Bayesian Sample Size Determination for Single-Particle Tracking of Pathogens in Biological Fluids

by

Bryan Yates

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Mathematics
in
Statistics

Waterloo, Ontario, Canada, 2018

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Single-particle tracking (SPT) experiments measure 2-dimensional particle position with a high-resolution digital camera, capturing microsecond motion. SPT has allowed novel investigation of membrane dynamics, enzymology, subdiffusion processes in proteins, and serves as a burgeoning application of statistical modelling. While particle tracking statistical methodology has shown great promise, the literature is comparatively scarce for methods that determine the necessary number of particles to track to assess a relevant scientific hypothesis. This work addresses this gap by providing a Bayesian sample size determination (SSD) algorithm. Namely, this work proceeds in two-stages, (1) model training and (2) the SSD algorithm. A single-trajectory location-scale model incorporating fractional Brownian motion is fit using maximum likelihood estimation for each 2-dimensional SPT trajectory. Subsequently, a multiple-trajectory hierarchical model is fit to capture different particle dynamics caused by fluid heterogeneity. A Bayesian SSD algorithm follows to evaluate scientific relevance based on population-level mean-squared displacement. The performance of the SSD algorithm is first studied under a simulation environment. Three simplified SSD algorithms are presented to accelerate computation. Following this, experimental data of 3,707 fluorescently labelled herpes-simplex virus trajectories are studied across five separate antibody concentrations ranging 0-1000 mg/L. A detailed analysis on the practical use of the SSD algorithm provides insight into virus dynamics as a function of antibody concentration.

Acknowledgements

Chiefly, I would like to thank my supervisor Professor Martin Lysy for the excellent guidance, instruction, and unparalleled opportunities that lead to this thesis. Martin truly engages is a variety of cross-disciplinary research which I am grateful to have had the opportunity to partake. Working with Martin not only motivated me to work on difficult problems, but also challenged me to view statistical problems from a fresh perspective.

Further, thank you to experimental collaborator Professor Sam Lai at the University of North Carolina at Chapel Hill. Many insightful conversations with Professor Lai contributed to the final focus of this thesis.

I would also like to thank the Department of Statistics and Actuarial Science faculty and staff that contributed to the superb environment at the University of Waterloo. Thank you to Professor Pengfei Li and Professor Wayne Oldford for continued guidance and advice. Thank you to Professor Greg Rice for outstanding teaching. Thank you to Mary Lou Dufton and Lisa Baxter for their continued support. Such a nice group of individuals made this thesis research a pleasure.

In addition, I would like to thank Natural Sciences and Engineering Research Council of Canada (NSERC), Canada's federal funding agency, for providing funding through the NSERC Canada's Graduate Scholarship - Masters (CGS-M). I would also like to thank the University of Waterloo for the President's Graduate Scholarship and particularly the Department of Statistics and Actuarial Science for conference travel funding.

Finally, I would like to thank the following people for their support throughout my thesis research: Petar Todorovic, Anthony Caterini, Josh Valchar, and Tong Zhan. I hope our careers lead us to work together in the future.

Dedication

This is dedicated to my mother who has given endless support and motivation.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

Nanotechnology offers revolutionary imaging methods capable of single-molecule motion measurement (Moerner, 2002). In the field of rheology, typical single-particle tracking (SPT) experiments record the position of various pathogens and foreign bodies in biological fluids over time using high-resolution cameras. SPT has allowed novel investigation of membrane dynamics (Saxton and Jacobson, 1997), enzymology (Xie and Lu, 1999), subdiffusion processes in proteins (Kou, 2008), and serves as a burgeoning application of statistical modelling (Lysy et al., 2016; Mellnik et al., 2016). An excellent review of important statistical developments in SPT can be found in Qian and Kou (2014).

Figure 1.1 show a simplified cartoon schematic of a typical SPT experiment capturing 2-dimensional particle movement. Typically, hundreds of single-particle trajectories are recorded concurrently using a high-resolution camera. This work will focus specifically on 2-dimensional trajectories. While 3-dimensional SPT experiments are feasible, the complexity of working in higher dimensions and the wider availability of 2-dimensional SPT data encourage use of 2-dimensional SPT trajectories. Some experimental challenges of capturing SPT trajectories will be discussed in Section 1.2 which further support the use of 2-dimensional trajectories for statistical analysis.

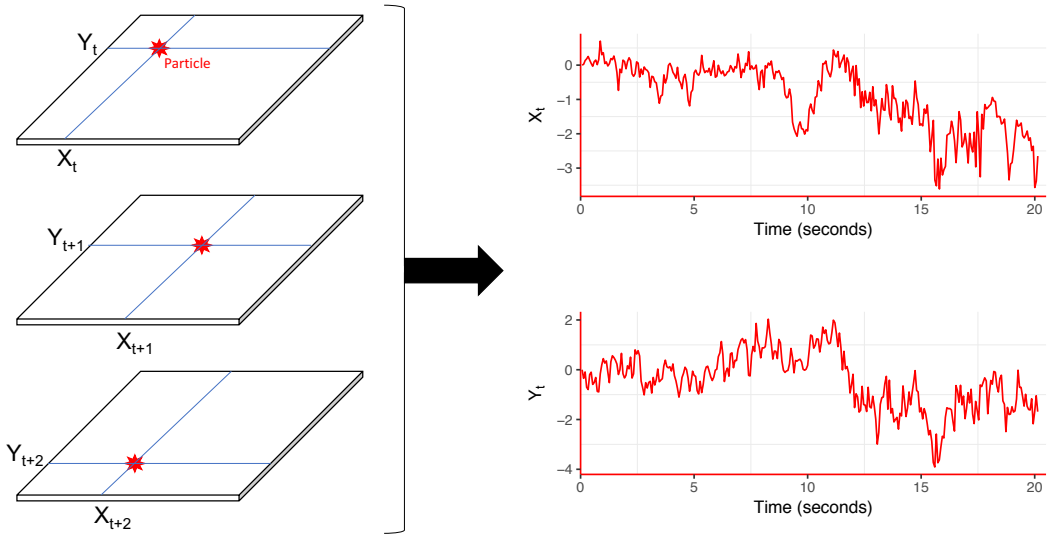

Figure 1.1: SPT procedure. Diffusion is monitored for individual particles over time.

1.1 Motivating Application

This work specifically focuses on the statistical analysis and study of virus diffusion. A subset of the immunology discipline, virology focuses on the structure, movements patterns, and ongoing adaptation of virus and virus-like agents. SPT has offered enormous insights into virus behaviour (Brandenburg and Zhuang, 2007), including how viral material transits across cell membranes (Babcock et al., 2004), lateral mobility of live-cell bound viruses (Ewers et al., 2005), and real-time imaging of infection pathways (Seisenberger et al., 2001).

Understanding virus movement is critical toward improved vaccine design (Nowak and May, 2000; Gottlieb and Johnston, 2017). Through understanding the motion and behaviour of viruses at the micrometer level, practitioners can develop improved methods to combat virus uptake and infiltration into the human body. One modern approach to vaccine development is to study the interaction of viruses and antibodies.

Common antibodies found in human blood and mucus are large protein molecules produced by the immune system to defend against pathogenic bacteria and viruses. Antibodies bind to a unique molecule, known as antigen, located on specific sites of pathogens. Upon binding, the deleterious effects of the pathogen are neutralized. The mechanism of which the pathogen is effectively deactivated depends on its type and structure, but generally an antibody bound pathogen will lose ability to begin any biological process that can cause disease or sickness. Hence, knowledge of virus motion is vital for immunologists to determine which antibodies are effective or possess great potential for new vaccination treatments.

The virus under study in this work is the herpes-simplex virus (HSV). The inimical culprit of cold sores and herpes, HSV is highly contagious with over 3 billion humans infected globally (Looker et al., 2015). Thus, there is considerable need for substantially improved vaccine design to protect future generations. The methodology developed in this work aims to accelerate SPT research and deployment by providing a fundamental approach to determine the required number of particles to capture in a given SPT experiment.

1.2 Heterogeneity and Particle Dynamics

This work examines 3,707 fluorescently labelled HSV trajectories in a mucus medium, at five separate antibody (Ab) concentrations ranging 0-1000 mg/L. Importantly, the 0 mg/L antibody concentration does not indicate no antibody is present in the mucus, but rather that no external antibody concentration has been added. Particle y-axis trajectories in the baseline environment are displayed in Figure 1.2. This restricted 1-dimensional view of the 2-dimensional experimental data offers enormous insight into some of the principle challenges in SPT experiments and experimental data characteristics, namely high-particle density, particle motion heterogeneity, and particle disappearance (Jaqaman et al., 2008). Each of these challenges are now discussed in terms of the current HSV data under study.

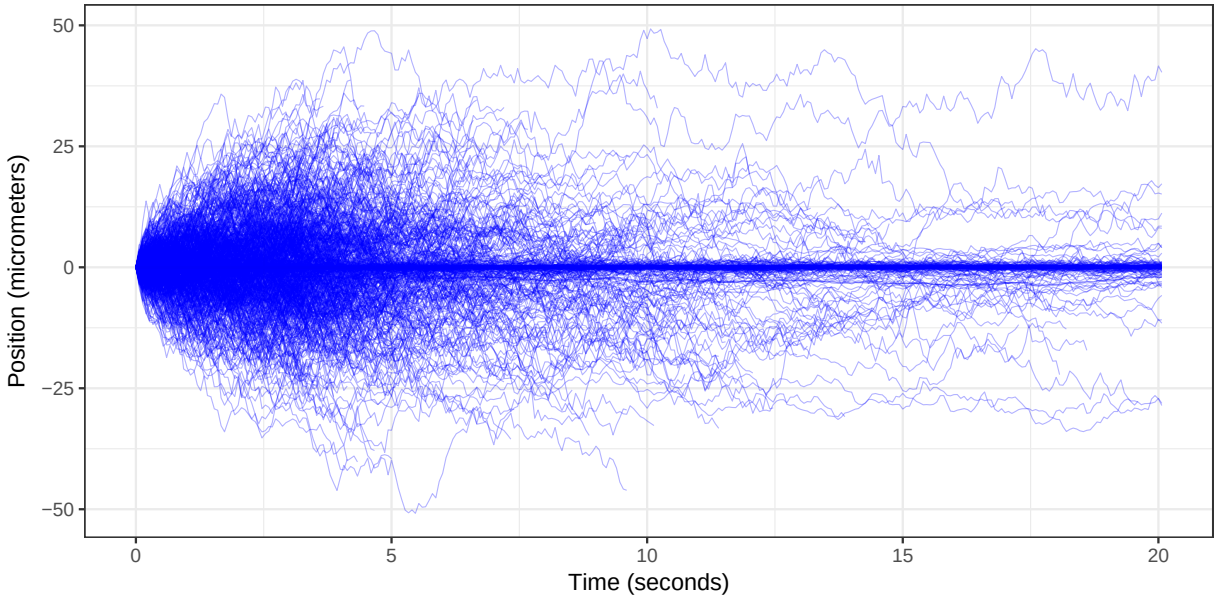

Figure 1.2: HSV particle diffusion 1-dimensional view of SPT trajectories for 0 mg/L antibody concentration. There is a clear variance of behaviour among different particle trajectories, some particle trajectories are cut short, and some have very slow movement while others show fast diffusion.

Firstly, high particle density in SPT experiments can cause neighbouring virus-virus interaction producing unnatural changes to virus motion. To combat this, SPT experiments usually require particles to have a certain separation distance to be recorded. This work assumes the virus particles under study are strictly interacting with the fluid medium, hence there are no virus-virus particle interactions.

A careful examination of Figure 1.2 illuminates that some SPT trajectories are cut short. That is, though the SPT experiment runs for 20 seconds, the data available for some virus particles in considerably shorter. This effect is primarily caused by the limited camera focal operating range. If particles begin in the focal range of the camera and diffuse out within the experiment time, only the time spent in the operating range of the camera is recorded. Hence, these trajectories will not have a full 20-seconds of position data. This is problematic for highly mobile particles since they have a higher chance to diffuse outside of the camera range than less mobile particles. Therefore, this work will distinguish between constrained (slow-moving) and unconstrained (fast-moving) particles. The physical mechanism that causes a virus particle to become constrained

in motion is primarily attributed to interaction with an antibody particle. This virus-antibody interaction hinders virus motion resulting in slow diffusion. This work will specifically study constrained HSV trajectories since these are of great interest when studying how the level of antibody concentration impacts virus motion. The majority of these constrained trajectories are observed in the central band visible in Figure 1.2. Despite studying only the constrained trajectories, there are still many HSV paths that exit the camera frame before 20 seconds. This work gives identical treatment to HSV paths, regardless of their length, when fitting the location-scale model of Section 2.2. While this assumption may be limiting, the shortest constrained paths still contain over 50 observations which amply allows for accurate parameter estimation.

Heterogeneity is a pervasive phenomena of SPT experimental data. In this work, heterogeneity refers to the varied behaviour of virus particles under a seemingly identical physical environment. For example, while Figure 1.2 shows many particles are clustered near the zero-position throughout the 20-second experiment, these trajectories do all behave identically. A detailed view of constrained path motion in Figure 1.3 illuminates the varied diffusion behaviour of constrained trajectories. Within these constrained paths, some particles move more slowly (trajectory number 19), while others exhibit faster motion (trajectory number 1). This heterogeneity of particle motion within constrained trajectories in the same experimental environment primarily arises due to the complex fluid that surrounds each virus particle. Varying fluid viscosity, temperature, and density all attribute to idiosyncratic particle behaviour depending on that particles exact location in the heterogeneous medium. Moreover, the virus particles under study are dispersed in mucus which is a viscous colloid containing salts, immunoglobulins, enzymes, and glycoproteins. Hence, the particular location of a virus in the medium can result in varied motion characteristic due to the structure of the surrounding local fluid environment.

