MICRORHEOLOGY AND HETEROGENEITY IN BIOLOGICAL FLUIDS: APPROACHES, MODELS AND APPLICATIONS

John William Ryckman Mellnik

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Curriculum for Bioinformatics and Computational Biology.

Chapel Hill
2015

Approved by:

M. Gregory Forest

Samuel K. Lai

Julie B. Dumond

Sean M. Gomez

Tim C. Elston

ABSTRACT

John William Ryckman Mellnik: Microrheology and Heterogeneity in Biological Fluids: Approaches, Models and Applications
(Under the direction of M. Gregory Forest)

Fluids play an important role in a wide range of biological processes. They facilitate cellular activities, protect us from infection and propagate nutrients throughout the body, to name a few. In each case, the properties of the fluid are finely tuned to the task at hand, and understanding those properties can afford a deeper understanding of the underlying biology. Furthermore, knowing how disease or environmental factors alter the properties of these fluids can provide a means to interpret, and forecast, downstream deleterious effects.

To this end, microrheology is an increasingly popular means of investigating biological fluids. This technique, whereby tracer particles are embedded in the fluid of interest and their diffusive movements are used to infer the viscous and elastic moduli of the surrounding fluid, offers insight into properties of the fluid at a spatial and temporal resolution unmatched by traditional macrorheology approaches.

Despite its benefits, the wider application of microrheology has been limited by the presence of two, frequently encountered, phenomena: the existence of an active driving force coupled to the stochastic movement of the tracer particles, and the presence of spatial, or temporal, heterogeneity in the fluid under investigation. This work proposes best practices for addressing each of these phenomena and demonstrates how they may be coupled to diffusion models to more accurately describe, and predict, the movement of micro- and nano-scale particles through biological fluids. We apply the methodology developed herein to the analysis of bronchoalveolar lavage fluid from a pediatric cystic fibrosis cohort as part of an ongoing effort to characterize pulmonary manifestations of the disease.

To my family,

past and present

William D. Ryckman, Jr.

1924 – 2003

Alexander Mellnik

1930 – 2003

ACKNOWLEDGEMENTS

This work represents a small facet of a much broader research effort, and would not have been possible without the collaborative spirit, and scientific selflessness, that I have encountered during my seven years at Carolina. It is to those individuals, male and female, whose names I do not know, who established this culture long before I came along, that I would like to offer thanks.

Few people have done more to uphold this culture than the individuals with whom I have been fortunate enough to work, and so, I would like to sincerely thank Paula Vasquez, who introduced me to this research, and David Hill, who introduced me to its applications. I would like to thank Martin Lysy, for his infectious enthusiasm, and Scott McKinley, for his thoughtful guidance. I am thankful for the opportunities afforded by my adviser, Greg Forest, who, despite me, has always demonstrated the most prolific patience and offered the most sage advice.

Early lessons often stick with us the longest, growing as we grow, and so I would like to thank Keith Hale and Mark Aspland, exemplar Corinthians.

Finally, I would like to thank my parents, always the lift below the layline and the leeway against the ebb.

CONFLICT OF INTEREST DISCLOSURE

John Mellnik, author of this dissertation, has equity ownership in, and serves on the board of directors of Path BioAnalytics, Inc., to which the following technologies evaluated in this paper have been licensed: methods, systems, and computer readable media for data analysis and inference of particle diffusion in target materials and target material simulants. Adviser M. Gregory Forest is also an inventor of this technology and Scientific Founder of Path BioAnalytics, Inc.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS AND SYMBOLS

aBm	Amnesic Brownian motion
API	Active pharmaceutical ingredient
BAL	Bronchoalveolar lavage
D	Diffusivity
fBm	Fractional Brownian motion
FPT	First passage time
GSER	Generalized Stokes-Einstein relation
H	Hurst parameter
HBE	Human bronchial epithelial (mucus)
iMSD	Individual mean squared displacement
mBm	Multifractional Brownian motion
MSD	Mean squared displacement
PPTM	Passive particle tracking microrheology
PC(A)	Principal component (analysis)
RL-fBm	Riemann-Liouville fractional Brownian motion
sBm	Standard Brownian motion
W-fBm	Weyl fractional Brownian motion

CHAPTER 1
Introduction

1.1 Rheology

Fluids are ubiquitous. Their physical properties impact a wide range of processes, from the quotidian (mixing cake batter), to the essential (preventing infections). Rheology is the study of the physical properties of fluids and how those properties respond to applied stress. Newtonian fluids, such as water, exhibit a purely viscous response to stress, dissipating energy, while solids exhibit an elastic response, storing energy. Between these two extremes, there are many synthetic and biological fluids that exhibit properties of both solids and liquids. These are referred to as viscoelastic fluids.

Often, the relative viscous and elastic response to an applied stress is dependent upon the frequency (ω) with which the stress is applied. For purely viscous fluids, the viscosity is the ratio of stress to strain upon deformation, while for viscoelastic fluids, the viscosity is replaced with the complex viscosity, η^*, or the complex shear modulus $G^*(s)$ that divides the frequency dependent stress-to-strain ratio into an in-phase storage component (G'), representing the elastic response, and an out-of-phase, loss component (G''), representing the viscous response. The complex viscosity is directly proportional to the complex shear modulus: $i\omega\eta^*(\omega) = G^*(\omega) = G'(\omega) + iG''(\omega)$. Fluids may exhibit shear-thickening or shear-thinning depending on which moduli dominates at the relevant frequencies. A mixture of cornstarch and water is a well known fluid that exhibits viscous properties at low frequency forcing and elastic properties at high frequency forcing.

In this chapter, we provide a brief presentation of the key differences between microrheology and macrorheology. We then present the foundational theory of Mason and Weitz [1] that gave rise to the modern field of microrheology. This chapter is concluded by a discussion of the motivations and objectives behind this research, and finally, a a brief guide to the subsequent chapters.

