DIFFERENTIATING BETWEEN DIFFUSIVE MOVEMENT PATTERNS IN SINGLE PARTICLE TRACKING EXPERIMENTS

AN HONORS THESIS

SUBMITTED ON THE 4TH DAY OF MAY, 2021

TO THE DEPARTMENT OF MATHEMATICS

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

OF THE HONORS PROGRAM

OF NEWCOMB TULANE COLLEGE

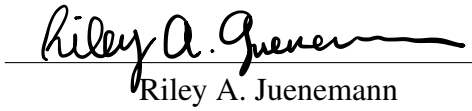
TULANE UNIVERSITY

FOR THE DEGREE OF

BACHELOR OF SCIENCE

WITH HONORS IN MATHEMATICS

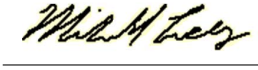
BY



Riley A. Juenemann

APPROVED:

Scott A. McKinley, Ph.D.
Director of Thesis



Michelle Lacey, Ph.D.
Second Reader

Jihun Hamm, Ph.D.
Third Reader

Riley Juenemann, Differentiating Between Diffusive Movement Patterns in Single Particle Tracking Experiments

(Professor Scott A. McKinley, Mathematics)

Categorizing the movement of organelles, nanoparticles, and other vesicles inside living cells is an important research problem for scientists conducting fluorescence microscopy and quantum dot experiments. Particle trajectories in these complex biological systems are incredibly diverse. The relative proportions of each movement type present in a dataset can be useful to making experimental conclusions. We focus on two of the main movement patterns: free diffusion, where the particle is passively fluctuating, and anchored diffusion, which exhibits similar fluctuations but is tethered to a fixed point. We model both movement types by approximations of a stochastic differential equation and use their properties to explore rigorous model selection methods in statistics and information theory. In particular, we analyze the performance of the Akaike Information Criterion (AIC) and Bayes Factor approaches. Conducting numerical simulation allows us to determine how well these methods work at identifying each movement type. We further characterize and compare the distribution of the AIC on a given trajectory of each movement type. In cases where the true and likelihood models disagree, this presents a challenge. We derive the expression for these mismatched cases, including maximum likelihood estimators. This allows us to determine the form of a likelihood ratio test between the two models.

ACKNOWLEDGEMENTS

I would like to extend my deepest gratitude to my advisor, Dr. Scott A. McKinley. Inviting me to join you in research after my freshman year has sparked my passion for both research and mathematical biology. This passion has greatly enriched my educational experience and shaped my life trajectory in ways I could have never anticipated. Thank you for your persistent intellectual insight, encouragement, and outstanding mentorship during my time at Tulane.

I also had the great pleasure of working with Dr. Michelle Lacey and Dr. Jihun Hamm on my thesis. Thank you for serving on my committee, teaching engaging courses, and supporting my goals.

I must thank Dr. Christine Payne, Dr. Veronica Ciocanel, Dr. Adriana Dawes, and Dr. Keisha Cook for being key mentors in my development as a researcher, introducing me to incredible conference opportunities, and deepening my understanding of biological processes.

I'd also like to extend my gratitude to Adriana Duncan, who has always been there for the challenges and celebrations in research. I am so glad that Math Club brought us together, and I am proud of our community.

I am grateful for all of the friends, family, peers, and teachers that have fostered my love for mathematics. I would especially like to thank my parents and brother, who have never wavered in their support. It is because of you that even my wildest dreams have come to fruition.

CONTENTS

List of Tables

List of Figures

1. INTRODUCTION

At the intersection of nanoscience and biology lies the question of precisely how particles move within cells. Recent developments in cellular imaging, including quantum dots and semiconductor nanoparticles, have made it possible for researchers to peek beneath the surface of living cells, tracking individual particles over significant timescales [21]. Within this research area, the use of mathematics has led to novel revelations about the heterogeneity that exists within cells. In contrast to in vitro particle tracking experiments, wherein there are great controls on particle and environmental homogeneity, live cell (in vivo) tracking exhibits tremendous diversity in particle movement. This is exemplified in a wide range of contexts, including Adenovirus-2 trajectories in live cells [13], endosome transport trajectories in HeLa cells [5], and actin dependent chromosome transport in starfish oocytes [12].

Before individual particle trajectories could be observed directly, mathematicians have worked with biologists under the assumption that the behavior of every particle could be modeled well by Brownian motion. Brownian motion is a random passive process: the particle just moves in response to thermal fluctuations and collisions with other molecules in the fluid. This process was discovered and named after Robert Brown, a Scottish botanist in the 19th century that was well respected for his work in classifying plants in Australia [2]. He first observed this behavior in pollens suspended in water. Curious if this observation extended beyond living matter, he suspended particles of dead pollen, rocks, and metals. Brown observed that it was the size of the material, and not the nature of the material, that mattered for these movement patterns. Unfortunately, he was not able to uncover the

precise reason why the particles were behaving in this manner. It was not until 1905 - annus mirabilis, miracle year in theoretical physics - that the puzzle was finally solved by Albert Einstein [9]. From a statistical mechanics point of view, Einstein proposed the idea that the particles are so small that they are subjected to thermal fluctuations and individual collisions with water molecules. In 1908, Langevin was able to explain these phenomena in terms of Newtonian physics, essentially writing the first stochastic differential equation [18].

Modern microscopy has revealed a much greater variety of movement within cells. Particle movement in cells can also exhibit other motion characteristics. In what we refer to as directed transport, biomolecules called molecular motors simultaneously bind to intracellular cargo and to individual filaments in the cytoskeleton called microtubules. These motors literally step along microtubules, producing directed motion that moves as much as 1 or 2 microns per second. By contrast, we observe many particles that do not appear to be moving much at all. We refer to these particles as experiencing "anchored diffusion" because they behave as though they are bound by a motor to a microtubule, but the motor is not stepping. Such particles behave as though they are tethered to a fixed location by a Hookean spring. In yet another behavior, there is no binding to the microtubule, yet where the particle can move is constrained by the environment. Such movement is called corralled diffusion. The scientific objective is to determine the statistical signatures of different movement types, and, given any particular path, to determine which of the different biophysical mechanisms is mostly to have produced the behavior.

This thesis is motivated by recent single particle tracking experiments by our collaborators, in which we must differentiate between free and anchored diffusion. In certain

conditions and when using techniques currently in common practice, these movement types can be hard to distinguish from one another. This is further complicated by experimental limits. In fluorescence microscopy, a popular method used in particle tracking experiments, one can only gather a certain number of data points before the fluorescent tag becomes overexposed. This leads to the questions "How many observations do we need to reliably tell the difference between these two movement types?" and "What is the frequency of observations that yields the most informative data?"

There have been several categories of proposed methods for differentiating between movement types. The first focuses on qualitative visual inspection. One example is mean squared displacement (MSD), which is often used to determine the movement pattern of a single particle trajectory [22]. Others have extended this and conducted Bayesian model selection on MSD curves [12, 19, 20, 21] or categorized based upon the slope of the moment scaling spectrum (MSS), a generalization of the MSD [10, 11, 23]. However, distinct stochastic processes can yield very similar MSD curves. There is not a standard way to determine if the trajectories are long enough, when to cut them off, and what deviations from a slope of one are statistically significant [25].

We use well-established tools from statistics and information theory to provide an alternative, more rigorous method for differentiating between biophysical models. One approach focuses on information criteria. Examples include the Akaike Information Criterion (AIC) [1] and the more general Watanabe–Akaike Information Criterion (WAIC) [24]. For these methods, a model receives a score for how likely that model is to produce the observed data, and that score is then penalized for complexity of the model. With scores that

can be computed for each candidate model, we can pick a "winner" for each path. However, a difficulty arises in deciding how large of a difference in scores is meaningful, and what kinds of differences can arise by chance alone.

This issue of determining what is a significant difference between models is naturally addressed by statistical techniques. From a frequentist viewpoint, this is accomplished via a likelihood ratio test [8]. Since we cannot express one model in terms of another by parameter choice, we call the models that we are considering non-nested. There is not a universal theory for applying likelihood ratio test methodology to non-nested models [7], so we will investigate this approach in our dichotomy. The Bayesian viewpoint uses a method called Bayes Factors for the goal of determining significant differences between models [4, 6, 17]. The trade-off we face is that the information criteria are easy to compute and analyze "by hand," but the statistical methods give more context. We will investigate the methods' relative effectiveness and attempt to determine which have the best combination of being easy to communicate to the engineering community while also being accurate in their assessments.

2. Models of Movement Types

We propose two models for a vector of position observations $\vec{X} = (X_0, X_1, ..., X_N)$ at times $\vec{t} = (t_0, t_1, ...t_N)$. We assume $t_i = i\Delta$ for some time increment size $\Delta > 0$.

2.1. **Free Diffusion.** Our model for free diffusion is Brownian motion, denoted as model 1. In this model, increments are independent and identically distributed. Let increments for $i = 0, 1, ..., n$ be represented as $\xi_i = X_i - X_{i-1}$, where $\xi_i \sim N(0, 2D\Delta)$ for diffusivity parameter, D. Define $\sigma_1^2 = 2D\Delta$. The likelihood for this model is

$$L_1(\sigma_1^2; \vec{\xi}) = \left(\frac{1}{2\pi\sigma_1^2}\right)^{\frac{N}{2}} e^{-\frac{1}{2\sigma_1^2}\sum_{i=1}^{N}\xi_i^2}. \tag{1}$$

For the methods in this thesis, we will typically use the log-likelihood:

$$\ell_1(\sigma_1^2; \vec{\xi}) = -\frac{N}{2}\ln(\sigma_1^2) - \frac{N}{2}\ln(2\pi) - \frac{1}{2\sigma_1^2}\sum_{i=1}^{N}\xi_i^2. \tag{2}$$

2.2. **Anchored Diffusion.** For anchored diffusion, we imagine that the particle is tethered to some anchor point. As detailed in Appendix A, we can use Langevin dynamics to model with a stochastic differential equation (SDE). By taking two of the parameters to zero, we obtain the following model. We approximate the behavior of that stochastic differential equation by taking every location of the particle as being drawn as independent and identically distributed samples from a Gaussian centered at the anchor point. We will denote this as model 2. In this model, the observed positions of the particle are independent and identically distributed. Let $X_i \sim N(0, \sigma_2^2)$ for $i = 0, 1, ..., N$, where $X_0 = 0$. The likelihood

for this model is

$$L_2(\sigma_2^2;\vec{X}) = \left(\frac{1}{2\pi\sigma_2^2}\right)^{\frac{N}{2}} e^{-\frac{1}{2\sigma_2^2}\Sigma_{i=0}^{N} X_i^2} \ . \tag{3}$$

For the methods in this thesis, we will typically use the log-likelihood:

$$\ell_2(\sigma_2^2;\vec{X}) = -\frac{N}{2}\ln(\sigma_2^2) - \frac{N}{2}\ln(2\pi) - \frac{1}{2\sigma_2^2}\sum_{i=1}^{N} X_i^2 \ . \tag{4}$$

3. Model Selection Methods

We define information criterion, likelihood ratio test, and Bayesian approaches for our models for Free and Anchored Diffusion. Further, we numerically compare the relative performance of the model selection techniques for simulated Free and Anchored Diffusion trajectories.

3.1. **Information Criterion Approach.**

Information criterion approaches account for how well a model fits the observed data and balances this with the complexity of the model, represented by a numerical score. After calculating the scores for candidate models, the one that maximizes or minimizes the score will be selected.

The Akaike Information Criterion (AIC) is one such method, named after its developer, Hirotugu Akaike [1]. This method chooses the model that minimizes

$$\text{AIC} = -2\ell(\hat{\theta}_{\text{MLE}}; \vec{X}) + 2(\text{number of parameters}) \,, \tag{5}$$

where $\ell(\hat{\theta}_{\text{MLE}}; \vec{X})$ is the log-likelihood of observing the data given that the maximum likelihood estimate is the correct choice for the parameter θ.

Before we apply this method to our models, we must determine the maximum likelihood estimator (MLE) for σ_1^2 and σ_2^2. Differentiating the log-likelihood equation for model 1, Equation 2, with respect to σ_1^2 yields $\ell_1'(\sigma_1^2; \vec{X}) = \frac{-N}{2\sigma_1^2} + \frac{1}{2(\sigma_1^2)^2} \sum_{i=1}^{\infty} \xi_i^2$. Setting $\ell_1'(\sigma_1^2; \vec{X}) = 0$ and solving for σ_1^2 results in the estimator

$$\hat{\sigma}_1^2 = \frac{\sum_{i=1}^{\infty} \xi_i^2}{N} \quad . \tag{6}$$

Using the First Derivative Test, we can confirm that this estimator is indeed a maximum:

$$\begin{aligned}
\ell_1'\left(\frac{1}{2}\hat{\sigma}_1^2;\vec{X}\right) &= \frac{-N}{2\frac{1}{2}\hat{\sigma}_1^2} + \frac{1}{2\left(\frac{1}{2}\hat{\sigma}_1^2\right)^2}\sum_{i=1}^{\infty}\xi_i^2 \\
&= \frac{-N^2}{\sum_{i=1}^{\infty}\xi_i^2} + \frac{2N^2}{\sum_{i=1}^{\infty}\xi_i^2} \\
&= \frac{N^2}{\sum_{i=1}^{\infty}\xi_i^2} > 0\ , \\
\ell_1'\left(2\hat{\sigma}_1^2;\vec{X}\right) &= \frac{-N}{4\hat{\sigma}_1^2} + \frac{1}{2\left(2\hat{\sigma}_1^2\right)^2}\sum_{i=1}^{\infty}\xi_i^2 \\
&= \frac{-N^2}{4\sum_{i=1}^{\infty}\xi_i^2} + \frac{N^2}{8\sum_{i=1}^{\infty}\xi_i^2} \\
&= \frac{-N^2}{8\sum_{i=1}^{\infty}\xi_i^2} < 0\ .
\end{aligned}$$

Since ℓ_1' changes from positive to negative at $\hat{\sigma}_1^2$, then ℓ_1 has a maximum at $\hat{\sigma}_1^2$.

We can similarly determine the MLE for σ_2^2 in model 2. Differentiating Equation 4 with respect to σ_2^2 yields $\ell_2'(\sigma_2^2;\vec{X}) = \frac{-N}{2\sigma_2^2} + \frac{1}{2(\sigma_2^2)^2}\sum_{i=1}^{\infty}\xi_i^2$. Setting $\ell_2'(\sigma_2^2;\vec{X}) = 0$ and solving for σ_2^2 gives us the estimator

$$\hat{\sigma}_2^2 = \frac{\sum_{i=1}^{\infty}X_i^2}{N}\ . \tag{7}$$

Again, using the First Derivative Test, we can confirm that this estimator is indeed a maximum. The derivation is identical to the previous case. Since ℓ_2' changes from positive to negative at $\hat{\sigma}_2^2$, then ℓ_2 has a maximum at $\hat{\sigma}_2^2$.

With these estimators, we are able to determine the AIC for our models, which both have just one parameter:

$$\mathrm{AIC}_1 = N\ln(\hat{\sigma}_1^2) + N\ln(2\pi) + \frac{1}{\hat{\sigma}_1^2}\sum_{i=1}^{\infty}\xi_i^2 + 2\ , \tag{8}$$

$$\mathrm{AIC}_2 = N\ln(\hat{\sigma}_2^2) + N\ln(2\pi) + \frac{1}{\hat{\sigma}_2^2}\sum_{i=1}^{\infty} X_i^2 + 2 \ . \tag{9}$$

We can use the MLE and these equations to compute the AIC assuming both models for the observed data.

3.2. **Connection Between Information Criterion and Likelihood Ratio Test.** The Neyman-Pearson Lemma tells us that the likelihood ratio test (LRT) is the most powerful method that could be used to reject a simple null model. Since our models are non-nested, in our case we cannot apply such a theorem. However, we can make a connection between the AIC and LRT.

