`TriOS` Schwarzschild Orbit Modeling: Robustness of Parameter Inference for Masses and Shapes of Triaxial Galaxies with Supermassive Black Holes

Jacob Pilawa[1] , Emily R. Liepold[1] , and Chung-Pei Ma[1,2]
[1] Department of Astronomy, University of California, Berkeley, CA 94720, USA; jacobpilawa@berkeley.edu
[2] Department of Physics, University of California, Berkeley, CA 94720, USA

Abstract

Evidence for the majority of the supermassive black holes in the local Universe has been obtained dynamically from stellar motions with the Schwarzschild orbit superposition method. However, there have been only a handful of studies using simulated data to examine the ability of this method to reliably recover known input black hole masses M_{BH} and other galaxy parameters. Here, we conduct a comprehensive assessment of the reliability of the triaxial Schwarzschild method at *simultaneously* determining M_{BH}, stellar mass-to-light ratio M^*/L, dark matter mass, and three intrinsic triaxial shape parameters of simulated galaxies. For each of 25 rounds of mock observations using simulated stellar kinematics and the `TriOS` code, we derive best-fitting parameters and confidence intervals after a full search in the 6D parameter space with our likelihood-based model inference scheme. The two key mass parameters, M_{BH} and M^*/L, are recovered within the 68% confidence interval, and other parameters are recovered between the 68% and 95% confidence intervals. The spatially varying velocity anisotropy of the stellar orbits is also well recovered. We explore whether the goodness-of-fit measure used for galaxy model selection in our pipeline is biased by variable complexity across the 6D parameter space. In our tests, adding a penalty term to the likelihood measure either makes little difference, or worsens the recovery in some cases.

Unified Astronomy Thesaurus concepts: Galaxy dynamics (591); Galaxy masses (607); Supermassive black holes (1663); Early-type galaxies (429); Galaxies (573); Galaxy dark matter halos (1880); Galaxy evolution (594); Galaxy kinematics (602)

1. Introduction

Since its introduction, the orbit superposition technique of Schwarzschild (1979) has become a useful tool for performing detailed dynamical modeling of the internal structures of galaxies. By integrating a set of representative stellar orbits in an assumed gravitational potential, a superposition of orbits is constructed to replicate the stellar kinematics and surface brightness profiles of a galaxy seen in projection. Schwarzschild's method was initially proposed to demonstrate the existence of self-consistent galaxies with triaxial mass distributions. It has since been extended to allow for fitting of kinematic and photometric observations (e.g., Pfenniger 1984; Richstone & Tremaine 1984, 1985; Rix et al. 1997) and is used to constrain properties of the host galaxy such as its supermassive black hole mass M_{BH}, stellar mass-to-light ratio M^*/L, dark matter content, intrinsic shape, and stellar orbital structure.

Despite the wide application of the Schwarzschild method to real data, its ability to recover known input parameters has only been tested in a handful of cases, each case with different underlying assumptions. Early tests of the Schwarzschild method on simulated data in the three-integral axisymmetric limit highlighted the advantages and challenges posed by the method when it was applied to long-slit kinematic data (e.g., Cretton & Emsellem 2004; Valluri et al. 2004; Magorrian 2006). The availability of integral-field spectroscopy (IFS) over the past two decades has led to significantly improved data quality and

number of kinematic constraints. An early application of axisymmetric orbit modeling to Sauron IFS data of M32 yielded well-constrained M_{BH}, M^*/L, as well as the inclination angle, $i = 70° \pm 5°$ (Verolme et al. 2002). A subsequent study of NGC 2974 (also using Sauron IFS data) obtained well-constrained M^*/L and i from axisymmetric modeling of real data, but tests on simulated data found the inclination to be marginally constrained; the authors cautioned the validity of the formal orbit model solution $i = 65° \pm 2°.5$ (Krajnović et al. 2005). Since then, other tests on simulated data using different axisymmetric orbit codes have found a range of results. For instance, Appendix A.3 of Siopis et al. (2009) and Appendix C of Quenneville et al. (2021) reported excellent recoveries of M_{BH} and halo circular velocity; Thomas et al. (2007) were able to use the orbit method to reconstruct the stellar masses and velocity anisotropies in *N*-body merger remnants with high accuracy; Vasiliev & Valluri (2020) recovered the true values of M^*/L but were unable to put strong constraints on M_{BH} and dark matter halos; Lipka & Thomas (2021) discussed the tendency of the inclination angle to be biased toward edge-on due to increased model flexibility.

When the axisymmetric assumption is removed, the Schwarzschild method becomes significantly more complicated and computationally expensive. Triaxial orbit modeling requires the specification of three unequal axes (or equivalently, three angles) and integration of new libraries of stellar orbits, so a full exploration of the galaxy parameter space is much more computationally intensive than for axisymmetric models. Accordingly, there have been fewer attempts at performing recovery tests on simulated data. Early tests on mock triaxial Abel galaxies recovered M^*/L and intrinsic axis lengths at 10%–20% accuracy (van den Bosch et al. 2008; van den Bosch & van de Ven 2009).

When the same triaxial code was applied to galaxies in hydrodynamical cosmological simulations, Jin et al. (2019) found that depending on the assumed halo profile, the mean stellar mass is underestimated by 13%–24% and the halo mass is overestimated by 18%–38%. All these tests, however, used the original 2008 code that had been shown to have 12 wrong signs in some components of the orbital velocities and various other issues (Quenneville et al. 2022). Neureiter et al. (2023) reported robust recovery of $M_{\rm BH}$ and M^*/L for several tested orientations of an N-body merger remnant.

In this paper, we conduct a comprehensive study of the ability of the triaxial orbit method to *simultaneously* recover the intrinsic shape and mass parameters of simulated triaxial galaxies containing SMBHs, stars, and dark matter. We have been working to reduce the computational cost of individual triaxial Schwarzschild models in the triaxial code TriOS (Quenneville et al. 2021, 2022), as well as improve the efficiency in exploring the mass and intrinsic shape parameter space via a non-grid-based likelihood approach to parameter inference. We apply this efficient parameter inference methodology to a suite of simulated triaxial galaxies, testing the ability of TriOS to recover known input parameters to our models while all parameters are varied simultaneously.

An additional aim of this paper is to investigate the extent to which our triaxial Schwarzschild orbit models are impacted by varying degrees of statistical complexity in the TriOS models. It has been a common practice while performing Schwarzschild modeling to compute a χ^2 value associated with a model's goodness of fit to a set of observed constraints, and select the *best-fit* model that minimizes the χ^2. While often effective, this strategy potentially introduces biases due to the variable complexity of the underlying models. Intuitively, if models in some region of the parameter space have more flexibility to fit the observations, those models may have lower χ^2 values as a simple consequence of that flexibility. A more prudent model selection procedure, therefore, should balance a combination of the fit quality and a measure of the complexity of a model.