1.1.1 A brief introduction to micro and macrorheology

Rheology may be divided into two main categories. The first, macrorheology, concerns a fluid's response to an applied stress at the length scales of the fluid volume. A rheometer, consisting of two horizontal plates between which a volume of fluid is placed, is used to apply a controlled stress to the trapped fluid and measure the resulting strain. While protocols for bulk characterization of many fluids have been rigorously established, this approach has three main limitations. First, microliters of the fluid are often required to engender a response above the noise floor of the rheometer while experimental conditions may limit the volume of fluid available for characterization. Second, a rheometer's noise floor and the fluid's surface tension, which acts as an elastic force at low driving stress, limits the types of fluids and stress regimes that may be investigated. Third, the physical geometry of the rheometer naturally measures a mean response, spatially averaged over the area of contact between the fluid and the rheometer's plates. This removes any heterogeneities within the fluid, potentially masking highly discerning information.

Microrheology is an alternative to bulk rheology that is well suited to the characterization of fluids that may be available in limited volumes or exhibit heterogeneity. Furthermore, as we will demonstrate, microrheology may be used to analyze the viscoelastic response of fluids to stresses below the noise floor of most rheometers. Microrheology focuses on the movement of tracer particles embedded in the fluid of interest, informing the linear viscoelastic response of the material across a wide range of frequencies [1–5]. In purely viscous fluids, the viscosity is related to the diffusivity of the embedded tracer particle through the Stokes-Einstein equation,

$$D = \frac{k_B T}{6\pi\eta r}, \tag{1.1}$$

where k_B is the Boltzman's constant, T is temperature and r is the radius of the embedded particle. This equation encompasses two key ideas, that the thermally activated motion of the particle is related to its mobility (the Einstein component), and that the mobility is related to the fluid's viscosity (the Stokes component)[6]. Application of the Stokes-Einstein relation was limited to purely viscous fluids until the generalized Stokes-Einstein relation (GSER) was derived in 1995 by Mason and Weitz [1, 6]. This generalization, allowing for the decomposition of a particle's diffusive movements into viscous

and elastic components, provides the foundation for all modern microrheological investigations. In the following section we present Mason and Weitz's derivation of the GSER.

1.1.2 Generalized Stokes-Einstein Relation

Diffusion refers to the stochastic movement of solute through a fluid and arises from the bombardment of the particles composing the solute by the molecules of the surrounding fluid. The Stokes-Einstein relation (1.1) formally relates the stochastic motion of an individual particle, represented by the diffusivity, D, to the viscosity η of the local fluid. We may think of viscosity as a dampening force, dissipating energy and decreasing the stochastic fluctuations of the particle. Conversely, an increase in the innate energy of the fluid, represented by k_BT, serves to increase the stochastic fluctuations. Following Newton's second law, the thermal and drag forces acting on a particle must balance, i.e.,

$$m\dot{V}(t) = F_{\text{drag}}(t) + F_{\text{thermal}}(t). \tag{1.2}$$

In their generalization of the Stokes-Einstein relation, Mason and Weitz begin with the generalized Langevin equation that presents $F_{\text{drag}}(t)$ in terms of a damping function with memory kernel $\zeta(t)$, and F_{thermal} as a stochastic term representing the random bombardment of the particle with the surrounding molecules [7, 8]:

$$m\dot{V}(t) = f_R(t) - \int_0^t \zeta(t-u)V(u)du. \tag{1.3}$$

where m and V are the probe mass and velocity, respectively. Stokes' law relates the viscosity of the the fluid to the memory kernel,

$$\zeta = 6\pi\eta r. \tag{1.4}$$

Letting $\mathcal{L}\{\text{-}\}$ be the standard Laplace transform such that

$$\tilde{f}(s) = \mathcal{L}\{f(t)\} = \int_0^\infty e^{-st}f(t)dt, \tag{1.5}$$

we may take the Laplace transform of (1.3) and solve for the velocity:

$$\tilde{V}(s) = \frac{mV(0) + \tilde{f}_R(s)}{ms + \tilde{\zeta}(s)} \tag{1.6}$$

$$\implies \tilde{V}(s)V(0) = V(0)\frac{mV(0) + \tilde{f}_R(s)}{ms + \tilde{\zeta}(s)}. \tag{1.7}$$

Taking the ensemble average, $\langle$-$\rangle$, and assuming we are in the zero-mass limit, we find,

$$\langle \tilde{V}(s)V(0)\rangle = \frac{k_B T}{\tilde{\zeta}(s)}. \tag{1.8}$$

In practice, the increment process is typically measured, not the velocity process. The time domain mean squared increments, $\langle x^2(t)\rangle$ are directly related to the velocity autocorrelation function through the Laplace transform,

$$\langle \tilde{V}(s)V(0)\rangle = \frac{s^2}{2}\mathcal{L}\langle x^2(s)\rangle := \frac{s^2}{2}\langle \tilde{x}^2(t)\rangle. \tag{1.9}$$

Thus, by substituting into (1.8), the memory kernel may be directly related to the observed increments,

$$\tilde{\zeta}(s) = \frac{2k_B T}{s^2\langle \tilde{x}^2(s)\rangle}, \tag{1.10}$$

which is in turn proportional to the complex viscosity of the fluid,

$$\tilde{\eta}(s) = \frac{\tilde{\zeta(s)}}{6\pi a}. \tag{1.11}$$

The complex shear modulus may then be related to the mean squared increments of the diffusing particle as,

$$\tilde{G}(s) = s\tilde{\eta}(s) = \frac{k_B T}{s\pi a\langle \tilde{x}^2(s)\rangle}. \tag{1.12}$$

In typical microrheology experiments, tracer particles are injected into the fluid of interest, video of particle movement is collected, and the position time series is extracted using tracking software. From the position time series, the mean squared displacements are calculated as a function of a lag

time and subsequently transformed according to (1.12) to estimate the viscous and elastic moduli of the fluid.