Suppose we are testing the null hypothesis, H_0, of free diffusion with $\sigma^2 \in \theta_0$, versus the alternative hypothesis, H_1, of anchored diffusion with $\sigma^2 \in \theta_0^C$. Then, using the framework of the generalized LRT, we would reject the null model if

$$\sup_{\sigma^2\in\theta_0} L(\sigma^2|\vec{X}) \leq k \sup_{\sigma^2\in\theta_0^C} L(\sigma^2|\vec{X}) \ ,$$

where k is a positive number. Note that this is equivalent to

$$L(\hat{\sigma}_1^2|\vec{X}) \leq k L(\hat{\sigma}_2^2|\vec{X}) \ .$$

Taking the natural logarithm of both sides we attain

$$\ln L(\hat{\sigma}_1^2|\vec{X}) \leq \ln(k) + \ln\left(L(\hat{\sigma}_2^2|\vec{X})\right) \ .$$

Adding and subtracting the appropriate terms yields

$$-2\ell(\hat{\sigma}_1^2|\vec{X}) + 2 \geq -2\ln(k) + \left(-2\ell(\hat{\sigma}_2^2|\vec{X}) + 2\right) ,$$

which by definition is

$$\mathrm{AIC}_1 \geq -2\ln(k) + \mathrm{AIC}_2 \ .$$

So altogether, the form of a rejection region for a likelihood ratio test would be

$$\left\{\vec{X} : \mathrm{AIC}_1(\vec{X}) - \mathrm{AIC}_2(\vec{X}) \geq C\right\} , \tag{10}$$

for some appropriately chosen critical value, C.

If the AIC assuming free diffusion is bigger than the AIC assuming anchored diffusion (by a margin greater than some critical value), then we will reject the null hypothesis that the diffusion is free.

This is consistent with the notion that the AIC method chooses the model that minimizes the score. In order to determine the critical value, C, we would need to understand the distribution of the test statistic $\mathrm{AIC}_1 - \mathrm{AIC}_2$. Our investigation in Chapter 5 proceeds with this as motivation.

3.3. **Bayesian Approach.** Hypothesis testing seeks to evaluate if the observations are significant evidence for rejecting a null hypothesis. By contrast, the Bayesian approach allows a way of evaluating if there is significant evidence in favor of a hypothesis. This general framework allows for the comparison of non-nested models, and was introduced by Jeffreys [14, 15].

$\log_{10}(B_{21})$	Evidence Against Model 1
< 1/2	Poor
1/2 to 1	Substantial
1 to 2	Strong
> 2	Decisive

TABLE 1. Criteria for determining significance when using Bayes Factors. We use evidence against a model, as it is more familiar, but we could reformulate these values easily in terms of evidence in favor of a model.

Bayes Factors are the standard method for this approach, expressing the likelihood of seeing the data for the model weighted by the prior distribution [17]. In other words: posterior odds = Bayes Factor × prior odds. For simplifying the integral evaluation, we rewrite our models equivalently in terms of other parameters, η_D and η_σ. For free diffusion, define $\eta_D = D^{-1}$. The likelihood for this model is then $L_1(\eta_D; \vec{\xi}) = (\frac{\eta_D}{\pi\Delta})^{\frac{n}{2}} e^{-\frac{\eta_D}{2}\Sigma_{i=1}^{\infty}\xi_i^2}$. We assume a prior distribution of $\eta_D \sim \text{Gamma}(\alpha_1, \beta_1)$. For anchored diffusion, define $\eta_\sigma = \frac{1}{\sigma_2^2}$. The likelihood for this model is then $L_2(\eta_\sigma; \vec{X}) = (\frac{\eta_\sigma}{2\pi})^{\frac{n}{2}} e^{-\frac{\eta_\sigma}{2}\Sigma_{i=0}^{\infty}X_i^2}$. We assume a prior distribution of $\eta_\sigma \sim \text{Gamma}(\alpha_2, \beta_2)$.

This yields the following equation for the Bayes Factor based upon models 1 and 2

$$B_{21} = \frac{\int_0^\infty L_2(\eta_\sigma; \vec{X})\pi_2(\eta_\sigma)d\eta_\sigma}{\int_0^\infty L_1(\eta_D; \vec{X})\pi_1(\eta_D)d\eta_D} := \frac{I_2}{I_1}\,, \tag{11}$$

where $\pi_2(\eta_\sigma)$ and $\pi_1(\eta_D)$ refer to the prior distribution of each parameter. Statistical significance can be determined using $\log_{10}(B_{21})$ based upon the criteria in Table 1 [17].

In some cases, the integrals can be solved directly. The models we consider are one of these special cases. First, we evaluate the numerator of the Bayes Factor:

$$
\begin{aligned}
I_2 &= \int_0^\infty L_2(\eta_\sigma;\vec{X})\pi_2(\eta_\sigma)d\eta_\sigma \\
&= \int_0^\infty (\frac{\eta_\sigma}{2\pi})^{\frac{n}{2}} e^{-\frac{\eta_\sigma}{2}\sum_{i=0}^\infty X_i^2} \frac{\beta_2^{\alpha_2}}{\Gamma(\alpha_2)} \eta_\sigma^{\alpha_2-1} e^{-\beta_2\eta_\sigma} d\eta_\sigma \\
&= (\frac{1}{2\pi})^{\frac{n}{2}} \frac{\beta_2^{\alpha_2}}{\Gamma(\alpha_2)} \int_0^\infty \eta_\sigma^{\frac{n}{2}+\alpha_2-1} e^{-\eta_\sigma(\frac{1}{2}\sum_{i=0}^\infty X_i^2+\beta_2)} d\eta_\sigma \\
&= (\frac{1}{2\pi})^{\frac{n}{2}} \frac{\beta_2^{\alpha_2}}{\Gamma(\alpha_2)} \frac{\Gamma(\alpha_2+\frac{n}{2})}{(\beta_2+\frac{1}{2}\sum_{i=0}^\infty X_i^2)^{\alpha_2+\frac{n}{2}}} \int_0^\infty \frac{(\beta_2+\frac{1}{2}\sum_{i=0}^\infty X_i^2)^{\alpha_2+\frac{n}{2}}}{\Gamma(\alpha_2+\frac{n}{2})} \eta_\sigma^{\frac{n}{2}+\alpha_2-1} e^{-\eta_\sigma(\frac{1}{2}\sum_{i=0}^\infty X_i^2+\beta_2)} d\eta_\sigma \\
&= (\frac{1}{2\pi})^{\frac{n}{2}} \frac{\beta_2^{\alpha_2}}{\Gamma(\alpha_2)} \frac{\Gamma(\alpha_2+\frac{n}{2})}{(\beta_2+\frac{1}{2}\sum_{i=0}^\infty X_i^2)^{\alpha_2+\frac{n}{2}}} .
\end{aligned}
$$

We can similarly evaluate the denominator of the Bayes Factor:

$$
\begin{aligned}
I_1 &= \int_0^\infty L_1(\eta_D;\vec{\xi})\pi_1(\eta_D)d\eta_D \\
&= (\frac{\eta_D}{\pi\Delta})^{\frac{n}{2}} e^{-\frac{\eta_D}{2}\sum_{i=1}^\infty \xi_i^2} \frac{\beta_1^{\alpha_1}}{\Gamma(\alpha_1)} \eta_D^{\alpha_1-1} e^{-\beta_1\eta_D} d\eta_D \\
&= (\frac{1}{\pi\Delta})^{\frac{n}{2}} \frac{\beta_1^{\alpha_1}}{\Gamma(\alpha_1)} \int_0^\infty \eta_D^{\frac{n}{2}+\alpha_1-1} e^{-\eta_D(\frac{1}{\Delta}\sum_{i=1}^\infty \xi_i^2+\beta_1)} d\eta_D \\
&= (\frac{1}{\pi\Delta})^{\frac{n}{2}} \frac{\beta_1^{\alpha_1}}{\Gamma(\alpha_1)} \frac{\Gamma(\alpha_1+\frac{n}{2})}{(\beta_1+\frac{1}{\Delta}\sum_{i=1}^\infty \xi_i^2)^{\alpha_1+\frac{n}{2}}} \int_0^\infty \frac{(\beta_1+\frac{1}{\Delta}\sum_{i=1}^\infty \xi_i^2)^{\alpha_1+\frac{n}{2}}}{\Gamma(\alpha_1+\frac{n}{2})} \eta_D^{\frac{n}{2}+\alpha_1-1} e^{-\eta_D(\frac{1}{\Delta}\sum_{i=1}^\infty \xi_i^2+\beta_1)} d\eta_D \\
&= (\frac{1}{\pi\Delta})^{\frac{n}{2}} \frac{\beta_1^{\alpha_1}}{\Gamma(\alpha_1)} \frac{\Gamma(\alpha_1+\frac{n}{2})}{(\beta_1+\frac{1}{\Delta}\sum_{i=1}^\infty \xi_i^2)^{\alpha_1+\frac{n}{2}}} .
\end{aligned}
$$

Notice that the numerator depends on the position observations, X_i, while the denominator depends on the increments, ξ_i. This is because the model for anchored diffusion, model 2, has normally distributed observed positions while in the model for free diffusion, model 1, the increments are normally distributed.

After combining the numerator and denominator, and simplifying, the analytical solution for the Bayes Factor of our models is

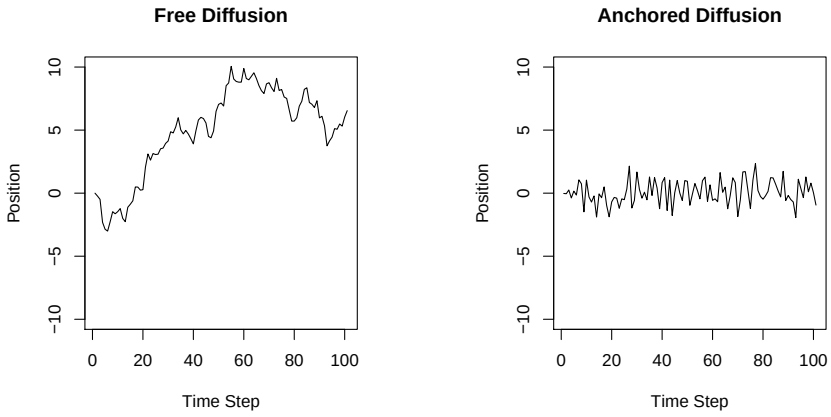
$$B_{21} = \left(\frac{\Delta}{2}\right)^{\frac{n}{2}} \frac{\beta_2^{\alpha_2}}{\beta_1^{\alpha_1}} \frac{\Gamma(\alpha_1)}{\Gamma(\alpha_2)} \frac{\Gamma(\alpha_2 + \frac{n}{2})}{\Gamma(\alpha_1 + \frac{n}{2})} \frac{(\beta_1 + \frac{1}{\Delta}\sum_{i=1}^{\infty} \xi_i^2)^{\alpha_1 + \frac{n}{2}}}{(\beta_2 + \frac{1}{2}\sum_{i=0}^{\infty} X_i^2)^{\alpha_2 + \frac{n}{2}}} \,. \tag{12}$$

We can use this equation to compute the Bayes Factor for a trajectory of observed particle positions. When $\alpha = \alpha_1 = \alpha_2$ and $\beta = \beta_1 = \beta_2$, notice that this reduces to

$$B_{21} = \left(\frac{\Delta}{2}\right)^{\frac{n}{2}} \frac{(\beta + \frac{1}{\Delta}\sum_{i=1}^{\infty} \xi_i^2)^{\alpha + \frac{n}{2}}}{(\beta + \frac{1}{2}\sum_{i=0}^{\infty} X_i^2)^{\alpha + \frac{n}{2}}} \,. \tag{13}$$

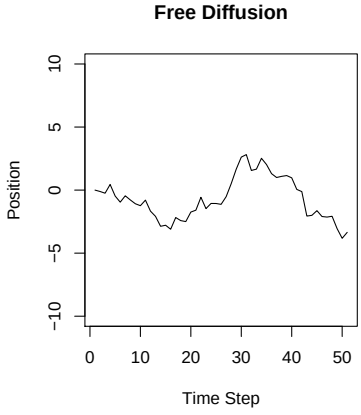
3.4. **Numerical Analysis.** We conduct numerical analysis using R, an open source programming language developed for statistical computing. We simulate trajectories from both models for three different lengths of observations: 100 points, 50 points, and 10 points. Sample trajectories are pictured in Figure 1. Notice that as the path length decreases, the free and anchored diffusion trajectories seem more similar. These simulated trajectories allow us to evaluate the performance of the AIC and Bayes Factor methods on detecting the "true" model and investigate how many observations are needed to reliably tell the difference between the two movement types.

For each number of observations, we simulate 100 of each model. In both models, we set $\Delta = 1$. For free diffusion, our model 1, we select $D = 1$, and thus $\eta_D = \frac{1}{D} = 1$. We simulate the independent and identically distributed increments from a normal distribution. Then, we set an initial position and take the cumulative sum of the increments at each time step to yield the observed position of the particle. For anchored diffusion, our model 2, we

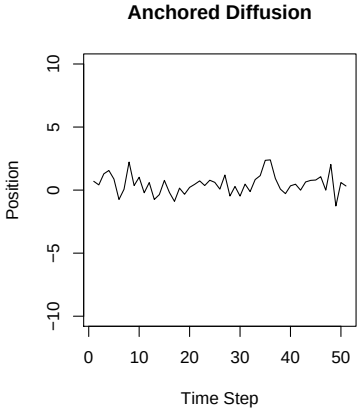


(A) 100 observations

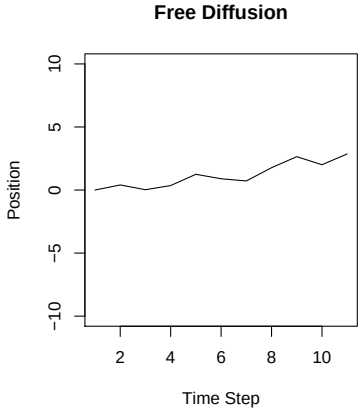
(B) 100 observations



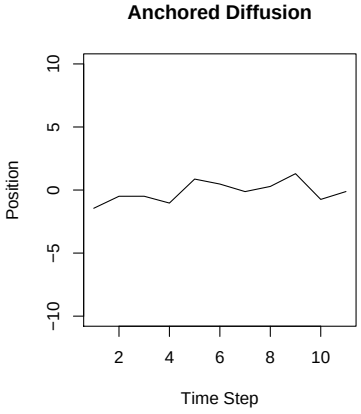
(C) 50 observations



(D) 50 observations



(E) 10 observations

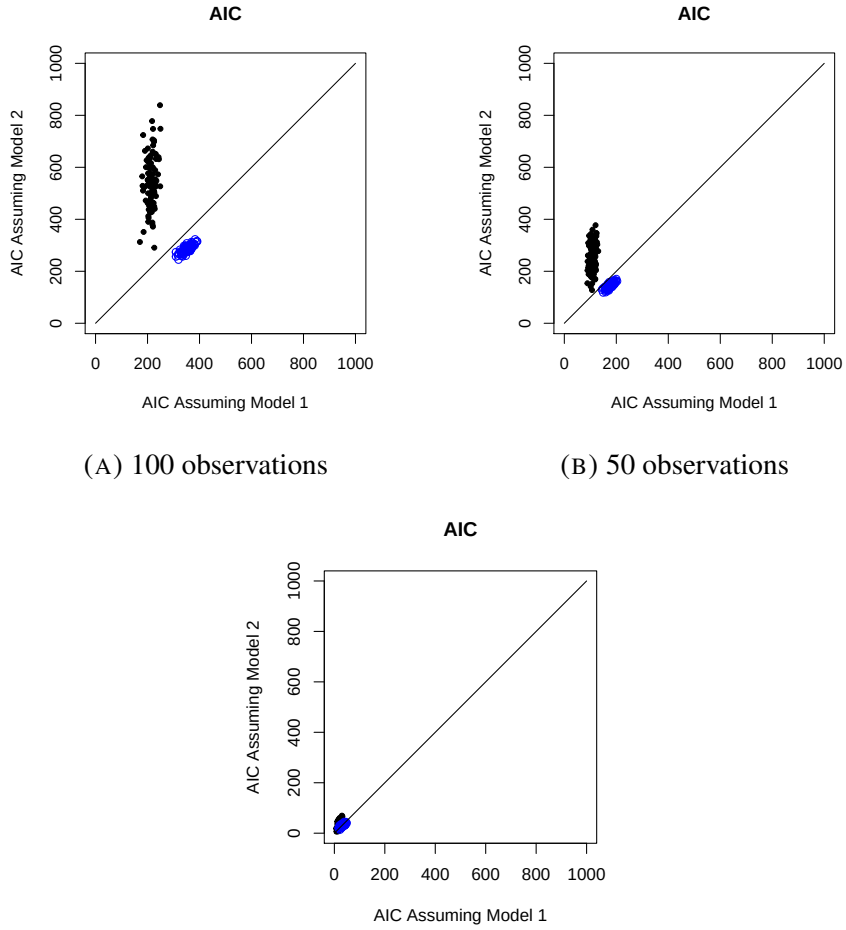


(F) 10 observations

FIGURE 1. Sample free and anchored diffusion trajectories simulated at different lengths of observations. Notice that as the path length decreases, the free and anchored diffusion trajectories seem more similar.

select $\sigma_2^2 = 1$, and thus $\eta_\sigma = \frac{1}{\sigma_2^2} = 1$. We then simulate the independent and identically distributed positions from a normal distribution. We take the differences between the positions at each time step to record the increments. We use both the increments and the positions in our calculations.