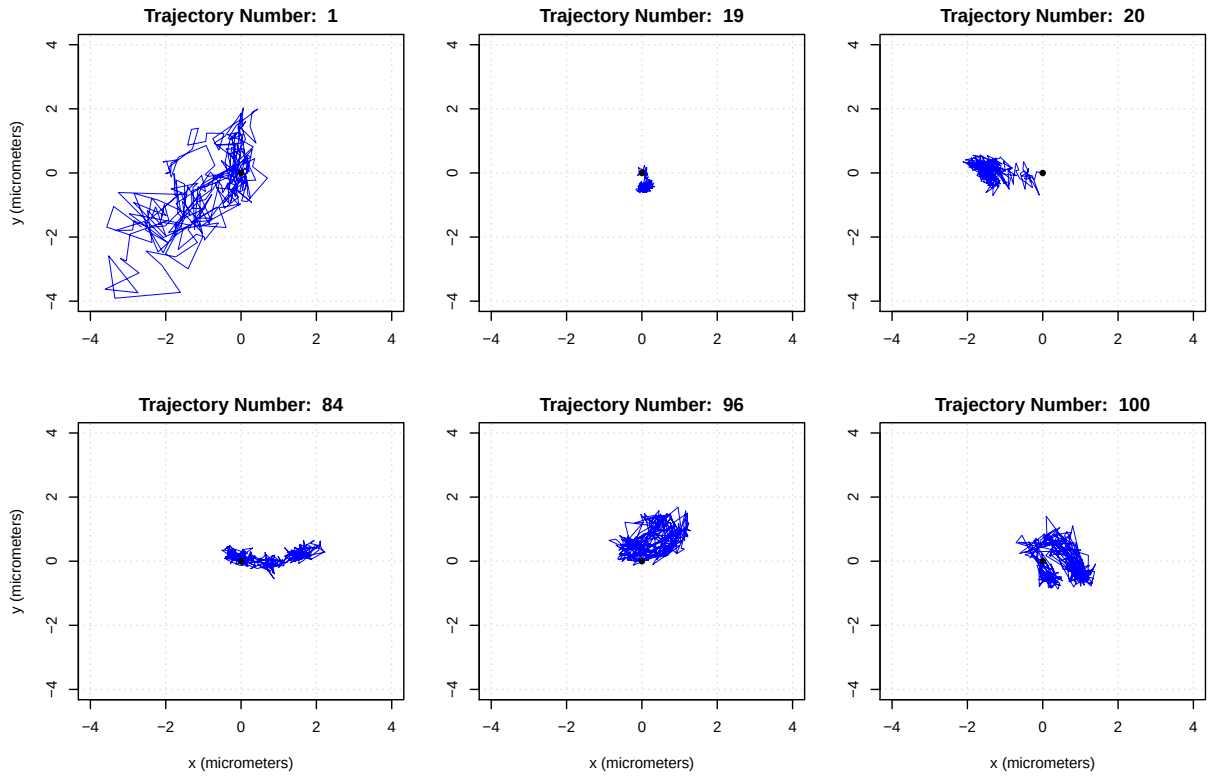

Figure 1.3: HSV 2-dimensional constrained trajectories under the 0 mg/L antibody environment. There is a clear variance of diffusion behaviour across paths. A black dot indicates the starting position of the particles at the origin.

This phenomena of heterogeneity motivates the use of a more complex multi-level model to understand population-level dynamics present in a given experimental environment. While this work will model single-path SPT trajectories using a single location-scale model (detailed in Section 2.2), heterogeneity will be captured by a hierarchical model discussed in Section 2.3.

1.3 Mean-Squared Displacement and Scientific Relevance

Modern SPT experiments routinely generate hundreds to a few thousand trajectories under a single experimental environment. A wide-spread technique to characterize the diffusive behaviour of a single particle is through mean-squared displacement (MSD)

analysis. The MSD of a particle's continuous-time trajectory $X(t)$ is defined as

$$MSD(t) = \langle X^2(t) \rangle = E\Big[(X(t) - X(0))^2\Big], \qquad t \geq 0.$$

The behaviour of the MSD over time is commonly employed to classify a particles behaviour, such as Brownian motion where $MSD(t) \propto t$. While Brownian motion is employed to model particle motion in many areas of physics, SPT experiments routinely reveal non-Brownian dynamics. Small particles such as viruses undergo complex interactions with their surrounding medium that results in motion known as anomalous subdiffusion. Anomalous subdiffusion is characterized through a particles MSD,

$$MSD(t) \propto t^{\alpha}, \qquad 0 < \alpha < 1, \qquad t_{min} < t < t_{max},$$

where α is the anomalous diffusion exponent. The anomalous diffusion exponent effectively represents the degree of subdiffusion exhibited by a particle. Lower values of α indicate a smaller area of space will be explored by a particle as time elapses.

A typical SPT experiment contains hundreds to a few thousand trajectories. Thus, a population-level MSD statistic is defined as

$$MSD(t) = E[D \cdot t^{\alpha}], \qquad t \geq 0.$$

where D represents a particles diffusion coefficient. A particle's diffusion coefficient coincides with the rate at which a particle diffuses in a fluid. The higher the value of a particle diffusion coefficient, the faster the particle diffuses. Further discussion of the diffusion coefficient with respect to single-trajectory modelling will commence in Section 2.2. This population-level MSD corresponds to the expectation taken across numerous single-particle trajectories MSD. This distinction between single-path MSD and multiple-path (i.e. population level) MSD becomes important to determine the scientific relevance of an experiment.

Scientific relevance in this work is determined through analysis of the population-level

MSD for a particular experimental environment. Namely, given some SPT experiment, practitioners wish to assess whether or not experimental findings are scientifically relevant. As a concrete example of scientific relevance, consider two SPT experiments A and B. For simplicity, assume all factors of the two experiments are identical except that experiment A has 0 mg/L antibody concentration and experiment B has 100 mg/L antibody concentration. A highly valuable scientific hypothesis to evaluate is whether the addition of 100 mg/L antibody in experiment B made a scientifically meaningful impact on particle motion. The definition of the scientific impact may take many forms. One definition could evaluate scientific relevance based on the mean first-passage time of experiment A compared to that of experiment B. A particles first-passage time is defined as the amount of time the particle will take to reach a defined distance from its starting origin. While this scientific hypothesis would prove very useful, it is not studied in this work due to the difficulty in modelling such a process for 2-dimensional trajectories. Rather, this work focuses on a more amenable definition of scientific relevance defined using the population-level MSD. Using the same example of experiments A and B, scientific relevance defined as a 20% reduction in population-level MSD between the two experiments is feasibly determined by the methodology developed in this work. This thesis employs a single-path location-scale model and a population-level hierarchical model to address whether an experiment is scientifically relevant. A formalized definition of scientific relevance is discussed in Chapter 3.

1.4 Thesis Contribution

In recent decades, statistics continues to have an immense impact on biological science. Stochastic processes readily lend to the analysis of biological particles (Allen, 2010; Cherstvy et al., 2013), and computational intensive methods such as Bayesian methods have been applied to model numerous biological phenomena (Huelsenbeck et al., 2001; Li and Zhang, 2010; Yau et al., 2011).

While research has predominantly focused on the stochastic dynamics of particle mo-

tion, literature is comparatively scarce for methods that determine necessary number of particles to track to assess a scientific hypothesis. This is a sample size determination (SSD) problem. Knowledge of the required number of experiments to run is extremely valuable for experimentalists since SPT experiments take considerable time and regularly require expensive resources. Moreover, accurate sample-size determination can also be important for ethical reasons depending on the scientific hypothesis under study which further supports the importance and practical significance of this work. This thesis offers a Bayesian SSD algorithm to specifically address the number of samples needed in an SPT experiment to assess whether an experiment is scientifically relevant based on its population-level MSD. Namely, the SSD algorithm presented in this work uses preliminary 2-dimensional particle trajectories from an experimental environment to determine the number of future samples needed to assess a whether a given difference in population-level MSD is present compared to some reference population-level MSD (set by the practitioner). While the population-level MSD is used to determine scientific relevance, the framework developed herein remains extremely flexible for future research to incorporate more complex scientific relevance criterion such as mean first-passage times.

The primary SSD algorithm presented is computationally intensive. This poses problems for practical implementation since SPT experiments are generally completed on the order of minutes to a few hours. The speed and efficiency of SSD technique will undoubtedly become increasingly important as SPT technology moves toward automated tracking. The proliferation of successful deep neural networks in image recognition (Krizhevsky et al., 2012; Wang and Yeung, 2013) has extended reach to the real-time particle tracking domain (Zhu et al., 2017; Newby et al., 2017). Thus, as automation of single-particle tracking trajectories becomes common there will be need for real-time methods to instruct acquisition programs to cease operation. Therefore, three accelerated SSD algorithms are explored, each making a set of simplifying assumptions. Chapter 3 delineates the main SSD algorithm and the three accelerated approximations. The accelerated SSD algorithms execute on the order of minutes and hence offer great potential for future use in a real-time SPT experiment.

The remainder of this thesis consists of five chapters. First, Chapter 2 presents the statistical modelling framework. Namely, a single-trajectory location-scale model is explored to model single virus diffusion, followed by a multiple-trajectory hierarchical model that captures heterogeneity. Chapter 3 discusses scientific relevance and delineates the primary SSD algorithm and the three accompanying accelerated approximations. A comprehensive simulation study follows in Chapter 4. Experimental HSV particle diffusion data is then carefully analyzed in Chapter 5. Closing remarks and future areas of research are discussed in Chapter 6.

Chapter 2

Model and Calibration

This chapter builds up to a model for particle trajectories in three steps. First, a fundamental model for a ubiquitous scientific property of particle trajectories is described. Next, this fundamental model is extended to single-particle 2-dimensional trajectories experiencing low-frequency drift noise which must be disentangled from the dynamical signal. Finally, the single-particle model is extended to multiple-trajectory experimental populations, which experience substantially more dynamic variability than a single, uniform-population model would predict.

2.1 Fractional Brownian Motion and Subdiffusion

Anomalous diffusion behaviour is captured in the anomalous exponent α. With a widespread presence in physics (Metzler and Klafter, 2000), anomalous diffusion is found in a variety of biological systems such as the transport of mammalian cells (Bronstein et al., 2009), lipid granule motion in yeast cells (Jeon et al., 2011), chaotic transport in laminar fluid flow (Solomon et al., 1993), and membrane dynamics (Saxton and Jacobson, 1997).

An extension to Brownian motion that incorporates a greater range of diffusion behaviour is known as fractional Brownian motion (fBM). The fBM process $B_\alpha(t)$ has properties:

1. Continuous-time stochastic process
2. Stationary increments: $\Delta B_\alpha(t_1) \overset{D}{=} \Delta B_\alpha(t_2)$, where $\Delta B_\alpha(t) = B_\alpha(t + \Delta t) - B_\alpha(t)$
3. Zero-mean Gaussian process

Importantly, fBM is the only stochastic process with properties (1)-(3) that exhibits uniform subdiffusion,

$$MSD(t) \propto t^\alpha, \qquad 0 < \alpha < 1, \qquad \forall t > 0.$$

The anomalous diffusion exponent of an fBM process is found in the covariance

$$cov(B_\alpha(t), B_\alpha(s)) = \frac{1}{2}(|t|^\alpha + |s|^\alpha - |t - s|^\alpha),$$

where subdiffusion occurs when $0 < \alpha < 1$. From a physical perspective, α can be seen as representing the degree of subdiffusion a particle exhibits, where a smaller α indicates stronger subdiffusive behaviour.

2.2 Single-Trajectory Location-Scale Model

Following the methods from Lysy et al. (2016), denote $\mathbf{X}(t) = (X_1(t), X_2(t))$ as the 2-dimensional particle position at time $t \geq 0$. A commonly employed subdiffusive location-scale model is

$$\mathbf{X}(t) = \boldsymbol{\mu}\mathbf{t} + \boldsymbol{\Psi}^{1/2}\mathbf{Z}(t).$$

In two-dimensions, $\boldsymbol{\mu} = (\mu_1, \mu_2)$, $\boldsymbol{\Psi} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$, and $\mathbf{Z}(t) = (Z_1(t), Z_2(t))$ are iid copies of the fBM model. Thus, each 2-dimensional SPT time series is mapped to a

1×6 parameter vector denoted as $\boldsymbol{\varphi}$, where $\boldsymbol{\varphi} = (\alpha, \sigma_1, \sigma_2, \rho, \mu_1, \mu_2)$. Furthermore, the path-wise MSD can be calculated,

$$MSD(t) = tr(\boldsymbol{\Psi}) \cdot t^{\alpha} = (\sigma_1^2 + \sigma_2^2) \cdot t^{\alpha} = D \cdot t^{\alpha},$$

where $tr(\boldsymbol{\Psi})$ denotes the trace of $\boldsymbol{\Psi}$, and the substitution $D = \sigma_1^2 + \sigma_2^2$ was made to introduce the diffusion coefficient D. Recall, a particle's diffusion coefficient coincides with the rate at which a particle diffuses in a fluid.

In discrete form, let $\boldsymbol{X} = (\boldsymbol{X}_0, \boldsymbol{X}_1, \cdots, \boldsymbol{X}_N)$ be the discretely recorded time series of length $N+1$ with sampling period Δt, with $\boldsymbol{X}_i = \boldsymbol{X}(t = i \cdot \Delta t)$. Then $\boldsymbol{x}_{N\times 2} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)$ is stationary where $\boldsymbol{x}_i = \boldsymbol{X}_i - \boldsymbol{X}_{i-1}$ are the increments of the discrete-time SPT data. The increments $\mathbf{x}$ then have a matrix normal distribution,

$$\mathbf{x} \sim \mathcal{MN}(\boldsymbol{\Delta t \mu}, \mathbf{V}_{\alpha}, \boldsymbol{\Psi})$$

where $\boldsymbol{\Delta t}_{N\times 1} = (\Delta t, \cdots, \Delta t)^T$. A generalization of the multivariate normal distribution to matrix-valued random variables, the matrix-normal distribution contains two scale matrices corresponding to the row covariance $\boldsymbol{V}_{\alpha}$, and column covariance $\boldsymbol{\Psi}$. By properties of the matrix-normal distribution,

$$\boldsymbol{x} \sim \mathcal{MN}(\boldsymbol{\Delta t \mu}, \mathbf{V}_{\alpha}, \boldsymbol{\Psi}) \iff vec(\boldsymbol{x}) \sim \mathcal{N}(vec(\boldsymbol{\Delta t \mu}), \mathbf{V}_{\alpha} \otimes \boldsymbol{\Psi}), \tag{2.1}$$

where $vec(\boldsymbol{x})$ denotes the vectorization of $\boldsymbol{x}$, and $\otimes$ denotes the Kronecker product. Here, $\mathbf{V}_{\alpha}$ is a Toeplitz matrix calculated using the autocorrelation of fBm increments. The profile likelihood for this multivariate regression in (2.1) is,

$$\ell_{prof}(\alpha|\boldsymbol{x}) = -\frac{1}{2}\left(2N + N\log\left(\frac{\mathbf{S}_{\alpha}}{N}\right) + 2\log|\boldsymbol{V}_{\alpha}|\right)$$

where,

$$\mathbf{S}_\alpha = (\boldsymbol{x} - \boldsymbol{\Delta t}\hat{\boldsymbol{\mu}}_\alpha)\mathbf{V}_\alpha^{-1}(\boldsymbol{x} - \boldsymbol{\Delta t}\hat{\boldsymbol{\mu}}_\alpha)\ , \qquad \hat{\boldsymbol{\mu}}_\alpha = (\boldsymbol{\Delta t}^T\mathbf{V}_\alpha^{-1}\boldsymbol{\Delta t})^{-1}\boldsymbol{\Delta t}^T\mathbf{V}_\alpha^{-1}\boldsymbol{x}.$$

The derivation of this log-likelihood can be found in Appendix A. Thus, working with the increments $\boldsymbol{x}$ instead of the position $\boldsymbol{X}$ is computationally motivated since working with a Toeplitz variance matrix $\boldsymbol{V}_\alpha$ reduces the cost of each log-likelihood evaluation from $\mathcal{O}(N^3)$ to $\mathcal{O}(N^2)$ via the Durbin-Levinson algorithm (Brockwell and Davis, 2009; Lysy et al., 2016). Hence, parameter estimation for the location-scale model results in a 1-dimensional optimization over α, for each 2-dimensional SPT time series in this work. Estimates for $\boldsymbol{\Psi}$ and $\boldsymbol{\mu}$ can then be determined by using the invariance property of the MLE.