Both the generalized (1.12) and non-generalized (1.1) Stokes-Einstein relations are based on two key assumptions. First, the Einstein component relating the thermally activated motion of the probe particle to its mobility requires that the system is in equilibrium, i.e., the particle is not experiencing a net force. This net force may arise from convective flow within the fluid volume, active manipulation of the particle, either by internal or external forces, or induced by technical deficiencies such as drift of the microscope stage [6, 9, 10]. Second, the Stokes component of (1.12) and (1.1) require that the fluid under investigation is homogeneous, isotropic and incompressible. The homogeneity requirement is often invalidated in gels and synthetic polymer solutions [11–15], as well as various biofluids [16–19], however Squires and Mason [6] point out in their excellent review that the conditions under which the Stokes component fails are themselves important indicators of a material's properties and are worth investigating. Importantly, heterogeneity may arise from three distinct sources, temporal changes in the fluid or probe particle, spatial variations in fluid constituents and length-scale dependent structural variations within the fluid. Incongruities in any of these variable, time, location, or length-scale, have the potential to engender disparate estimations of a heterogeneous fluid's viscoelastic properties.

1.2 Motivations and objectives

Transmucosal drug delivery, the primary motivation of this work, is an increasingly active area of investigation, with relevance for a wide range of conditions. Every organ of the human body not covered by skin is coated with a viscoelastic fluid, known as the mucosal layer, that protects the underlying organ from pathogens and other foreign insults. Many medical and biological scenarios involve the diffusion of foreign bodies, such as micron-scale drug delivery particles, or infectious xenobiotic and viral agents, through the mucosal layer. Through its ubiquitous nature, the mucosal layer is forced to take on dual, and often contradictory roles, hindering diffusion of malevolent species, while allowing for the unfettered absorption of beneficial therapeutics. Although it is increasingly recognized that transmucosal delivery is a promising pathway for treating disease throughout the entire body [20–22], mucus presents a formidable barrier to the distribution of some therapeutics,

particularly micron-scale drug delivery particles and dry powder formulations that are large enough to interact with the elastic network of mucin proteins that make up mucus, and endow it with its viscoelastic properties.

The impact of the mucosal layer on the distribution of micron-scale drug delivery particles and dry powder formulations is a complex function of the compound's surface chemistry and size, as well as the physical properties of the mucosal layer itself. The mucosal layer dictates the movement of these compounds as they dissolve or decompose, significantly influencing the concentration time profile of the embedded active pharmaceutical ingredients (API). It is increasingly recognized that the diffusive properties of microscopic particles in viscoelastic fluids are not only anomalous, and poorly described by Brownian motion [23–28], thus invalidating a central assumption of most pharmacokinetic models, but they exhibit potentially biologically relevant inter-patient variability due to age, genetics, disease progression and environmental conditions [22, 23, 29, 30]. The design of therapeutics for controlled delivery of an API through the mucosal layer requires statistical and mathematical tools to analyze experimental data, perform model selection, fit model parameters, and forecast outcomes. No current methodology adequately address the unique and highly variability properties of the mucosal layer that can significantly influence diffusion, API uptake and subsequent bioavailability of therapeutic compounds. Here, we seek to provide a first step toward such a methodology.

The application of this research to pulmonary drug delivery is of particular interest because pulmonary drug delivery has been shown to lead to a reduction in side effects and faster drug onset times [31, 32]. Pulmonary drug delivery has been identified as a potentially superior method of drug delivery for a range of conditions such as chronic obstructive pulmonary disease, asthma and cystic fibrosis. Inhalation has also been proposed as a delivery route for vaccines, gene therapies, insulin treatments and cancer treatments [31]. The rigorous analysis of the diffusive movements of *specific* therapeutic compounds through the pulmonary mucosal layer of *specific* patients will facilitate personalized medical treatments for a wide range of conditions. The ability to perform these evaluations transiently holds great promise to couple patient characteristics and drug treatments, and to modify treatments in real time.

The use of microrheology techniques to guide therapeutic design and dosing protocols requires a rigorous understanding of both the innate viscoelastic properties of the fluid under investigation

and how those properties change depending upon the characteristics of the therapeutic compound. The GSER significantly advanced the field of microrheology beyond its initial applications and the seminal work of Mason and Weitz has been been adapted to the study of a wide variety of viscoelastic fluids across a range scientific disciplines [33–36]. However, as microrheology techniques are applied to increasingly complex fluids such as mucus, two hurdles must be addressed: the differentiation of thermally-activated motion from driven motion in the probe particle's trajectory, and the presence of spatial or temporal heterogeneity due to changes in the local physical properties of the fluid, or changes in the surface chemistry or physical size of the particle itself. Furthermore, to make the this methodology practical, more accurate models for particle diffusion must be developed, and a theoretical framework is required to predict particle passage times, and subsequent bioavailability, beyond experimentally-observable time scales.

1.3 Overview

In the remaining chapters, we propose best practices addressing the hurdles outlined above. Chapter 2 presents existing Brownian models for the diffusive movements of particles in homogeneous and heterogeneous, viscous and viscoelastic fluids, with a focus on simulation algorithms for each model. We conclude with a novel model based on fractional Brownian motion for diffusion in heterogeneous viscoelastic fluids and demonstrate how its simulation algorithm may be adapted to address the inverse problem of parameter estimation given a single realization of such a stochastic process.

In Chapter 3 we discuss approaches to diffusive parameter estimation when driven motion is coupled to a particle's thermally-activated movements. We demonstrate the impact of driven motion on traditional parameter estimation techniques, as well as the recovery of the storage and loss moduli, and viscosity of the fluid.

Chapter 4 addresses technical challenges inherent in the analysis of microparticle tracking data from heterogeneous fluids. An algorithm based on clustering techniques adapted from the Machine Learning community is presented for the quantification of fluid heterogeneity, under the assumption that each tracer particle has probed a homogeneous sub-region of the fluid. Application of the algorithm to simulated homogeneous and heterogeneous, viscous and viscoelastic data suggest that

the proposed algorithm is able to accurately and consistently quantify heterogeneity in complex fluids.

