the former are easily obtained, measuring the latter often requires years of observation. Subsequently, more than half of the satellites around the Milky Way are missing proper motion measurements, making these data points incomplete. Previous studies have dealt with this problem by either a) using only the line-of-sight measurements, or b) using only satellites for which proper motions have already been measured. Either way, information contained within the available data is being thrown away.

Eadie et al. (2015) addressed the problem of incomplete data by introducing a Bayesian method that estimates the mass and cumulative mass profile of the Milky Way using complete and incomplete data simultaneously. In their preliminary study, they assumed a simple, spherical DM halo by applying the Hernquist (1990) model, and used kinematic data from 24 dwarf galaxies and 64 globular clusters. They found the mass of the Milky Way within 260 kpc to be $1.37 \times 10^{12} M_{\odot}$ with a 95% credible interval of $(1.27, 1.51) \times 10^{12} M_{\odot}$. Their results, which fall within the range of masses found by many other studies that use different methods, suggest further development of the method is worthwhile (for a list of results from other studies, see Wang et al. 2015).

The author's goal is to improve upon the method in Eadie et al. (2015) by incorporating the measurement uncertainties associated with the line-of-sight velocities and proper motions. In the current proceedings, we discuss this next step, and show some very preliminary results using the same data as in the aforementioned paper.

2. Terminology and Models

Before presenting details about how we include measurement uncertainties in a Bayesian analysis, it is important to distinguish between the Heliocentric and Galactocentric reference frames. The Heliocentric coordinate system is centered around the Sun, whereas the Galactocentric coordinate system has its origin at the center of the Galaxy. The notation used for the distance from the Earth to the satellite in the Heliocentric frame is R, whereas the distance from the center of the Galaxy to the satellite in the Galactocentric frame is r.

We would also like to emphasize that the terms *line-of-sight velocity* (v_{los}) and *proper motion* ($\boldsymbol{\mu}$[1]), have different meanings than two other terms commonly used in the literature: the *radial velocity* (v_r) and *tangential velocity* (v_t). Whereas v_{los} and $\boldsymbol{\mu}$ refer to velocities in the Heliocentric reference frame, v_r and v_t refer to velocities in the Galactocentric reference frame. In the Galactocentric reference frame, v_r is along the line-of-sight from the

[1] We emphasize that $\boldsymbol{\mu}$ here is not the mean of a distribution, but the 2-dimensional velocity vector. Using $\boldsymbol{\mu}$ in this way is the convention in astrophysics literature.

centre of the Galaxy to the object, and v_t is perpendicular to v_r in the same way that $\boldsymbol{\mu}$ is perpendicular to v_{los}.

Another point of note is that the proper motion is split up into two components: μ_δ, the velocity in declination δ (north or south of the celestial equator), and $\mu_\alpha \cos\delta$, the velocity in right ascension α (east or west from the vernal equinox[2]).

Almost all Galactic models are mathematically simpler in the Galactocentric reference frame, and require kinematic data in the form (r, v_r, v_t). However, transforming velocities from the Heliocentric to the Galactocentric frame is complicated and non-linear— one must take into account the motion of the Sun around the Galaxy, the rotation of the Galactic disk, the object's distance from the Sun, and the object's coordinates on the sky (α, δ). Furthermore, the transformation can only be completed when both v_{los} and $\boldsymbol{\mu}$ of the satellite are known.

Eadie et al. (2015) used kinematic data of satellites that had already been transformed to the Galactocentric frame, and assumed (r, v_r, v_t) as fixed and known. Consequently, they had to make approximations of v_r for satellites without proper motion measurements, i.e. satellite data that was incomplete could only be used if it satisfied certain geometric conditions (see Eadie et al. 2015). Through a sensitivity analysis they also showed that uncertainties in v_r and v_t may contribute up to half of the model uncertainty. Thus, our next goal is to not only limit the geometric condition, but to also include the measurement uncertainties in the analysis.