A simple transformation of the parameters will aide in computations going forward,

$$\boldsymbol{\varphi} = \left(logit(\frac{\alpha}{2}), log(\sigma_1^2 + \sigma_2^2), log(\sigma_1^2 - \sigma_2^2), logit(\frac{\rho + 1}{2}), \mu_1, \mu_2\right).$$

Namely, this transformation aids in the calculation of scientific relevance since the first two parameters of $\boldsymbol{\varphi}$ are sufficient for calculating path-wise MSD. In practice each 2-dimensional SPT trajectory can be fit to the location-scale model in parallel which greatly accelerates computation for a large number of time series under study.

2.3 Multiple-Trajectory Hierarchical Model

Hierarchical modelling features prominently in the analysis of multi-level data processes such as in mechanical experiments (Qian and Wu, 2008), ecology (Clark, 2005; Royle and Dorazio, 2008; Cressie et al., 2009), nanowire growth processes (Huang, 2010), and single-particle tracking (Lysy et al., 2016).

Consider M particle trajectories $\boldsymbol{X}_1(t), \cdots, \boldsymbol{X}_M(t)$ recorded for a given antibody concentration. Ideally, all trajectories would have the same location-scale subdiffusive

parameters,

$$\boldsymbol{X}_i(t) \overset{\text{iid}}{\sim} f(\boldsymbol{X}(t) \,|\, \boldsymbol{\varphi}_i).$$

However, this is not the case in practice. As discussed in the introductory chapter, particles experience different dynamics due to heterogeneity of the fluid and local environment effects. This behaviour suggests the need for a more complex multi-level model that can take into account individual particle behaviour as well as population level effects. The following hierarchical model captures such heterogeneity among particle trajectories in a given experimental environment,

$$\boldsymbol{X}_i(t) \,|\, \boldsymbol{\varphi}_i \overset{\text{ind}}{\sim} f(\boldsymbol{X}(t) \,|\, \boldsymbol{\varphi}_i), \qquad \boldsymbol{\varphi}_i \overset{\text{iid}}{\sim} \mathcal{N}(\boldsymbol{\lambda}, \boldsymbol{\Sigma}),$$

where $\boldsymbol{\varphi}_i$ corresponds to the i^{th} particles location-scale subdiffusive parameters. The hyperparameters of this hierarchical model are $\boldsymbol{\theta} = (\boldsymbol{\lambda}, \boldsymbol{\Sigma})$. Moving beyond a single-path MSD presented in Section 2.2, the population-level MSD is

$$MSD(t) = E[D \cdot t^{\alpha} \,|\, \boldsymbol{\lambda}, \boldsymbol{\Sigma}] \tag{2.2}$$

This population-level MSD will be used to evaluate scientific relevance in the subsequent chapters. The likelihood function for hyperparameters $\boldsymbol{\theta} = (\boldsymbol{\lambda}, \boldsymbol{\Sigma})$ is

$$\mathcal{L}(\boldsymbol{\theta} \,|\, \boldsymbol{X}) = \prod_{i=1}^{M} p(\boldsymbol{X}_i \,|\, \boldsymbol{\theta}) = \prod_{i=1}^{M} \int p(\boldsymbol{X}_i \,|\, \boldsymbol{\varphi}_i) p(\boldsymbol{\varphi}_i \,|\, \boldsymbol{\theta}) d\boldsymbol{\varphi}_i$$

which is intractable. Exact parameter inference would typically require MCMC on high-dimensional parameter set $(\boldsymbol{\theta}, \boldsymbol{\varphi}_1, \cdots, \boldsymbol{\varphi}_M)$. Therefore, the following approximation (Lysy et al., 2016) is employed,

$$\ell(\boldsymbol{\varphi}_i \,|\, \boldsymbol{X}_i) \approx -\frac{1}{2}(\boldsymbol{\varphi}_i - \hat{\boldsymbol{\varphi}}_i)' \boldsymbol{V}_i^{-1} (\boldsymbol{\varphi}_i - \hat{\boldsymbol{\varphi}}_i),$$

where $\hat{\boldsymbol{\varphi}}_i$ is MLE from the location-scale model and $\boldsymbol{V}_i$ is the corresponding Fisher information calculated from the hessian of the matrix-normal log-likelihood evaluated at the MLE. Hence, the following likelihood $\ell(\boldsymbol{\lambda}, \boldsymbol{\Sigma}, \boldsymbol{\varphi}_1, \cdots, \boldsymbol{\varphi}_M \,|\, \boldsymbol{X})$ is approximately

the same as the likelihood from the following "normal-normal" model,

$$\begin{aligned}\hat{\boldsymbol{\varphi}}_i \,|\, \boldsymbol{\varphi}_i, \boldsymbol{V}_i &\overset{\text{ind}}{\sim} \mathcal{N}(\boldsymbol{\varphi}_i, \boldsymbol{V}_i) \\ \boldsymbol{\varphi}_i \,|\, \boldsymbol{\lambda}, \boldsymbol{\Sigma} &\overset{\text{iid}}{\sim} \mathcal{N}(\boldsymbol{\lambda}, \boldsymbol{\Sigma}).\end{aligned}$$

Therefore, this hierarchical model effectively pools together the location-scale model MLE $\hat{\boldsymbol{\varphi}}_i$ and the corresponding Fisher information $\boldsymbol{V}_i$. Approximate Bayesian inference can then be performed where the parameter prior is Matrix-Normal Inverse-Wishart (MNIW),

$$(\boldsymbol{\lambda}, \boldsymbol{\Sigma}) \sim \textsc{mniw}(\boldsymbol{\Lambda}_0, \mathbf{Y}_0, \boldsymbol{\Omega}_0, \nu_0) \iff \begin{aligned}\boldsymbol{\Sigma} &\sim \text{Inv-}\mathcal{W}(\boldsymbol{\Omega}_0, \nu_0) \\ \boldsymbol{\lambda} \,|\, \boldsymbol{\Sigma} &\sim \mathcal{MN}(\boldsymbol{\Lambda}_0, \mathbf{Y}_0^{-1}, \boldsymbol{\Sigma}).\end{aligned}$$

The approximate posterior $p(\boldsymbol{\lambda}, \boldsymbol{\Sigma}, \boldsymbol{\varphi}_1, \cdots, \boldsymbol{\varphi}_M)$ can be efficiently sampled from with a Gibbs sampler. Further details on the Gibbs sampling procedure in the context of a normal-normal hierarchical model can be found in Appendix B.

Chapter 3

Bayesian Sample Size Determination

Let $\boldsymbol{X}_i(t)$ denote a 2-dimensional trajectory of the ith particle under a single experimental environment. The model for this data given parameter $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ is

$$\boldsymbol{X}_i(t) \overset{\text{ind}}{\sim} p(\boldsymbol{X}(t) \mid \boldsymbol{\theta}).$$

The goal is to determine whether the particle dynamics in the experiment is scientifically *meaningful*. That is, the null hypothesis of *no meaningfulness* is formulated as

$$H_0 : \boldsymbol{\theta} \in \mathcal{S}_0, \tag{3.1}$$

where $\boldsymbol{\theta}$ and $\mathcal{S}_0$ are determined by the scientist. Under this framework, rejecting the null hypothesis indicates scientific relevance. This work uses the population-level MSD in (2.2) to evaluate scientific relevance, and $\boldsymbol{\theta} = (\boldsymbol{\lambda}, \boldsymbol{\Sigma})$ from the multiple-trajectory hierarchical model covered in the preceding Chapter. Namely, we consider null hypotheses of the form

$$H_0 : \tau(\boldsymbol{\theta}) \in \mathcal{S}_0 = [\tau_0, \infty),$$

where $\tau(\boldsymbol{\theta}) = E[D \cdot t_0^{\alpha} \mid \boldsymbol{\theta}]$ is the population-level MSD at a specific time t_0 determined by the scientist, where $t_0 = 2$ seconds in this work. Through appropriately setting $\mathcal{S}_0$ practitioners can assess scientific relevance, such as a 20% reduction in population-level

MSD. A Monte Carlo approach is used to evaluate $Pr(\tau(\boldsymbol{\theta}) \in \mathcal{S}_0) = Pr(H_0)$.

Next, let $\boldsymbol{X}_i = (\boldsymbol{X}_i(t_0), \boldsymbol{X}_i(t_1), \ldots, \boldsymbol{X}_i(t_{N_i}))$ denote the discretely recorded observations for the i^{th} particle, where $t_n = n\Delta t$. Given N two-dimensional particle trajectories $\boldsymbol{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N)$, the posterior parameter distribution is

$$p(\boldsymbol{\theta} \,|\, \boldsymbol{X}) = \pi(\boldsymbol{\theta}) \prod_{i=1}^{N} p(\boldsymbol{X}_i \,|\, \boldsymbol{\theta}_i).$$

We then have the posterior distribution of no scientific meaningfulness, i.e. scientific irrelevance, as defined by (3.1):

$$\Pr(H_0 \,|\, \boldsymbol{X}) = \Pr\{\boldsymbol{\theta} \in \mathcal{S}_0 \,|\, \boldsymbol{X}\} = \int_{\mathcal{S}_0} p(\boldsymbol{\theta} \,|\, \boldsymbol{X}) \, \mathrm{d}\boldsymbol{\theta}. \tag{3.2}$$

Thus, conceptually, given a particular dataset $\boldsymbol{X}$, we can evaluate whether the experiment is scientifically meaningful based on based on how far $Pr(H_0 \,|\, \boldsymbol{X})$ is from $1/2$. While the calculation of (3.2) provides a decision rule for assessing scientific relevance, values of $\Pr(H_0 \,|\, \boldsymbol{X})$ close to $\frac{1}{2}$ indicate low confidence in the given decision.

The Bayesian SSD problem can be formulated as follows. Suppose that we have already obtained a set of N_{tr} particle trajectories from a single experiment, denoted as our training data $\boldsymbol{X}_{\text{tr}}$. We wish to determine the sample size N_{ts} of additional samples required such that the probability $p_0 = \Pr(H_0 \,|\, \boldsymbol{X}_{\text{tr}}, \boldsymbol{X}_{\text{ts}})$ based on all the data is sufficiently far from $\frac{1}{2}$. That is, let

$$q_0(\boldsymbol{X}) = \max\big\{ \Pr(H_0 \,|\, \boldsymbol{X}), 1 - \Pr(H_0 \,|\, \boldsymbol{X}) \big\}.$$

Namely, $q_0(\boldsymbol{X})$ represents the probability of being scientifically relevant or scientifically irrelevant. Hence, experimentalists require $q_0(\boldsymbol{X})$ be close to 1 to make a decision on whether an experiment is scientifically meaningful. Then the SSD problem is to find $N = N_{\text{ts}}$ such that

$$Q_N = \Pr\left\{ q_0(\boldsymbol{X}_{\text{tr}}, \boldsymbol{X}_{\text{ts}}) > \varepsilon \,|\, \boldsymbol{X}_{\text{tr}} \right\} > 1 - \phi,$$

for given $\varepsilon \in (0.5, 1)$ and $\phi \in (0, 1)$. Thus, as $N \to \infty$ we expect $Q_N \to 1$. The following is a general Monte Carlo algorithm for estimating Q_N.

Algorithm 1 Monte Carlo Approximation to Q_N

procedure GETQN($\boldsymbol{X}_{tr}, N_{ts}, B, C, \varepsilon$)
  $\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(B)} \sim \text{post}(\boldsymbol{X}_{\text{tr}})$
  for $b = 1 : B$ **do**
    $\boldsymbol{\theta}_1^{(b)}, \ldots, \boldsymbol{\theta}_C^{(b)} \sim$ POST_UPDATE$(\boldsymbol{X}_{tr}, \boldsymbol{\theta}^{(b)}, N_{ts})$
    for $c = 1 : C$ **do**
      $\delta_c^{(b)} \leftarrow$ IS_RELEVANT$(\boldsymbol{\theta}_c^{(b)}, S_0)$
    end for
    $\hat{p}_0^{(b)} \leftarrow \frac{1}{C} \sum_{c=1}^{C} \delta_c^{(b)}$
    $\hat{q}_0^{(b)} \leftarrow max(\hat{p}_0^{(b)}, 1 - \hat{p}_0^{(b)})$
  end for
  $\hat{Q}_N \leftarrow \frac{1}{B} \sum_{b=1}^{B} \mathbb{I}[\hat{q}_0^{(b)} > \varepsilon]$
  return $\hat{Q}_N$
end procedure

procedure POST_UPDATE($\boldsymbol{X}_{tr}, \boldsymbol{\theta}, N_{ts}$)
  generate $\boldsymbol{X}_{ts}$
  return $p(\boldsymbol{\theta} \mid \boldsymbol{X}_{tr}, \boldsymbol{X}_{ts})$
end procedure

procedure IS_RELEVANT($\boldsymbol{\theta}, S_0$)
  $M \leftarrow 1e5$
  $\boldsymbol{\varphi}_1, \ldots, \boldsymbol{\varphi}_M \sim p(\boldsymbol{\varphi} \mid \boldsymbol{\theta})$
  for $m = 1 : M$ **do**
    $\eta_m \leftarrow D_m t^{\alpha_m}$
  end for
  $\bar{\eta} \leftarrow \frac{1}{M} \sum_{m=1}^{M} \eta_m$
  return $\mathbb{I}[\bar{\eta} \in S_0]$
end procedure

The bottleneck in Algorithm 1 is the Monte Carlo or MCMC sampling from B posterior distributions, each of the form

$$p(\boldsymbol{\theta} \mid \boldsymbol{X}_{\text{tr}}, \boldsymbol{X}_{\text{ts}}) \propto \left[\prod_{i=1}^{N_{\text{tr}}} p(\boldsymbol{X}_i \mid \boldsymbol{\theta})\right] \cdot \left[\prod_{j=1}^{N_{\text{ts}}} p(\boldsymbol{X}_j \mid \boldsymbol{\theta})\right] \cdot \pi(\boldsymbol{\theta}).$$

Due to experimental time restrictions in actual laboratory settings, three methods to accelerate computation are explored.

3.1 Bootstrap-Variance Approximation

Maximum likelihood estimation of the location-scale model parameters for each new time series takes considerable time for a large number of trajectories and can be avoided through the use of bootstrapping. Bootstrapping is an incredibly powerful method in statistics and is considered an outstanding breakthrough in the statistical community (Efron, 1992).