We calculate the AIC for each trajectory twice - once assuming that model 1 is the underlying "true" model, and once assuming model 2. Taking the minimum of the two calculations determines which model has been selected as the best fit by the AIC. In Figure 2, we plot the AIC score assuming model 1 (free diffusion) and the AIC score assuming model 2 (anchored diffusion) for each trajectory at each observation length. The trajectories simulated from model 1 are in black, while the trajectories simulated from model 2 are in blue. What qualifies as a significant difference between the AIC values for distinct models is debated. The closer the values are, the more difficult it is for the AIC to distinguish between the models. On our plots in Figure 2, the closer the AIC score is for the two models, the closer the point is to the diagonal line. Notice that as the number of observations decreases, the AIC for each group becomes closer to the diagonal, and it becomes harder to determine which model is the correct one. Additionally, in each plot, it is relatively easier for the AIC to correctly identify trajectories simulated from model 1 compared to those simulated from model 2.



(A) 100 observations

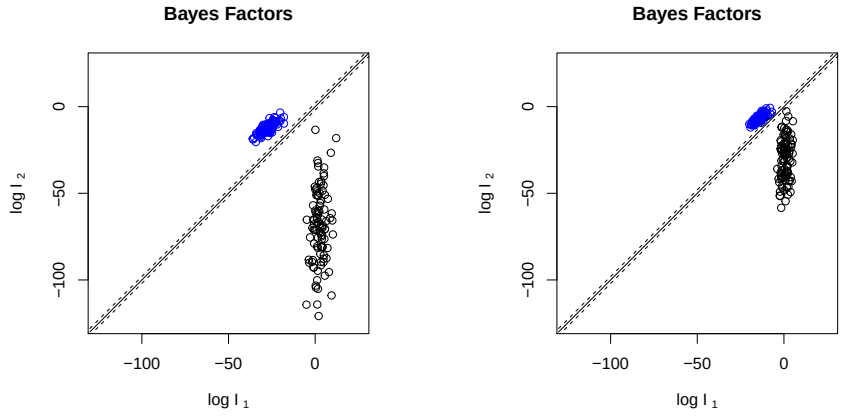
(B) 50 observations

(C) 10 observations

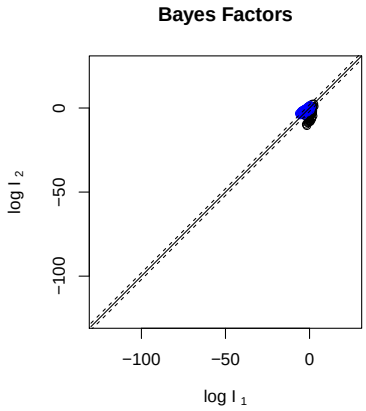
FIGURE 2. The AIC evaluated for 100 freely diffusing (model 1, black) and anchored diffusing (model 2, blue) trajectories at different lengths of observations.

We also compute the Bayes Factor for each trajectory, choosing our priors with $\alpha_1 = \alpha_2 = 1, \beta_1 = \beta_2 = 1$. Since Bayes Factors evaluate if there is significant evidence in favor of a hypothesis, as opposed to evidence for rejecting a null hypothesis, we do not have to perform multiple calculations where we assume that each model is the null model. We use the analytical solution from Equation 12 to determine the Bayes Factor for each trajectory, taking the log so that we may use Table 1 to determine significance. We have plotted the

numerator and the denominator of the Bayes Factors for simulations at each observation length in Figure 3. The trajectories simulated from model 1 (free diffusion) are in black, while the trajectories simulated from model 2 (anchored diffusion) are in blue. The X axis represents the denominator of the Bayes Factor while the Y axis represents the numerator. Note that both axes are on log scale. We expect points below the diagonal to favor model 1, while points above the diagonal favor model 2. The dashed lines indicate the Bayes Factor needed to have significant, decisive evidence for model selection based upon Table 1. For 100 and 50 observations, the free diffusion trajectories correctly show that there is poor evidence to reject model 1. Similarly, the Bayes Factors for the anchored diffusion trajectories correctly represent decisive evidence against model 1. For 10 observations, we see performance deteriorate for the Bayes Factors. For the free diffusion paths, 97 percent are still showing poor evidence against model 1. This is only a minor dip. However, for the anchored diffusion trajectories, we drop from 100 to 32 percent that have decisive evidence against model 1, with 18 percent even indicating poor evidence against model 1.



(A) 100 observations (B) 50 observations



(C) 10 observations

FIGURE 3. The Bayes Factors evaluated for 100 free diffusion (model 1, black) and anchored diffusion (model 2, blue) trajectories at different lengths of observations. The X axis represents the denominator of the Bayes factor while the Y axis represents the numerator.

Table 2 summarizes Figures 2 and 3. For each path length, it gives the percentage of the simulated paths that were on the correct side of the diagonal line for both movement types in the AIC and Bayes Factors cases. For the shortest path length, the AIC is more accurate than the Bayes Factors on true Anchored Diffusion paths. For Free Diffusion paths, the Bayes Factors are slightly more accurate than the AIC.

Path Length	AIC on Free Diffusion	AIC on Anchored Diffusion	BF on Free Diffusion	BF on Anchored Diffusion
10	92%	81%	98%	66%
50	100%	100%	100%	100%
100	100%	100%	100%	100%

TABLE 2. Numerical analysis of performance of AIC and Bayes Factors (BF) on simulated data. The AIC outperforms BF on Anchored Diffusion trajectories, while BF is slightly more accurate than the AIC for Free Diffusion trajectories.

Path Length	AIC on Free Diffusion	AIC on Anchored Diffusion	BF on Free Diffusion	BF on Anchored Diffusion
10	93.4%	85.0%	98.6%	72.6%
20	97.3%	97.4%	98.5%	94.7%
30	99.4%	99.6%	99.9%	99.5%
40	99.9%	99.8%	100%	99.6%
50	100%	100%	100%	100%

TABLE 3. Numerical analysis of performance of AIC and Bayes Factors (BF) on simulated data with shorter simulated trajectories. 1000 trajectories of each... smaller path lenghts...

We support that observation for the shortest path length by conducting further numerical analysis. We simulate 1000 free diffusion and 1000 anchored diffusion trajectories with the same parameters as before, but this time with path lengths 10, 20, 30, 40, and 50. The results are summarized in Table 3 Again, we observe that for shorter path lengths, the AIC is more accurate than the Bayes Factors on true anchored diffusion paths, while the Bayes Factors are slightly more accurate than the AIC on true free diffusion paths. For path lengths around 30 and above, there is not a significant difference in performance.

3.5. **Discussion.** We expect that it would be harder to differentiate between movement types with fewer data points. For both AIC and Bayes Factors, we see a general decrease in performance with fewer position observations for each trajectory. Both perform effectively for 100 observations, still well for 50 observations, and in some cases poorly at 10 observations. This decline is significantly faster for the anchored diffusion movement types. In other words, as the paths become shorter, it is harder to determine the model that best fits

an anchored diffusion trajectory than it is in the case of a free diffusion trajectory. We see this with both the AIC and Bayes Factor methods.

Neither method is without its limitations. Bayes Factors can be sensitive to the choice of prior, which is not information that you have to know or specify for the AIC. Thus, a full investigation of Bayes Factors would involve an analysis of sensitivity to the prior. On the other hand, Bayes Factors are able to convey a notion of significance and level of uncertainty that are not as intuitive for AIC. It is hard to determine what a significant difference in AIC values for different models should be, and to what extent a greater difference reflects increased confidence in the model chosen. This is further complicated by the fact that the multiplier on the number of parameters, 2, is fixed regardless of the parameters of interest. Other information criterion approaches, such as the Watanabe–Akaike Information Criterion (WAIC), have different penalties for complexity. This could be a topic of exploration in future work.

4. An Exploration of Akaike Information Criterion (AIC) Assuming Model Parameters are Known

In the numerical analysis from the previous chapter, we saw that as the trajectories become shorter, it is harder to determine the model that best fits an anchored diffusion trajectory than it is in the case of a free diffusion trajectory. In this chapter, we seek to investigate this more rigorously using theoretical methods. In particular, we want to answer:

- For a free diffusion trajectory, what is the expected difference between AIC_1 and AIC_2?
- For an anchored diffusion trajectory, what is the expected difference between AIC_1 and AIC_2?

We can answer these questions by determining the distribution of the AIC in four cases:

- Case 1: The trajectory is truly free diffusion and we use free diffusion as the likelihood model
- Case 2: The trajectory is truly anchored diffusion and we use free diffusion as the likelihood model
- Case 3: The trajectory is truly free diffusion and we use anchored diffusion as the likelihood model
- Case 4: The trajectory is truly anchored diffusion and we use anchored diffusion as the likelihood model

Cases 1 and 3 are relevant to that first question, while Cases 2 and 4 can provide insight for the second question.

We first compute the expected value of the AIC, denoted $\mathbb{E}(\text{AIC})$, with the simplification that σ_1^2 and σ_2^2 are known. Notice that in Cases 2 and 3, we have model misspecification. Here, we need to re-express the trajectory in the language of the other model.

Case 1: True Free Diffusion, Likelihood Free Diffusion

Recall our model for free diffusion, where increments for $i = 0, 1, ..., N$ are represented as $\xi_i = X_i - X_{i-1}$ and $\xi_i \sim N(0, \sigma_1^2)$. For this model, $\mathbb{E}_1(\xi_i) = 0$ and $\text{Var}_1(\xi_i) = \sigma_1^2 = \mathbb{E}_1(\xi_i^2) - \mathbb{E}_1(\xi_i)^2 = \mathbb{E}_1(\xi_i^2)$. Our log-likelihood model for free diffusion is Equation 2. Thus,

$$\text{AIC}_1 = -2\ell_1(\sigma_1^2; \vec{\xi}) + 2 = N\ln(\sigma_1^2) + N\ln(2\pi) + \frac{1}{\sigma_1^2}\sum_{i=1}^{N}\xi_i^2 + 2\ ,$$

as in Equation 8. Taking the expected value:

$$\begin{aligned}
\mathbb{E}_1(\text{AIC}_1) &= \mathbb{E}_1(N\ln(\sigma_1^2) + N\ln(2\pi) + \frac{1}{\sigma_1^2}\sum_{i=1}^{N}\xi_i^2 + 2) \\
&= N\ln(\sigma_1^2) + N\ln(2\pi) + \frac{1}{\sigma_1^2}\sum_{i=1}^{N}\mathbb{E}_1(\xi_i^2) + 2 \\
&= N\ln(\sigma_1^2) + N\ln(2\pi) + \frac{1}{\sigma_1^2}N\sigma_1^2 + 2 \\
&= N\ln(\sigma_1^2) + N\ln(2\pi) + N + 2
\end{aligned}$$

Case 4: True Anchored Diffusion, Likelihood Anchored Diffusion

Recall our model for anchored diffusion, where the observed positions are distributed as $X_i \sim N(0, \sigma_2^2)$ for $i = 0, 1, ..., N$. For this model, $\mathbb{E}_2(X_i) = 0$ and $\text{Var}_2(X_i) = \sigma_2^2$. Our log-likelihood model for anchored diffusion is Equation 4. Thus,

$$\text{AIC}_2 = -2\ell_2(\sigma_2^2; \vec{X}) + 2 = N\ln(\sigma_2^2) + N\ln(2\pi) + \frac{1}{\sigma_2^2}\sum_{i=1}^{N}X_i^2 + 2\ ,$$

as in Equation 9. Taking the expected value:

$$\begin{aligned}\mathbb{E}_2(\mathrm{AIC}_2) &= \mathbb{E}_2(N\ln(\sigma_2^2) + N\ln(2\pi) + \frac{1}{\sigma_2^2}\sum_{i=1}^{N} X_i^2 + 2) \\ &= N\ln(\sigma_2^2) + N\ln(2\pi) + \frac{1}{\sigma_2^2}\sum_{i=1}^{N} \mathbb{E}_2(X_i^2) + 2 \\ &= N\ln(\sigma_2^2) + N\ln(2\pi) + \frac{1}{\sigma_2^2} N\sigma_2^2 + 2 \\ &= N\ln(\sigma_2^2) + N\ln(2\pi) + N + 2\ .\end{aligned}$$

Case 2: True Anchored Diffusion, Likelihood Free Diffusion

In this case, we again have the anchored diffusion model, $X_i \sim N(0,\sigma^2)$ for $i = 0,1,...,N$, coupled with the free diffusion log-likelihood model, Equation 2, and AIC, Equation 8. Taking the expected value yields

$$\begin{aligned}\mathbb{E}_2(\mathrm{AIC}_1) &= \mathbb{E}_2(N\ln(\sigma_1^2) + N\ln(2\pi) + \frac{1}{\sigma_1^2}\sum_{i=1}^{N} \xi_i^2 + 2) \\ &= N\ln(\sigma_1^2) + N\ln(2\pi) + \frac{1}{\sigma_1^2}\sum_{i=1}^{N} \mathbb{E}_2(\xi_i^2) + 2\ .\end{aligned}$$

In order to evaluate $\mathbb{E}_2(\sum_{i=1}^{N}\xi_i^2)$, we need to re-express $\sum_{i=1}^{N}\xi_i^2$ in terms of the positions, X_i for $i = 0,1,...,N$, rather than the increments, ξ_i, for $i = 1,...,N$. We evaluate $\mathbb{E}_2(\sum_{i=1}^{N}\xi_i^2)$ for several small values of N, and then use these to determine a general solution.