No explicit formula currently exists for the first passage time of a fractional Brownian process from the unit interval. In Chapter 5 we motivate a functional form for this quantity through analysis of the simulation algorithms discussed in Chapter 2. This analysis highlights an initially unintuitive feature of fractional Brownian motion–that subdiffusive processes travel faster over short distances than superdiffusive processes. These results emphasize the importance of accurately estimating diffusive parameter values and quantifying the spacial dimensions over which diffusion occurs.

In Chapter 6 we use the tools presented in the previous chapters to investigate and compare the viscoelastic properties of mucus harvested from primary human bronchial epithelial cell cultures and clinical mucus samples collected from a pediatric cystic fibrosis (CF) population via bronchoalveolar lavage. We illustrate how the results afforded by these tools may be used to estimate clinically relevant parameters critical to the understanding and treatment of early CF lung disease and other pulmonary pathologies.

CHAPTER 2

Brownian Diffusion Models and Simulation Algorithms

In this chapter we discuss several models for the diffusive movements of nano- and micro-scale particles. We begin with Brownian motion, a well-known model appropriate for homogeneous, purely viscous fluids, and progress toward models suitable for diffusion in heterogeneous, viscoelastic fluids. Simulation techniques for each model are detailed.

2.1 Introduction

Following its formal mathematical description by Einstein in 1905 [37], Brownian motion has become a foundational model, linking the mobility exhibited by a particle to the physical properties of the surrounding fluid via the Stokes-Einstein relation (1.1). However, similar to the breakdown of the Stokes-Einstein relation in the presence of both viscous and elastic components, requiring the GSER (1.12), a generalization of Brownian motion, or an alternative model, is required to adequately describe the movement of particles in viscoelastic fluids. To this end, several Brownian processes have been proposed to address diffusion in viscoelastic, and heterogeneous viscoelastic fluids. Developing appropriate descriptions for the diffusive movement of particles is essential for the accurate forecasting of desired statistics such as, in the context of transmucosal drug delivery, the concentration time profile of an inhaled formulation at the mucus-epithelial interface, or the time course of systemic exposure. The absence of closed-form diffusive transport equations applicable to heterogeneous viscoelastic fluids requires the development of efficient methods to generate diffusive trajectories of individual particles and the pursuit of the desired statistics through rigorous simulations.

Brownian processes can be divided into three categories, standard Brownian motion (sBm), fractional Brownian motion (fBm) and multifractional Brownian motion (mBm). fBm is the generalization of standard Brownian motion for diffusion in viscoelastic fluids. fBm may be divided into two

subcategories, Weyl (Section 2.3) and Riemann-Liouville (Section 2.4) based on the definitions of their respective fractional derivatives. In viscoelastic fluids, the elastic modulus introduces memory, i.e., correlation, in the probe particle's increment process. For fBm processes of both types, this memory is characterized by a unitless constant known as the Hurst parameter, $H \in [0, 1]$ where $H > 0.5$ and $H < 0.5$ correspond to positive correlation (superdiffusion) and negative correlation (subdiffusion) in the particle's increments. When $H = 0.5$, the increments are uncorrelated and sBm dynamics are recovered. mBm is the generalization of fBm for time-varying H caused by the movement of the probe particle between regions of a fluid with different elastic moduli (Section 2.5). In this chapter, we present strengths, weaknesses and simulation techniques for each model. We conclude by introducing amnesiac Brownian motion, an alternative model for diffusion in heterogeneous viscoelastic fluids (Section 2.6).

2.2 Standard Brownian motion

sBm, the most common Brownian model, was first formulated by Einstein based on his observations of pollen particles in water [37]. While sBm is a ubiquitous model for diffusion, it is best suited for "ideal" diffusion through homogeneous, purely viscous, i.e., non-viscoelastic, fluids. A sBm process, denoted $\{S(t)\}$, is a Gaussian process with stationary increments. The following properties also hold:

(i) $S(0) = 0$

(ii) $\mathbf{E}[S(t)] = 0$

(iii) $\mathbf{E}[S^2(t)] \sim t$

A continuous-time sBm process may be written as

$$S(t) = \int_0^t \xi(u)du, \tag{2.1}$$

where $\xi(u)$ is uncorrelated white noise. Since most experimental data is composed of discrete measurements of continuous-time processes, it is often advantageous to consider the discrete representation $S_n = S(n\Delta t)$ where Δt is the inter-observational time. The sBm increments $s_n = S_{n+1} - S_n$

are Gaussian and uncorrelated. sBm may be easily simulated by computing the cumulative sum of normally distributed iid numbers. A particle position process with inter-observational time Δt is

$$X_n = X(n\Delta t) = \sum_{i=0}^{n} \xi_i, \quad \xi \sim \mathcal{N}(0, \Delta t), \tag{2.2}$$

where $\mathcal{N}(0, \Delta t)$ is the normal distribution with mean zero and variance Δt.

Because the increments are uncorrelated, sBm is not well suited for modeling subdiffusive processes that exhibit correlation in the increment process. A common scenario engendering subdiffusion is the passive, thermally activated movement of particles in viscoelastic fluids where the elasticity of the fluid's gel-structure introduces correlation in the particle's movement. To model such phenomena for which sBm is poorly suited, we turn to fBm.