The measurement uncertainties $(\Delta v_{los}, \Delta\mu_\delta, \Delta\mu_\alpha \cos\delta)$ are very well understood by astronomers. However, transforming these uncertainties to the Galactocentric reference frame is non-linear and difficult, and the resulting uncertainties in the Δv_r, and Δv_t will be dependent and non-Gaussian. In the Bayesian paradigm, it is easier to incorporate measurement uncertainties that are independent and normally distributed. Therefore, we will work in the Heliocentric reference frame so we may use Δv_{los}, $\Delta\mu_\delta$, and $\Delta(\mu_\alpha \cos\delta)$, which are independent and approximately Gaussian.

3. Methods, Models, and Data

3.1 Methods: Incorporating Measurement Uncertainty in the Bayesian Framework

A measurement of a random variable such as v_{los} or μ_δ is inherently uncertain, and this is precisely why measurement uncertainties are reported. The measurement uncertainties signify that we do not know the true value of a quantity. Therefore, we now treat the quantities $(r, v_{los}, \mu_\delta, \mu_\alpha \cos\delta)$ as data drawn from a distribution centered on the parameters $(\color{blue}{r}, \color{blue}{v_{los}}, \color{blue}{\mu_\delta}, \color{blue}{\mu_\alpha \cos\delta})$. We assume that the data are normally distributed about their corresponding parameter values, with standard deviation equal to the measurement uncertainty. That is, $\sigma_R = \Delta R$, $\sigma_{v_{los}} = \Delta v_{los}$, $\sigma_{\mu_\delta} = \Delta\mu_\delta$ and $\sigma_{\mu_\alpha \cos\delta} = \Delta\mu_\alpha \cos\delta$. The likelihood is then

$$\mathcal{L} = p(r|\color{blue}{r}, \color{red}{\Delta r})p(v_{los}|\color{blue}{v_{los}}, \color{red}{\Delta v_{los}})p(\mu_\delta|\color{blue}{\mu_\delta}, \color{red}{\Delta\mu_\delta})p(\mu_\alpha \cos\delta|\color{blue}{\mu_\alpha \cos\delta}, \color{red}{\Delta\mu_\alpha \cos\delta}) \quad (1)$$

where the blue characters denote parameters and the red characters denote the known measurement uncertainties.

In equation 1, note that we are treating the Galactocentric distance r, not the Heliocentric distance R, as a parameter. Measurement uncertainties for R exist, but incorporating them is much more difficult. The transformation matrix necessary to change a satellite's velocity $\boldsymbol{v} =< v_{los}, \boldsymbol{\mu} >$ to the Galactocentric frame requires the R value of the satellite, as well as the distance of the Sun from the center of the Galaxy, $R_\odot$. For simplicity's sake, and to avoid having to recalculate the transformation matrix at every step in the Markov

[2]The factor of $\cos\delta$ comes from spherical symmentry.

Chain, for every satellite, we fix R for each satelltie and treat $R_\odot$ as a known constant. We do, however, assign an uncertainty of $\Delta r = 10\%$ for each satellite's r value. In future work we plan to allow R and $R_\odot$ to vary as parameters instead of r (although we expect it to be much more computationally expensive).

The likelihood in equation 1 is used in a hierarchical Bayesian model,

$$p(\boldsymbol{\theta}|r, v_{los}, \mu_\delta, \mu_\alpha \cos\delta, \Delta) \propto \mathcal{L}(r, v_{los}, \mu_\delta, \mu_\alpha \cos\delta, \Delta|\vartheta)p(h(\vartheta)|\boldsymbol{\theta})p(\boldsymbol{\theta}) \tag{2}$$

where Δ is the vector of known uncertainties, ϑ represents the parameters $(r, v_{los}, \mu_\delta, \mu_\alpha \cos\delta)$ in equation 1, and h is the transformation function to go from the Heliocentric to the Galactocentric reference frame. The transformation of Heliocentric velocities to Galactocentric ones follows the method outlined in Johnson and Soderblom (1987), but with updated values for the J2000 epoch. The term $p(h(\vartheta)|\boldsymbol{\theta})$ in equation 2 is the probability density function for the Galaxy model, given a vector of parameters $\boldsymbol{\theta}$. In the case of having N satellites, equation 2 is written more compactly as

$$p(\boldsymbol{\theta}|\boldsymbol{y}, \Delta) \propto \prod_i^N \mathcal{L}(\boldsymbol{y_i}, \Delta_i|\vartheta_i)p(h(\vartheta_i)|\boldsymbol{\theta})p(\boldsymbol{\theta}). \tag{3}$$