It is, however, nontrivial to quantify the complexity of a model. The number of degrees of freedom (DOF) is often used for linear models without constraints because the number of nonredundant free parameters in a model provides a natural measure of the model complexity, or flexibility in overfitting. While subtleties exist in the use of DOF for unconstrained linear models, the problem of how to capture complexity is even thornier for linear models with constraints or priors (and hence with reduced model flexibility), or for nonlinear models (Andrae et al. 2010). Different forms of complexity measure, or information criterion, for the effective number of parameters in a model, have been proposed, e.g., Akaike information criterion (Akaike 1998), Takeuchi information criterion (Shibata 1989), Bayesian information criterion (Schwarz 1978), network information criterion (Murata & Yoshizawa 1994), and a Bayesian measure (Spiegelhalter et al. 2002). Ye (1998) suggests a generalized DOF that measures how sensitive the model predictions are to perturbations in the model constraints. A model with higher complexity or flexibility is one whose predictions are more responsive to those perturbations. Following Ye (1998), Lipka & Thomas (2021) constructed an estimate of the complexity of axisymmetric Schwarzschild orbit models. Here, we examine the role of penalty terms for triaxial orbit models in simulated galaxies.

The paper is laid out as follows. Section 2 summarizes the Schwarzschild orbit method and our procedure for constructing simulated galaxy kinematics. Section 3 outlines the main results of the paper, including a description of the parameter inference scheme used to compute our parameter estimates and corresponding confidence intervals. In Section 3, we briefly discuss estimating the statistical complexity of triaxial Schwarzschild models, and the extent to which taking into account the model flexibility changes our parameter estimation. Section 5 summarizes the results and the outlook for future triaxial stellar dynamical modeling efforts.

2. Dynamical Modeling and Simulated Data

2.1. Triaxial Schwarzschild Orbit Modeling

Throughout this work, we perform triaxial Schwarzschild orbit modeling with the TriOS code (Quenneville et al. 2021, 2022), which is based on an earlier unnamed triaxial modeling code (van den Bosch et al. 2008). In this method, a stationary gravitational potential is proposed, and a library of orbits spanning the phase space is integrated. A set of weights is assigned to these orbits such that the linear superposition of the orbits can both reproduce the mass distribution associated with the gravitational potential and fit a set of kinematic observations. This procedure enforces self-consistency between the mass distribution and gravitational potential without requiring specific assumptions about the form of the distribution function or velocity anisotropy.

The mass components of a galaxy in TriOS can consist of a central SMBH with mass $M_{\rm BH}$, a stellar component with a mass-to-light ratio M^*/L, and a dark matter halo. Most prior work using the Schwarzschild orbit method has assumed a spherical or axisymmetric potential; TriOS relaxes these assumptions and allows triaxially shaped 3D stellar mass densities. The stellar mass density is determined by first modeling the 2D (observed) surface brightness distribution of a galaxy as a multi-Gaussian expansion (MGE; Cappellari 2002). For a given set of angles θ, ϕ, and ψ that relate the intrinsic (3D) and projected (2D) coordinate systems, we deproject each component of the MGE and sum them to create the 3D stellar density. Each MGE component (labeled by subscript j) is allowed to have its own 2D axis ratio, q'_j, which is the ratio of the lengths of its semiminor and semimajor axes. Through the relations in the appendix of Quenneville et al. (2022), the intrinsic axis ratios are determined: $p_j = b_j/a_j$ is the intrinsic middle-to-long axis ratio; $q_j = c_j/a_j$ is the intrinsic short-to-long axis ratio; and u_j is the apparent-to-intrinsic long axis ratio. When performing parameter searches and inferences below, we follow Quenneville et al. (2022) and use an alternative set of shape parameters, T, $T_{\rm maj}$, and $T_{\rm min}$. The definitions and advantages of these parameters are articulated in Section 3 of Quenneville et al. (2022).

For each galaxy model, the TriOS code samples the phase space with a set of representative stellar orbits and integrates their trajectories. The orbits are initiated in two separate spaces: the x–z and the stationary start spaces, where the x-, y-, and z-axes are directed along the intrinsic major, intermediate, and minor axes of the galaxy, respectively. In each space, N_E values of the orbital energy are sampled. In the x–z start space, orbits of a given energy are initialized on a polar grid in the x–z plane along N_{I_2} rays uniformly spaced in angle from the z-axis to the x-axis, and at N_{I_3} positions along each ray (see Figure 4 of

Quenneville et al. 2022). Each orbit has initial velocities of $v_x = v_z = 0$ and $v_y > 0$. In addition, a set of *retrograde* orbits is constructed by simply inverting the velocities of orbits in the prograde x–z orbit library, and seven additional copies of each orbit are generated through the mirroring scheme described in Quenneville et al. (2022). In the stationary start space, orbits of a given energy are initiated with zero velocities at positions on the equipotential surface sampled on a uniform grid over the two spherical angles θ and ϕ. To enhance phase-space sampling, the code allows for *dithering* the initial conditions, by which the number of sample points along each of the three dimensions of the start space is increased by a factor of $N_{\rm dither}$, resulting in an overall factor of $N_{\rm dither}^3$ increase in the phase-space sampling. The properties of bundles of $N_{\rm dither}^3$ adjacent orbits are averaged and each bundle of orbits is given a single orbital weight when constructing the superposition.

Each orbit is projected onto the sky, and its line-of-sight velocity distribution (LOSVD) within $N_{\rm bin}$ spatial apertures is stored. Its mass distribution in a 3D grid is also stored. We perform a non-negative least squares (NNLS) optimization to determine a set of orbit weights such that the superposition of orbits reproduces both the 3D stellar mass distribution (obtained from deprojection of observed surface brightness distribution) and the observed LOSVD in each spatial bin. The LOSVD is characterized by a set of Gauss–Hermite moments (van der Marel & Franx 1993), and the NNLS objective function consists primarily of the χ^2 associated with the reproduction of those observed moments (see Equation (1)). The models contain $N_{\rm kin}$ kinematic constraints, which are equal to the product of the number of spatial apertures and the number of Gauss–Hermite moments used to characterize the LOSVDs. The mass self-consistency is enforced by adding additional χ^2-like terms to the NNLS objective function that are associated with the mismatch between the input's and model's mass in each of the 3D grid cells and within each of the 2D spatial apertures with a 1% error imposed on each mass. The phase-space distribution function for the galaxy model can be understood as being composed of this set of orbit weights.