2.3 Weyl fractional Brownian motion

2.3.1 Theory

Weyl fractional Brownian motion (W-fBm) is a generalization of sBm based on the Weyl fractional integral and allows for correlation in the increment process. W-fBm was first formulated by Kolmogorov in 1940 [38] and later popularized by Mandelbrot and Van Ness [39]. A W-fBm continuous time position process $\{B(t)\}$ may be write as,

$$B_H(t) = \frac{1}{\Gamma(H + 1/2)} \left[\int_{-\infty}^{0} \left\{ (t-u)^{H-1/2} - (-u)^{H-1/2} \right\} \xi(u) du + \int_{0}^{t} (t-u)^{H-1/2} \xi(u) du \right]. \tag{2.3}$$

Similar to sBm, W-fBm is a Gaussian process and exhibits the following properties:

(i) $B_H(0) = 0$

(ii) $\mathbf{E}[B_H(t)] = 0$

(iii) $\mathbf{E}[B_H^2(t)] \sim t^{2H}$

(iv) $B_H(ct) \sim c^H B_H(t)$

where H is the Hurst Parameter. As noted earlier, W-fBm processes may exhibit negative correlation ($H < 0.5$), giving rise to sub-diffusion, or positive correlation ($H > 0.5$), giving rise to super-diffusion, i.e., persistent motion, and when $H = 0.5$ the increment process is uncorrelated and standard Brownian dynamics are recovered. From (2.3) we see that W-fBm may be thought of as an infinite history, represented by the integral from $-\infty$ to 0, combined with a finite moving average, represented by the integral from 0 to t. While the increment process resulting from only the later integral is non-stationary, the summation of both integrals leads to stationarity of the increments [40].

The correlation in the position process for W-fBm is given by

$$\text{ACF}_B(t_1, t_2) = \mathbf{E}[B_H(t_1)B_H(t_2)] = \frac{1}{2}\left(t_1^{2H} + t_2^{2H} - |t_2 - t_1|^{2H}\right) \tag{2.4}$$

and autocorrelation of the discrete increments $b_n = B_{n+1} - B_n$ is

$$\text{ACF}_b(n) = \text{cov}(b_k, b_{k+n}) = \frac{1}{2}\sigma^2\Delta t^{2H}\left(|n+1|^{2H} + |n-1|^{2H} - 2|n|^{2H}\right) \tag{2.5}$$

where the subscript H has been dropped for clarity.

W-fBm has been used to model processes as diverse as river flow [41–44] and network traffic [45, 46]. In a biological setting, it has been used to describe the movements of 1 μm diameter particles in human bronchial epithelial mucus [23] and intra-cellular biopolymers [47], among other applications. For particle diffusion, the following stochastic model for the particle's position process $X(t)$ has been proposed [25]:

$$X(t) = \mu t + \sigma B_H(t). \tag{2.6}$$

Here, $B_H(t)$ is a zero-mean Gaussian process specified by (2.3) and σ is a scaling factor that may be related to the viscosity of the fluid. μ is the "drift parameter" and takes into account systematic or driven motion. In chapter 3 we discuss factors that give rise to non-zero values of μ and approaches to, and complications regarding, parameter estimation when $\mu \neq 0$.

2.3.2 Simulation techniques

2.3.2.1 Direct algorithm

Direct algorithms for W-fBm generate sample paths by multiply a decomposition of the correlation matrix specified by (2.4) with vectors of normally distributed random numbers [48, 49]. This approach may be implemented using the square root decomposition of the correlation matrix, however it is more computationally efficient to consider the Cholesky decomposition. To generate a position process of length N, we first build the correlation matrix $\mathbf{\Lambda}$. Entry i, j of $\mathbf{\Lambda}$ is

$$\mathbf{\Lambda}_{i,j} = \frac{1}{2}\left(t_i^{2H} + t_j^{2H} - |t_i - t_j|^{2H}\right), \tag{2.7}$$

for $i, j = 1, ..., N$. Let $\mathbf{L}$ be the lower triangular matrix resulting from the Cholesky decomposition of $\mathbf{\Lambda}$ such that $\mathbf{L}\mathbf{L}' = \mathbf{\Lambda}$. The position process of a particle diffusing via W-fBm is then generated as $X = \sigma^{1/2}(\mathbf{L}\xi)$, where $\xi \overset{iid}{\sim} \mathcal{N}(0, \Delta t)$ is a $N \times 1$ vector. The matrix $\mathbf{L}$ serves to weight the white noise terms in order to produce a process X_t with the desired correlation structure. In general we have,

$$X_j = \sigma^{1/2}\left[\xi_1 \mathbf{L}_{1,j} + \xi_2 \mathbf{L}_{2,j} + ... + \xi_j \mathbf{L}_{j,j}\right]. \tag{2.8}$$

Figure 2.1 shows the Cholesky weights of paths of length $N = 500$ for three values of H at various points in the path.

The Direct W-fBm simulation algorithm is straight forward, however, in its current presentation, the length of the path must be specified a priori. This makes the algorithm poorly suited for adaptive implementations where the simulated path is extended until a stopping condition is met. Furthermore, Cholesky decomposition is computationally expensive since most algorithms exhibit $\mathcal{O}(N^3)$ complexity [50].

2.3.2.2 Frugal Cholesky updates

To address these limitations, the Direct algorithm may be adapted to an iterative framework by taking advantage of the fact that the correlation matrix for a path of length N is a subset of the correlation matrix for a path of length $N + 1$. Indeed, the construction of the $\widetilde{\mathbf{\Lambda}} \in \mathbb{R}^{(N+1)\times(N+1)}$