Eadie et al. (2015) showed how a hybrid-Gibbs sampler is used to treat the unknown v_t's as parameters in the model. They constructed Markov chains that were proportional to the posterior distribution

$$p(\boldsymbol{\theta}|\boldsymbol{y}) \propto \prod_i^N p(r_i, v_{r,i} v_{t,i}|\boldsymbol{\theta})p(\boldsymbol{\theta}) \tag{4}$$

We continue to use the hybrid-Gibbs sampling, but instead extend it to sampling the parameters ϑ and hyperparameters $\boldsymbol{\theta}$ in equation 3. In this way, the method has gone from a basic Bayesian model to a hierarchical Bayesian model.

3.2 Model: Hernquist probability density function

To test the hierarchical method outlined above and for easy comparison of results, we use the same Galaxy model and the same data as Eadie et al. (2015). In the previous paper, the model from Hernquist (1990) is used in three different forms, corresponding to three different velocity anisotropies. Here, we focus on the isotropic case.

The Hernquist model is a self-consistent model of a spherical mass distribution with a gravitational potential

$$\Phi(r) = -\frac{M_{\text{tot}}}{r+a} \tag{5}$$

and mass density profile

$$\rho(r) = \frac{aM_{tot}}{2\pi r\,(r+a)^3} \tag{6}$$

where M_{tot} is the total mass of the system, and a is the scale radius. The Hernquist cumulative mass profile is found by integrating equation 6, and is given by

$$M(r) = M_{\text{tot}}\frac{r^2}{(r+a)^2}. \tag{7}$$

The probability density function (pdf) for the Hernquist model was found and presented by Hernquist (1990). In general, armed with a gravitational potential and mass density profile, one can follow the mathematical framework outlined in Binney and Tremaine (2008) to derive the pdf for the position $\boldsymbol{r}$ and velocity $\boldsymbol{v}$ of a particle in a such a model:

$$f(\mathcal{E}) = \frac{1}{\sqrt{8}\pi^2} \int_0^{\mathcal{E}} \frac{1}{\sqrt{\mathcal{E} - \psi}} \left(\frac{d^2\rho}{d\psi^2} \right) d\psi + \frac{1}{\sqrt{\mathcal{E}}} \left(\frac{d\rho}{d\psi} \right)_{\psi=0}, \tag{8}$$

where $\mathcal{E}$ is the energy per unit mass,

$$\mathcal{E}(r, v_r, v_t) = -\frac{v_r^2 + v_t^2}{2} + \Psi(r) \tag{9}$$

and where $\Psi(r)$ is the relative gravitational potential. Satellites that are not gravitationally bound to the Galaxy have $\mathcal{E} \leq 0$, and in this case $f(\mathcal{E} \leq 0) \equiv 0$.

Equation 8 is for a special case in which the particles in the model have an isotropic velocity dispersion. When the model has an anisotropic distribution of velocities, then the pdf is a function of both $\mathcal{E}$ and the angular momentum per unit mass $L = rv_t$ (Binney and Tremaine 2008). Finally, the pdf must also satisfy the condition that

$$\int f(\mathbf{r}, \mathbf{v}) d^3\mathbf{r} d^3\mathbf{v} = 1. \tag{10}$$

It should also be noted that in the physics and astronomy literature, f is sometimes referred to simply as the *distribution function* (DF) (e.g. see Binney and Tremaine 2008). The isotropic Hernquist DF is what we use as the term $p(h(\vartheta) = r, v_r, v_t|\boldsymbol{\theta})$ in equation 3 for this work.

3.3 Data: Dwarf Galaxies and Globular Clusters

The present paper uses the same satellite data as Eadie et al. (2015), which consists of kinematic data from 24 dwarf galaxies and 64 globular clusters. However, as mentioned in Section 2, we now use the Heliocentric data (R, v_{los}, $\boldsymbol{\mu}$). With 88 satellites, there are 354 parameters ϑ, in addition to the 2 Hernquist model parameters $\boldsymbol{\theta} = (a, M_{tot})$.