2.2. Generation of Simulated Galaxy Data

We use the Schwarzschild method to create simulated stellar kinematics for a suite of model galaxies, each of known mass components and a known intrinsic shape. For the simulated data to provide useful insight into actual observations, we opt to create stellar kinematics that mimic those observed in real galaxies. We achieve this by building the simulated galaxy kinematics using the measured kinematics of a real galaxy as a template. Here, we use the kinematics of NGC 2693 presented in Pilawa et al. (2022), which are typical of those known for massive elliptical galaxies in the MASSIVE survey (Ma et al. 2014). Since these kinematics are only used as a starting template, we do not expect the main conclusions of this paper to depend strongly on the particular galaxy choice. We note that while the velocity anisotropy profiles investigated here tend to be tangential at small radii and radial at large radii (Figure 6 of Pilawa et al. 2022; Figure 3 here), this profile is not unique to NGC 2693. Similar profiles are found for many massive galaxies that have been studied with the orbit modeling method (e.g., Extended Data Figure 6 of Thomas et al. 2016). A different procedure would be needed to produce a qualitatively different velocity anisotropy profile.

Table 1
Mass and Shape Parameters for the Five Triaxial Galaxy Models Tested in This Work

Model	G_1	G_2	G_3	G_4	G_5
$M_{\rm BH}$ [$10^9 \, M_\odot$]	2.1	1.7	1.0	2.4	1.3
M^*/L [$M_\odot/L_\odot$]	2.23	2.31	2.43	2.25	2.28
M_{15} [$10^{11} \, M_\odot$]	5.5	8.9	10.2	5.5	6.0
T	0.39	0.47	0.46	0.49	0.41
$T_{\rm maj}$	0.04	0.17	0.26	0.25	0.10
$T_{\rm min}$	0.07	0.09	0.06	0.05	0.09
u	0.997	0.982	0.971	0.970	0.991
p	0.909	0.886	0.885	0.878	0.902
q	0.746	0.733	0.729	0.731	0.739

Note. $M_{\rm BH}$ is the SMBH mass, M^*/L is the stellar mass-to-light ratio, M_{15} is the dark matter mass enclosed within a radius of 15 kpc, and T, $T_{\rm maj}$, and $T_{\rm min}$ are the three shape parameters specifying the triaxial potential. The axial ratios u, p, and q are related to the shape parameters and are computed here by taking luminosity-weighted averages over the MGE components (see the text).

Six parameters were determined for NGC 2693: SMBH mass $M_{\rm BH}$, the stellar mass-to-light ratio M^*/L, the dark matter mass enclosed within a radius of 15 kpc M_{15}, and three shape parameters (T, $T_{\rm maj}$, $T_{\rm min}$), or equivalently, the luminosity-averaged axis ratios (p, q, u). The dark matter halo is logarithmic with a fixed scale radius of $R_c = 15$ kpc. To generate simulated stellar kinematics for a given galaxy model G_i listed in Table 1, we first run TriOS to compute the orbit libraries representing this galaxy and determine the set of orbit weights with a regularization scheme (see below) such that the model's projected kinematics best reproduce those of real observations of NGC 2693. We then perturb this model's noise-free projected kinematics with Gaussian noise with seed r and amplitude set by the observed uncertainties on each measurement.

While determining the orbital weights of the simulated galaxies described above, we incorporate the distribution function regularization scheme outlined in Section 5.2 of van den Bosch et al. (2008). This scheme penalizes the goodness-of-fit statistic when fitting the orbital weight distribution by a term proportional to the second derivative of the orbital weights with respect to the orbits' indices in their start space. This procedure yields distribution functions that are substantially smoother in phase space than can be obtained in models without regularization. The stellar masses in regularized models are represented by many more orbits than in nonregularized models. For example, for the galaxy model that best reproduced the data in Pilawa et al. (2022), nonregularized models had 99%, 99.9%, and 99.99% of the mass in 298, 365, and 397 of the 14,040 orbits in the library, respectively; while regularized models have the same percentages of the mass represented by ~5100, ~6400, and ~7200 orbits, respectively. Here, we use regularization as a way to generate simulated galaxy kinematics that are represented by a wide variety of orbital weights. In the tests below, we will assess the ability of the TriOS code to recover the input galaxy parameters without using regularization, as is commonly done when the Schwarzschild method is applied to real data to measure $M_{\rm BH}$ and other parameters.

In this work, we generate simulated kinematics for five different locations in the 6D parameter space, which are labeled G_1–G_5 and listed in Table 1. These parameter values are chosen to span a broad range in each dimension, but narrow enough that the models can be explored with a reasonably sized

collection of orbit models. This allows us to test the ability of the orbit code to recover model parameters over a wide range of physically reasonable values. We recall that the axis ratios (p, q, u) must obey the inequality $0 \leqslant q \leqslant uq' \leqslant p \leqslant u \leqslant 1$, where $q' = b'/a'$ is the observed axis ratio of the 2D projected galaxy shape (e.g., Quenneville et al. 2022), and only a subspace of (p, q, u) has 3D shapes that could produce the required flattening q' upon projection. The range of allowed (p, q, u) is therefore typically quite narrow.

For each galaxy model G_i, we perform five draws of random Gaussian noise and add it to the galaxy model's noise-free projected kinematics. The amplitudes of the injected noise are chosen to resemble those inherent in real data, so the five realizations mimic repeated observations of the same galaxy and allow us to assess the uncertainties in the recovery test for each of the simulated galaxies. The different noise realizations are labeled r_1–r_5.

In total, we have 25 sets of simulated stellar kinematics for five models of realistic galaxies. Next, we treat each of the 25 sets of kinematics as a mock observation of a galaxy, and perform recovery tests over the full 6D parameter space to search for the best-fit parameters.

3. Parameter Inference and Recovery Test

In this section, we present the results of the 25 recovery tests and compare the inferred model parameters against the known input parameters of each simulated galaxy to assess how well the procedure recovers the model parameters. The computational cost is substantial and is akin to performing triaxial Schwarzschild orbit modeling and parameter inference for 25 galaxies.

For each of the five realizations for a given galaxy model in Table 1, we treat the mock galaxy's projected kinematics as simulated data and perform a 6D parameter space search following the grid-free procedure we have developed for recent analyses of real data (e.g., Pilawa et al. 2022; Quenneville et al. 2022; Liepold et al. 2023). In this procedure, a grid-free Latin hypercube scheme is used to choose sampling points in the galaxy model parameter space (about 3000 models in this work). For each sampled model, we run TriOS and compute a goodness-of-fit metric associated with the model predictions compared to the input kinematics. To mimic the procedure used in orbit modeling of real data, we do not use regularization in this step (see discussion in Section 2.2). The metric, hereafter referred to as the log-likelihood, is given by

$$-2 \ln \mathcal{L}(\boldsymbol{\mu}) = \chi^2_{\text{kin}}(\boldsymbol{\mu}) = \sum_j^{N_{\text{bin}}} \sum_i^{N_{\text{GH}}} \frac{(h_{ij}(\boldsymbol{\mu}) - h_{ij,\text{input}})^2}{\Delta h^2_{ij,\text{input}}}, \quad (1)$$

where h_{ij} is the ith Gauss–Hermite moments of the stellar LOSVDs in the jth spatial bin, Δh_{ij} is the associated measurement uncertainty, and $\boldsymbol{\mu}$ is the set of six parameters describing the galaxy's potential. For the template galaxy NGC 2693 used in this work, Gemini GMOS IFS yielded 60 spatial bins, each with eight Gauss–Hermite moments, for the central $5'' \times 7''$ region of the galaxy, and the McDonald–Mitchell IFS yielded 29 bins, each with six moments, for the outer part of the galaxy (Pilawa et al. 2022); together, the total number of kinematic constraints is $N_{\text{kin}} = 654$.