Consider N_{tr} location-scale MLE parameter fits $\hat{\boldsymbol{\varphi}}_i, \quad i \in \{1, \cdots N_{tr}\}$, and their corresponding Fisher information matrices $\boldsymbol{V}_i$. Define $\boldsymbol{S}_i = (\hat{\boldsymbol{\varphi}}_i, \boldsymbol{V}_i)$. Then further location-scale parameter fitting for new test particle trajectories can be eliminated by drawing *with replacement* from $\boldsymbol{S}_1, \ldots, \boldsymbol{S}_{N_{\text{tr}}}$. Namely, bootstrapping N_{ts} samples from the training data acts as acquiring N_{ts} new experimental trajectories. Updating the posterior with the original training samples and the bootstrap samples permits sampling from $p(\boldsymbol{\theta} \mid \boldsymbol{X}_{\text{tr}}, \boldsymbol{X}_{\text{ts}})$.

While bootstrapping circumvents location-scale model fitting in this approximation, it does not allow one to avoid MCMC posterior training. This bootstrap-variance approximation (BV) is expected to perform well as long as N_{tr} is large enough that the training samples represent the population. Hence, if a large amount of heterogeneity is expected, it can be beneficial to increase N_{tr} if one uses this method to ensure the training samples capture a large number of diverse particle behaviours.

3.2 Common-Variance Approximation

A faster but less accurate approximation is to assume a common variance approximation (CV). Recall the first two terms of the transformed location-scale parameter vector, $\boldsymbol{\varphi}^{(1)} = logit\left(\frac{\alpha}{2}\right)$ and $\boldsymbol{\varphi}^{(2)} = log(\sigma_1^2 + \sigma_2^2)$. Then the 5-dimensional parameter set

sufficient for MSD calculation, denoted here by $\mathbf{\Gamma}$, is:

$$\mathbf{\Gamma}_{1\times 5} = \left(\boldsymbol{\varphi}^{(1)}, \boldsymbol{\varphi}^{(2)}, log(var(\boldsymbol{\varphi}^{(1)})), log(var(\boldsymbol{\varphi}^{(2)})), logit(\frac{cov(\boldsymbol{\varphi}^{(1)}, \boldsymbol{\varphi}^{(2)})}{\sqrt{var(\boldsymbol{\varphi}^{(1)})var(\boldsymbol{\varphi}^{(2)})}})\right)$$

Training on $\boldsymbol{X}_{\text{tr}}$, we calculate $\hat{\mathbf{\Gamma}}_{\text{tr}}$ by fitting the location-scale model to each trajectory, and subsequently calculate $\hat{\boldsymbol{V}}_{\text{tr}} = Var(\hat{\mathbf{\Gamma}}_{\text{tr}} \mid \boldsymbol{X}_{\text{tr}})$. The following approximation then allows for accelerated simulation of new sufficient parameter sets $\hat{\mathbf{\Gamma}}_{\text{ts}}$,

$$\widetilde{\mathbf{\Gamma}} \mid \hat{\mathbf{\Gamma}}_{\text{tr}}, \hat{\boldsymbol{V}}_{\text{tr}} \sim \mathcal{N}(\hat{\mathbf{\Gamma}}_{\text{tr}}, (1 + \frac{N_{\text{tr}}}{N_{\text{ts}}})\hat{\boldsymbol{V}}_{\text{tr}}),$$

$$\hat{\mathbf{\Gamma}}_{\text{ts}} \mid \widetilde{\mathbf{\Gamma}}, \hat{\mathbf{\Gamma}}_{\text{tr}}, \hat{\boldsymbol{V}}_{\text{tr}} \overset{\text{ind}}{\sim} \mathcal{N}\left(\frac{N_{\text{tr}}\hat{\mathbf{\Gamma}}_{\text{tr}} + N_{\text{ts}}\widetilde{\mathbf{\Gamma}}}{N_{\text{tr}} + N_{\text{ts}}}, \frac{N_{\text{tr}}}{N_{\text{tr}} + N_{\text{ts}}}\hat{\boldsymbol{V}}_{\text{tr}}\right).$$

That is, sampling $\hat{\mathbf{\Gamma}}_{\text{ts}}$ corresponds to generating new samples of the parameter set sufficient for MSD calculation with the same uncertainty $\hat{\boldsymbol{V}}_{\text{tr}}$, relative to sample size. We can then use these samples $\hat{\mathbf{\Gamma}}_{\text{ts}}$ to calculate the population-level MSD at a significantly accelerated pace.

3.3 Zero-Variance Approximation

Another simplification to consider is to assume $\boldsymbol{V}_i = \mathbf{0}$, denoted as the zero-variance approximation (ZV). While this assumption effectively removes a level of the hierarchical model, it nonetheless gives reasonable results as will be presented in the simulation section. Furthermore, the speed of ZV is comparable to CV.

The approximation is as follows. If length N_j of each trajectory is long enough, then $\hat{\boldsymbol{\varphi}}_j \approx \boldsymbol{\varphi}_j$, such that

$$\hat{\boldsymbol{\varphi}}_j \overset{\text{iid}}{\sim} \mathcal{N}(\boldsymbol{\lambda}, \mathbf{\Sigma}).$$

Then for the noninformative prior

$$\pi(\boldsymbol{\lambda}, \boldsymbol{\Sigma}) \propto \exp\left\{ -\tfrac{\nu+2}{2} \cdot \log|\boldsymbol{\Sigma}| \right\},$$

the posterior distribution given $\boldsymbol{X} = (\boldsymbol{X}_{\text{tr}}, \boldsymbol{X}_{\text{ts}})$ and $N = N_{\text{tr}} + N_{\text{ts}}$ is

$$\begin{aligned} \boldsymbol{\Sigma} \,|\, \boldsymbol{X} &\sim \text{Inv-}\mathcal{W}(\boldsymbol{S}, \nu + N - 1) \\ \boldsymbol{\lambda} \,|\, \boldsymbol{\Sigma}, \boldsymbol{X} &\sim \mathcal{N}\big(\hat{\boldsymbol{\lambda}}, \tfrac{1}{N}\boldsymbol{\Sigma}\big), \end{aligned}$$

where $\hat{\boldsymbol{\lambda}} = \frac{1}{N}\sum_{j=1}^{N} \hat{\boldsymbol{\varphi}}_j$ and $\boldsymbol{S} = \sum_{j=1}^{N} (\hat{\boldsymbol{\varphi}}_j - \hat{\boldsymbol{\lambda}})(\hat{\boldsymbol{\varphi}}_j - \hat{\boldsymbol{\lambda}})'$. While a stronger approximation than the CV method, simulation results to follow indicate reasonable performance.

Chapter 4

Simulation Study

To evaluate the performance of the SSD algorithms presented in Chapter 3, it is necessary to explore a controlled simulation environment. While the SSD methodology has broad applicability in a variety of scientific domains, this work focuses on a simulation study for the motivating SPT application.

The simulation environment is modelled from the 0 mg/L antibody concentration experiment in Chapter 5, focusing solely on the constrained trajectories. Hence, a single-training dataset is used for this Chapter. True parameter values are presented in Table 4.1. The simulation Monte Carlo parameters were set to $B = 250$, $C = 500$, and the interval threshold for Q_N was set to $\varepsilon = 0.8$. Excellent convergence of the Gibbs sampler can be found in Appendix C.

Table 4.1: Simulation Environment

Variable	Value	Description
τ_0	3.57	True Population-Level MSD
N_{tr}	200	Number of Training Samples
L	300	Trajectory Length
Δt	1/15	Frame Rate (Hz)
t_0	2	MSD Reference Time (seconds)

Firstly, we look at the Bayesian SSD algorithms while varying the null hypothesis of no meaningfulness, $H_0 : \boldsymbol{\theta} \in \mathcal{S}_0$, where $\mathcal{S}_0 = [\tau_0, \infty)$ with $\tau_0 = E[D \cdot t_0^{\alpha}]$ evaluated at $t_0 = 2$ seconds. By changing $\mathcal{S}_0$ (i.e. changing τ_0) we can explore the effect of the null hypothesis on $\hat{q}_0(\boldsymbol{X}_{\text{tr}}, \boldsymbol{X}_{\text{ts}})$. Figure 4.1 presents the $\hat{q}_0$ estimates with increasing N_{ts}. Each row uses a different $\mathcal{S}_0$ set. The closer the box plots to 1, the higher the confidence in the decision of scientific relevance.

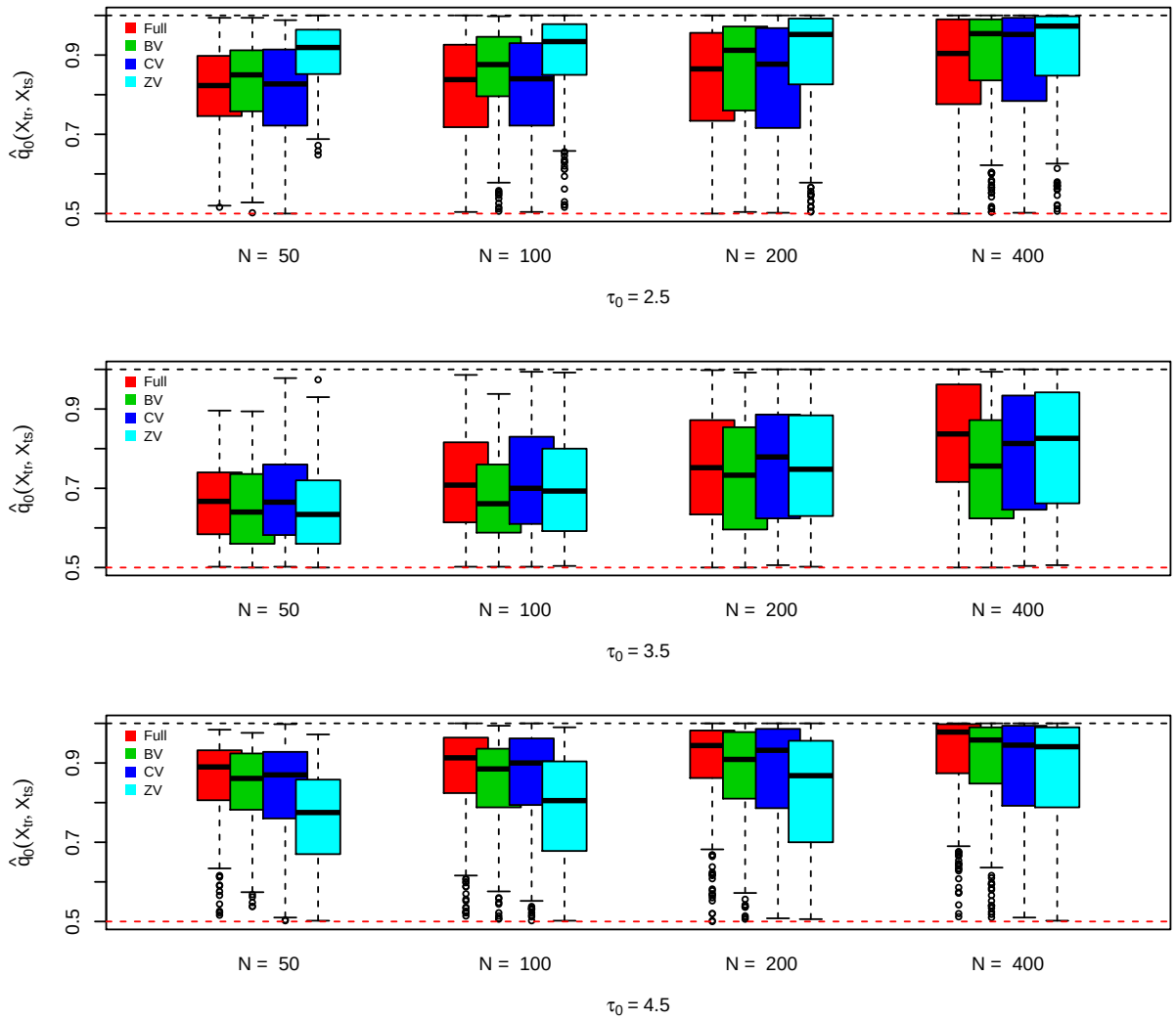

Figure 4.1: Effects of $\mathcal{S}_0 = [\tau_0, \infty)$ on $\hat{q}_0(\boldsymbol{X}_{\text{tr}}, \boldsymbol{X}_{\text{ts}})$ as a function of increasing N_{ts}. Note that the true population-level MSD is close to 3.5, hence the middle box plots have values of $\hat{q}_0$ closer to 0.5 for any given N_{ts}. Furthermore, as N_{ts} increased from 50 to 400, the expected behaviour of increasing $\hat{q}_0$ is observed. This indicates that with more paths practitioners can gain confidence on the conclusion of whether or not the SPT trajectories lead to a scientifically relevant conclusion.

Two immediate conclusions can be drawn from Figure 4.1. Firstly, the scientifically relevant region impacts the spread of the $\hat{q}_0$ estimates. This asymmetry is observed in Figure 4.1 as the top-row box plots show less spread than the bottom-row box plots. The most compelling reason for this behaviour is that $\hat{q}_0$ is being calculated using the expected value of path-wise MSDs, which exhibit a non-symmetric and strictly positive distribution. While the effects of this asymmetry will be overlooked in this work, it poses an interesting area for future research on how the structure of the scientifically relevant decision criterion impacts inference.

The second take-away from Figure 4.1 is that there still remains open areas to explore regarding the current SSD approximations. Among the four SSD algorithms, ZV seems to underestimate $\hat{q}_0$ when τ_0 is set above the true population-level MSD, while it overestimates $\hat{q}_0$ when τ_0 is set below the true population-level MSD. However, when τ_0 is near the true population-level MSD (middle row in Figure 4.1), then ZV gives relatively similar results for $\hat{q}_0$ compared to the other three SSD algorithms. This behaviour could suggest that the assumption underlying the ZV algorithm is too strong. Hence, a fruitful area of further research entails a deeper analysis of the approximation assumptions and how inference is impacted based on the structure of the null hypothesis.

4.1 Effect of Heterogeneity

Discussed in Chapter 1, heterogeneity is pervasive in SPT experiments. The level of heterogeneity differ for a number of factors such as different experimental conditions (e.g. temperature, fluid viscosity, local fluid density). Thus, it is of interest to explore the impact of varying levels of heterogeneity on the Bayesian SSD algorithms. Figure 4.2 shows $\hat{q}_0$ estimates as heterogeneity is increased from 10% to 100%. Here, 100% heterogeneity corresponds to the mean posterior variance hyperparameter, denoted in this section as $\mathbf{\Sigma}_{ref}$, from the reference baseline simulation hierarchical model fit (i.e. the 0 mg/L antibody concentration environment). In Figure 4.2, each level of heterogeneity corresponds to a modified variance matrix, i.e. 30% heterogeneity corresponds

to an environment with a mean posterior variance hyperparameter of $0.3 \cdot \mathbf{\Sigma}_{ref}$. All box plots correspond to $N_{\text{tr}} = 200$ training paths, and $N_{\text{ts}} = 500$ test paths, where each path is of length $L = 300$. $B = 250$ estimates were simulated at each heterogeneity level.