4. Results

Figure 1 shows preliminary results using our hierarchical Bayesian approach, compared to the results from the original method outlined in Eadie et al. (2015). The results from Eadie et al. (2015) are shown in faded colours, and the results from the current paper in vivid colours. The Bayesian credible regions in the analysis by Eadie et al. (2015) do not overlap with the current results beyond approximately 20 kpc. The M_{tot} estimate is $0.78 \times 10^{12} M_{\odot}$, with a 95% credible region of (0.69, 0.90)$\times 10^{12} M_{\odot}$. The credible regions for the mass profile are represented by different shades of teal in Figure 1.

5. Discussion

The preliminary results presented in Figure 1 are noticeably different from previous efforts. Interestingly, Eadie et al. (2015) showed that high-velocity objects, such as the globular cluster Pal 3, can have a significant affect on the mass estimate of the Galaxy. When they removed Pal 3 from the data and performed their analysis again, the mass estimate of the Galaxy decreased significantly. Pal 3's measurement uncertainty in $\mu_\alpha \cos\delta$ is 70% of its measured proper motion in right ascension, and its measurement uncertainty in μ_δ is 103% of its measured proper motion in declination. Therefore, it would not be surprising if Pal 3 carried less weight in the analysis once its measurement uncertainties were included.

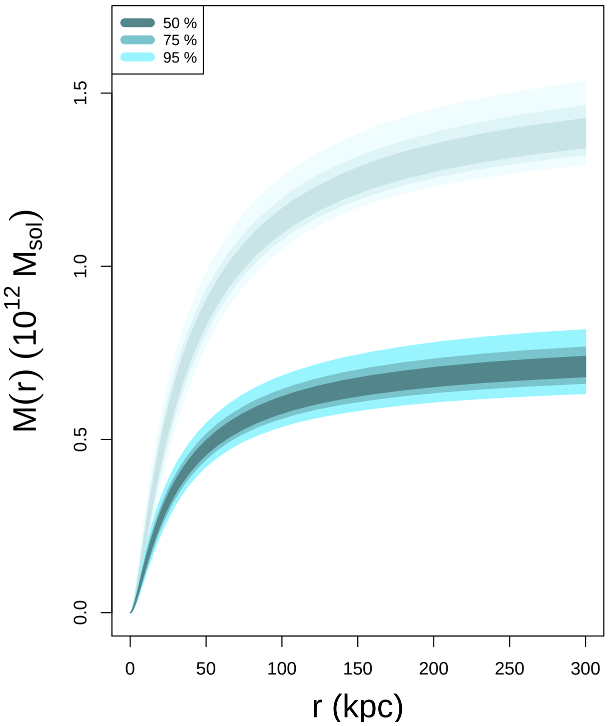

Figure 1: Comparison of the results from Eadie et al. (2015) (upper curve) with the results found here (lower curve), which include measurement uncertainties. Shown are the 95, 75, and 50 percent credible regions for each.

However, it should also be noted that the reduced mass profile shown in Figure 1 is the cumulative effect of using all measurement uncertainties in the analysis.

The transformations from the Heliocentric to Galactocentric reference frame may also be playing a role in the mass estimate. Eadie et al. (2015) used pre-determined Galactocentric measurements from many different studies. Although differences in $R_\odot$ are small between studies, and should not substantially affect these transformations, it is worth investigating to make sure this isn't the case.

The hierarchical method presented here could be improved. Although the measurement uncertainties in v_{los} and $\boldsymbol{\mu}$ are independent because they are measured using entirely different methods, $\Delta\mu_\delta$ and $\Delta\mu_\alpha \cos\delta$ are probably dependent. The measurements μ_δ and $\mu_\alpha \cos\delta$ are taken from the same set of images, and so the measurement uncertainties between μ_δ and $\mu_\alpha \cos\delta$ will be correlated. It would be beneficial to include their dependence in the analysis.

There are also other important Galaxy models to test that are popular in the literature, such as the empirical model proposed by Navarro et al. (1996). However, some theoretical problems remain with these empirical models (see Eadie et al. 2015, for more details).