Adopting the parameter inference procedure used in our prior work (Pilawa et al. 2022; Quenneville et al. 2022; Liepold et al. 2023), we first construct an interpolated log-likelihood

surface from the discrete set of evaluated models using Gaussian process regression (Rasmussen & Williams 2005) with a Matérn covariance kernel with $\nu = 3/2$. The resulting Gaussian process mean function is used as a smooth surrogate function for the true log-likelihood surface. The dynamic nested sampler dynesty (Speagle 2020) is then used to sample this smooth log-likelihood function and to produce a posterior assuming uniform priors on each parameter.

A typical outcome of our recovery tests is shown in Figure 1, where the 1D and 2D marginalized posteriors for the six model parameters in realization r_5 of model G_4 are plotted. The true input values of the model parameters (orange circles) are well recovered, lying within the 95% confidence interval of the best-fit model for all six parameters.

The results for all 25 recovery tests are summarized in Figure 2. The input parameter values are shown in the legend of each panel and in Table 1. Overall, these results show an excellent recovery of the input parameters for (i) individual mock realizations (small dots), (ii) when averaging over five realizations for each galaxy model G_i (big circles), and (iii) when averaging over all 25 tests (big square). In particular, the two key mass parameters—black hole mass and stellar mass-to-light ratios—are consistently recovered within the 68% confidence interval. For other parameters, while there are some variations across different realizations and galaxy models, the gray squares indicate that the *average* bias is consistent with 0 for all parameters.

In addition to recoveries of galaxy model parameters, we can also assess the ability of the TriOS code to recover the intrinsic properties of the simulated galaxies. One such key property is the velocity anisotropy, typically parameterized by $\beta \equiv 1 - \sigma_t^2/\sigma_r^2$, where σ_r is the radial velocity dispersion, and σ_t is the tangential velocity dispersion with $\sigma_t^2 \equiv (\sigma_\theta^2 + \sigma_\phi^2)/2$. The value $\beta = 0$ indicates velocity isotropy, while a negative (positive) β indicates tangential (radial) anisotropy.

Figure 3 shows the velocity anisotropy of the five simulated galaxies (black-dashed curve in each panel) is preferentially tangential ($\beta < 0$) near the centers of the galaxies and preferentially radial ($\beta > 0$) beyond ~1 kpc. The β profiles from the five lowest χ^2 models (one for each r_i) in our recovery tests are shown as a solid color band in each panel, with the band corresponding to the mean and standard deviation in β at a given radius. We include $\beta = 0$ as the dotted gray line to guide the eye. Our tests recover the simulated β profiles quite well, with the largest discrepancies near the innermost and outermost portions of the galaxy. This trend in the innermost region has been noticed in anisotropy recovery studies before, but has been attributed to a number of different factors, including the relative paucity of orbits/mass in this innermost region (Breddels et al. 2013; Kowalczyk et al. 2017).

One feature in Figure 2 is the tendency toward over-estimation of the dark matter mass parameter M_{15} in three of five galaxy models (G1, G4, G5) at the ~1σ level. While this trend could be simply due to the small number of statistics, there are factors that could make the halo parameter less well constrained than other parameters. In general, the constraint on the halo is driven by the outermost portion of the kinematic data since this is the region in which the enclosed dark matter mass starts to become comparable to that of the stars. Due to the rapidly decreasing surface brightness in the outer regions of galaxies, we must coadd stellar spectra over a much larger area of the sky to achieve a reasonable signal-to-noise ratio. Furthermore, our outer data were obtained from the wide-field Mitchell IFS on a 2.7 m
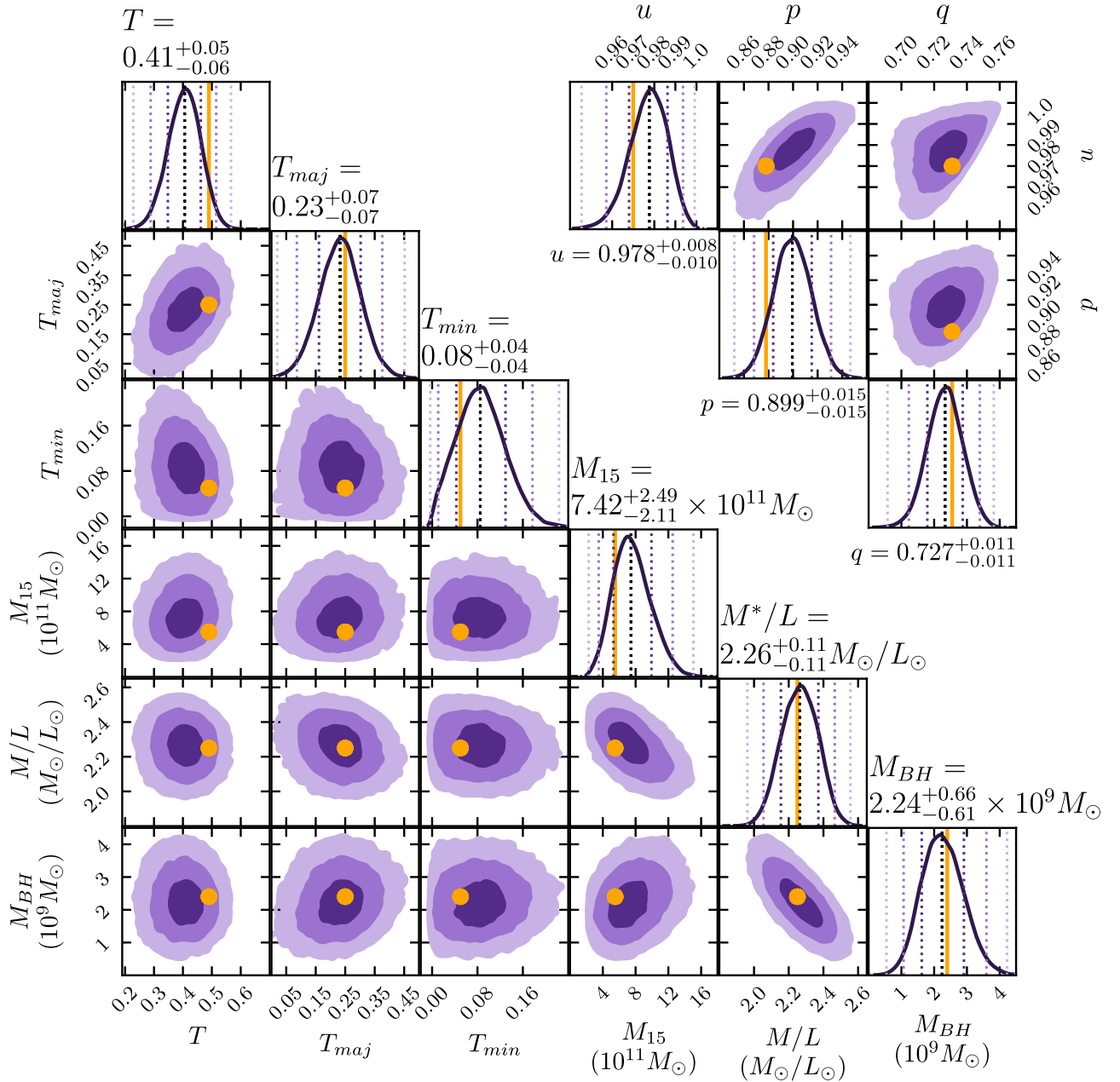
Figure 1. (Lower left) 1D and 2D marginalized posteriors of the six model parameters in an example recovery test (realization r_5 of model G_4). The 68%, 95%, and 99% confidence interval contours are represented by the different shades of purple in the 2D panels, and the 1D marginalized posteriors are shown in the 1D panels. The true input value for each parameter is represented by the orange-filled circle and orange vertical line in each panel. (Upper right) 1D and 2D marginalized posteriors in the axis–ratio space of (p, q, u), computed from the (luminosity-weighted) posteriors of $(T, T_{\mathrm{maj}}, T_{\mathrm{min}})$.