N=1: Expanding ξ_1 as $X_1 - X_0$, we have

$$\xi_1^2 = (X_1 - X_0)^2 = (X_1 - 0)^2 = X_1^2\ .$$

So

$$\mathbb{E}_2(\xi_1^2) = \mathbb{E}_2(X_1^2) = \sigma_2^2 \ .$$

N=2: Expanding the increments, ξ_1 and ξ_2, in terms of the positions X_0, X_1, and X_2, gives

$$\xi_1^2 + \xi_2^2 = (X_1 - X_0)^2 + (X_2 - X_1)^2 = X_1^2 + (X_2 - X_1)^2 = 2X_1^2 - 2X_1X_2 + X_2^2 \ .$$

Taking the expected value, we attain

$$\begin{aligned}
\mathbb{E}_2(\xi_1^2 + \xi_2^2) = \mathbb{E}_2(2X_1^2 - 2X_1X_2 + X_2^2) &= 2\,\mathbb{E}_2(X_1^2) - 2\,\mathbb{E}_2(X_1X_2) + \mathbb{E}_2(X_2^2) \\
&= 2\sigma_2^2 - 2\,\mathbb{E}_2(X_1)\,\mathbb{E}_2(X_2) + \sigma_2^2 \\
&= 3\sigma_2^2 - 2(0)(0) \\
&= 3\sigma_2^2 \ .
\end{aligned}$$

N=3: Again, we expand the increments in terms of the positions to determine

$$\begin{aligned}
\xi_1^2 + \xi_2^2 + \xi_3^2 &= (X_1 - X_0)^2 + (X_2 - X_1)^2 + (X_3 - X_2)^2 \\
&= 2X_1^2 - 2X_1X_2 + X_2^2 + (X_3 - X_2)^2 \\
&= 2X_1^2 - 2X_1X_2 + 2X_2^2 - 2X_2X_3 + X_3^2 \ .
\end{aligned}$$

Taking the expected value yields

$$\begin{aligned}\mathbb{E}_2(\xi_1^2+\xi_2^2+\xi_3^2) &= \mathbb{E}_2(2X_1^2-2X_1X_2+2X_2^2-2X_2X_3+X_3^2)\\ &= 2\,\mathbb{E}_2(X_1^2)-2\,\mathbb{E}_2(X_1)\mathbb{E}_2(X_2)+2\,\mathbb{E}_2(X_2^2)-2\,\mathbb{E}_2(X_2)\mathbb{E}_2(X_3)+\mathbb{E}_2(X_3^2)\\ &= 2\,\sigma_2^2+2\,\sigma_2^2+\sigma_2^2\\ &= 5\,\sigma_2^2\ .\end{aligned}$$

N=4: We expand the increments in terms of the positions to see that

$$\begin{aligned}\xi_1^2+\xi_2^2+\xi_3^2+\xi_4^2 &= (X_1-X_0)^2+(X_2-X_1)^2+(X_3-X_2)^2+(X_4-X_3)^2\\ &= 2X_1^2-2X_1X_2+2X_2^2-2X_2X_3+2X_3^2-2X_3X_4+X_4^2\ .\end{aligned}$$

We take the expected value to determine

$$\begin{aligned}\mathbb{E}_2(\xi_1^2+\xi_2^2+\xi_3^2+\xi_4^2) &= \mathbb{E}_2(2X_1^2-2X_1X_2+2X_2^2-2X_2X_3+2X_3^2-2X_3X_4+X_4^2)\\ &= 2\,\mathbb{E}_2(X_1^2)-2\,\mathbb{E}_2(X_1)\mathbb{E}_2(X_2)+2\,\mathbb{E}_2(X_2^2)-2\,\mathbb{E}_2(X_2)\mathbb{E}_2(X_3)\\ &\quad +2\,\mathbb{E}_2(X_3^2)-2\,\mathbb{E}_2(X_3)\mathbb{E}_2(X_4)+\mathbb{E}_2(X_4^2)\\ &= 2\,\sigma_2^2+2\,\sigma_2^2+2\,\sigma_2^2+\sigma_2^2\\ &= 7\,\sigma_2^2\ .\end{aligned}$$

The results for these values of N are summarized in Table 4. This pattern suggests that $\mathbb{E}_2(\sum_{i=1}^{N}\xi_i^2)=(2N-1)\sigma_2^2$. We can prove this using induction. We have already shown that this is true for the base case of $N=1$. Assume the pattern holds for any given case $N=k$.

N	$\mathbb{E}_2(\sum_{i=1}^{N} \xi_i^2)$
1	σ_2^2
2	$3\sigma_2^2$
3	$5\sigma_2^2$
4	$7\sigma_2^2$

TABLE 4. Summary of $\mathbb{E}_2(\sum_{i=1}^{N} \xi_i^2)$ for Case 2, where σ_2^2 is known.

We must show the pattern holds for $N = k+1$:

$$\begin{aligned}
\mathbb{E}_2\left(\sum_{i=1}^{k+1} \xi_i^2\right) &= \mathbb{E}_2\left(\sum_{i=1}^{k} \xi_i^2 + \xi_{k+1}^2\right) \\
&= \mathbb{E}_2\left(\sum_{i=1}^{k} \xi_i^2\right) + \mathbb{E}_2(\xi_{k+1}^2) \\
&= (2k-1)\sigma_2^2 + \mathbb{E}_2(\xi_{k+1}^2) \\
&= (2k-1)\sigma_2^2 + \mathbb{E}_2((X_{k+1} - X_k)^2) \\
&= (2k-1)\sigma_2^2 + \mathbb{E}_2(X_{k+1}^2 - 2X_kX_{k+1} + X_k^2) \\
&= (2k-1)\sigma_2^2 + \mathbb{E}_2(X_{k+1}^2) - 2\mathbb{E}_2(X_kX_{k+1}) + \mathbb{E}_2(X_k^2) \\
&= (2k-1)\sigma_2^2 + \mathbb{E}_2(X_{k+1}^2) - 2\mathbb{E}_2(X_k)\mathbb{E}_2(X_{k+1}) + \mathbb{E}_2(X_k^2) \\
&= (2k-1)\sigma_2^2 + \sigma_2^2 + \sigma_2^2 \\
&= (2k-1+2)\sigma_2^2 \\
&= (2(k+1)-1)\sigma_2^2 \quad .
\end{aligned}$$

Thus, we have proven by induction that $\mathbb{E}_2(\sum_{i=1}^{N}\xi_i^2) = (2N-1)\sigma_2^2$. We can substitute this in to evaluate the expected value of the AIC,

$$\begin{aligned}\mathbb{E}_2(\mathrm{AIC}_1) &= N\ln(\sigma_1^2) + N\ln(2\pi) + \frac{1}{\sigma_1^2}\sum_{i=1}^{N}\mathbb{E}_2(\xi_i^2) + 2 \\ &= N\ln(\sigma_1^2) + N\ln(2\pi) + \frac{1}{\sigma_1^2}(2N-1)\sigma_2^2 + 2 \ .\end{aligned}$$

Case 3: True Free Diffusion, Likelihood Anchored Diffusion

We use the free diffusion model, where increments for $i = 0,1,...,N$ are represented as $\xi_i = X_i - X_{i-1}$ and $\xi_i \sim N(0,\sigma_1^2)$ in conjunction with the anchored diffusion log-likelihood model, Equation 4, and AIC, Equation 9. Taking the expected value yields

$$\begin{aligned}\mathbb{E}_1(\mathrm{AIC}_2) &= \mathbb{E}_1(N\ln(\sigma_2^2) + N\ln(2\pi) + \frac{1}{\sigma_2^2}\sum_{i=1}^{N}X_i^2 + 2) \\ &= N\ln(\sigma_2^2) + N\ln(2\pi) + \frac{1}{\sigma_2^2}\sum_{i=1}^{N}\mathbb{E}_1(X_i^2) + 2 \ .\end{aligned}$$

In order to evaluate $\mathbb{E}_1(\sum_{i=1}^{N}X_i^2)$, we need to re-express $\sum_{i=1}^{N}X_i^2$ in terms of the increments, ξ_i, for $i = 1,2,...,N$, rather than the positions, X_i, for $i = 0,1,2,...,N$. In doing so, we need to use the fact that $X_i = \sum_{j=1}^{i}\xi_j$. We evaluate $\mathbb{E}_1(\sum_{i=1}^{N}X_i^2)$ for several small values of N, and then use these to determine a general solution.

N=1: Here, $X_0 = 0$ and we rewrite the position X_1 as ξ_1:

$$X_1^2 + X_0^2 = (\xi_1)^2 + 0 = \xi_1^2 \ .$$

So the expected value is

$$\mathbb{E}_1(X_1^2 + X_0^2) = \mathbb{E}_1(\xi_1^2) = \sigma_1^2 \ .$$

N=2: Rewriting the positions in terms of the increments and expanding gives

$$X_2^2 + X_1^2 + X_0^2 = (\xi_1 + \xi_2)^2 + (\xi_1)^2 + 0 = 2\xi_1^2 + 2\xi_1\xi_2 + \xi_2^2 \, .$$

Then, we take the expected value to see that

$$\begin{aligned}
\mathbb{E}_1(X_2^2 + X_1^2 + X_0^2) &= \mathbb{E}_1(2\xi_1^2 + 2\xi_1\xi_2 + \xi_2^2) \\
&= 2\,\mathbb{E}_1(\xi_1^2) + 2\,\mathbb{E}_1(\xi_1\xi_2) + \mathbb{E}_1(\xi_2^2) \\
&= 2\sigma_1^2 + 2\,\mathbb{E}_1(\xi_1)\,\mathbb{E}_1(\xi_2) + \sigma_1^2 \\
&= 2\sigma_1^2 + 0 + \sigma_1^2 \\
&= 3\sigma_1^2 \, .
\end{aligned}$$

N=3: We expand the positions in terms of their increments to give

$$\begin{aligned}
X_3^2 + X_2^2 + X_1^2 + X_0^2 &= (\xi_1 + \xi_2 + \xi_3)^2 + (\xi_1 + \xi_2)^2 + (\xi_1)^2 + 0 \\
&= \xi_1^2 + 2\xi_1\xi_2 + 2\xi_1\xi_3 + \xi_2^2 + 2\xi_2\xi_3 + \xi_3^2 + 2\xi_1^2 + 2\xi_1\xi_2 + \xi_2^2 \\
&= 3\xi_1^2 + 4\xi_1\xi_2 + 2\xi_1\xi_3 + 2\xi_2^2 + 2\xi_2\xi_3 + \xi_3^2 \, .
\end{aligned}$$

This means the expected value is

$$\begin{aligned}\mathbb{E}_1(X_3^2+X_2^2+X_1^2+X_0^2) &= \mathbb{E}_1(3\xi_1^2+4\xi_1\xi_2+2\xi_1\xi_3+2\xi_2^2+2\xi_2\xi_3+\xi_3^2)\\ &= 3\,\mathbb{E}_1(\xi_1^2)+4\,\mathbb{E}_1(\xi_1)\mathbb{E}_1(\xi_2)+2\,\mathbb{E}_1(\xi_1)\mathbb{E}_1(\xi_3)+2\,\mathbb{E}_1(\xi_2^2)\\ &\quad +2\,\mathbb{E}_1(\xi_2)\mathbb{E}_1(\xi_3)+\mathbb{E}_1(\xi_3^2)\\ &= 3\sigma_1^2+2\sigma_1^2+\sigma_1^2\\ &= 6\sigma_1^2\ .\end{aligned}$$

N=4: Expanding the positions in terms of their increments yields

$$\begin{aligned}X_4^2+X_3^2+X_2^2+X_1^2+X_0^2 &= (\xi_1+\xi_2+\xi_3+\xi_4)^2+(\xi_1+\xi_2+\xi_3)^2+(\xi_1+\xi_2)^2+(\xi_1)^2+0\\ &= \xi_1^2+2\xi_1\xi_2+2\xi_1\xi_3+2\xi_1\xi_4+\xi_2^2+2\xi_2\xi_3+2\xi_2\xi_4+\xi_3^2+2\xi_3\xi_4+\xi_4^2\\ &\quad +3\xi_1^2+4\xi_1\xi_2+2\xi_1\xi_3+2\xi_2^2+2\xi_2\xi_3+\xi_3^2\\ &= 4\xi_1^2+6\xi_1\xi_2+4\xi_1\xi_3+2\xi_1\xi_4+3\xi_2^2+4\xi_2\xi_3+2\xi_2\xi_4+2\xi_3^2+2\xi_3\xi_4+\xi_4^2\ .\end{aligned}$$

N	$\mathbb{E}_1(\sum_{i=1}^{N} X_i^2)$
1	σ_2^2
2	$3\sigma_2^2$
3	$6\sigma_2^2$
4	$10\sigma_2^2$

TABLE 5. Summary of $\mathbb{E}_1(\sum_{i=1}^{N} X_i^2)$ for Case 3, where σ_1^2 is known.

So taking the expected value gives

$$\begin{aligned}
\mathbb{E}_1(X_4^2 + X_3^2 + X_2^2 + X_1^2 + X_0^2) &= \mathbb{E}_1(4\xi_1^2 + 6\xi_1\xi_2 + 4\xi_1\xi_3 + 2\xi_1\xi_4 + 3\xi_2^2 + 4\xi_2\xi_3 + 2\xi_2\xi_4 + 2\xi_3^2 + 2\xi_3\xi_4 + \xi_4^2) \\
&= 4\mathbb{E}_1(\xi_1^2) + 6\mathbb{E}_1(\xi_1)\mathbb{E}_1(\xi_2) + 4\mathbb{E}_1(\xi_1)\mathbb{E}_1(\xi_3) + 2\mathbb{E}_1(\xi_1)\mathbb{E}_1(\xi_4) + 3\mathbb{E}_1(\xi_2^2) \\
&\quad + 4\mathbb{E}_1(\xi_2\xi_3) + 2\mathbb{E}_1(\xi_2)\mathbb{E}_1(\xi_4) + 2\mathbb{E}_1(\xi_3^2) + 2\mathbb{E}_1(\xi_3)\mathbb{E}_1(\xi_4) + \mathbb{E}_1(\xi_4^2) \\
&= 4\sigma_1^2 + 3\sigma_1^2 + 2\sigma_1^2 + \sigma_1^2 \\
&= 10\sigma_1^2 \ .
\end{aligned}$$

The results for these values of N are summarized in Table 5. This pattern suggests that $\mathbb{E}_1(\sum_{i=1}^{N} X_i^2) = \frac{N(N+1)}{2}\sigma_1^2$. We can prove this using induction. We have already shown that this is true for the base case of $N = 1$. Assume the pattern hold for any given case $N = k$. We must show the pattern holds for $N = k + 1$:

$$
\begin{aligned}
\mathbb{E}_1\left(\sum_{i=1}^{k+1} X_i^2\right) &= \mathbb{E}_1\left(\sum_{i=1}^{k} X_i^2 + X_{k+1}^2\right) \\
&= \mathbb{E}_1\left(\sum_{i=1}^{k} X_i^2\right) + \mathbb{E}_1(X_{k+1}^2) \\
&= \frac{k(k+1)}{2}\sigma_1^2 + \mathbb{E}_1(X_{k+1}^2) \\
&= \frac{k(k+1)}{2}\sigma_1^2 + \mathbb{E}_1((\xi_1 + \xi_2 + \ldots + \xi_k + 1)^2) \\
&= \frac{k(k+1)}{2}\sigma_1^2 + \mathbb{E}_1(\xi_1^2 + 2\xi_1\xi_2 + 2\xi_1\xi_3 + \ldots + 2\xi_1\xi_{k+1} + \xi_2^2 + \ldots + \xi_{k+1}^2) \\
&= \frac{k(k+1)}{2}\sigma_1^2 + \mathbb{E}_1\left(\xi_1^2 + \xi_2^2 + \ldots + \xi_{k+1}^2 + 2\sum_{j=i+1}^{k+1}\sum_{i=1}^{k}\xi_i\xi_j\right) \\
&= \frac{k(k+1)}{2}\sigma_1^2 + \mathbb{E}_1(\xi_1^2) + \mathbb{E}_1(\xi_2^2) + \ldots + \mathbb{E}_1(\xi_{k+1}^2) + \mathbb{E}_1\left(2\sum_{j=i+1}^{k+1}\sum_{i=1}^{k}\xi_i\xi_j\right) \\
&= \frac{k(k+1)}{2}\sigma_1^2 + \mathbb{E}_1(\xi_1^2) + \mathbb{E}_1(\xi_2^2) + \ldots + \mathbb{E}_1(\xi_{k+1}^2) + 2\sum_{j=i+1}^{k+1}\sum_{i=1}^{k}\mathbb{E}_1(\xi_i)\mathbb{E}_1(\xi_j) \\
&= \frac{k(k+1)}{2}\sigma_1^2 + (k+1)\sigma_1^2 \\
&= \frac{k(k+1)}{2}\sigma_1^2 + \frac{2(k+1)}{2}\sigma_1^2 \\
&= \frac{k(k+1) + 2(k+1)}{2}\sigma_1^2 \\
&= \frac{(k+1)(k+2)}{2}\sigma_1^2 \ .
\end{aligned}
$$

Case	True Model	Likelihood Model	$\mathbb{E}$(AIC)
1	Free	Free	$N\ln(\sigma_1^2)+N\ln(2\pi)+N+2$
2	Anchored	Free	$(2N-1)\left(\frac{\sigma_2^2}{\sigma_1^2}\right)+N\ln(\sigma_1^2)+N\ln(2\pi)+2$
3	Free	Anchored	$\frac{N^2+N}{2}\left(\frac{\sigma_1^2}{\sigma_2^2}\right)+N\ln(\sigma_2^2)+N\ln(2\pi)+2$
4	Anchored	Anchored	$N\ln(\sigma_2^2)+N\ln(2\pi)+N+2$

TABLE 6. Expected value of the AIC based upon the specified true model and likelihood model for each case.