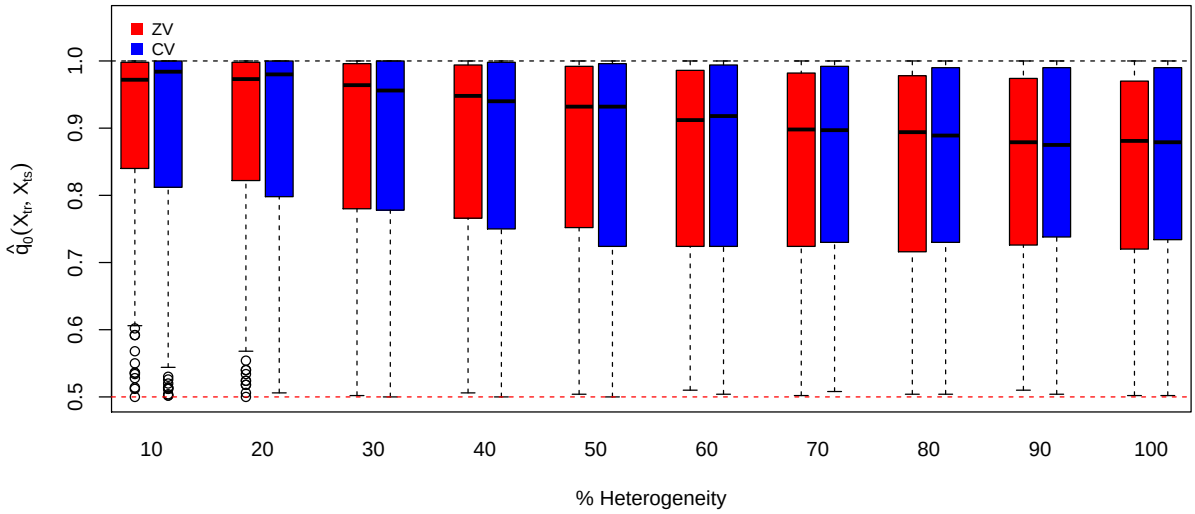

Figure 4.2: Simulated $\hat{q}_0(\boldsymbol{X}_{\text{tr}}, \boldsymbol{X}_{\text{ts}})$ as a function of increasing heterogeneity. Increased heterogeneity naturally leads to a decreasing confidence of the null hypothesis for a fixed sample size $N_{\text{ts}} = 500$. Namely, increasing the population-level heterogeneity at a fixed sample size causes $\hat{q}_0(\boldsymbol{X}_{\text{tr}}, \boldsymbol{X}_{\text{ts}})$ estimates to fall, suggesting you need more paths at higher levels of heterogeneity to retain a given $\hat{q}_0$ level.

4.2 Performance Evaluation

This section explores whether the reported confidence in scientific relevance indeed correlates well with the actual magnitude of the scientific response. Through simulating N_{ts} trajectories, the percentage of time the resulting population-level MSD actually falls in $\mathcal{S}_0 = [\tau_0, \infty)$ is recorded to assess $P(H_0|\boldsymbol{X}_{\text{tr}}, \boldsymbol{X}_{\text{ts}})$. Figure 4.3 displays the effect of $\mathcal{S}_0$ on $\hat{p}_0$. By using simulated data, the reflected sigmoid-shaped plots can be used as a visual check of the performance as $\mathcal{S}_0$ is varied. Here, τ_{true} denotes the true population-level MSD from Table 4.1. It is clear that increasing N_{ts} pushes the $\hat{p}_0$ esti-

mates closer to 0 ($\tau_0 < \tau_{true}$) or 1 ($\tau_0 > \tau_{true}$) for a given $\mathcal{S}_0$. This is expected since a larger number of test paths will result in higher confidence in the decision of whether an experiment is scientifically meaningful.

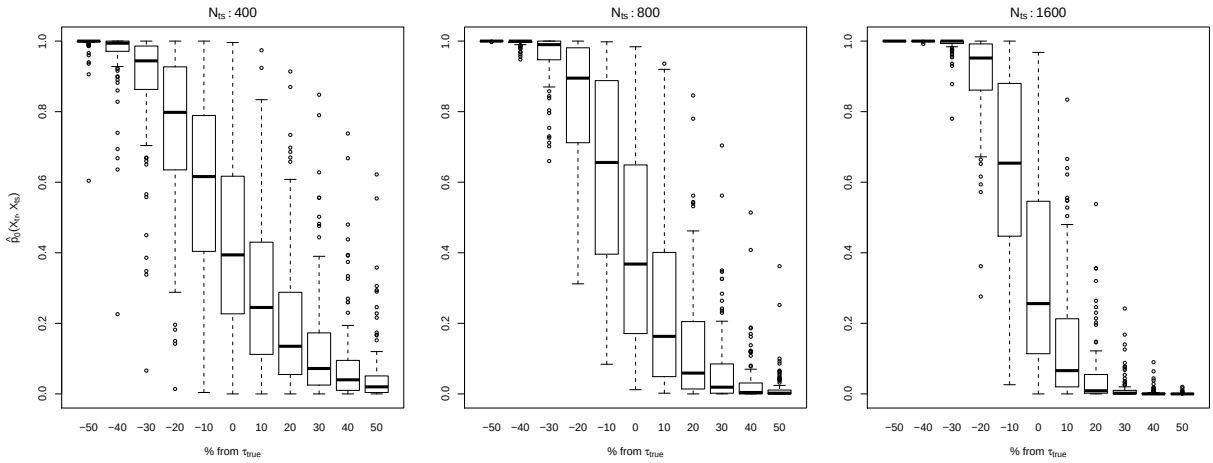

Figure 4.3: $P(H_0|\boldsymbol{X}_{\text{tr}}, \boldsymbol{X}_{\text{ts}})$ as $\mathcal{S}_0 = [\tau_0, \infty)$ is varied. The reflected-sigmoidal shape becomes more apparent as N_{ts} increases. This indicates that larger N_{ts} provides increased evidence to make a conclusion regarding scientific relevance. As τ_0 moves away from the true population-level MSD (i.e. as τ_0 moves farther from τ_{true}) the $\hat{p}_0$ estimates approach 0 or 1. Each box corresponds to 100 simulations.

Furthermore, Table 4.2 displays Q_N for varying levels of N_{ts} for each the CV and ZV methods for τ_0 set 20% below τ_{true}. The threshold $\varepsilon = 0.8$ was chosen, but may be any value between 0.5 and 1. However the closer ε is set to 1, the larger number of N_{ts} samples will be needed to reach a given decision confidence.

Table 4.2: Simulation Results for CV and ZV Methods

N_{ts}	100	400	800	1600	3200
CV: P[$\hat{q}_0 > 0.8$]	0.30	0.57	0.75	0.82	0.85
ZV: P[$\hat{q}_0 > 0.8$]	0.40	0.63	0.71	0.82	0.87

Furthermore, SPT laboratory experiments routinely have time restrictions on the order of minutes to a few hours. It is therefore crucial that an SSD algorithm be efficient for practical implementation in real-time experiments. Table 4.3 displays the computing

time require to run a standard simulation to estimate $\hat{q}_0$. The rapid speed of ZV and CV are significantly smaller than that of BV or the full algorithm.

Table 4.3: Simulation Speed Comparison ($N_{\text{ts}} = 400$, $L = 300$, $B = 10$)

SSD Method	Run Time (min)
Full	9.7
BV	5.8
CV	0.4
ZV	0.4

Given the results in the Section 4.1, ZV and CV offer very similar $\hat{q}_0$ estimates to that of the full algorithm, and therefore should be used instead of the full algorithm in this case since they offer approximately a 20-times speed improvement. Regarding how to select whether to use ZV or CV, this is an open problem. Due to the stronger assumption underlying the ZV algorithm, it is recommended to use CV if results are similar for the two algorithms. Future research studies, including those of different particles and mediums, may reveal one of the four presented algorithms as superior depending on the experimental setup. Hence, experimentalists should run a simulation similar to the one presented in Section 4.1 anytime they are beginning an entirely new experiment (i.e. new scientific study). This way, the behaviour of the four SSD algorithms presented in this work can be analyzed and the algorithm giving the optimal performance in terms of speed and accuracy can be selected for moving forward in an experimental study.

4.3 Conclusion

This simulation section has addressed a wide-variety of scenarios regarding the effectiveness of the proposed Bayesian SSD algorithms. Firstly, the full SSD algorithm can be well-approximated by simpler algorithms achieving a large reduction in computation

time. The ZV and CV algorithms achieve about a 20-fold reduction in computation time compared to the full SSD algorithm. This is an outstanding result and further strengthens the adoption and use of these algorithms in practice.

The impact of increasing levels of heterogeneity on inference was also explored. Larger heterogeneity among SPT sample paths requires an increased number of test samples N_{ts} to achieve the same spread in $\hat{q}_0$ estimates as that of lower heterogeneity environments. Therefore, practitioners must be knowledgeable of the amount of heterogeneity expected in an experiment, and adjust their expectations accordingly for the order of the recommended sample size.

Furthermore, the performance assessment in this section indicates an acceptable match between the Bayesian SSD algorithms sample size recommendations and the actual results of a practitioner acquiring the suggested number of sample paths. The reflected-sigmoidal shape of $\hat{p}_0$ as N_{ts} increases was clearly evident and supports the proper behaviour of the SSD algorithms. Moreover, the fact that this simulation study focused on a baseline simulation with high heterogeneity present in the SPT sample paths illustrates that high-heterogeneity experiments may require a large number of test samples to determine scientific relevance. This poses two experimental issues. Firstly, thousands of paths may not be experimentally feasible if resources are significantly constrained. Secondly, if experiments are run sequentially, acquisition of thousands of experiments may be lengthy and thus practitioners must be careful to keep experimental conditions consistent across the same experiment.

This simulation section has explored an approach to SPT sample size determination problem to assess scientific relevance based on population-level MSD. Nonetheless, the scientific relevance definition is open to adjustment. This opens up a variety of future research avenues such as scientific relevance based on statistics other than the population-level MSD. Some innovative frontiers with significant practical applications include first-passage time behaviour based scientific relevance. First-passage time was not used in this work to evaluate scientific relevance due to the difficulty in modelling the first-passage time of 2-dimensional fractional Brownian motion, however this poses

an open area for future research. Understanding the length of time a particle takes to diffuse a certain prescribed length may hold novel insights experimentalists can exploit to accelerate SPT research and advance their application under study.

Chapter 5

Application

This section extends the previous simulation study to real SPT experimental data. Analyzing experimental data is critical in the development of practical algorithms since statisticians and experimental collaborators must work together to solve real-world complex problems. This chapter begins with a detailed analysis of constrained versus unconstrained trajectories, followed by a MSD ratio analysis to emphasize differences in virus behaviour across varying antibody concentrations. The chapter closes with illustrated use of the SSD algorithm in practice.

To illustrate use of the Bayesian SSD algorithm, this section compares results across five experimental environments. Hundreds of 2-dimensional HSV trajectories were recorded at 15 frames per second at five different levels of antibody concentration. A summary of the data is displayed in Table 5.1.

Table 5.1: HSV Experimental Data Profile

Ab Concentration	0 mg/L	33 mg/L	100 mg/L	333 mg/L	1000 mg/L
Total Number of Paths	775	724	737	737	734
Number of Constrained Paths	141	283	432	593	574
Number of Unconstrained Paths	634	441	305	144	160
Average Length	130 (50, 302)	178 (50, 302)	214 (50, 302)	259 (50, 302)	253 (51, 302)

5.1 Constrained and Unconstrained Trajectories

The 3,707 trajectories under study in this work were classified as either constrained or unconstrained based on their trajectory increment standard deviation, denoted here by Δx and Δy, where x and y correspond to the 2 dimensions of movement under study. Using the increment standard deviations for each dimensional 2-dimensional SPT trajectory, virus trajectories were labelled as constrained if $log(\sigma_{\Delta x}) < -0.3$ and $log(\sigma_{\Delta y}) < -0.3$. The threshold was chosen based on Figure 5.1 to achieve reasonable separation between the two visible clusters. Thus there are 2023 constrained SPT paths, and 1684 unconstrained SPT paths.

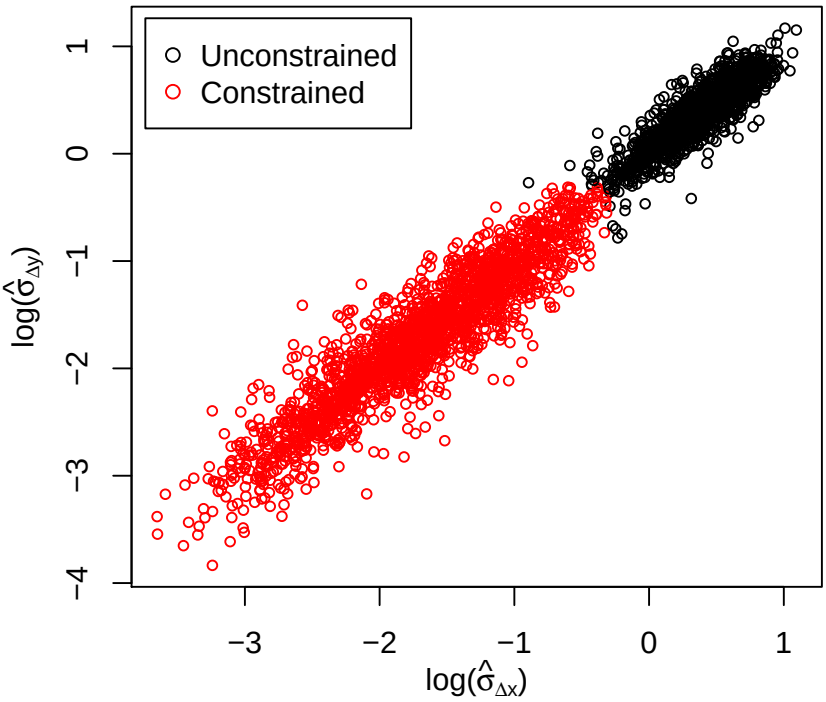

Figure 5.1: Labelling trajectories based on increment standard deviations. Each point is calculated from a single virus trajectory, resulting in 2023 constrained paths, and 1684 unconstrained paths. A simple cut-off for identification of constrained paths was chosen as $log(\hat{\sigma}_{\Delta x}) < -0.3$ and $log(\hat{\sigma}_{\Delta y}) < -0.3$.

Three of the location-scale parameter estimates for every SPT trajectory (both constrained and unconstrained) are shown in Figure 5.2. Recall, α is the anomalous diffusion exponent, D is the diffusion coefficient, and μ_1 is the linear drift term in the first

dimension of the location-scale model. There are significant differences among varying level of antibody concentration, such as the negative shift in parameter estimates as the antibody concentration increases. This effect is expected as higher antibody concentration should cause more HSV viruses to become constrained in motion (recall virus-antibody interaction hinders virus movement). However, it is important to notice that even at 1000 mg/L antibody concentration there remain virus trajectories that are unconstrained - seen by the 1000 mg/L histograms where there are paths that retain large $logit(\alpha/2)$ and $log(D)$. In physical terms this suggests that there are some viruses in solution that plainly do not interact with the antibody. While outside the scope of this work, an open area of research in virology could pose and analyze the physical mechanism of why some virus particles do not interact with the corresponding antibody.