telescope, while the inner data were from the high-resolution GMOS on an 8.1 m telescope. The larger uncertainties in the outermost kinematic data points thus allow for larger perturbations when we generate the simulated galaxy kinematics. In addition, our recipe for computing the simulated Gauss–Hermite moments thus far does not take into account the correlations in the *uncertainties* in the moments (Houghton et al. 2006). These factors together can potentially bias the halo parameter more than the other parameters in our model.

4. Model Complexity and Impact of Penalty Terms

As discussed in Section 1, in the framework of classical statistic modeling, increasing the complexity of a model generally results in a better goodness-of-fit statistic. If the complexity is variable across models, the apparent goodness of fit may be biased by the variable complexity. In this section, we examine this issue by adding a term to our log-likelihood function of Equation (1) in the form of

$$-2\ln\mathcal{L}(\boldsymbol{\mu}) \rightarrow -2\ln\mathcal{L}(\boldsymbol{\mu}) + \kappa(\boldsymbol{\mu}), \qquad (2)$$

where κ is a *penalty* term used to capture a model's complexity. Various forms of κ have been proposed in the literature (see Section 1) to penalize the models in the parameter space with more complexity and alleviate any resultant biases in the inferred model parameters.

Intuitively, κ can be interpreted as serving a similar role as the Bayesian prior probability distribution. Just as the prior changes the shape of the posterior probability distribution given
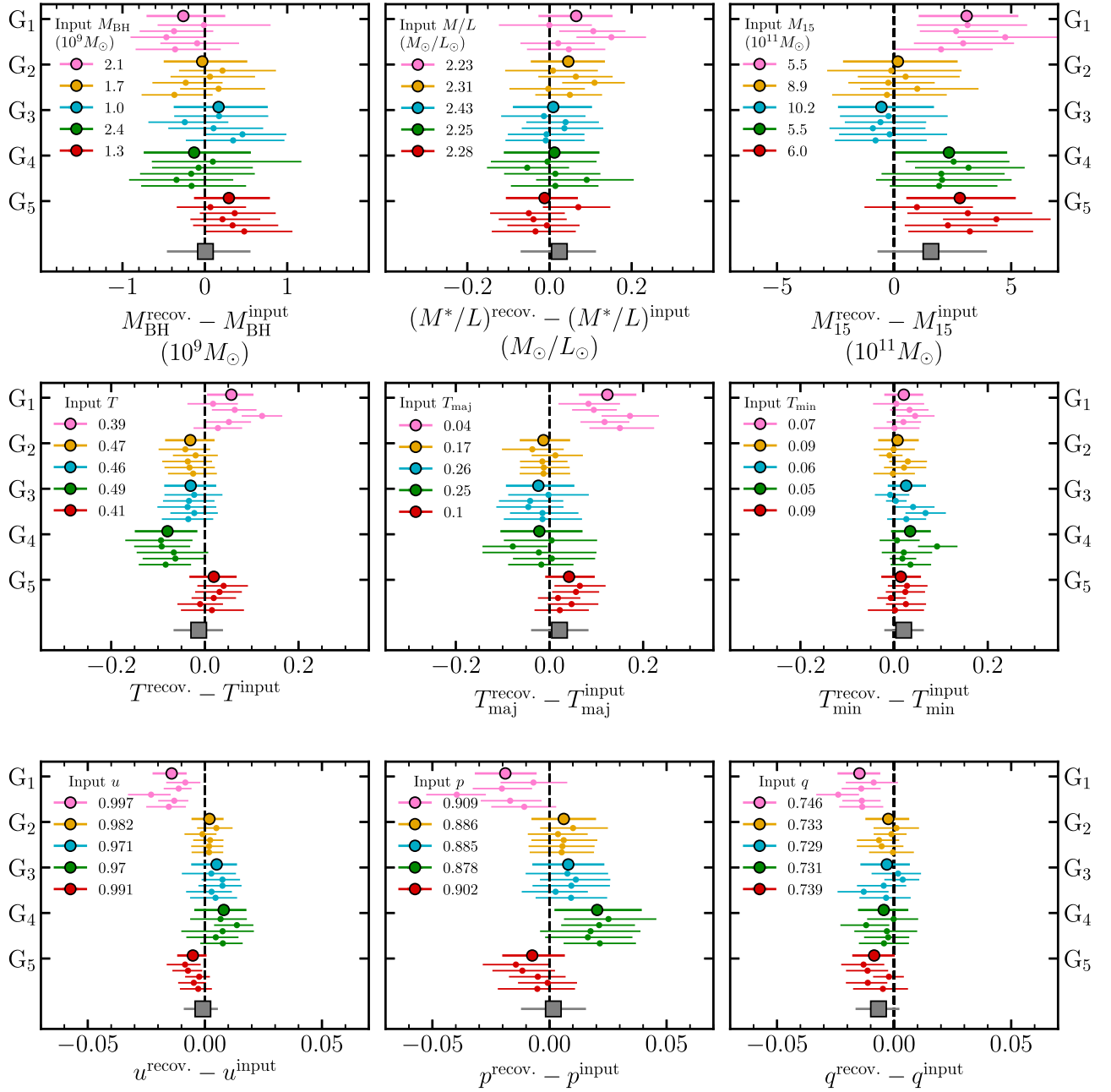
Figure 2. Results of the 25 rounds of recovery tests for the galaxy model parameters. The upper three panels denote the supermassive black hole mass M_{BH}, stellar mass-to-light ratio M^*/L, and enclosed dark matter mass at 15 kpc. The middle panels show the recovered intrinsic shape parameters T, T_{maj}, and T_{min}, and the bottom panels show the same shape recovery but for the more intuitive luminosity-averaged axis ratios u, p, and q. In each panel, the difference between the input parameter value (listed in the legend) and the recovered value is shown; the black-dashed vertical line indicates the exact recovery of that parameter. The five galaxy models G_1–G_5 are grouped by colors. Within each color group, the small dot denotes the recovery result for one noise realization of model G_i, where the error bars represent the 68% credible region of the posterior samples (see Figure 1 for an example). The larger filled circle for each G_i model is the mean recovered value for the five realizations, where the error bars denote the median uncertainty of the realizations. The large gray square at the bottom of each panel shows the mean over all 25 recovery tests, where the error bars denote the median value of the uncertainties of the 25 tests.