Thus, as proved by induction, $\mathbb{E}_1(\sum_{i=1}^N X_i^2) = \frac{N(N+1)}{2}\sigma_1^2$. We can substitute this in to evaluate the expected value of the AIC,

$$\begin{aligned}\mathbb{E}_1(\mathrm{AIC}_2) &= N\ln(\sigma_2^2)+N\ln(2\pi)+\frac{1}{\sigma_2^2}\sum_{i=1}^N \mathbb{E}_1(X_i^2)+2\\ &= N\ln(\sigma_2^2)+N\ln(2\pi)+\frac{1}{\sigma_2^2}\left(\frac{N(N+1)}{2}\right)\sigma_1^2+2\\ &= N\ln(\sigma_2^2)+N\ln(2\pi)+\frac{1}{\sigma_2^2}\frac{N^2+N}{2}\sigma_1^2+2\ .\end{aligned}$$

Discussion

Table 6 summarizes the expected value of the AIC in all four cases. We have determined the expected value of the difference between AIC_1 and AIC_2 in both cases:

$$\begin{aligned}\mathbb{E}_1(\mathrm{AIC}_2-\mathrm{AIC}_1) &= \frac{N^2+N}{2}\left(\frac{\sigma_1^2}{\sigma_2^2}\right)+N\ln\left(\frac{\sigma_2^2}{\sigma_1^2}\right)-N\ ,\\ \mathbb{E}_2(\mathrm{AIC}_1-\mathrm{AIC}_2) &= (2N-1)\left(\frac{\sigma_2^2}{\sigma_1^2}\right)+N\ln\left(\frac{\sigma_1^2}{\sigma_2^2}\right)-N\ .\end{aligned}$$

While the forms are similar, note that the expected value of the difference in the AIC's assuming model 1 (free diffusion) has an additional power of N in the leading term. Further, the "true" values for both σ_1^2 and σ_2^2 are in the expressions. This is due to our simplifying

assumption that the values of σ_1^2 and σ_2^2 are known, but in conducting model selection on a trajectory, we would only have one "true value," which will be unknown. This inconsistency is addressed in the next section.

5. Main Result: Convergence of the AIC for Estimated Model Parameters

We seek to understand the distribution of the AIC for estimated values of σ_1^2 and σ_2^2, because the trajectories we analyze will only have one of these values, and the true value will be unknown. We use maximum likelihood estimation on the observed increments or positions from the particle trajectories to estimate σ_1^2 and σ_2^2 as $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$ respectively. From these estimators, we can determine the expected value and variation of the AIC in each case.

We use the MLE for each case to revisit the AIC and better understand its distribution. Specifically, we will focus on $\frac{1}{N}\mathrm{AIC}_1$ and $\frac{1}{N}\mathrm{AIC}_2$. We rescale by $\frac{1}{N}$ in this manner for a stable outcome:

$$\frac{1}{N}\mathrm{AIC}_1 = \ln\left(\hat{\sigma}_1^2\right) + \ln(2\pi) + \frac{1}{N\hat{\sigma}_1^2}\sum_{i=1}^{N}\xi_i^2 + \frac{2}{N}\,, \tag{14}$$

$$\frac{1}{N}\mathrm{AIC}_2 = \ln\left(\hat{\sigma}_2^2\right) + \ln(2\pi) + \frac{1}{N\hat{\sigma}_2^2}\sum_{i=1}^{N}X_i^2 + \frac{2}{N}\,. \tag{15}$$

Using these equations, we arrive at the following results:

Theorem 5.1 (Case 1)**.** *Let $X_0, X_1, X_2, ..., X_N$ be the position observations of a free diffusion trajectory at times $t_0, t_1, t_2, ..., t_N$, where $t_i = \Delta i$. Define increments $\xi_1, \xi_2, ..., \xi_N$ as $\xi_i = X_i - X_{i-1}$ for $i = 1, ..., N$. Then*

$$\sqrt{N}\left(\frac{1}{N}AIC_1 - \left[\ln(\sigma_1^2) + \ln(2\pi) + 1\right]\right) \to N(0,2)$$

in distribution.

Theorem 5.2 (Case 4)**.** *Let $X_0, X_1, X_2, ..., X_N$ be the position observations of an anchored diffusion trajectory at times $t_0, t_1, t_2, ..., t_N$, where $t_i = \Delta i$. Then*

$$\sqrt{N}\left(\frac{1}{N}AIC_2 - \left[\ln(\sigma_2^2) + \ln(2\pi) + 1\right]\right) \to N(0,2)$$

in distribution.

Theorem 5.3 (Case 2)**.** *Let $X_0, X_1, X_2, ..., X_N$ be the position observations of an anchored diffusion trajectory at times $t_0, t_1, t_2, ..., t_N$, where $t_i = \Delta i$. Then*

$$\sqrt{N}\left(\frac{1}{N}AIC_1 - \left[\ln(2) + \ln(\sigma_2^2) + \ln(2\pi) + 1\right]\right) \to N(0,8)$$

in distribution.

We prove these theorems in the sections that follow, and make some remarks on why it is hard to make a similar statement for Case 3.

5.1. **Proof of Theorem 5.1 for Case 1: True Free Diffusion, Likelihood Model Free Diffusion.** Let $X_0, X_1, X_2, ..., X_N$ be the position observations of a free diffusion trajectory at times $t_0, t_1, t_2, ..., t_N$, where $t_i = \Delta i$. Define increments $\xi_1, \xi_2, ..., \xi_N$ as $\xi_i = X_i - X_{i-1}$ for $i = 1, ..., N$.

Note that

$$s_1^2 = \frac{1}{N-1}\sum_{i=1}^{N}\xi_i^2$$

is an unbiased estimator of σ_1^2, meaning that $\mathbb{E}_1(s_1^2) = \sigma_1^2$. We can further express s_1^2 as

$$s_1^2 = \frac{N}{N-1}\hat{\sigma}_1^2 \ ,$$

where $\hat{\sigma}_1^2$ is the MLE estimator for Case 1 from Table 9. This implies that

$$\hat{\sigma}_1^2 = \frac{N-1}{N}s_1^2 \ .$$

Since regularity conditions are met (refer to Appendix C), by the asymptotic efficiency of MLE's, Theorem B.5,

$$\sqrt{N}\left(\hat{\sigma}_1^2 - \sigma_1^2\right) \rightarrow N\left(0, v\left(\sigma_1^2\right)\right) \ ,$$

where $v\left(\sigma_1^2\right)$ is the Cramér-Rao Lower Bound for a single observation from Theorem B.4. Applied to our case of iid $\xi_1, \xi_2, ...\xi_N$, Corollary B.1 gives

$$v\left(\sigma_1^2\right) = \frac{1}{I\left(\sigma_1^2\right)} \ , \tag{16}$$

where $I\left(\sigma_1^2\right)$ is the Fisher Information. Since $f(\xi|\sigma_1^2)$ is an exponential family, by Lemma B.2,

$$\begin{aligned} I\left(\sigma_1^2\right) &= \mathbb{E}_{\sigma_1^2}\left(\left(\frac{\partial}{\partial \sigma_1^2}\ln f\left(\xi|\sigma_1^2\right)\right)^2\right) \\ &= -\mathbb{E}_{\sigma_1^2}\left(\frac{\partial^2}{\partial {\sigma_1^2}^2}\ln f\left(\xi|\sigma_1^2\right)\right) \\ &= -\mathbb{E}_{\sigma_1^2}\left(\frac{1}{2\left(\sigma_1^2\right)^2} - \frac{1}{\left(\sigma_1^2\right)^3}\xi^2\right) \\ &= \frac{-1}{2\left(\sigma_1^2\right)^2} + \frac{1}{\left(\sigma_1^2\right)^3}\sigma_1^2 \\ &= \frac{1}{2\left(\sigma_1^2\right)^2}\ . \end{aligned}$$

Thus,

$$\sqrt{N}\left(\hat{\sigma}_1^2 - \sigma_1^2\right) \to N\left(0, 2\left(\sigma_1^2\right)^2\right)\ ,$$

in distribution.

Next, we can apply the Delta Method, Theorem B.3. For a differentiable function A_1:

$$\sqrt{N}\left[A_1\left(\hat{\sigma}_1^2\right) - A_1\left(\sigma_1^2\right)\right] \to N\left(0, 2\left(\sigma_1^2\right)^2\left[A_1'\left(\sigma_1^2\right)\right]^2\right)\ .$$

Suppose $A_1(\sigma_1^2) = \ln(\sigma_1^2)$. Then $A_1'(\sigma_1^2) = \frac{1}{\sigma_1^2}$, and thus

$$\sqrt{N}\left[\ln\left(\hat{\sigma}_1^2\right) - \ln\left(\sigma_1^2\right)\right] \to N(0, 2)\ ,$$

in distribution.

Let $Y_N = \ln(2\pi) + 1 + \frac{2}{N}$. Observe that $Y_N \to \ln(2\pi) + 1$ in probability. Then, by Theorem B.2, Slutsky's Theorem,

$$\sqrt{N}\left(\frac{1}{N}\,\mathrm{AIC}_1 - \left[\ln(\sigma_1^2) + \ln(2\pi) + 1\right]\right) \to N(0,2)$$

in distribution, as desired.

5.2. **Proof of Theorem 5.2 for Case 4: True Anchored Diffusion, Likelihood Model Anchored Diffusion.** Recall that in our model for Free Diffusion, the increments are independently and identically distributed: $\xi_i \sim N(0,\sigma_1^2)$, where $i = 1,2,...,N$. Our model for Anchored Diffusion differs in that it is the positions, not the increments, that are independently and identically distributed: $X_i \sim N(0,\sigma_2^2)$, where $i = 0,1,2,...,N$ and $X_0 = 0$. Due to the way we have written the parameters of the models and their similar normal distributions, the logic of the proof of Theorem 5.2 is identical to the proof of Theorem 5.1. We simply replace the ξ_i terms with $\boldsymbol{X}_i$ terms, σ_1^2 with σ_2^2, and $\hat{\sigma}_1^2$ with $\hat{\sigma}_2^2$.

5.3. **Lemma: Maximum Likelihood Estimation (MLE) in Model Misspecification Cases.** Recall that the Maximum Likelihood Estimators for the free diffusion and anchored diffusion models are:

$$\hat{\sigma}_1^2 = \frac{1}{N}\sum_{i=1}^{N} \xi_i^2 \,,$$

$$\hat{\sigma}_2^2 = \frac{1}{N}\sum_{i=1}^{N} X_i^2 \,.$$

These estimators can be applied directly to Case 1 (True Free Diffusion, Likelihood Free Diffusion) and Case 4 (True Anchored Diffusion, Likelihood Anchored Diffusion). However, for Cases 2 and 4, where we have model misspecification, we need to re-express these estimators.

Case 2: True Anchored Diffusion, Likelihood Free Diffusion

We need to re-express $\hat{\eta}_D = \left(\frac{1}{N}\sum_{i=1}^{N}\xi_i^2\right)^{-1}$ in terms of our positions $X_0, X_1, X_2, ..., X_N$. We rewrite $\sum_{i=1}^{N}\xi_i^2$ for several small values of N, and then use these to determine a general solution.

N=1: We rewrite the first increment, ξ_1 in terms of the positions, $X_1 - X_0$, where $X_0 = 0$, which yields

$$\xi_1^2 = (X_1 - X_0)^2 = (X_1 - 0)^2 = X_1^2 \, .$$

N=2: Rewriting increments in terms of positions and expanding gives

$$\xi_1^2 + \xi_2^2 = (X_1 - X_0)^2 + (X_2 - X_1)^2 = X_1^2 + (X_2 - X_1)^2$$

$$= 2X_1^2 - 2X_1X_2 + X_2^2 \, .$$

N=3: We expand the increments in terms of the positions to see that

$$\begin{aligned}\xi_1^2+\xi_2^2+\xi_3^2 &= (X_1-X_0)^2+(X_2-X_1)^2+(X_3-X_2)^2\\ &= 2X_1^2-2X_1X_2+X_2^2+(X_3-X_2)^2\\ &= 2X_1^2-2X_1X_2+2X_2^2-2X_2X_3+X_3^2\ .\end{aligned}$$

N=4: Expanding the increments in terms of the positions gives

$$\begin{aligned}\xi_1^2+\xi_2^2+\xi_3^2+\xi_4^2 &= (X_1-X_0)^2+(X_2-X_1)^2+(X_3-X_2)^2+(X_4-X_3)^2\\ &= 2X_1^2-2X_1X_2+2X_2^2-2X_2X_3+2X_3^2-2X_3X_4+X_4^2\ .\end{aligned}$$

The results for these values of N are summarized in Table 7.

N	$\sum_{i=1}^{N}\xi_i^2$
1	X_1^2
2	$2X_1^2-2X_1X_2+X_2^2$
3	$2X_1^2-2X_1X_2+2X_2^2-2X_2X_3+X_3^2$
4	$2X_1^2-2X_1X_2+2X_2^2-2X_2X_3+2X_3^2-2X_3X_4+X_4^2$

TABLE 7. Rewriting $\sum_{i=1}^{N}\xi_i^2$ in terms of positions, X_i, for Case 2.

This pattern suggests that

$$\sum_{i=1}^{N}\xi_k^2 = X_N^2+2\sum_{i=1}^{N-1}X_i^2-2\sum_{j=1}^{N-1}X_jX_{j+1}\ .$$

We can prove this using induction. We have already shown that this is true for the base case of $N = 1$. Assume the pattern holds for any given case $N = k$. We must show the pattern holds for $N = k + 1$:

$$\begin{aligned}
\sum_{i=1}^{k+1} \xi_i^2 &= \sum_{i=1}^{k} \xi_i^2 + \xi_{k+1}^2 \\
&= X_k^2 + 2\sum_{i=1}^{k-1} X_i^2 - 2\sum_{j=1}^{k-1} X_j X_{j+1} + \xi_{k+1}^2 \\
&= X_k^2 + 2\sum_{i=1}^{k-1} X_i^2 - 2\sum_{j=1}^{k-1} X_j X_{j+1} + (X_{k+1} - X_k)^2 \\
&= X_k^2 + 2\sum_{i=1}^{k-1} X_i^2 - 2\sum_{j=1}^{k-1} X_j X_{j+1} + X_{k+1}^2 - 2X_k X_{k+1} + X_k^2 \\
&= X_{k+1}^2 + 2X_k^2 + 2\sum_{i=1}^{k-1} X_i^2 - 2\sum_{j=1}^{k-1} X_j X_{j+1} - 2X_k X_{k+1} \\
&= X_{k+1}^2 + 2\sum_{i=1}^{k} X_i^2 - 2\sum_{j=1}^{k} X_j X_{j+1} \, .
\end{aligned}$$

Thus, we have proven by induction $\sum_{i=1}^{k} \xi_k^2 = X_N^2 + 2\sum_{i=1}^{N-1} X_i^2 - 2\sum_{j=1}^{N-1} X_j X_{j+1}$. Substituting this into the expression for $\hat{\sigma}_1^2$ yields

$$\hat{\sigma}_1^2 = \frac{1}{N}\sum_{i=1}^{N} \xi_i^2 = \frac{1}{N}\left[X_N^2 + 2\sum_{i=1}^{N-1} X_i^2 - 2\sum_{j=1}^{N-1} X_j X_{j+1}\right] \tag{17}$$

as the MLE for the free diffusion likelihood on true anchored diffusion.

Case 3: True Free Diffusion, Likelihood Anchored Diffusion

We need to re-express $\hat{\sigma}_2^2 = \frac{1}{N}\sum_{i=1}^{N} X_i^2 s$ in terms of our increments $\xi_1, \xi_2, ..., \xi_N$. We rewrite $\sum_{i=1}^{N} X_i^2$ for several small values of N, and then use these to determine a general solution.