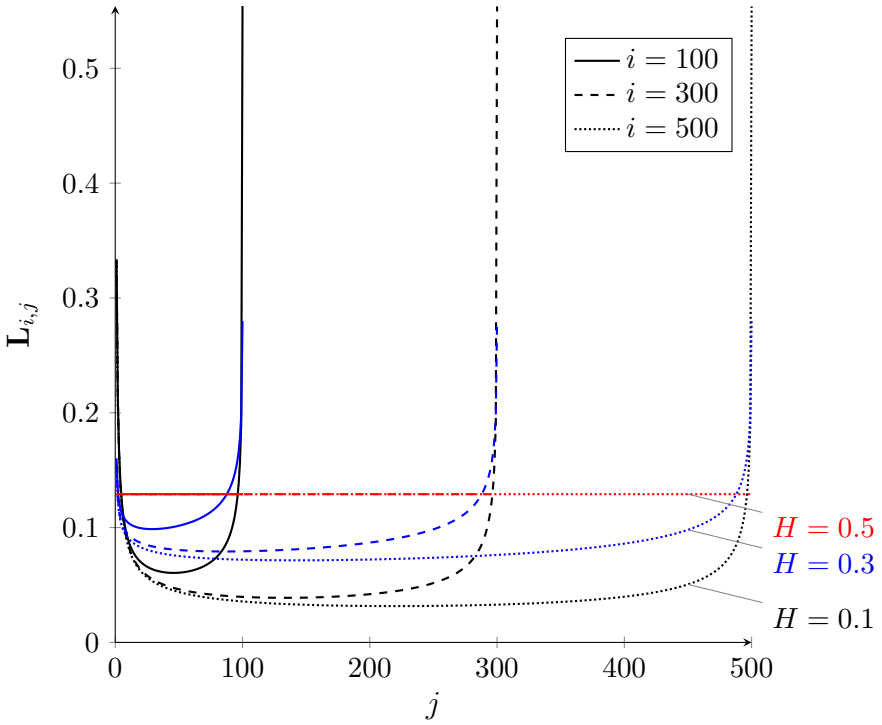


Figure 2.1: Cholesky weights of three $N = 500$ step paths ($\Delta t = 1/60$) based on the Weyl-fBm covariance (2.7). The Hurst parameter of the paths are $H = 0.1$ (black), $H = 0.3$ (blue) and $H = 0.5$ (red).The $i \leq j$ weights $\mathbf{L}_{i,j}$ used to construct X_j are shown for three points in the path, $j = 100$, $j = 300$ and $j = 500$. For the Brownian path ($H = 0.5$), all weights are equivalent.

given $\mathbf{\Lambda} \in \mathbb{R}^{N \times N}$ only requires the computation of one additional row. $\widetilde{\mathbf{L}}$, the resulting Cholesky decomposition of $\widetilde{\mathbf{\Lambda}}$, is a lower triangular matrix of size $(N+1) \times (N+1)$ such that $\widetilde{\mathbf{L}}\widetilde{\mathbf{L}}' = \widetilde{\mathbf{\Lambda}}$. Similar to the relationship between $\mathbf{\Lambda}$ and $\widetilde{\mathbf{\Lambda}}$, $\widetilde{\mathbf{L}}$ and $\mathbf{L}$ differ only in their last row, i.e. $\widetilde{\mathbf{L}}_{i,j} = \mathbf{L}_{i,j}$ for $i, j = 1, 2, ...N$. Therefore, calculating $\widetilde{\mathbf{L}}$, the new Cholesky matrix, based on $\mathbf{L}$, the old Cholesky matrix, and $\widetilde{\mathbf{\Lambda}}$, the correlation matrix at the current time step, is $\mathcal{O}(N)$ times more efficient than directly computing the full Cholesky decomposition of $\widetilde{\mathbf{\Lambda}}$. The elements in the N^{th} row of $\widetilde{\mathbf{L}}$ are given by

$$\widetilde{\mathbf{L}}_{N,j} = \frac{1}{\widetilde{\mathbf{L}}_{j,j}} \left[\widetilde{\mathbf{\Lambda}}_{N,j} - \sum_{k=1}^{j-1} \widetilde{\mathbf{L}}_{N,k} \widetilde{\mathbf{L}}_{j,k} \right], \tag{2.9}$$

$$\widetilde{\mathbf{L}}_{N,N} = \left[\widetilde{\mathbf{\Lambda}}_{N,N} - \sum_{k=1}^{N-1} \widetilde{\mathbf{L}}_{N,k}^2 \right]^{1/2}. \tag{2.10}$$

Importantly, the Direct algorithm has the property that row $j-1$ of $\mathbf{L}$ is not a subset of row j of $\mathbf{L}$, therefore at each stage of the algorithm, j new weights must be explicitly calculated leading to $\mathcal{O}(N^2)$ complexity.

2.3.2.3 Hypergeometric Discrete algorithm

Discrete algorithms are based on discrete approximations to the continuous time position process (2.3). They have the form,

$$X_N = \sum_{i=1}^{N} \int_{(i-1)\Delta t}^{i\Delta t} f(u,t,H)\xi_i du, \tag{2.11}$$

where the weighting function $f(u,t,H)$ depends on the version of the algorithm. The general approach is to find a set of weights that produce W-fBm when multiplied by the noise terms ξ_i. The primary benefit of Discrete algorithms compared to Direct algorithms is that they are well suited for adaptive simulation schemes since the weights for X_N are a subset of the weights for X_{N-1} and new positions can be appended to the existing position process simply by evaluating the appropriate weight function.

The hypergeometric Discrete algorithm is based on an alternative representation of the continuous time W-fBm process described by (2.3). It is written as [51]:

$$B_H(t) = \int_0^t K^H(t,u)\xi(u)du \tag{2.12}$$

where

$$K^H(t,u) = \frac{(t-u)^{H-1/2}}{\Gamma(H+1/2)} \times {}_2F_1\left(H-\frac{1}{2}, \frac{1}{2}-H; H+\frac{1}{2}; 1-\frac{t}{u}\right), \quad t>u. \tag{2.13}$$

As an aside, the above equation is presented as it is in Carmona et al. [52] who cite the work of Decreusefond [51]. Decreusefond however, defines the Gaussian hypergeometric function as

$${}_2F_1\left(a,b;c;z\right), \quad |z|<1. \tag{2.14}$$

The condition on z is clearly not satisfied in the presentation given by Carmona et al. if $t>u$.

${}_2F_1$ is the Gaussian hypergeometric function,

$$ {}_2F_1(a,b;c;d) = \sum_{n=0}^{\infty} \frac{(a,n)(b,n)}{(c,n)} \frac{d^n}{n!} \tag{2.15} $$

where

$$ (a,d) = \frac{\Gamma(a+d)}{\Gamma(d)}. \tag{2.16} $$

The discrete approximation is thus

$$ X_N = \sum_{i=1}^{N} \int_{(i-1)\Delta t}^{i\Delta t} K^H(t_N,u)\xi(u)du. \tag{2.17} $$

Exact calculation of the above integral is inefficient to compute, but for sufficiently small Δt, the integral may be approximated with techniques such as the Mid Point method, the Trapezoidal method or Gaussian quadrature. For the simulations presented here, we utilize the Trapezoidal method.