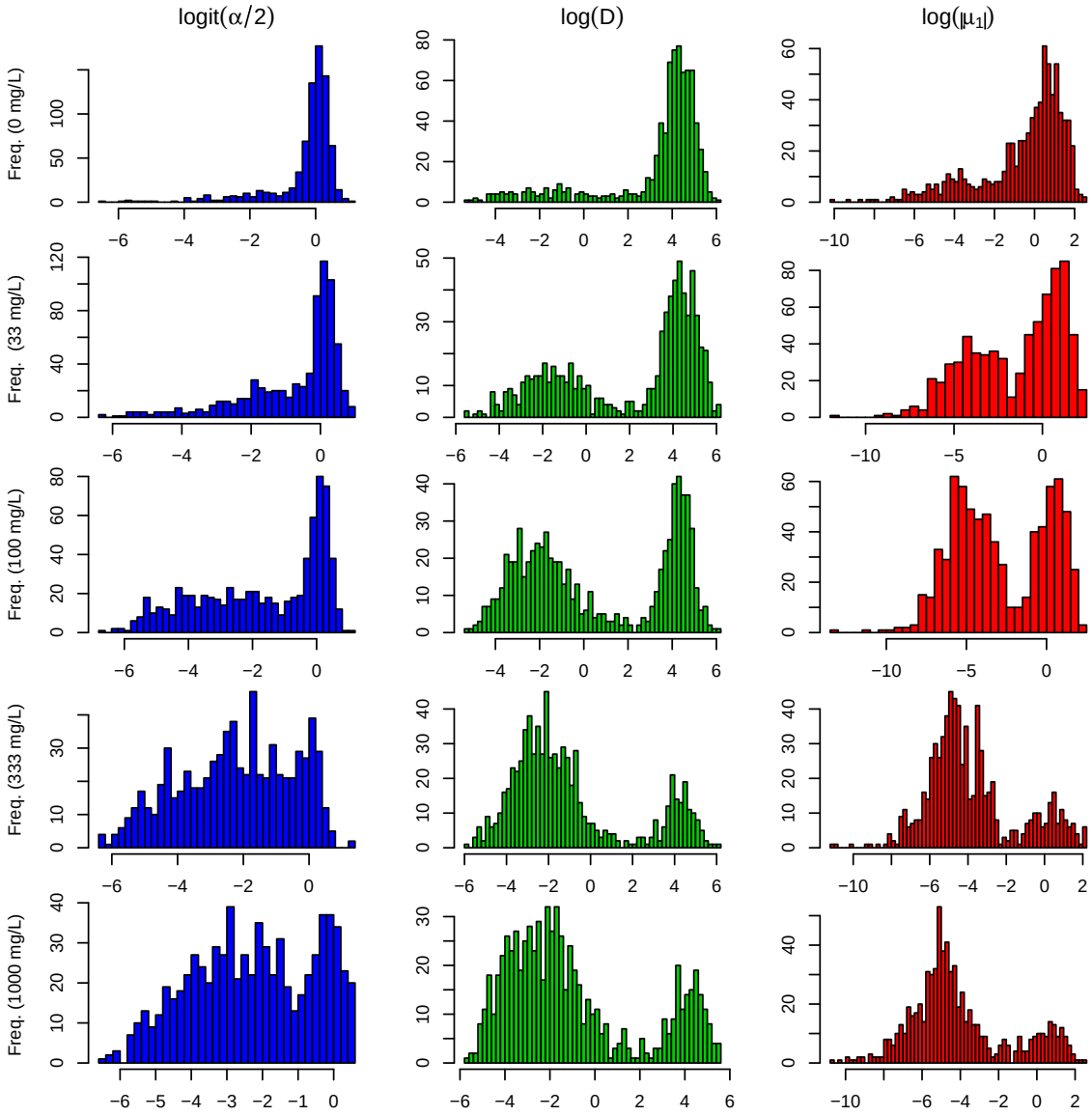

Figure 5.2: Differences across three location-scale parameters are directly visible as the antibody concentration increases from 0 mg/L to 1000 mg/L. A drift toward smaller values is seen across the three parameters as the antibody concentration increases.

Extending the analysis to the joint distribution of the location-scale parameters further illustrates the differences between constrained and unconstrained labelled trajectories. Figure 5.3 displays joint plots for three of the parameters from the location-scale model. Constrained and unconstrained trajectories are again colour-labelled. Unconstrained paths clearly exhibit larger anomalous diffusion exponent, larger diffusion coefficient, and larger linear drift. Based on the separation between the two cluster formations

in Figure 5.3, the simplistic method of labelling trajectories as constrained and unconstrained reasonably partitions the location-scale model parameters. Namely, constrained trajectories arise as slow moving particles exhibiting low anomalous diffusion exponent, low diffusion coefficient, and a small linear drift. An increasingly complicated separation technique for labelling constrained and unconstrained trajectories such as using a mixture-model is an open area to explore.

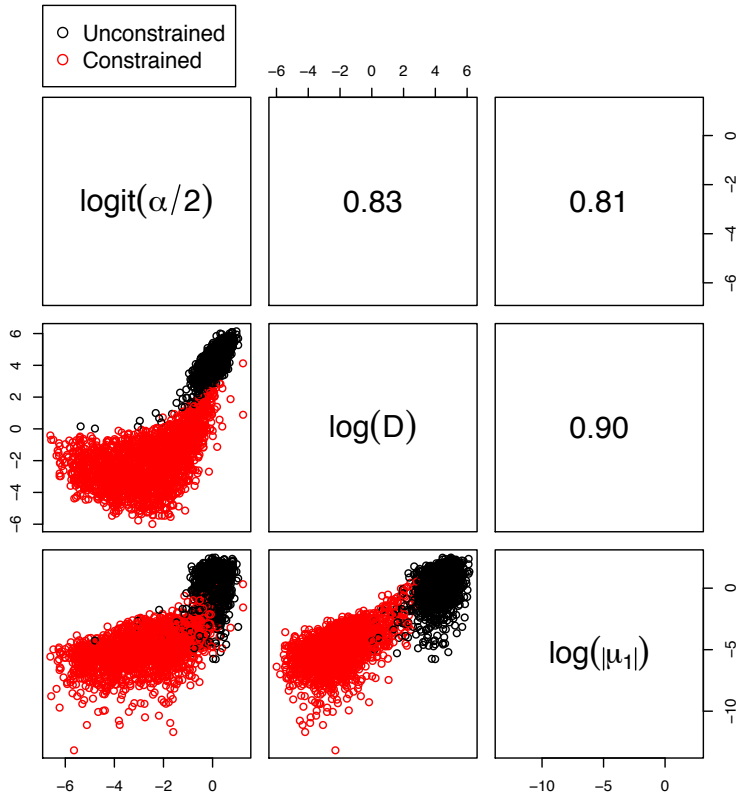

Figure 5.3: Joint distribution view of three of the locaiton-scale parameters for all trajectories across the five antibody concentrations. Upper diagonal values correspond to the Pearson correlation between each parameter pair. Each trajectory is colour-labelled as constrained or unconstrained.

Furthermore, analysis of MSD behaviour offers insight into the differences between constrained and unconstrained motion. Following the location-scale model fit and subsequently fitting a hierarchical model to each experimental environment, population-level MSD values were explored for three antibody concentrations (0, 100, and 1000 mg/L). Figure 5.4 displays these MSD values over time.

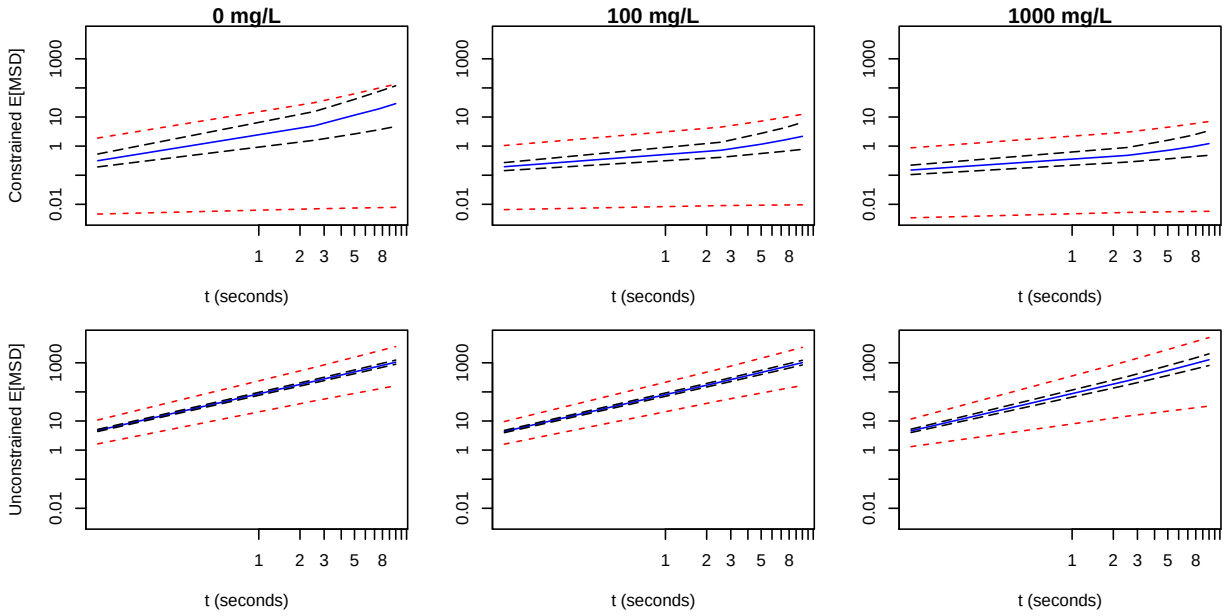

Figure 5.4: Significant differences in population-level MSD (blue line) are seen at low and high antibody concentrations for constrained HSV trajectories (top row). Unconstrained trajectories do not follow the strong decreasing pattern seen in constrained trajectories. The red lines indicate upper (97.5%) and lower (2.5%) quantiles for single-path MSDs. Upper and lower quantiles on the population-level MSD are indicated by the black dashed lines. The increase in interval width is expected as time increases where particles have time to explore a larger area.

Two conclusions can be immediately drawn from Figure 5.4, (1) there is a significant difference in MSD upon addition of antibody for constrained trajectories, and (2) there is sizeable heterogeneity present at all concentrations. The first conclusion stems from the downward moving line as the antibody concentration increases for constrained trajectories. The latter conclusion follows from the significant spread of the lower and upper intervals (red dashed lines) at all time steps.

This section has highlighted two categories of particle motion - constrained and unconstrained motion. While these two categories of motion must be subjectively labelled, the analysis presented in this section offers evidence on why and how these labels are determined. The remaining analysis in this section will focus solely on constrained trajectories for two reasons. Firstly, constrained trajectories indicate that the antibody-virus interaction is likely present as discussed in Section 1.2. Since a main application of this work is for improved vaccine design, analyzing constrained trajectories allows

one to focus on particles that have likely interacted with the added antibody solution. Secondly, constrained trajectories exhibit clear patterns as antibody concentration increases - a phenomenon that has potential to provide important biological insights. While unconstrained trajectories are still important, they do not exhibit significant differences as the antibody concentration increases. Thus, constrained trajectories will remain the focus moving forward.

5.2 Posterior Analysis

One method to explore differences across antibody concentrations is through population-level MSD ratios. Given two separate SPT experiments (e.g. low and high antibody concentration), one can train the posterior distribution $p(\boldsymbol{\lambda}, \boldsymbol{\Sigma} \mid \boldsymbol{X}_{\text{tr}})$ and subsequently analyze population-level MSD ratios from experiments i and j, that is τ_i/τ_j with $\tau_i = E[D \cdot t_0^{\alpha} \mid \boldsymbol{X}_{\text{tr}}^{(i)}]$, where $\boldsymbol{X}_{\text{tr}}^{(i)}$ corresponds to the SPT data from the i^{th} experimental environment (i.e. at a specific antibody concentration). Here t_0 is set to 2 seconds and $i, j \in \{1, 2, 3, 4, 5\}$ correspond to antibody concentrations 0, 33, 100, 333, and 1000 mg/L, respectively. Table 5.2 displays summary statistics for each cross-comparison of the 5 antibody concentrations for constrained trajectories. The diagonal entires are for of each individual τ_i after sampling from the hierarchical model posterior. The off-diagonal entires correspond to simulated MSD ratios, i.e. τ_i/τ_j. The numbers in brackets correspond to the lower (2.5%) and upper (97.5%) quantiles.

Table 5.2: MSD Ratios for Constrained Trajectories

$p(\tau_i \ / \ \tau_j)$	0 mg/L	33 mg/L	100 mg/L	333 mg/L	1000 mg/L
0 mg/L	3.85 (1.92,7.59)	3.27 (1.46,6.77)	6.27 (2.94,12.80)	9.81 (4.78,19.68)	9.14 (4.33,18.44)
33 mg/L		1.21 (0.87,1.71)	1.98 (1.28,2.97)	3.09 (2.07,4.53)	2.88 (1.91,4.28)
100 mg/L			0.62 (0.49,0.81)	1.59 (1.15,2.18)	1.48 (1.05,2.06)
333 mg/L				0.4 (0.33,0.48)	0.94 (0.70,1.24)
1000 mg/L					0.43 (0.34,0.53)

There are large MSD ratios when comparing low and high antibody concentration experiments, however comparing the two largest antibody concentration environments (333 and 1000 mg/L) the values are centered close to 1. This indicates that increasing the antibody concentration past 333 mg/L does not cause any further reduction in population-level MSD values. This finding may have important cost savings significance in practice since it may suggest using more than a 333 mg/L antibody concentration is unneeded since there appears to be a saturation effect beyond this concentration.

5.3 Sample Size Considerations

Briefly discussed in the introduction, the literature on the required number of particle trajectories to acquire for a particular SPT experimental hypothesis is scarce. Often, the number of recorded SPT trajectories are a function of the equipment cost and experiment time, with thought given to establishing scientific relevance only after trajectory data-collection. The presented SSD algorithms have addressed this gap and offer a computationally efficient method to accept or reject scientific hypotheses based on population-level MSD. Figure 5.5 gives a simplified schematic to illustrate the methods discussed in this work.