N=1: We write the positions X_1 and X_0 in terms of the increment between them, ξ_1, to give

$$X_1^2 + X_0^2 = (\xi_1)^2 + 0 = \xi_1^2 \, .$$

N=2: Re-expressing the positions in terms of increments and expanding yields

$$X_2^2 + X_1^2 + X_0^2 = (\xi_1 + \xi_2)^2 + (\xi_1)^2 + 0 = 2\xi_1^2 + 2\xi_1\xi_2 + \xi_2^2 \, .$$

N=3: We expand the positions in terms of the increments to determine that

$$\begin{aligned} X_3^2 + X_2^2 + X_1^2 + X_0^2 &= (\xi_1 + \xi_2 + \xi_3)^2 + (\xi_1 + \xi_2)^2 + (\xi_1)^2 + 0 \\ &= \xi_1^2 + 2\xi_1\xi_2 + 2\xi_1\xi_3 + \xi_2^2 + 2\xi_2\xi_3 + \xi_3^2 + 2\xi_1^2 + 2\xi_1\xi_2 + \xi_2^2 \\ &= 3\xi_1^2 + 4\xi_1\xi_2 + 2\xi_1\xi_3 + 2\xi_2^2 + 2\xi_2\xi_3 + \xi_3^2 \, . \end{aligned}$$

N=4: Again, we rewrite the positions as sums of the proceeding increments and expand to give

$$\begin{aligned} X_4^2 + X_3^2 + X_2^2 + X_1^2 + X_0^2 &= (\xi_1 + \xi_2 + \xi_3 + \xi_4)^2 + (\xi_1 + \xi_2 + \xi_3)^2 + (\xi_1 + \xi_2)^2 + (\xi_1)^2 + 0 \\ &= \xi_1^2 + 2\xi_1\xi_2 + 2\xi_1\xi_3 + 2\xi_1\xi_4 + \xi_2^2 + 2\xi_2\xi_3 + 2\xi_2\xi_4 + \xi_3^2 + 2\xi_3\xi_4 + \xi_4^2 \\ &\quad + 3\xi_1^2 + 4\xi_1\xi_2 + 2\xi_1\xi_3 + 2\xi_2^2 + 2\xi_2\xi_3 + \xi_3^2 \\ &= 4\xi_1^2 + 6\xi_1\xi_2 + 4\xi_1\xi_3 + 2\xi_1\xi_4 + 3\xi_2^2 + 4\xi_2\xi_3 + 2\xi_2\xi_4 + 2\xi_3^2 + 2\xi_3\xi_4 + \xi_4^2 \, . \end{aligned}$$

The results for these values of N are summarized in Table 8.

N	$\sum_{i=1}^{N} X_i^2$
1	ξ_1^2
2	$2\xi_1^2 + 2\xi_1\xi_2 + \xi_2^2$
3	$3\xi_1^2 + 4\xi_1\xi_2 + 2\xi_1\xi_3 + 2\xi_2^2 + 2\xi_2\xi_3 + \xi_3^2$
4	$4\xi_1^2 + 6\xi_1\xi_2 + 4\xi_1\xi_3 + 2\xi_1\xi_4 + 3\xi_2^2 + 4\xi_2\xi_3 + 2\xi_2\xi_4 + 2\xi_3^2 + 2\xi_3\xi_4 + \xi_4^2$

TABLE 8. Rewriting $\sum_{i=1}^{N} X_i^2$ in terms of increments, ξ_i, for Case 3.

This pattern suggests that, for $N > 1$,

$$\sum_{i=1}^{N} X_i^2 = \sum_{p=1}^{N} (N - p + 1)\xi_p^2 + 2 \sum_{j=i+1}^{N} \sum_{i=1}^{N-1} (N - j + 1)\xi_i\xi_j \, .$$

We can prove this using induction. We have already shown that this is true for the base case of $N = 2$. Assume the pattern holds for any given case $N = k$. We must show the pattern

holds for $N = k+1$:

$$\begin{aligned}
\sum_{i=1}^{k+1} X_i^2 &= \sum_{i=1}^{k} X_i^2 + X_{k+1}^2 \\
&= \sum_{p=1}^{k} (k-p+1)\xi_p^2 + 2\sum_{j=i+1}^{k}\sum_{i=1}^{k-1}(k-j+1)\xi_i\xi_j + X_{k+1}^2 \\
&= \sum_{p=1}^{k} (k-p+1)\xi_p^2 + 2\sum_{j=i+1}^{k}\sum_{i=1}^{k-1}(k-j+1)\xi_i\xi_j + (\xi_1+\xi_2+\ldots+\xi_{k+1})^2 \\
&= \sum_{p=1}^{k} (k-p+1)\xi_p^2 + 2\sum_{j=i+1}^{k}\sum_{i=1}^{k-1}(k-j+1)\xi_i\xi_j + \xi_1^2 + 2\xi_1\xi_2 + 2\xi_1\xi_3 + \ldots + 2\xi_1\xi_{k+1} + \xi_2^2 + \ldots + \xi_{k+1}^2 \\
&= \sum_{p=1}^{k} (k-p+1)\xi_p^2 + 2\sum_{j=i+1}^{k}\sum_{i=1}^{k-1}(k-j+1)\xi_i\xi_j + \xi_1^2 + \xi_2^2 + \ldots + \xi_{k+1}^2 + 2\sum_{j=i+1}^{k+1}\sum_{i=1}^{k}\xi_i\xi_j \\
&= \sum_{p=1}^{k} ((k+1)-p+1)\xi_p^2 + 2\sum_{j=i+1}^{k}\sum_{i=1}^{k-1}((k+1)-j+1)\xi_i\xi_j + \xi_{k+1}^2 + 2\xi_k\xi_{k+1} \\
&= \sum_{p=1}^{k+1} ((k+1)-p+1)\xi_p^2 + 2\sum_{j=i+1}^{k+1}\sum_{i=1}^{k}((k+1)-j+1)\xi_i\xi_j \ .
\end{aligned}$$

Thus, we have proven by induction that $\sum_{i=1}^{N} X_i^2 = \sum_{p=1}^{N}(N-p+1)\xi_p^2 + 2\sum_{j=i+1}^{N}\sum_{i=1}^{N-1}(N-j+1)\xi_i\xi_j$ for $N > 1$. Substituting this into the expression for $\hat{\sigma}_2^2$ yields

$$\begin{aligned}
\hat{\sigma}_2^2 &= \frac{1}{N}\sum_{i=1}^{N} X_i^2 \\
&= \frac{1}{N}\Big[\sum_{p=1}^{N}(N-p+1)\xi_p^2 + 2\sum_{j=i+1}^{N}\sum_{i=1}^{N-1}(N-j+1)\xi_i\xi_j\Big]
\end{aligned}$$

as the MLE for the anchored diffusion likelihood on true free diffusion.

Table 9 summarizes the MLE in all four cases. Note that in each case, the estimator is in terms of the observations from the trajectory itself - the observed increments in Case 1 and Case 3 for true free diffusion paths, and the observed positions in Case 2 and Case

Case	True Model	Likelihood Model	MLE
1	Free	Free	$\hat{\sigma}_1^2 = \frac{1}{N}\sum_{i=1}^{N}\xi_i^2$
2	Anchored	Free	$\hat{\sigma}_1^2 = \frac{1}{N}\left[X_N^2 + 2\sum_{i=1}^{N-1}X_i^2 - 2\sum_{j=1}^{N-1}X_jX_{j+1}\right]$
3	Free	Anchored	$\hat{\sigma}_2^2 = \frac{1}{N}\left[\sum_{p=1}^{N}(N-p+1)\xi_p^2 + 2\sum_{j=i+1}^{N}\sum_{i=1}^{N-1}(N-j+1)\xi_i\xi_j\right]$
4	Anchored	Anchored	$\hat{\sigma}_2^2 = \frac{1}{N}\sum_{i=1}^{N}X_i^2$

TABLE 9. Maximum Likelihood Estimator (MLE) based upon the specified true model and likelihood model for each case.

4 for true anchored diffusion paths. Computing these estimators allows us to explore the distributions of $\frac{1}{N}$AIC in the model misspecification cases.

5.4. **Proof of Theorem 5.3 for Case 2: True Anchored Diffusion, Likelihood Model Free Diffusion.** Let $X_0, X_1, X_2, ..., X_N$ be the position observations of an anchored diffusion trajectory at times $t_0, t_1, t_2, ..., t_N$, where $t_i = \Delta i$. Define increments $\xi_1, \xi_2, ..., \xi_N$ as $\xi_i = X_i - X_{i-1}$ for $i = 1, ..., N$. We will be working with $\frac{1}{N}\text{AIC}_1$ from Equation 14, but we need to determine its expected value in terms of the true parameter, σ_2^2. To accomplish this, we focus on $\ln\left(\hat{\sigma}_1^2\right)$.

We begin with the MLE for this case from Table 9:

$$
\begin{aligned}
\hat{\sigma}_1^2 &= \frac{1}{N}\sum_{i=1}^{N}\xi_i^2 \\
&= \frac{1}{N}\left[X_N^2 + 2\sum_{i=1}^{N-1}X_i^2 - 2\sum_{j=1}^{N-1}X_jX_{j+1}\right] \\
&= \frac{1}{N}\left[X_N^2\right] + \frac{2}{N}\left[\sum_{i=1}^{N-1}X_i^2\right] - \frac{2}{N}\left[\sum_{j=1}^{N-1}X_jX_{j+1}\right] \\
&= \frac{1}{N}\left[X_N^2\right] + \frac{1}{N}\left[X_N^2\right] - \frac{1}{N}\left[X_N^2\right] + \frac{2}{N}\left[\sum_{i=1}^{N-1}X_i^2\right] - \frac{2}{N}\left[\sum_{j=1}^{N-1}X_jX_{j+1}\right] \\
&= \frac{2}{N}\left[X_N^2\right] + \frac{2}{N}\left[\sum_{i=1}^{N-1}X_i^2\right] - \frac{1}{N}\left[X_N^2\right] - \frac{2}{N}\left[\sum_{j=1}^{N-1}X_jX_{j+1}\right] \\
&= \frac{2}{N}\left[\sum_{i=1}^{N}X_i^2\right] - \frac{1}{N}\left[X_N^2\right] - \frac{2}{N}\left[\sum_{j=1}^{N-1}X_jX_{j+1}\right] \\
&= 2\hat{\sigma}_2^2 - \frac{1}{N}\left[X_N^2\right] - \frac{2}{N}\left[\sum_{j=1}^{N-1}X_jX_{j+1}\right] .
\end{aligned}
$$

Taking the expected value yields

$$\begin{aligned}
\mathbb{E}_{\sigma_2^2}\left(\hat{\sigma}_1^2\right) &= \mathbb{E}_{\sigma_2^2}\left(2\hat{\sigma}_2^2 - \frac{1}{N}\left[X_N^2\right] - \frac{2}{N}\left[\sum_{j=1}^{N-1} X_j X_{j+1}\right]\right) \\
&= 2\mathbb{E}_{\sigma_2^2}\left(\hat{\sigma}_2^2\right) - \frac{1}{N}\mathbb{E}_{\sigma_2^2}\left(X_N^2\right) - \frac{2}{N}\sum_{j=1}^{N-1}\mathbb{E}_{\sigma_2^2}\left(X_j\right)\mathbb{E}_{\sigma_2^2}\left(X_{j+1}\right) \\
&= 2\mathbb{E}_{\sigma_2^2}\left(\frac{N-1}{N}s_2^2\right) - \frac{1}{N}\sigma_2^2 \\
&= 2\left(\frac{N-1}{N}\right)\mathbb{E}_{\sigma_2^2}\left(s_2^2\right) - \frac{1}{N}\sigma_2^2 \\
&= 2\left(\frac{N-1}{N}\right)\sigma_2^2 - \frac{1}{N}\sigma_2^2 \\
&= \left(\frac{2N-3}{N}\right)\sigma_2^2 \\
&= \left(2 - \frac{3}{N}\right)\sigma_2^2 \ .
\end{aligned}$$

Let $Z_N = \frac{1}{N}\left[X_N^2\right] + \frac{2}{N}\left[\sum_{j=1}^{N-1} X_j X_{j+1}\right]$. By Chebyshev's Inequality, Theorem B.1, since

$$\begin{aligned}
\mathbb{E}_{\sigma_2^2}\left(Z_N^2\right) &= \frac{1}{N^2}\mathbb{E}_{\sigma_2^2}\left(X_N^4\right) + \frac{2}{N^2}\mathbb{E}_{\sigma_2^2}\left(\boldsymbol{X}_N^2\left[\sum_{j=1}^{N-1} X_j X_{j+1}\right]\right) + \frac{4}{N^2}\mathbb{E}_{\sigma_2^2}\left(\sum_{j=1}^{N-1} X_j^2 X_{j+1}^2\right) \\
&= \frac{1}{N^2}\mathbb{E}_{\sigma_2^2}\left(X_N^4\right) + \frac{2}{N^2}\mathbb{E}_{\sigma_2^2}\left(\boldsymbol{X}_N^2\right)\sum_{j=1}^{N-1}\mathbb{E}_{\sigma_2^2}\left(X_j\right)\mathbb{E}_{\sigma_2^2}\left(X_{j+1}\right) + \frac{4}{N^2}\sum_{j=1}^{N-1}\mathbb{E}_{\sigma_2^2}\left(X_j^2\right)\mathbb{E}_{\sigma_2^2}\left(X_{j+1}^2\right) \\
&= \frac{3\left(\sigma_2^2\right)^2}{N^2} + \frac{4N\left(\sigma_2^2\right)^2}{N^2} \\
&= \frac{3\left(\sigma_2^2\right)^2}{N^2} + \frac{4\left(\sigma_2^2\right)^2}{N} \ ,
\end{aligned}$$

then $Z_N \to 0$ in probability.

Recall that

$$\sqrt{N}\left(\hat{\sigma}_2^2 - \sigma_2^2\right) \to N\left(0, 2\left(\sigma_2^2\right)^2\right),$$

from the Proof of Theorem 5.2. Then, since $Z_N \to 0$ in probability, by Theorem B.2 (Slutsky's Theorem),

$$\sqrt{N}\left(\hat{\sigma}_1^2 - 2\sigma_2^2\right) \to N\left(0, 8\left(\sigma_2^2\right)^2\right).$$

Applying the Delta Method, Theorem B.3, for a differentiable function A_2:

$$\sqrt{N}\left[A_2\left(\hat{\sigma}_1^2\right) - A_2\left(2\sigma_2^2\right)\right] \to N\left(0, 8\left(\sigma_2^2\right)^2\left[A_2'\left(2\sigma_2^2\right)\right]^2\right).$$

Suppose $A_2(2\sigma_2^2) = \ln(2\sigma_2^2)$. Then $A_2'(2\sigma_2^2) = \frac{1}{\sigma_2^2}$ and

$$\sqrt{N}\left[\ln\left(\hat{\sigma}_1^2\right) - \ln\left(2\sigma_2^2\right)\right] \to N(0, 8)$$

in distribution. Let $Y_N = \ln(2\pi) + 1 + \frac{2}{N}$. Observe that $Y_N \to \ln(2\pi) + 1$ in probability. Then, by Theorem B.2 (Slutsky's Theorem),

$$\sqrt{N}\left(\frac{1}{N}\,\mathrm{AIC}_1 - \left[\ln(2) + \ln(\sigma_2^2) + \ln(2\pi) + 1\right]\right) \to N(0, 8)$$

in distribution, as desired.

5.5. **Remarks about Case 3: True Free Diffusion, Likelihood Model Anchored Diffusion.** Let $X_0, X_1, X_2, ..., X_N$ be the position observations of a free diffusion trajectory at

times $t_0, t_1, t_2, ..., t_N$, where $t_i = \Delta i$. Define increments $\xi_1, \xi_2, ..., \xi_N$ as $\xi_i = X_i - X_{i-1}$ for $i = 1, ..., N$. We will be working with $\frac{1}{N}\mathrm{AIC}_2$ from Equation 15, but we need to determine its expected value in terms of the true parameter, σ_1^2. To accomplish this, we focus on $\ln(\hat{\sigma}_2^2)$.