Finally, it is important to consider the correlation between distance and measurement uncertainty. In general, the further away a satellite is, the more uncertain its distance and velocity measurements. Furthermore, satellites that are far away are more likely to be missing proper motion measurements, because more time is required to observe the motion across the plane of the sky. Thus, there is a selection bias in the proper motion measurements, and its effect on the mass estimate should be investigated with simulations.

6. Conclusion

We have built upon the work of Eadie et al. (2015) and introduced a hierarchical Bayesian technique that estimates the mass and mass profile of the Milky Way Galaxy while incorporating the measurement uncertainties of the data as well as incomplete data. The preliminary results presented here suggest that including measurement uncertainties can substantially change the mass estimate and mass profile of the Galaxy, and warrant further investigation.

7. Acknowledgement

The auhor thanks J. Cisewski for the invitation to present in the Astrostatistics session at JSM 2015. The author acknowledges funding from the Natural Sciences and Engineering Research Council of Canada, and from McMaster University.

References

An, J. H. and Evans, N. W. (2011), "Modified virial formulae and the theory of mass estimators," *Monthly Notices of the Royal Astronomical Society*, 413, 1744–1752.

Bahcall, J. N. and Tremaine, S. (1981), "Methods for determining the masses of spherical systems. I - test particles around a point mass," *The Astrophysical Journal*, 244, 805–819.

Battaglia, G., Helmi, A., Morrison, H., Harding, P., Olszewski, E. W., Mateo, M., Freeman, K. C., Norris, J., and Shectman, S. A. (2005), "The radial velocity dispersion profile of the Galactic halo: constraining the density profile of the dark halo of the Milky Way," *Monthly Notices of the Royal Astronomical Society*, 364, 433–442.

Binney, J. and Tremaine, S. (2008), *Galactic Dynamics*, Princeton, 2nd ed.

Boylan-Kolchin, M., Bullock, J. S., Sohn, S. T., Besla, G., and van der Marel, R. P. (2013), "The Space Motion of Leo I: The Mass of the Milky Way's Dark Matter Halo," *The Astrophysical Journal*, 768, 140.

Eadie, G. M., Harris, W. E., and Widrow, L. M. (2015), "Estimating the Galactic Mass Profile in the Presence of Incomplete Data," *The Astrophysical Journal*, 806, 54.

Gelman, A. and Rubin, D. (1992), "Inference from iterative simulation using multiple sequences," *Statistical science*, 7, 457–472.

Hernquist, L. (1990), "An analytical model for spherical galaxies and bulges," *The Astrophysical Journal*, 356, 359–364.

Johnson, D. R. H. and Soderblom, D. R. (1987), "Calculating galactic space velocities and their uncertainties, with an application to the Ursa Major group," *The Astronomical Journal*, 93, 864–867.

Little, B. and Tremaine, S. (1987), "Distant satellites as probes of our Galaxy's mass distribution," *The Astrophysical Journal*, 320, 493–501.

Navarro, J. F., Frenk, C. S., and White, S. D. M. (1996), "The structure of cold dark matter halos," *The Astrophysical Journal*, 462, 563–575.

Rubin, V. C., Ford, W. K. J., and . Thonnard, N. (1980), “Rotational properties of 21 SC galaxies with a large range of luminosities and radii, from NGC 4605 /R = 4kpc/ to UGC 2885 /R = 122 kpc/,” *The Astrophysical Journal*, 238, 471–487.

Wang, W., Han, J., Cooper, A. P., Cole, S., Frenk, C., and Lowing, B. (2015), “Estimating the dark matter halo mass of our Milky Way using dynamical tracers,” *Monthly Notices of the Royal Astronomical Society*, 453, 377–400.

Wilkinson, M. I., Evans, N. W., Perrett, K. M., and Bridges, T. M. (2003), “New mass estimators for tracer populations,” *The Astrophysical Journal*, 583, 752–757.

Zwicky, F. (1933), “Die Rotverschiebung von extragalaktischen Nebeln,” *Helvetica Physica Acta*, 6, 110–127.

— (1937), “On the Masses of Nebulae and of Clusters of Nebulae,” *The Astrophysical Journal*, 86, 217.