some previous knowledge of the parameter space, the shape of the penalty term $\kappa(\boldsymbol{\mu})$ adjusts the goodness-of-fit landscape in our study, $\chi^2_{\mathrm{kin}}(\boldsymbol{\mu})$, to correct for variability in model complexity as a function of the parameter space location $\boldsymbol{\mu}$. In regions in which the model complexity κ is high, the goodness-of-fit χ^2_{kin} should tend to be lower, and the relative trade-off of these two quantities yields the final inferred locations by subtly displacing the minimum of the modified log-likelihood landscape.

Here, we consider the complexity measure proposed in Ye (1998) and adopted in Lipka & Thomas (2021). Those works

compute a generalized DOF that is related to the ability of the model to respond to small perturbations in the model constraints. Intuitively, a model with a higher complexity (or flexibility) will be more responsive to those small perturbations, while a simple (or inflexible) model will be less responsive. In our application, the constraints are given by the observed stellar kinematic moments and the goodness of fit is described by the log-likelihood. To measure the flexibility of a given model, we perturb the stellar kinematic predictions from that model a refit the model given those perturbed kinematics. The complexity (or flexibility) of the model is related to the
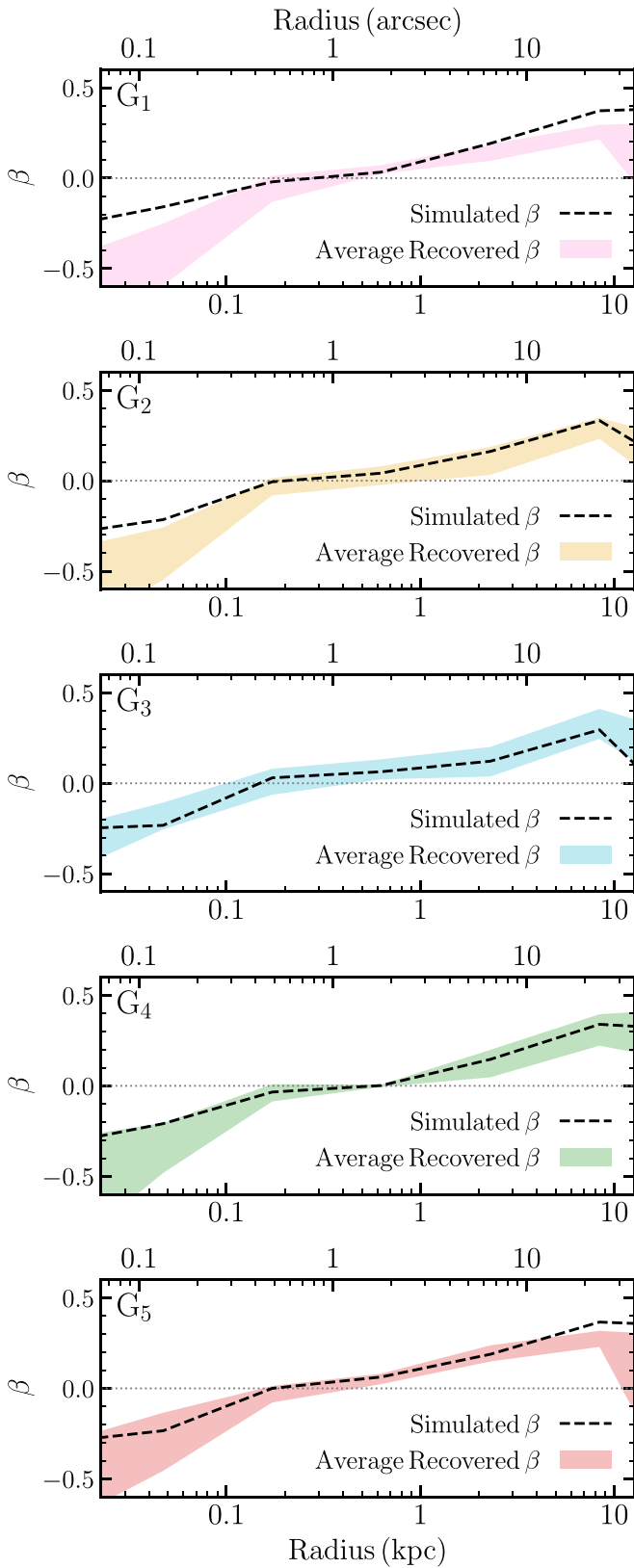
Figure 3. Profile of the velocity anisotropy parameter, $\beta \equiv 1 - \sigma_t^2/\sigma_r^2$, for the five simulated galaxies G_1–G_5 (top to bottom; black-dashed curves). At each radius, the colored band indicates the standard deviation of β from the recovery tests of the five realizations for the given galaxy model.

empirically connects variations in the data directly to variations in the goodness of fit on a model-by-model basis without making significant assumptions about the form of the modeling.