We begin with the MLE for this case from Table 9:

$$
\begin{aligned}
\hat{\sigma}_2^2 &= \frac{1}{N}\sum_{i=1}^{N} X_i^2 \\
&= \frac{1}{N}\left[\sum_{p=1}^{N}(N-p+1)\xi_p^2 + 2\sum_{j=i+1}^{N}\sum_{i=1}^{N-1}(N-j+1)\xi_i\xi_j\right] \\
&= \frac{1}{N}\left[\sum_{k=1}^{N}\xi_k^2 + \sum_{p=1}^{N}(N-p)\xi_p^2 + 2\sum_{j=i+1}^{N}\sum_{i=1}^{N-1}(N-j+1)\xi_i\xi_j\right] \\
&= \frac{1}{N}\sum_{k=1}^{N}\xi_k^2 + \frac{1}{N}\sum_{p=1}^{N}(N-p)\xi_p^2 + \frac{2}{N}\sum_{j=i+1}^{N}\sum_{i=1}^{N-1}(N-j+1)\xi_i\xi_j \\
&= \hat{\sigma}_1^2 + \frac{1}{N}\sum_{p=1}^{N}(N-p)\xi_p^2 + \frac{2}{N}\sum_{j=i+1}^{N}\sum_{i=1}^{N-1}(N-j+1)\xi_i\xi_j \ .
\end{aligned}
$$

Taking the expected value yields

$$
\begin{aligned}
\mathbb{E}_{\sigma_1^2}\left(\hat{\sigma}_2^2\right) &= \mathbb{E}_{\sigma_1^2}\left(\hat{\sigma}_1^2 + \frac{1}{N}\sum_{p=1}^{N}(N-p)\xi_p^2 + \frac{2}{N}\sum_{j=i+1}^{N}\sum_{i=1}^{N-1}(N-j+1)\xi_i\xi_j\right) \\
&= \mathbb{E}_{\sigma_1^2}\left(\hat{\sigma}_1^2\right) + \mathbb{E}_{\sigma_1^2}\left(\frac{1}{N}\sum_{p=1}^{N}(N-p)\xi_p^2\right) + \mathbb{E}_{\sigma_1^2}\left(\frac{2}{N}\sum_{j=i+1}^{N}\sum_{i=1}^{N-1}(N-j+1)\xi_i\xi_j\right) \\
&= \mathbb{E}_{\sigma_1^2}\left(\frac{N-1}{N}s_1^2\right) + \frac{1}{N}\sum_{p=1}^{N}(N-p)\mathbb{E}_{\sigma_1^2}\left(\xi_p^2\right) + \frac{2}{N}\sum_{j=i+1}^{N}\sum_{i=1}^{N-1}(N-j+1)\mathbb{E}_{\sigma_1^2}(\xi_i)\mathbb{E}_{\sigma_1^2}\left(\xi_j\right) \\
&= \mathbb{E}_{\sigma_1^2}\left(\frac{N-1}{N}s_1^2\right) + \frac{1}{N}\sum_{p=1}^{N}(N-p)\mathbb{E}_{\sigma_1^2}\left(\xi_p^2\right) \\
&= \left(\frac{N-1}{N}\right)\sigma_1^2 + \frac{1}{N}\left[(N-1)\mathbb{E}_{\sigma_1^2}\left(\hat{\sigma}_{N-1}^2\right) + (N-2)\mathbb{E}_{\sigma_1^2}\left(\hat{\sigma}_{N-2}^2\right) + \ldots + \mathbb{E}_{\sigma_1^2}\left(\hat{\sigma}_1^2\right)\right] \\
&= \left(\frac{N-1}{N}\right)\sigma_1^2 + \frac{1}{N}\left[(N-1)\mathbb{E}_{\sigma_1^2}\left(\frac{1}{N-1}\sum_{i=1}^{N-1}\xi_i^2\right) + (N-2)\mathbb{E}_{\sigma_1^2}\left(\frac{1}{N-2}\sum_{i=1}^{N-2}\xi_i^2\right) + \ldots + \mathbb{E}_{\sigma_1^2}\left(\xi_1^2\right)\right] \\
&= \left(\frac{N-1}{N}\right)\sigma_1^2 + \frac{1}{N}\left[(N-1)\mathbb{E}_{\sigma_1^2}\left(\xi_1^2\right) + (N-2)\mathbb{E}_{\sigma_1^2}\left(\xi_1^2\right) + \ldots + \mathbb{E}_{\sigma_1^2}\left(\xi_1^2\right)\right] \\
&= \left(\frac{N-1}{N}\right)\sigma_1^2 + \frac{1}{N}\left[(N-1)\sigma_1^2 + (N-2)\sigma_1^2 + \ldots + \sigma_1^2\right] \\
&= \left(\frac{N-1}{N}\right)\sigma_1^2 + \frac{1}{N}\left[\sigma_1^2\sum_{i=1}^{N-1} i\right] \\
&= \left(\frac{N-1}{N}\right)\sigma_1^2 + \frac{1}{N}\left[\sigma_1^2\frac{(N-1)((N-1)+1)}{2}\right] \\
&= \left(\frac{N-1}{N}\right)\sigma_1^2 + \frac{1}{N}\left[\sigma_1^2\frac{(N-1)(N)}{2}\right] \\
&= \left(\frac{N-1}{N}\right)\sigma_1^2 + \left(\frac{N-1}{2}\right)\sigma_1^2 \\
&= \frac{2(N-1)+N(N-1)}{2N}\sigma_1^2 \\
&= \frac{N^2+N-2}{2N}\sigma_1^2 \\
&= \left(\frac{1}{2}N + \frac{1}{2} - \frac{1}{N}\right)\sigma_1^2 \ .
\end{aligned}
$$

Due to the N term in the expression for $\mathbb{E}_{\sigma_1^2}\left(\hat{\sigma}_2^2\right)$, we are not able to use the same methods as in proving Theorem 5.3. While we were not able to determine the distribution of $\frac{1}{N}\,\mathrm{AIC}_2$ in this case, we were able to make a contribution by unpacking the MLE to determine the above pattern.

6. CONCLUSION

We focused on two specific movement types - free diffusion and anchored diffusion- and used well-established tools from statistics and information theory to perform rigorous model selection. We developed models for both movement types based upon independent and identically distributed increments and positions respectively. We have shown that it is more difficult for these methods to discern the true model when the truth is an anchored diffusion trajectory in two ways. First, we numerically compared the performance of the AIC and Bayes Factors on simulated trajectories. Then, we explored the distribution for the AIC more in depth, analyzing four cases which paired different true and likelihood models.

While we were not able to determine the distribution of the test statistic for the likelihood ratio test, $\mathrm{AIC}_1 - \mathrm{AIC}_2$, we do know the (preconvergence) form. For true free diffusion trajectories (Cases 1 and 3):

$$\begin{aligned}
\mathrm{AIC}_1 - \mathrm{AIC}_2 &= -(\mathrm{AIC}_2 - \mathrm{AIC}_1) \\
&= -N\left[\ln\left(\hat{\sigma}_1^2 + \frac{1}{N}\sum_{p=1}^{N}(N-p)\xi_p^2 + \frac{2}{N}\sum_{j=i+1}^{N}\sum_{i=1}^{N-1}(N-j+1)\xi_i\xi_j\right) - \ln\left(\hat{\sigma}_1^2\right)\right] \\
&= -N\ln\left(\frac{\hat{\sigma}_1^2 + \frac{1}{N}\sum_{p=1}^{N}(N-p)\xi_p^2 + \frac{2}{N}\sum_{j=i+1}^{N}\sum_{i=1}^{N-1}(N-j+1)\xi_i\xi_j}{\hat{\sigma}_1^2}\right) \\
&= -N\ln\left(1 + \frac{\frac{1}{N}\sum_{p=1}^{N}(N-p)\xi_p^2}{\hat{\sigma}_1^2} + \frac{\frac{2}{N}\sum_{j=i+1}^{N}\sum_{i=1}^{N-1}(N-j+1)\xi_i\xi_j}{\hat{\sigma}_1^2}\right) \\
&= -N\ln\left(1 + \frac{\sum_{p=1}^{N}(N-p)\xi_p^2}{\sum_{k=1}^{N}\xi_k^2} + \frac{2\sum_{j=i+1}^{N}\sum_{i=1}^{N-1}(N-j+1)\xi_i\xi_j}{\sum_{k=1}^{N}\xi_k^2}\right).
\end{aligned}$$

So altogether, rewriting Equation 10, the form of a rejection region for a likelihood ratio test would be

$$\left\{ \vec{X} : \frac{\sum_{k=1}^{N}(N-k)\xi_k^2}{\sum_{k=1}^{N}\xi_k^2} + \frac{2\sum_{j=i+1}^{N}\sum_{i=1}^{N-1}(N-j+1)\xi_i\xi_j}{\sum_{k=1}^{N}\xi_k^2} \geq C_1 \right\}, \tag{18}$$

for some appropriately chosen critical value, C_1.

For true anchored diffusion trajectories (Cases 2 and 4):

$$\begin{aligned}
\mathrm{AIC}_1 - \mathrm{AIC}_2 &= N\ln\left(2\hat{\sigma}_2^2 - \frac{1}{N}\left[X_N^2\right] - \frac{2}{N}\left[\sum_{j=1}^{N-1} X_j X_{j+1}\right]\right) - N\ln\left(\hat{\sigma}_2^2\right) \\
&= N\ln\left(\frac{2\hat{\sigma}_2^2 - \frac{1}{N}\left[X_N^2\right] - \frac{2}{N}\left[\sum_{j=1}^{N-1} X_j X_{j+1}\right]}{\hat{\sigma}_2^2}\right) \\
&= N\ln\left(2 - \frac{\frac{1}{N}\left[X_N^2\right]}{\hat{\sigma}_2^2} - \frac{\frac{2}{N}\left[\sum_{j=1}^{N-1} X_j X_{j+1}\right]}{\hat{\sigma}_2^2}\right) \\
&= N\ln\left(2 - \frac{X_N^2}{\sum_{k=1}^{N} X_i^2} - \frac{2\left[\sum_{j=1}^{N-1} X_j X_{j+1}\right]}{\sum_{k=1}^{N} X_i^2}\right).
\end{aligned}$$

So altogether, rewriting Equation 10, the form of a rejection region for a likelihood ratio test would be

$$\left\{ \vec{X} : \frac{X_N^2}{\sum_{k=1}^{N} X_i^2} + \frac{2\left[\sum_{j=1}^{N-1} X_j X_{j+1}\right]}{\sum_{k=1}^{N} X_i^2} \leq C_2 \right\}, \tag{19}$$

for some appropriately chosen critical value, C_2.

6.1. **Future Work.** One direction of future work would be to continue the in-depth analysis of the AIC, try to find a lower bound for Case 3, and take further steps towards understanding the distribution of the test statistic $\mathrm{AIC}_1 - \mathrm{AIC}_2$. Further, a similar theoretical

analysis of other information criterion methods, such as the Watanabe-Akaike Information Criterion (WAIC), or Bayesian methods, like Bayes Factors, could be explored.

The authors hope that one day this analysis may be useful to include in complementary work, an automated dashboard to categorize different types of particle movement in cells. In that work, we developed six statistical features and used them as inputs to a supervised machine learning algorithm that had been trained on simulated data. Statistics like the AIC are candidates for improved statistical feature extraction. However, one would need to explore their performance on movement types beyond free and anchored diffusion. A natural next step would be to expand this investigation to compare free diffusion (where the particle is not attached to a molecular motor) to directed motion (where the particle is attached to a molecular motor that is stepping along a microtubule with some velocity).

REFERENCES

[1] Hirotugu Akaike. A new look at the statistical model identification. IEEE Transactions on Automatic Control, 19(6):716–723, 1974.

[2] Robert Brown. XXVII. A brief account of microscopical observations made in the months of June, July and August 1827, on the particles contained in the pollen of plants; and on the general existence of active molecules in organic and inorganic bodies. The Philosophical Magazine, 4(21):161–173, 1828.

[3] George Casella and Roger L Berger. Statistical Inference. Cengage Learning, 2021.

[4] Debashis Chatterjee, Trisha Maitra, and Sourabh Bhattacharya. A short note on almost sure convergence of Bayes Factors in the general set-up. The American Statistician, 74(1):17–20, 2020.

[5] Kejia Chen, Bo Wang, Juan Guan, and Steve Granick. Diagnosing heterogeneous dynamics in single-molecule/particle trajectories with multiscale wavelets. ACS Nano, 7(10):8634–8644, 2013.

[6] Siddhartha Chib. Marginal likelihood from the Gibbs output. Journal of the American Statistical Association, 90(432):1313–1321, 1995.

[7] Basilio de Bragança Pereira and Carlos Alberto de Bragança Pereira. Model Choice in Non-nested Families. Springer, 2016.

[8] Arthur P Dempster. The direct use of likelihood for significance testing. Statistics and Computing, 7(4):247–252, 1997.

[9] Albert Einstein. On the theory of the Brownian movement. Ann. Phys, 19(4):371–381, 1906.

[10] Helge Ewers, Alicia E Smith, Ivo F Sbalzarini, Hauke Lilie, Petros Koumoutsakos, and Ari Helenius. Single-particle tracking of murine polyoma virus-like particles on live cells and artificial membranes. Proceedings of the National Academy of Sciences, 102(42):15110–15115, 2005.

[11] R Ferrari, AJ Manfroi, and WR Young. Strongly and weakly self-similar diffusion. Physica D: Nonlinear Phenomena, 154(1-2):111–137, 2001.

[12] Syuan-Ming Guo, Jun He, Nilah Monnier, Guangyu Sun, Thorsten Wohland, and Mark Bathe. Bayesian approach to the analysis of fluorescence correlation spectroscopy data II: application to simulated and in vitro data. Analytical Chemistry, 84(9):3880–3888, 2012.

[13] Jo A Helmuth, Christoph J Burckhardt, Petros Koumoutsakos, Urs F Greber, and Ivo F Sbalzarini. A novel supervised trajectory segmentation algorithm identifies distinct types of human adenovirus motion in host cells. Journal of Structural Biology, 159(3):347–358, 2007.

[14] Harold Jeffreys. Some tests of significance, treated by the theory of probability. In Mathematical Proceedings of the Cambridge Philosophical Society, volume 31, pages 203–222. Cambridge University Press, 1935.

[15] Harold Jeffreys. The Theory of Probability. OUP Oxford, 1998.

[16] Melanie Jensen. Inference of biophysical states of microparticles from particle tracking data. PhD thesis, Tulane University, 2019.

[17] Robert E Kass and Adrian E Raftery. Bayes Factors. Journal of the American Statistical Association, 90(430):773–795, 1995.

[18] Don S Lemons and Anthony Gythiel. Paul Langevin's 1908 paper "on the theory of Brownian motion"["sur la théorie du mouvement brownien," cr acad. sci.(paris) 146, 530–533 (1908)]. American Journal of Physics, 65(11):1079–1081, 1997.

[19] Nilah Monnier. Bayesian Inference Approaches for Particle Trajectory Analysis in Cell Biology. PhD thesis, 2013.

[20] Nilah Monnier, Zachary Barry, Hye Yoon Park, Kuan-Chung Su, Zachary Katz, Brian P English, Arkajit Dey, Keyao Pan, Iain M Cheeseman, Robert H Singer, et al. Inferring transient particle transport dynamics in live cells. Nature Methods, 12(9):838, 2015.

[21] Fredrik Persson, Irmeli Barkefors, and Johan Elf. Single molecule methods with applications in living cells. Current Opinion in Biotechnology, 24(4):737–744, 2013.

[22] Hong Qian, Michael P Sheetz, and Elliot L Elson. Single particle tracking: Analysis of diffusion and flow in two-dimensional systems. Biophysical Journal, 60(4):910–921, 1991.

[23] Ivo F Sbalzarini and Petros Koumoutsakos. Feature point tracking and trajectory analysis for video imaging in cell biology. Journal of Structural Biology, 151(2):182–195, 2005.

[24] Sumio Watanabe and Manfred Opper. Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. Journal of Machine Learning Research, 11(12), 2010.

[25] Kui Zhang, Katelyn PR Crizer, Mark H Schoenfisch, David B Hill, and Gustavo Didier. Fluid heterogeneity detection based on the asymptotic distribution of the time-averaged mean squared displacement in single particle tracking experiments. Journal

of Physics A: Mathematical and Theoretical, 51(44):445601, 2018.