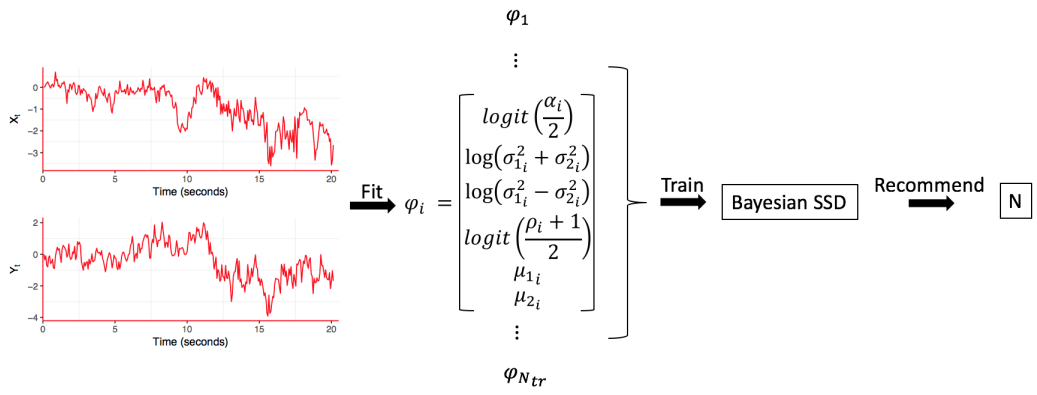

Figure 5.5: First obtain N_{tr} 2-dimensional single-particle trajectories, calculate single-path location-scale parameter estimates, subsequently feed this to train a multiple-path hierarchical model, followed by execution of the Bayesian SSD algorithm. Details such as experimental conditions and SPT data acquisition have been omitted for clarity.

In practice, the search for a recommended sample size N can be accomplished through a binary-search algorithm. The practitioner will input both a recommended sample size to initialize the search in addition to a maximum allowable number of samples (e.g. a value of which anything higher is experimentally impractical due to time or cost). The SSD algorithm can then commence at the recommended starting value, doubling or halving the sample size at each step until convergence or until the algorithm states N greater than the maximum allowable sample size is needed. Through this binary-search, finding the require sample size for the CV and ZV methods are feasible in an SPT experimental setting due to their significantly lower computational cost compared to the full and BV methods.

To illustrate use of the SSD procedure, the constrained trajectories for each of the five antibody concentration environments are analyzed using the CV algorithm. Figure 5.6 displays results for testing each of the five antibody experiments for scientific relevance, where $\mathcal{S}_0 = [\tau_0, \infty)$ uses $\tau_0 = 0.8 \cdot \tau_{ref}$, where τ_{ref} is the population-level MSD of the listed reference experiment in each row. In other words, each row in Figure 5.6 sets τ_0 to 20% less than the reference experiment indicated for that row. Hence, scientific relevance is defined in this experimental study as at least a 20% reduction in population-level MSD. This exploratory analysis plots $\hat{p}_0(\boldsymbol{X}_{\text{tr}}, \boldsymbol{X}_{\text{ts}})$ to emphasize whether an experiment is scientifically relevant or scientifically irrelevant. In Figure 5.6, box plots close to 0 indicate high confidence of scientific relevance, while box plots close to 1 indicate high confidence of scientific irrelevance, i.e. scientifically meaningless.

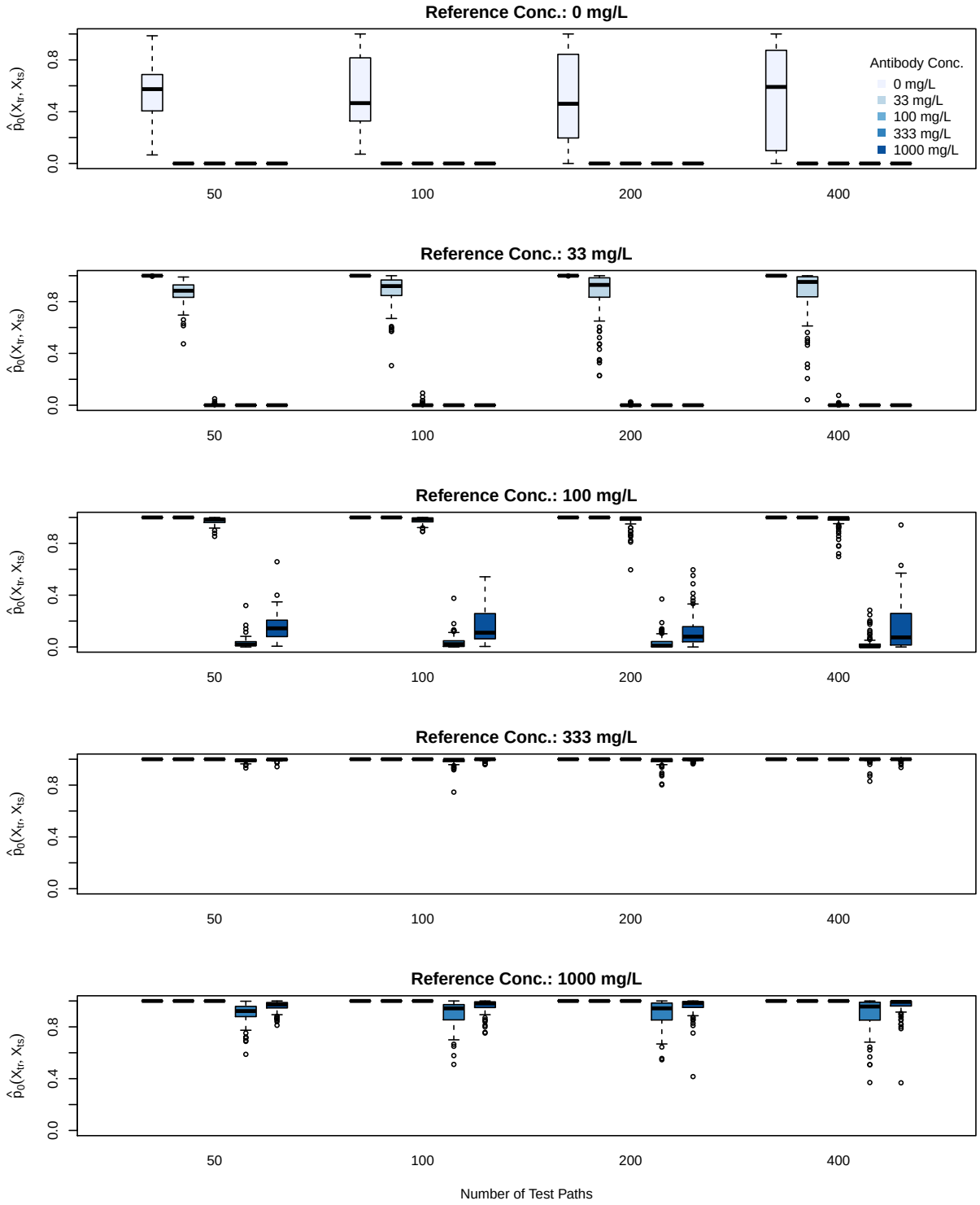

Figure 5.6: $\hat{p}_0(\boldsymbol{X}_{\text{tr}}, \boldsymbol{X}_{\text{ts}})$ for the CV algorithm as N_{ts} increase for different $\mathcal{S}_0$. Scientific relevance is defined here as at least a 20% reduction in population-level MSD compared to the listed reference experiment population-level MSD. Box plots close to 0 indicate high confidence of scientific relevance, while box plots close to 1 indicate high confidence of scientific irrelevance, i.e. scientifically meaningless.

Importantly, $\hat{p}_0(\boldsymbol{X}_{\text{tr}}, \boldsymbol{X}_{\text{ts}})$ is specifically used in this exploratory analysis in Figure 5.6 as it provides information on whether an experiment is scientifically meaningful or meaningless. Observing the first row of box plots in Figure 5.6, all experimental environments except the 0 mg/L confidently have a population-level MSD that is 20% less than the reference population-level MSD. Results for the second row indicate that the 100, 333, and 1000 mg/L experiments are scientifically relevant (i.e. have at least a 20% reduction their population-level MSD compared to 33 mg/L reference experiment). Moreover, the 0 mg/L box plots in the second row are all near 1 indicating that the 0 mg/L experiment is scientifically meaningless in this case. This is expected since scientific relevance is defined as at least a 20% decrease in population-level MSD, and 0 mg/L has a considerably higher population-level MSD compared to the 33 mg/L reference experiment. Continuing analysis in this manner, experimentalists can make decisions on whether or not experiments are scientifically meaningful based on how far $\hat{p}_0(\boldsymbol{X}_{\text{tr}}, \boldsymbol{X}_{\text{ts}})$ estimates are from $1/2$. Depending on the confidence the experimentalist requires for evaluation of scientific relevance, the number of test samples recommended by the SSD algorithm will vary. Naturally, requiring a greater proportion of $\hat{p}_0(\boldsymbol{X}_{\text{tr}}, \boldsymbol{X}_{\text{ts}})$ estimates close to 0 or 1 will result in a larger required sample size. SPT experiments often deal with hundreds or a few thousand single-particle trajectories, thus sample recommendations of this order are should be considered feasible in practice.

The analysis presented in this section offers experimentalists an actionable approach to analyze differences across SPT experiments in terms of scientific relevance based on population-level MSD. This analysis should prove extremely useful for application areas of SPT such as vaccine design, where experiments must be compared against control experiments to indicate the efficacy of a particular vaccine in hindering virus motion.

Chapter 6

Discussion

Sample size determination plays a key role in developing cost-effective SPT experiments and accelerating vaccine design. This work has developed a two-stage modelling process, followed by a Bayesian SSD algorithm for assessing scientific relevance of an SPT experiment based on population-level MSD. A computationally intensive SSD algorithm was first presented. Since SPT experiments occur in real-time on the order of minutes to hours, three simplified algorithms were offered to considerably accelerate computation while retaining very similar statistical results as the full algorithm. The latter of the two simplified algorithms, CV and ZV, are recommended for use in practice as their computations can be completed on the order of minutes.

Through a careful simulation study, effects of heterogeneity were explored in relation to the performance of the Bayesian SSD algorithm. Higher levels of heterogeneity naturally lead to higher sample sizes to achieve a given level of confidence regarding scientific relevance. Experimentalists have the ability to adjust the level of confidence desired in practice, a desired flexibility built-in to the presented Bayesian SSD framework.

A detailed analysis of HSV 2-dimensional particle trajectories across five antibody concentration levels provided numerous insights, both statistical and biological. From

location-scale parameter analysis, a distinct trend toward increasingly confined motion was seen as the antibody concentration increased. This result follows intuition as high quantities of antibody should increase the chance a virus interacts with antibody, where the virus motion becomes hindered.

MSD ratios were calculated for each cross-pair of the five experiment environments. Low antibody concentrations were shown to have large differences among MSD values, while higher antibody concentrations (333 mg/L and 1000 mg/L) gave similar population-level MSD values (MSD ratios were near 1). This suggests that there may be a saturation point at which further increasing the antibody concentration has negligible effects. This finding has potential to offer experimentalists important cost savings as they can objectively determine the maximum concentration of antibody needed in an experiment. Research into whether this occurs for viruses other than HSV requires exploration.

Overall, the algorithms presented in this work offer practitioners the ability to significantly accelerate SPT analysis. Through statistically modelling scientific relevance, experiments can be designed to use the minimum number of resources to asses the scientific hypothesis under study. While this work focused on scientific hypotheses involving population-level MSD, the SSD algorithms can readily extend to incorporate more complex scientific relevance criterions such as first-passage times.

Finally, as real-time SPT experiments continue to proliferate, the need for fast and efficient SSD algorithms will only increase. The ability to concurrently run real-time SSD and SPT experiments offers enormous practical opportunity. A promising future direction is to employ the SSD algorithms developed in this work in a real-time experimental system, with the goal of increasing laboratory cost-efficiency and bolstering scientific output.

References

Allen, L. J. (2010), *An Introduction to Stochastic Processes with Applications to Biology*, CRC Press.

Babcock, H. P., Chen, C., and Zhuang, X. (2004), "Using single-particle tracking to study nuclear trafficking of viral genes," in *Biophysical Journal*, 87 (4), 2749–2758.

Brandenburg, B., and Zhuang, X. (2007), "Virus trafficking–learning from single-virus tracking," in *Nature Reviews Microbiology*, 5 (3), 197–208.

Brockwell, P. J., and Davis, R. A. (2009), *Time Series: Theory and Methods*, Springer.

Bronstein, I., Israel, Y., Kepten, E., Mai, S., Shav-Tal, Y., Barkai, E., and Garini, Y. (2009), "Transient Anomalous Diffusion of Telomeres in the Nucleus of Mammalian Cells," in *Physical Review Letters*, 103 (1), 018102.

Cherstvy, A. G., Chechkin, A. V., and Metzler, R. (2013), "Anomalous Diffusion and Ergodicity Breaking in Heterogeneous Diffusion Processes," in *New Journal of Physics*, 15 (8), 083039.

Clark, J. S. (2005), "Why Environmental Scientists are Becoming Bayesians," in *Ecology Letters*, 8 (1), 2–14.

Cressie, N., Calder, C. A., Clark, J. S., Hoef, J. M. V., and Wikle, C. K. (2009), "Accounting for uncertainty in ecological analysis: the strengths and limitations of hierarchical statistical modeling," in *Ecological Applications*, 19 (3), 553–570.

Efron, B. (1992), "Bootstrap methods: another look at the jackknife," in *Breakthroughs in statistics*, Springer, pp. 569–593.

Ewers, H., Smith, A. E., Sbalzarini, I. F., Lilie, H., Koumoutsakos, P., and Helenius, A. (2005), "Single-particle tracking of murine polyoma virus-like particles on live cells and artificial membranes," in *Proceedings of the National Academy of Sciences of the United States of America*, 102 (42), 15110–15115.

Gottlieb, S. L., and Johnston, C. (2017), "Future Prospects for New Vaccines Against Sexually Transmitted Infections," in *Current Opinion in Infectious Diseases*, 30 (1), 77.

Huang, Q. (2010), "Physics-driven Bayesian hierarchical modeling of the nanowire growth process at each scale," in *IIE transactions*, 43 (1), 1–11.

Huelsenbeck, J. P., Ronquist, F., Nielsen, R., and Bollback, J. P. (2001), "Bayesian Inference of Phylogeny and its Impact on Evolutionary Biology," in *Science*, 294 (5550), 2310–2314.

Jaqaman, K., Loerke, D., Mettlen, M., Kuwata, H., Grinstein, S., Schmid, S. L., and Danuser, G. (2008), "Robust single-particle tracking in live-cell time-lapse sequences," in *Nature methods*, 5 (8), 695–702.

Jeon, J.-H., Tejedor, V., Burov, S., Barkai, E., Selhuber-Unkel, C., Berg-Sørensen, K., Oddershede, L., and Metzler, R. (2011), "In vivo Anomalous Diffusion and Weak Ergodicity Breaking of Lipid Granules," in *Physical Review Letters*, 106 (4), 048103.

Kou, S. C. (2008), "Stochastic modeling in nanoscale biophysics: subdiffusion within proteins," in *The Annals of Applied Statistics*, 501–535.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012), "Imagenet Classification with Deep Convolutional Neural Networks," in *Advances in Neural Information Processing Systems*, pp. 1097–1105.