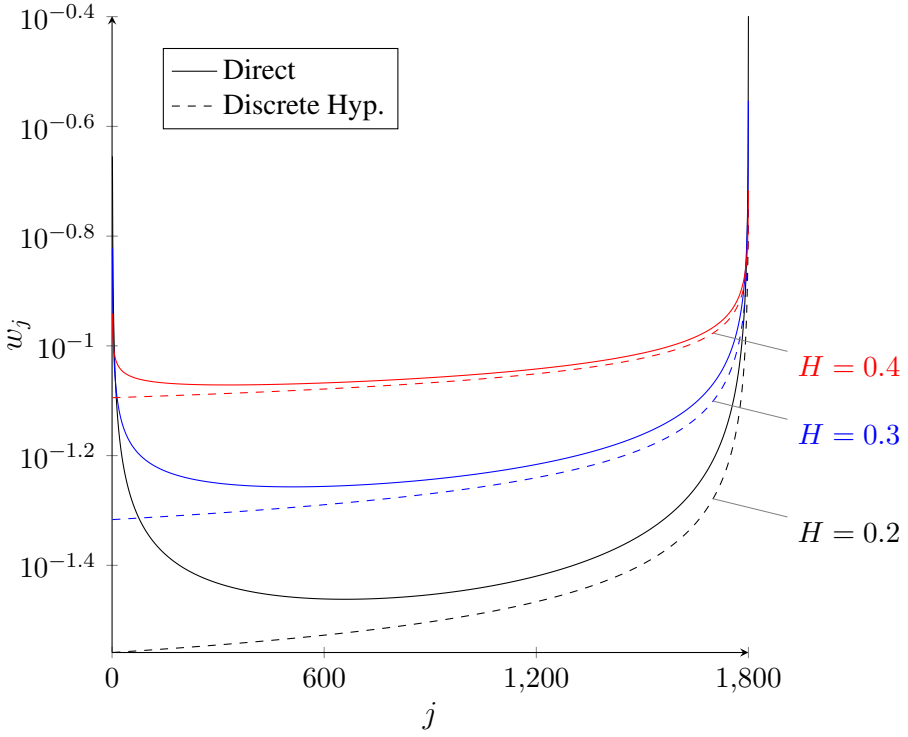


Figure 2.2: Comparison of weights for the Direct and Hypergeometric Descrete W-fBm algorithms for $H = 0.2$ (black), $H = 0.3$ (blue) and $H = 0.4$ (red). The weights are for a path of temporal length $T = 30$ s, inter-observational time $\Delta t = 1/60$ and $N = 1,800$ steps.

The weights for the Direct and Hypergeometric Discrete W-fBm algorithms are presented in Fig. 2.2. We make two observations: First, the Direct weights for points early in the time series are larger than the weights for points in the middle of the time series and thus have a larger influence on X_N. Second, the Hypergeometric Discrete approximation does not capture this trend. The pronounced influence of the initial noise terms for W-fBm processes is intimately linked to the representation of W-fBm as an infinite memory process. In simulating such processes, it is assumed that the first data point incorporates the infinite "pre-history". This initial pre-history, whatever it may be, is then perpetuated throughout the length of the simulated path via the large weights assigned to the initial noise components. While this may seem like a reasonable approach in theory, many physical phenomena, such as diffusion processes, have a well defined commencement at some $t = 0$, at which point there is no memory in the process. W-fBm as a model for diffusion incorporates infinite memory at $t = 0$ that we do not expect to exist in the physical phenomena itself. We will refer to this as the Starting Point Problem. As noted by others [40, 53–55], this fundamental difference between the model and the physical phenomena it is tasked with describing, calls into question the use of W-fBm in this context. To address this miss-match caused by the Starting Point Problem, we turn to Riemann-Liouville fractional Brownian motion.

2.4 Riemann-Liouville fractional Brownian motion

2.4.1 Theory

Riemann-Liouville fractional Brownian motion (RL-fBm), is a an alternative representation of fBm based on the Riemann-Liouville fractional integral [56]. A RL-fBm continuous time position process $\{B(t)\}$ may be written as,

$$B_H(t) = \frac{1}{\Gamma(H + 1/2)} \int_0^t (t - u)^{H-1/2} \xi(u) du, \qquad t \geq 0. \tag{2.18}$$

Comparing this representation to that of W-fBm presented in (2.3), we see RL-fBm does not exhibit infinite memory at $t = 0$, effectively abrogating the Starting Point Problem, but does so at the expense of stationarity of the increments [40]. The increment process however, converges to *local stationarity* when the sampling frequency Δt is significantly smaller than the observation time, i.e. when $t/\Delta t \gg 1$ [40].

Correlation of the RL-fBm position process [54] is given by

$$\mathbf{E}[B_H(t_1)B_H(t_2)] = \frac{t_1^{H+1/2}t_2^{H-1/2}}{(H+1/2)[\Gamma(H+1/2)]^2} \times {}_2F_1\left(\frac{1}{2}-H, 1, H+\frac{3}{2}, \frac{t_1}{t_2}\right), \tag{2.19}$$

where $0 < t_1 < t_2$ and ${}_2F_1$ is the Gaussian hypergeometric function defined in (2.15).