APPENDIX A. JUSTIFICATION OF MODELING CHOICES

The following argument was provided by my advisor, Prof. Scott A. McKinley. It will appear in a forthcoming paper, but we have included the text here for reader convenience:

In both active transport and anchored diffusion models it is assumed that the particle is bound to some other structure and its movement is dictated by the properties of that biophysical interaction. It is natural to model the response of a Brownian particle to external force through the Langevin equation. Let $\{V(t)\}_{t\in\mathbb{R}_+}$ denote the velocity of the particle at time t. Suppose that the particle is subjected to an external force that is well-modeled by a time-dependent potential energy well $\Phi(X(t),t)$. Then the Newton's second law implies that

$$m\mathrm{d}V(t) = -\gamma V(t) - \nabla\Phi(X(t),t)\mathrm{d}t + \sqrt{2k_BT\gamma}\mathrm{d}W(t)\,. \tag{20}$$

It is common to take the mass m of the microparticle to be small compared to other parameters, which allows us to reduce the system to a stochastic differential equation (SDE) depending only on the position $X(t)$. This is called the overdamped Langevin equation:

$$\gamma\,\mathrm{d}X(t) = -\nabla\Phi(X(t),t)\mathrm{d}t + \sqrt{2k_BT\gamma}\,\mathrm{d}W(t)\,. \tag{21}$$

In the context of molecular-motor-based intracellular transport, it is useful to write the external force in terms of the location where the molecular motor is bound to a microtubule. We denote this "anchor location" by $\{Z(t)\}_{t\in\mathbb{R}_+}$. If we further assume that the microparticle is bound by a tether that is well-modeled by a Hookean spring, then the potential energy can be written $\Phi(X(t),Z(t)) = \frac{\kappa}{2}|X(t)-Z(t)|^2$. The dynamics can then be

expressed as a linear SDE

$$\mathrm{d}X(t) = -\widetilde{\kappa}(X(t) - Z(t))\mathrm{d}t + \sqrt{2D}\mathrm{d}W(t)\,, \tag{22}$$

where $\widetilde{\kappa} = \kappa/\gamma$ is the ratio of the spring and fluid drag forces, and $D = k_B T/\gamma$ is the diffusivity of the particle. Of course, observations of the microparticle occur at discrete times. Suppose that the location of a particle at time 0 is $X(0) = x$ and the location of the anchor through the next time step is $\{Z(t); 0 \le t \le \Delta\}$. (We remind the reader, though, that that for current experimental techniques, the location of the anchor is unknown and must be inferred.) Then the solution to (22) is [16]

$$X(t) = e^{-\widetilde{\kappa}t}x + \int_0^t e^{-\widetilde{\kappa}(t-s)}Z(s)\mathrm{d}s + \sqrt{2D}\int_0^t e^{-\widetilde{\kappa}(t-s)}\mathrm{d}W(s)\,. \tag{23}$$

If we take the anchor movement to be a straight line with speed v (taking $Z(t) = z$), then this can be written

$$X(t) = (z + vt) + (x - z)e^{-\widetilde{\kappa}t} - \frac{v}{\widetilde{\kappa}}\left(1 - e^{-\widetilde{\kappa}t}\right) + \sqrt{2D}\int_0^t e^{-\widetilde{\kappa}(t-s)}\mathrm{d}W(s)\,. \tag{24}$$

We can take one final limit, now assuming the fluid drag γ is small compared to the time increment that has elapsed and κ. In this limit $\widetilde{\kappa} \to \infty$, which sends the terms $(x-z)e^{-\widetilde{\kappa}t}$ and $\frac{v}{\widetilde{\kappa}}(1-e^{-\widetilde{\kappa}t})$ to zero. Meanwhile, since stochastic integral term has a deterministic integrand, it is normally distributed with mean zero and variance

$$\mathrm{Var}\left(\sqrt{2D}\int_0^t e^{-\widetilde{\kappa}(t-s)}\mathrm{d}W(s)\right) = 2D\int_0^t e^{-2\widetilde{\kappa}(t-s)}\mathrm{d}s = \frac{D}{\widetilde{\kappa}}\left(1 - e^{-2\widetilde{\kappa}t}\right)\,. \tag{25}$$

Noting that γ appears in both the diffusivity D and the constants $\widetilde{\kappa}$, we have $D/\widetilde{\kappa} = k_B T/\kappa$. Taking $\gamma \to 0$ rest of the equation yields a normal random variable with mean 0 and variance $k_B T/\kappa$. It follows that as a numerical scheme or a statistical inference scheme, we can think of the cargo positions $(x_0, x_1, \ldots, x_n)$ at times $t_0, t_1, \ldots, t_n$ as

$$x_n = z + v t_n + \sqrt{\frac{k_B T}{\kappa}} \eta_n \tag{26}$$

where the $\{\eta_n\}$ are independent and identically distributed standard normal random variables.

Our model for anchored diffusion simply has $v = 0$.

APPENDIX B. STATISTICAL BACKGROUND

We refer to George Casella and Roger Lee Berger's *Statistical Inference* text for relevant statistical background [3]. The following concepts and theorems were used in the Honors Thesis.

Theorem B.1 (Chebychev's Inequality)**.** *Let X be a random variable and let $g(x)$ be a nonnegative function. Then, for any $r > 0$,*

$$P(g(X) \geq r) \leq \frac{\mathbb{E}g(X)}{r} .$$

Theorem B.2 (Slutsky's Theorem)**.** *If $X_n \to X$ in distribution and $Y_n \to a$, where a is constant, in probability, then*

- *$Y_n X_n \to aX$ in distribution.*
- *$X_n + Y_n \to X + a$ in distribution*

Theorem B.3 (Delta Method)**.** *Let Y_n be a sequence of random variables that satisfies $\sqrt{n}(Y_n - \theta) \to N(0, \sigma^2)$ in distribution. For a given function g and specific value of θ, suppose that $g'(\theta)$ exists and is not 0. Then $\sqrt{n}(g(Y_n) - g(\theta)) \to N(0, \sigma^2[g'(\theta)]^2)$ in distribution.*

Theorem B.4 (Cramér-Rao Inequality)**.** *Let $X_1, X_2, ..., X_n$ be a sample with pdf $f(\mathbf{x}|\theta)$, and let $W(\mathbf{X}) = W(X_1, X_2, ..., X_N)$ be any estimator satisfying*

$$\frac{d}{d\theta}\mathbb{E}_\theta W(\mathbf{X}) = \int_\chi \frac{\partial}{\partial\theta}[W(\mathbf{X})f(\mathbf{x}|\theta)]d\mathbf{x}$$

and $Var_\theta W(\mathbf{X}) < \infty$. *Then*

$$Var_\theta W(\mathbf{X}) \geq \frac{\left(\frac{d}{d\theta}\mathbb{E}_\theta W(\mathbf{X})\right)^2}{\mathbb{E}_\theta\left(\left(\frac{\partial}{\partial\theta}\log f(\mathbf{X}|\theta)\right)^2\right)}$$

Corollary B.1 (Cramér-Rao Inequality, iid case)**.** *If the assumptions of Theorem B.4 are satisfied and, additionally, if* $X_1, X_2, ..., X_n$ *are iid with pdf* $f(x|\theta)$*, then*

$$Var_\theta W(\mathbf{X}) \geq \frac{\left(\frac{d}{d\theta}\mathbb{E}_\theta W(\mathbf{X})\right)^2}{n\,\mathbb{E}_\theta\left(\left(\frac{\partial}{\partial\theta}\log f(X|\theta)\right)^2\right)} \quad .$$

Lemma B.2. *If* $f(x|\theta)$ *satisfies*

$$\frac{d}{d\theta}\mathbb{E}_\theta\left(\frac{\partial}{\partial\theta}\log f(X|\theta)\right) = \int \frac{\partial}{\partial\theta}\left[\left(\frac{\partial}{\partial\theta}\log f(x|\theta)\right)f(x|\theta)\right]dx$$

(true for an exponential family), then

$$\mathbb{E}_\theta\left(\left(\frac{\partial}{\partial\theta}\log f(X|\theta)\right)^2\right) = -\,\mathbb{E}_\theta\left(\frac{\partial^2}{\partial\theta^2}\log f(X|\theta)\right) \quad .$$

Regularity Conditions:

Assumption 1. *We observe* $X_1, ..., X_n$*, where* $X_i \sim f(x|\theta)$ *are iid.*

Assumption 2. *The parameter is identifiable; that is, if* $\theta \neq \theta'$*, then* $f(x|\theta) \neq f(x|\theta')$*.*

Assumption 3. *The densities* $f(x|\theta)$ *have common support, and* $f(x|\theta)$ *is differentiable in* θ*.*

Assumption 4. *The parameter space* Ω *contains an open set* ω *of which the true parameter value* θ_0 *is an interior point.*

Assumption 5. *For every $x \in \chi$, the density $f(x|\theta)$ is three times differentiable with respect to θ, the third derivative is continuous in θ, and $\int f(x|\theta)dx$ can be differentiated three times under the integral sign.*

Assumption 6. *For any $\theta_0 \in \Omega$, there exists a positive number c and a function $M(x)$ (both of which may depend on θ_0) such that*

$$\left|\frac{\partial^3}{\partial\theta^3}\log f(x|\theta)\right| \leq M(x)$$

for all $x \in \chi$, $\theta_0 - c < \theta < \theta_0 + c$, with $\mathbb{E}_{\theta_0}[M(x)] < \infty$.

These conditions are sufficient to prove the following theorem.

Theorem B.5 (Asymptotic Normality and Efficiency of MLEs)**.** *Let $X_1, X_2, ..., X_n$ be iid $f(x|\theta)$, let $\hat{\theta}$ denote the MLE of θ, and let $\tau(\theta)$ be a continuous function of θ. Under the regularity conditions above on $f(x|\theta)$, and hence, $L(\theta|\vec{x})$,*

$$\sqrt{n}[\tau(\hat{\theta}) - \tau(\theta)] \to N(0, \nu(\theta)),$$

where $\nu(\theta)$ is the Cramér-Rao Lower Bound. That is, $\tau(\hat{\theta})$ is a consistent and asymptotically efficient estimator of $\tau(\theta)$.

APPENDIX C. CHECKING REGULARITY CONDITIONS FOR PROOF OF THEOREM 5.1

The following is a check of the regularity conditions for applying Theorem B.5, Asymptotic Normality and Efficiency of MLEs, in the proof of Theorem 5.1.

Checking Assumption 1: The random variables, $\xi_1, \xi_2, ..., \xi_N$, are iid.

Checking Assumption 2: Recall that

$$f\left(\xi|\sigma_1^2\right) = \left(\frac{1}{\sigma_1^2 2\pi}\right)^{\frac{1}{2}} e^{-\frac{\xi^2}{2\sigma_1^2}} .$$

Suppose there exists a $\tilde{\sigma}_1^2$ such that $\tilde{\sigma}_1^2 \neq \sigma_1^2$, then

$$f\left(\xi|\tilde{\sigma}_1^2\right) = \left(\frac{1}{\tilde{\sigma}_1^2 2\pi}\right)^{\frac{1}{2}} e^{-\frac{\xi^2}{2\tilde{\sigma}_1^2}} \neq f\left(\xi|\sigma_1^2\right) .$$

Checking Assumption 3: Since $e^{-\frac{\xi^2}{2\sigma_1^2}} \neq 0$ for any value of ξ or σ_1^2, then $f\left(\xi|\sigma_1^2\right) \neq 0$ for any value of ξ or σ_1^2, and $f\left(\xi|\sigma_1^2\right)$ has common support. Further, $f\left(\xi|\sigma_1^2\right)$ is differentiable in σ_1^2:

$$\frac{\partial}{\partial \sigma_1^2} f\left(\xi|\sigma_1^2\right) = \frac{1}{2\sqrt{2\pi}} \left(\sigma_1^2\right)^{\frac{-3}{2}} e^{\frac{-\xi^2}{2\sigma_1^2}} + \frac{\xi^2}{2\sqrt{2\pi}} \left(\sigma_1^2\right)^{\frac{-5}{2}} e^{\frac{-\xi^2}{2\sigma_1^2}} .$$

Checking Assumption 4: We assume $\sigma_1^2 > 1$. This means the parameter space, Ω, is the positive real line, and there is always an open set of which the true parameter value is an interior point.

Checking Assumption 5: The density, $f\left(\xi|\sigma_1^2\right)$ is three times differentiable with respect to σ_1^2:

$$f\left(\xi|\sigma_1^2\right) = \left(\frac{1}{\sigma_1^2 2\pi}\right)^{\frac{1}{2}} e^{-\frac{\xi^2}{2\sigma_1^2}},$$

$$f'\left(\xi|\sigma_1^2\right) = \frac{1}{2\sqrt{2\pi}}\left(\sigma_1^2\right)^{\frac{-3}{2}} e^{\frac{-\xi^2}{2\sigma_1^2}} + \frac{\xi^2}{2\sqrt{2\pi}}\left(\sigma_1^2\right)^{\frac{-5}{2}} e^{\frac{-\xi^2}{2\sigma_1^2}},$$

$$f''\left(\xi|\sigma_1^2\right) = \frac{1}{4\sqrt{2\pi}}\left(\sigma_1^2\right)^{\frac{-5}{2}} e^{\frac{-\xi^2}{2\sigma_1^2}}\left[3 - 6\xi^2\left(\sigma_1^2\right)^{-1} + \xi^4\left(\sigma_1^2\right)^{-2}\right],$$

$$f^{(3)}\left(\xi|\sigma_1^2\right) = \frac{1}{8\sqrt{2\pi}}\left(\sigma_1^2\right)^{\frac{-7}{2}} e^{\frac{-\xi^2}{2\sigma_1^2}}\left[-15 + 45\xi^2\left(\sigma_1^2\right)^{-1} - 15\xi^4\left(\sigma_1^2\right)^{-2} + \xi^6\left(\sigma_1^2\right)^{-3}\right].$$

Since we assume σ_1^2 is strictly positive, $f\left(\xi|\sigma_1^2\right)$ and each derivative are continuous in σ_1^2.

Next, we check that $\int f\left(\xi|\sigma_1^2\right)dx$ can be differentiated three times under the integral sign. By the Leibniz Integral Rule, for arbitrary region $a \le \sigma_1^2 \le b, \xi_a \le \xi \le \xi_b$:

$$\frac{\partial}{\partial \sigma_1^2}\left(\int_a^b f\left(\xi|\sigma_1^2\right)dx\right) = \int_a^b \frac{\partial}{\partial \sigma_1^2} f\left(\xi|\sigma_1^2\right)dx.$$

Further, since the derivatives are continuous, you can repeatedly apply this rule for the integral on the right hand side. Thus, $\int f\left(\xi|\sigma_1^2\right)dx$ can be differentiated three times under the integral sign.

Checking Assumption 6: First, determine $\frac{\partial^3}{\partial \sigma_1^{2^3}} \ln f\left(\xi | \sigma_1^2\right)$:

$$f\left(\xi | \sigma_1^2\right) = \left(\frac{1}{\sigma_1^2 2\pi}\right)^{\frac{1}{2}} e^{-\frac{\xi^2}{2\sigma_1^2}} ,$$

$$\ln f\left(\xi | \sigma_1^2\right) = -\frac{1}{2}\ln(\sigma_1^2) - \frac{1}{2}\ln(2\pi) - \frac{\xi^2}{2\sigma_1^2} ,$$

$$\frac{\partial}{\partial \sigma_1^2} \ln f\left(\xi | \sigma_1^2\right) = -\frac{1}{2\sigma_1^2} + \frac{\xi^2}{2\left(\sigma_1^2\right)^2} = -\frac{1}{2}\left(\sigma_1^2\right)^{-1} + \frac{\xi^2}{2}\left(\sigma_1^2\right)^{-2} ,$$

$$\frac{\partial^2}{\partial \sigma_1^{2^2}} \ln f\left(\xi | \sigma_1^2\right) = \frac{1}{2}\left(\sigma_1^2\right)^{-2} - \xi^2\left(\sigma_1^2\right)^{-3} ,$$

$$\frac{\partial^3}{\partial \sigma_1^{2^3}} \ln f\left(\xi | \sigma_1^2\right) = -\left(\sigma_1^2\right)^{-3} + 3\xi^2\left(\sigma_1^2\right)^{-4} = \frac{-1}{\left(\sigma_1^2\right)^3} + \frac{3\xi^2}{\left(\sigma_1^2\right)^4} .$$

Then, for a given value of σ_1^2 and positive number c,

$$\left|\frac{-1}{\left(\sigma_1^2\right)^3} + \frac{3\xi^2}{\left(\sigma_1^2\right)^4}\right| \leq \frac{3\xi^2}{\left(\sigma_1^2 - c\right)^4} = M(\xi) ,$$

and $\mathbb{E}_{\sigma_1^2}[M(\xi)]$ is finite since $\mathbb{E}_{\sigma_1^2}\left(\xi^2\right) = \sigma_1^2$, which is finite.