The specific steps used to compute the penalty term are as follows. We first fit a set of galaxy kinematics using a Schwarzschild model with parameters $\boldsymbol{\mu}$. We then perturb this model's kinematic moment predictions using Gaussian-distributed random values whose amplitudes are equal to the original measurement errors. We generate 20 realizations of these perturbations. For each realization, a χ^2-like term $\chi^2_{\mathrm{prior,i}}$ is computed, which describes the difference between the perturbed and unperturbed moments. This quantity only depends on the number of kinematic moments that have been perturbed, with $\langle \chi^2_{\mathrm{prior,i}} \rangle = N_{\mathrm{kin}}$. We then reoptimize the orbital weights in the Schwarzschild model to fit these perturbed kinematics, resulting in a goodness-of-fit $\chi^2_{\mathrm{posterior,i}}(\boldsymbol{\mu})$. The difference in these two quantities is taken to be a measure of the complexity with $\chi^2_{\mathrm{prior}} - \chi^2_{\mathrm{posterior}}(\boldsymbol{\mu}) \equiv m_{\mathrm{eff}}(\boldsymbol{\mu})$, appropriately averaged across the 20 realizations. For a given set of galaxy model parameters, this procedure is identical to generating a simulated galaxy with those parameters, and then fitting the simulated data with the same set of parameters. This is repeated for each of the ~3000 models in our sample of orbit models used for the analyses in the previous section.

The recovered parameters for the five simulated galaxies using the original likelihood $\mathcal{L}$ (circles; same as in Figure 2) versus the penalized likelihood (squares) are shown in Figure 4. Each symbol represents the mean value (relative to the true value) over the five noise realizations, and the uncertainty is taken to be the median uncertainty of the realizations. Overall, there are only small differences in the resulting best-fit parameters and confidence intervals, and the average recovered values and median uncertainties are all consistent with one another at the 68% confidence level. The surface of our penalty term as a function of galaxy parameters is generally quite flat and shows no consistent structure. In some cases in Figure 4, adding the penalty term, in fact, moves the recovered median value away from the true value. This first exploration of model complexity for triaxial models, therefore, indicates that both the mass and shape parameters of a simulated triaxial galaxy are well recovered using our likelihood function without a penalty term.

When axisymmetry is imposed in orbit models, Lipka & Thomas (2021) find their penalty term to be essentially monotonically increasing with increasing galaxy inclination angle i. The source of this dependence was attributed to the orbital structures of an axisymmetric potential as follows. A fair phase-space sampling of an axisymmetric galaxy includes both a prograde and a retrograde copy of every orbit. In the edge-on axisymmetric limit, these two orbits have opposite signs in the line-of-sight velocities and can contribute maximally to the model LOSVDs. In the face-on limit, on the other hand, the velocities of these two orbits are perpendicular to the line of sight, and the prograde and retrograde orbits contribute essentially identically to the resultant LOSVD. The inclination-dependent degeneracy between the prograde and retrograde orbits yields an effectively smaller number of distinct orbits for oblique and maximally face-on inclinations than edge-on inclinations, resulting in models with edge-on inclinations having more flexibility to produce good fits to the input data.

Triaxial potentials admit a larger suite of orbital types with more complicated structures and less symmetry than
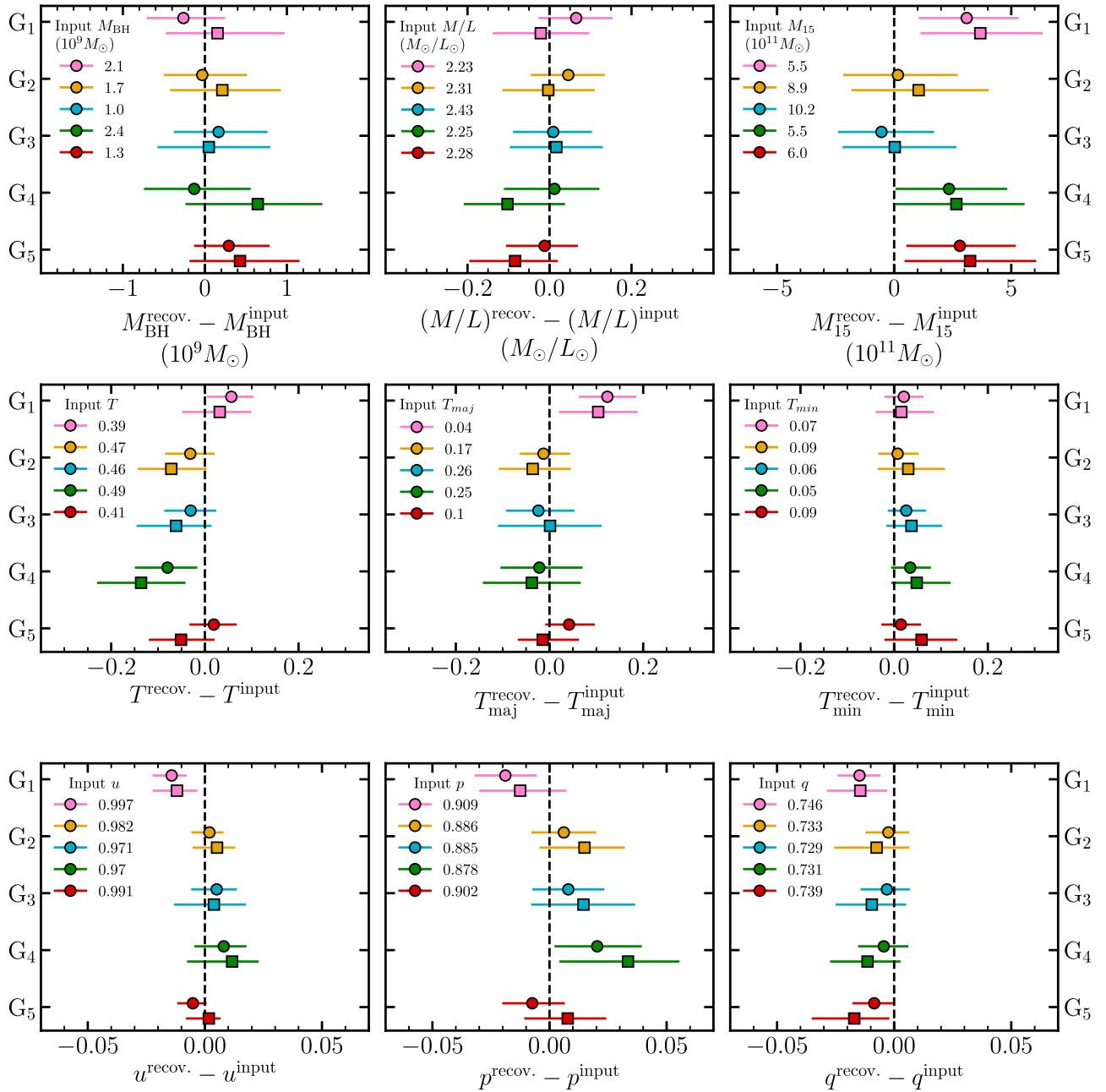
amount of improvement in the log-likelihood after this refitting. For flexible models, the improvement is large, while for inflexible models, the improvement is smaller. This scheme

Figure 4. Effect of adding a penalty term to the log-likelihood function (see Equation (2)) in the recovery tests. The filled circles are identical to those in Figure 2 and denote the mean difference between the recovered and true parameter values of the five noise realizations for each galaxy model G_i. The corresponding result after including a penalty term is shown as squares. A penalty term leads to no improvement in parameter recovery.

axisymmetric potentials. The inclination angle for an axisymmetric galaxy is replaced by three angles, or equivalently, three shape parameters (e.g., van den Bosch et al. 2008; Quenneville et al. 2022). While adding a penalty term improved the recovery of the inclination angle in axisymmetric mock models, we do not find such a term to be necessary for triaxial models.