Li, F., and Zhang, N. R. (2010), "Bayesian Variable Selection in Structured High-dimensional Covariate Spaces with Applications in Genomics," in *Journal of the American Statistical Association*, 105 (491), 1202–1214.

Looker, K. J., Magaret, A. S., May, M. T., Turner, K. M., Vickerman, P., Gottlieb, S. L.,

and Newman, L. M. (2015), "Global and Regional Estimates of Prevalent and Incident Herpes Simplex Virus Type 1 infections in 2012," in *PloS one*, 10 (10), e0140765.

Lysy, M., Pillai, N. S., Hill, D. B., Forest, M. G., Mellnik, J. W., Vasquez, P. A., and McKinley, S. A. (2016), "Model comparison and assessment for single particle tracking in biological fluids," in *Journal of the American Statistical Association*, 111 (516), 1413–1426.

Mellnik, J. W., Lysy, M., Vasquez, P. A., Pillai, N. S., Hill, D. B., Cribb, J., McKinley, S. A., and Forest, M. G. (2016), "Maximum likelihood estimation for single particle, passive microrheology data with drift," in *Journal of Rheology*, 60 (3), 379–392.

Metzler, R., and Klafter, J. (2000), "The random walk's guide to anomalous diffusion: a fractional dynamics approach," in *Physics reports*, 339 (1), 1–77.

Moerner, W. (2002), "A dozen years of single-molecule spectroscopy in physics, chemistry, and biophysics," .

Newby, J. M., Schaefer, A. M., Lee, P. T., Forest, M. G., and Lai, S. K. (2017), "Deep Neural Networks Automate detection for Tracking of Submicron Scale Particles in 2D and 3D," in *arXiv preprint arXiv:1704.03009*.

Nowak, M., and May, R. M. (2000), *Virus Dynamics: Mathematical Principles of Immunology and Virology: Mathematical Principles of Immunology and Virology*, Oxford University Press, UK.

Qian, H., and Kou, S. C. (2014), "Statistics and related topics in single-molecule biophysics," in *Annual review of statistics and its application*, 1, 465–492.

Qian, P. Z., and Wu, C. J. (2008), "Bayesian hierarchical modeling for integrating low-accuracy and high-accuracy experiments," in *Technometrics*, 50 (2), 192–204.

Royle, J. A., and Dorazio, R. M. (2008), *Hierarchical modeling and inference in ecology: the analysis of data from populations, metapopulations and communities*, Academic Press.

Saxton, M. J., and Jacobson, K. (1997), "Single-particle tracking: applications to mem-

brane dynamics," in *Annual review of biophysics and biomolecular structure*, 26 (1), 373–399.

Seisenberger, G., Ried, M. U., Endress, T., Büning, H., Hallek, M., and Bräuchle, C. (2001), "Real-time single-molecule imaging of the infection pathway of an adeno-associated virus," in *Science*, 294 (5548), 1929–1932.

Solomon, T., Weeks, E. R., and Swinney, H. L. (1993), "Observation of Anomalous Diffusion and Lévy Flights in a Two-Dimensional Rotating Flow," in *Physical Review Letters*, 71 (24), 3975.

Wang, N., and Yeung, D.-Y. (2013), "Learning a Deep Compact Image Representation for Visual Tracking," in *Advances in Neural Information Processing Systems*, pp. 809–817.

Xie, X. S., and Lu, H. P. (1999), "Single-molecule enzymology," in *Journal of Biological Chemistry*, 274 (23), 15967–15970.

Yau, C., Papaspiliopoulos, O., Roberts, G. O., and Holmes, C. (2011), "Bayesian Non-parametric Hidden Markov Models with Applications in Genomics," in *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73 (1), 37–57.

Zhu, Y., Ouyang, Q., and Mao, Y. (2017), "A Deep Convolutional Neural Network Approach to Single-Particle Recognition in Cryo-electron Microscopy," in *BMC Bioinformatics*, 18 (1), 348.

Appendices

A Derivation of Profile Likelihood

Consider the following model,

$$\mathbf{x} \sim \mathcal{MN}_{N\times 2}(\mathbf{\Delta t}\boldsymbol{\mu}, \mathbf{V}_\alpha, \mathbf{\Psi}) \iff vec(\boldsymbol{x}) \sim \mathcal{N}_{2N}(vec(\mathbf{\Delta t}\boldsymbol{\mu}), \boldsymbol{V}_\alpha \otimes \mathbf{\Psi}).$$

where $vec(\boldsymbol{x})$ denotes the vectorization of $\boldsymbol{x}$, and $\otimes$ denotes the Kronecker product. Following the supplementary material of Lysy et al. (2016), the log-likelihood that follows from this regression model is,

$$\ell(\alpha, \boldsymbol{\mu}, \mathbf{\Psi}|\boldsymbol{x}) = -\frac{1}{2}\Big\{ tr\Big[\mathbf{\Psi}^{-1}(\mathbf{x} - \mathbf{\Delta t}\boldsymbol{\mu})\mathbf{V}_\alpha^{-1}(\mathbf{x} - \mathbf{\Delta t}\boldsymbol{\mu})\Big] + N\log|\mathbf{\Psi}| + 2\log|\mathbf{V}_\alpha|\Big\}.$$

Through setting

$$\mathbf{S}_\alpha = (\boldsymbol{x} - \mathbf{\Delta t}\hat{\boldsymbol{\mu}}_\alpha)\mathbf{V}_\alpha^{-1}(\boldsymbol{x} - \mathbf{\Delta t}\hat{\boldsymbol{\mu}}_\alpha)\ , \qquad \hat{\boldsymbol{\mu}}_\alpha = (\mathbf{\Delta t}^T\mathbf{V}_\alpha^{-1}\mathbf{\Delta t})^{-1}\mathbf{\Delta t}^T\mathbf{V}_\alpha^{-1}\boldsymbol{x},$$

the likelihood $\ell(\alpha, \boldsymbol{\mu}, \mathbf{\Psi}|\boldsymbol{x})$ at a fixed α is maximized at $\hat{\boldsymbol{\mu}}(\alpha) = \hat{\boldsymbol{\mu}}_\alpha$ and $\hat{\mathbf{\Psi}}(\alpha) = \frac{1}{N}\boldsymbol{S}_\alpha$.

Upon substituting the conditional MLEs for $\boldsymbol{\mu}$ and $\mathbf{\Psi}$ into the full likelihood function, we arrive at the so-called profile likelihood

$$\ell_{prof}(\alpha\,|\,\boldsymbol{x}) = \ell(\alpha, \boldsymbol{\mu} = \hat{\boldsymbol{\mu}}(\alpha), \mathbf{\Psi} = \hat{\mathbf{\Psi}}(\alpha)\,|\,\boldsymbol{x}) = -\frac{1}{2}\left(2N + N\log\left(\frac{S_\alpha}{N}\right) + 2\log|\boldsymbol{V}_\alpha|\right)$$

B Gibbs Sampling from a Hierarchical Model

Consider the "normal-normal" hierarchical model

$$y_i \,|\, \mu_i, V_i \overset{\text{ind}}{\sim} \mathcal{N}_q(\mu_i, V_i), \qquad \mu_i \overset{\text{iid}}{\sim} \mathcal{N}_q(x_i'\beta, \Sigma),$$

where the V_i and x_i are known matrices and vectors of size $q \times q$ and $p \times 1$, and μ_i are unknown random effects, and $\beta_{p\times q}$ and $\Sigma_{q\times q}$ are the model parameters under study. Setting $Y = (y_1, \ldots, y_n)$, $X = (x_1, \ldots, x_n)$ and $\mu = (\mu_1, \ldots, \mu_n)$, we have

$$\mu_i \,|\, Y, \beta, \Sigma \overset{\text{ind}}{\sim} \mathcal{N}_q\Big(B_i(x_i'\beta - y_i) + y_i, (I - B_i)V_i\Big),$$

where $B_i = V_i(V_i + \Sigma)^{-1}$. However, note that

$$\ell(\beta, \Sigma \,|\, Y, \mu) = \ell(\beta, \Sigma \,|\, \mu),$$

such that conditioned on everything else, we can consider the likelihood for β and Σ as being equivalent to the likelihood for

$$\mu = X\beta + [\epsilon_{ij}]_{n\times q}, \quad (\epsilon_{i1}, \ldots, \epsilon_{iq}) \overset{\text{iid}}{\sim} \mathcal{N}_q(0, \Sigma),$$

which is precisely the multivariable regression model. Thus, changing form to the commonly seen regression framework $Y \leftarrow \mu$,

$$Y = X\beta + E, \quad (\epsilon_{i1}, \ldots, \epsilon_{iq}) \overset{\text{iid}}{\sim} \mathcal{N}_q(0, \Sigma),$$

The conjugate prior for this model is $\beta \sim \text{MNIW}_{p,q}(\Lambda, \Omega, \Psi, \nu)$, i.e.,

$$\begin{aligned} \Sigma &\sim \text{Inv-}\mathcal{W}_q(\Psi, \nu) \\ \beta \,|\, \Sigma &\sim \mathcal{MN}_{p,q}\left\{\Lambda, \Omega^{-1}, \Sigma\right\}. \end{aligned}$$

For this choice of prior, the posterior distribution is

$$\begin{aligned}\Sigma \mid Y, X &\sim \text{Inv-}\mathcal{W}_q(\Psi + S + C, \nu + n) \\ \beta \mid \Sigma, Y, X &\sim \mathcal{MN}_{p,q}\left\{\hat{\Lambda}, (X'X + \Omega)^{-1}, \Sigma\right\},\end{aligned}$$

where

$$\begin{aligned}\hat{\beta} &= (X'X)^{-1}X'Y \\ S &= (Y - X\hat{\beta})'(Y - X\hat{\beta}) = Y'(I - H)Y \\ \hat{\Lambda} &= (X'X + \Omega)^{-1}\Omega(\Lambda - \hat{\beta}) + \hat{\beta} \\ &= (X'X + \Omega)^{-1}(X'Y + \Omega\Lambda) \\ C &= \hat{\beta}'(X'X)\hat{\beta} + \Lambda'\Omega\Lambda - (X'X\hat{\beta} + \Omega\Lambda)'(X'X + \Omega)^{-1}(X'X\hat{\beta} + \Omega\Lambda) \\ &= Y'HY + \Lambda'\Omega\Lambda - (X'Y + \Omega\Lambda)'(X'X + \Omega)^{-1}(X'Y + \Omega\Lambda)\end{aligned}$$

Thus, the Gibbs Sampler alternates between conditional draws of β, Σ, and μ.

C MCMC Convergence

Gibbs sampling is an iterative algorithm. Thus, checking convergence is necessary to ensure the posterior samples are meaningful. Figure C.1 displays 10000 iterations for each of the six elements of $\boldsymbol{\lambda}$. While not shown here, excellent convergence is also achieved for the variance matrix $\boldsymbol{\Sigma}$ components.

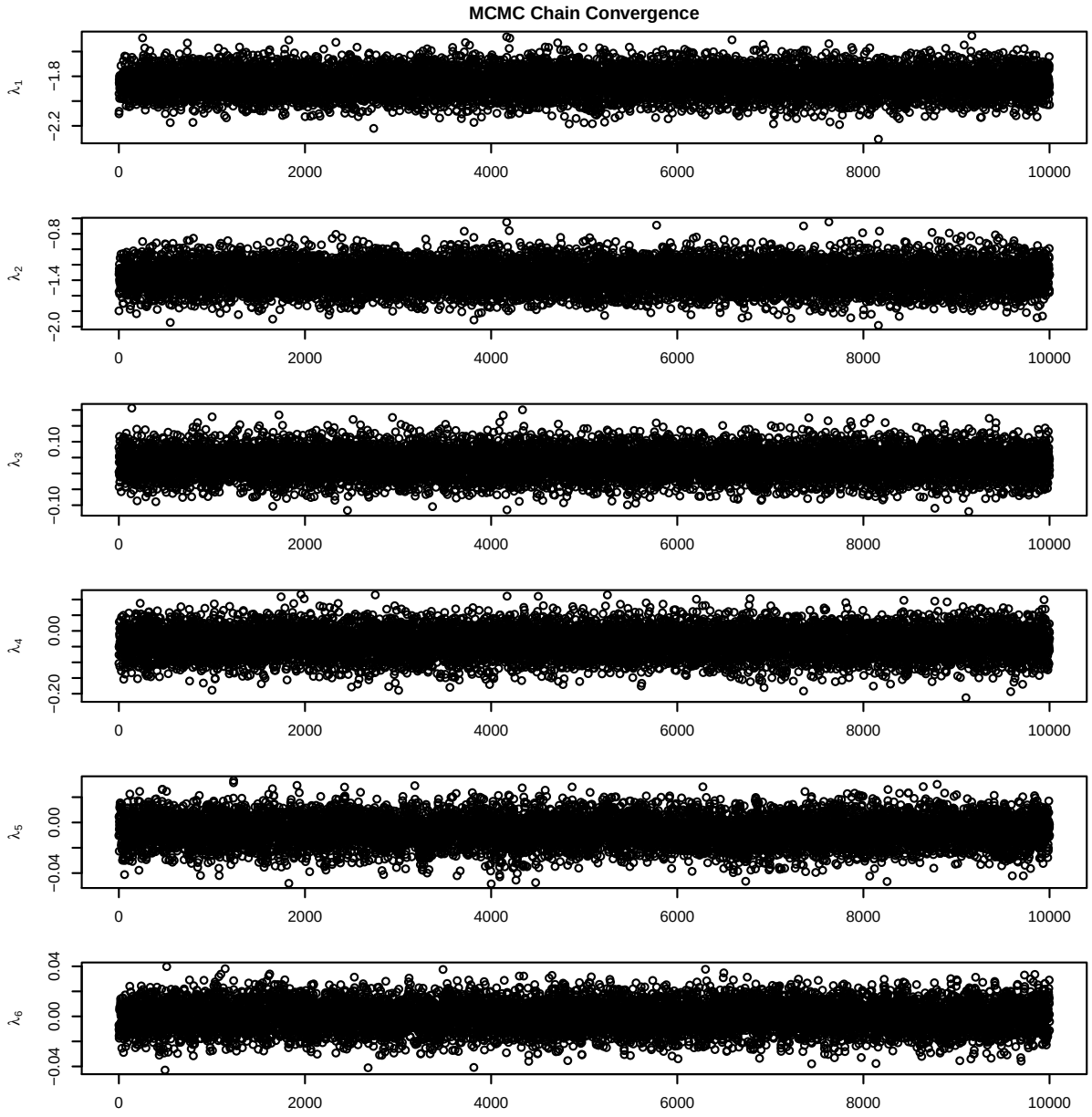

Figure C.1: Each element of $\boldsymbol{\lambda}$ plotted throughout sampling. There appears to be no convergence issues (not plotting burn-in), hence these posterior results are stable to use in simulation.