5. Conclusion

We have presented a study on the accuracy and reliability of recovering known sets of input galaxy mass and shape parameters using the triaxial Schwarzschild orbit modeling code TriOS. We generated 25 sets of simulated galaxy kinematics and performed a full search in the 6D parameter space using our likelihood-based inference scheme to derive best-fitting parameters and confidence intervals. Treating the goodness of fit as proportional to a Bayesian log-likelihood, we produced posterior probability distributions for each parameter using dynamic nested sampling. The computational cost of this study is similar to that of performing full-scale triaxial Schwarzschild orbit modeling and parameter inference for 25 galaxies.

The outcome of this study indicates robust recovery of both the parameters defining the galaxy potential (Figures 1 and 2) and internal orbital structures such as the stellar velocity anisotropy profiles (Figure 3). In particular, the two key mass parameters in the galaxy models—black hole mass and the stellar mass-to-light ratio—are always recovered within the 68% confidence interval of the true values.

We additionally estimated the statistical model complexity of the triaxial models and tested the impact of this penalty term on the resulting parameter estimation. In contrast to recent tests of axisymmetric orbit models, we find that penalty terms quantifying the varying model-to-model complexity do not appreciably alter the shape of the likelihood landscape, nor do they significantly impact the location of the best-fitting models (Figure 4). At best, the posteriors obtained with the inclusion of a penalty term reproduce the posteriors obtained without a penalty term; at worst, the addition of a penalty term artificially increases the sizes of the confidence intervals, weakening the overall statistical power of the models and parameter inference scheme.

Taken together, our results suggest that triaxial Schwarzschild orbit modeling with the `TriOS` code, when combined with dense sampling of the 6D parameter space and our parameter inference scheme, can *simultaneously* recover known mass parameters (for black holes, stars, and dark matter) as well as triaxial galaxy shape parameters without the need of a penalty term to compensate for model complexities.

Acknowledgments

ORCID iDs

Jacob Pilawa ⓘ https://orcid.org/0000-0001-7040-9117
Emily R. Liepold ⓘ https://orcid.org/0000-0002-7703-7077
Chung-Pei Ma ⓘ https://orcid.org/0000-0002-4430-102X

References

Akaike, H. 1998, in Selected Papers of Hirotugu Akaike, ed. E. Parzen, K. Tanabe, & G. Kitagawa (New York: Springer), 199
Andrae, R., Schulze-Hartung, T., & Melchior, P. 2010, arXiv:1012.3754
Astropy Collaboration, Price-Whelan, A. M., Sipőcz, B. M., et al. 2018, AJ, 156, 123
Astropy Collaboration, Robitaille, T. P., Tollerud, E. J., et al. 2013, A&A, 558, A33
Breddels, M. A., Helmi, A., van den Bosch, R. C. E., van de Ven, G., & Battaglia, G. 2013, MNRAS, 433, 3173
Cappellari, M. 2002, MNRAS, 333, 400
Cretton, N., & Emsellem, E. 2004, MNRAS, 347, L31
Harris, C. R., Millman, K. J., van der Walt, S. J., et al. 2020, Natur, 585, 357
Houghton, R. C. W., Magorrian, J., Sarzi, M., et al. 2006, MNRAS, 367, 2
Hunter, J. D. 2007, CSE, 9, 90
Jin, Y., Zhu, L., Long, R. J., et al. 2019, MNRAS, 486, 4753
Kowalczyk, K., Łokas, E. L., & Valluri, M. 2017, MNRAS, 470, 3959
Krajnović, D., Cappellari, M., Emsellem, E., McDermid, R. M., & de Zeeuw, P. T. 2005, MNRAS, 357, 1113
Liepold, E. R., Ma, C.-P., & Walsh, J. L. 2023, ApJL, 945, L35
Lipka, M., & Thomas, J. 2021, MNRAS, 504, 4599
Ma, C.-P., Greene, J. E., McConnell, N., et al. 2014, ApJ, 795, 158
Magorrian, J. 2006, MNRAS, 373, 425
Murata, N., Yoshizawa, S., & Amari, S.-I. 1994, ITNN, 5, 865
Neureiter, B., de Nicola, S., Thomas, J., et al. 2023, MNRAS, 519, 2004
Pfenniger, D. 1984, A&A, 134, 373
Pilawa, J. D., Liepold, E. R., Delgado Andrade, S. C., et al. 2022, ApJ, 928, 178
Quenneville, M. E., Liepold, E. R., & Ma, C.-P. 2021, ApJS, 254, 25
Quenneville, M. E., Liepold, E. R., & Ma, C.-P. 2022, ApJ, 926, 30
Rasmussen, C. E., & Williams, C. K. I. 2005, Gaussian Processes for Machine Learning (Cambridge, MA: The MIT Press),
Richstone, D. O., & Tremaine, S. 1984, ApJ, 286, 27
Richstone, D. O., & Tremaine, S. 1985, ApJ, 296, 370
Rix, H.-W., de Zeeuw, P. T., Cretton, N., van der Marel, R. P., & Carollo, C. M. 1997, ApJ, 488, 702
Schwarz, G. 1978, AnSta, 6, 461
Schwarzschild, M. 1979, ApJ, 232, 236
Shibata, R. 1989, in From Data to Model, ed. J. C. Willems (Berlin: Springer), 215
Siopis, C., Gebhardt, K., Lauer, T. R., et al. 2009, ApJ, 693, 946
Speagle, J. S. 2020, MNRAS, 493, 3132
Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. 2002, J. R. Stat. Soc. B, 64, 583
Thomas, J., Ma, C.-P., McConnell, N. J., et al. 2016, Natur, 532, 340
Thomas, J., et al. 2007, MNRAS, 382, 657
Valluri, M., Merritt, D., & Emsellem, E. 2004, ApJ, 602, 66
van den Bosch, R. C. E., van de Ven, G., Verolme, E. K., Cappellari, M., & de Zeeuw, P. T. 2008, MNRAS, 385, 647
van den Bosch, R. C. E., & van de Ven, G. 2009, MNRAS, 398, 1117
van der Marel, R. P., & Franx, M. 1993, ApJ, 407, 525
Vasiliev, E., & Valluri, M. 2020, ApJ, 889, 39
Verolme, E. K., Cappellari, M., Copin, Y., et al. 2002, MNRAS, 335, 517
Ye, J. 1998, J. Am. Stat. Assoc., 93, 120