To get the variance of the RL-fBm position process, we set $t_1 = t_2$, thus

$$\mathbf{E}[B_H^2(t)] = \frac{2t^{2H}}{(H+1/2)[\Gamma(H+1/2)]^2} \times {}_2F_1\left(\frac{1}{2}-H, 1, H+\frac{3}{2}, 1\right) \tag{2.20}$$

$$= \frac{2t^{2H}}{(H+1/2)[\Gamma(H+1/2)]^2}\left[\frac{\Gamma(H+3/2)\Gamma(2H)}{\Gamma(2H+1)\Gamma(H+1/2)}\right] \tag{2.21}$$

$$= \frac{t^{2H}}{H[\Gamma(H+1/2)]^2}, \tag{2.22}$$

and we recognize that the variance scales with t^{2H} as expected.

Given a RL-fBm process with total path length t and inter-observational time Δt such that we may assume the increments are locally asymptotically stationary, their correlation is given by [40]

$$\text{ACF}_b(n) = \left[\sigma^2 + \left(\{2H(\Gamma(H+1/2)\}^2\right)^{-1}\right](n\Delta t)^{2H} \tag{2.23}$$

where $\sigma^2 = \langle B_H^2(1)\rangle$.

2.4.2 Simulation techniques

2.4.2.1 Direct algorithm

We can easily implement a Direct algorithm for the simulation of RL-fBm using the same method described in Section 2.3.2.1 for W-fBm. Whereas for W-fBm the covariance matrix was defined by (2.4), here we simply substitute (2.19) to form a covariance matrix for RL-fBm, denoted $\mathbf{\Phi}$. The i, j element of $\mathbf{\Phi}$ is

$$\mathbf{\Phi}_{i,j} = \frac{t_i^{H+1/2}t_j^{H-1/2}}{(H+1/2)[\Gamma(H+1/2)]^2} \times {}_2F_1\left(\frac{1}{2}-H, 1, H+\frac{3}{2}, \frac{t_j}{t_i}\right), \quad t_i \geq t_j. \tag{2.24}$$

2.4.2.2 Exact Discrete algorithm

The Exact Discrete RL-fBm algorithm is based on the discrete position process [57]

$$X_N = \frac{1}{\Gamma(H+1/2)} \sum_{i=1}^{N} \int_{(i-1)\Delta t}^{i\Delta t} (t_N - u)^{H-1/2} \left(\frac{\xi_i}{\sqrt{\Delta t}} \right) du. \tag{2.25}$$

The integral can be solved exactly to yield,

$$X_N = \sum_{i=1}^{N} w_{N-i+1} \xi_i, \tag{2.26}$$

$$w_i = \frac{1}{\Gamma(H+1/2)} \left[\frac{t_i^{H+1/2} - (t_i - \Delta t)^{H+1/2}}{\sqrt{\Delta t}(H+1/2)} \right] \tag{2.27}$$

Figure 2.3 shows the weights w_{N-i+1} for three paths with three different Hurst parameters.

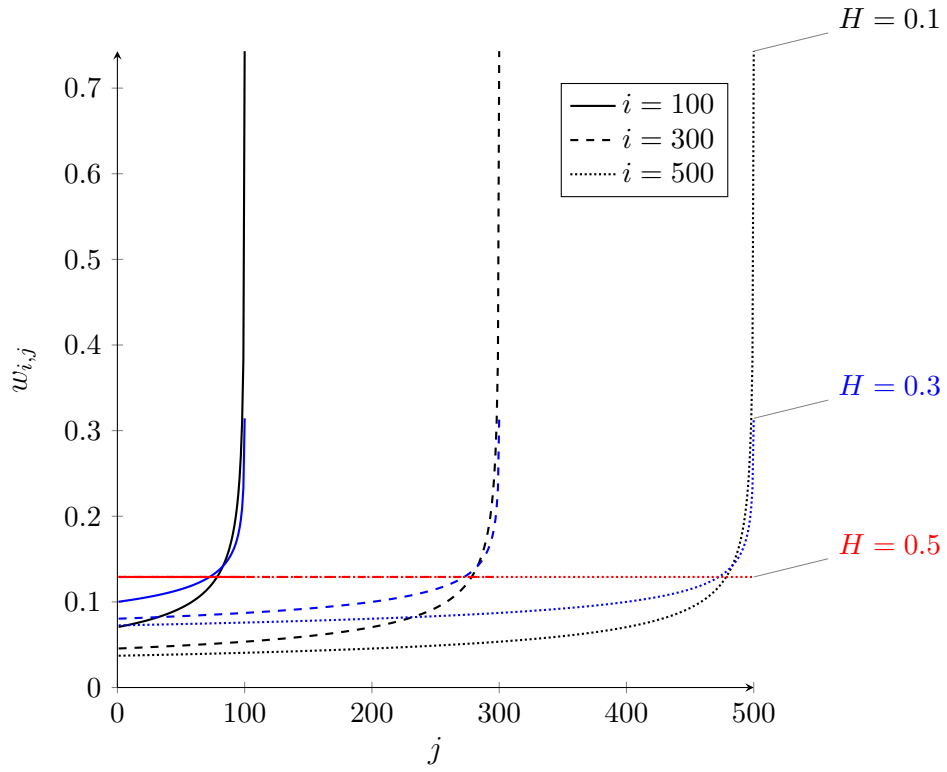


Figure 2.3: Weights for three $N = 500$ step paths ($\Delta t = 1/2$) generated via the Exact Discrete RL-fBm algorithm. The paths have Hurst parameter $H = 0.1$ (black), $H = 0.3$ (blue) and $H = 0.5$ (red). The weights used to construct X_N are shown for three points within each path, $j = 100$, $j = 300$ and $j = 500$.

2.4.2.3 Improved Discrete algorithm

Rambaldi et al. [57] and Muniandy et al. [54] utilize an improved weighting method that exactly satisfies the local scaling behavior of the covariance. Specifically,

$$X_N = \sum_{i=1}^{N} \xi_i \widetilde{w}_{N-i+1} \tag{2.28}$$

where,

$$\widetilde{w}_i = \frac{1}{\Gamma(H + 1/2)} \left[\frac{t_i^{2H} - (t_i - \Delta t)^{2H}}{2H\Delta t} \right]^{1/2}. \tag{2.29}$$

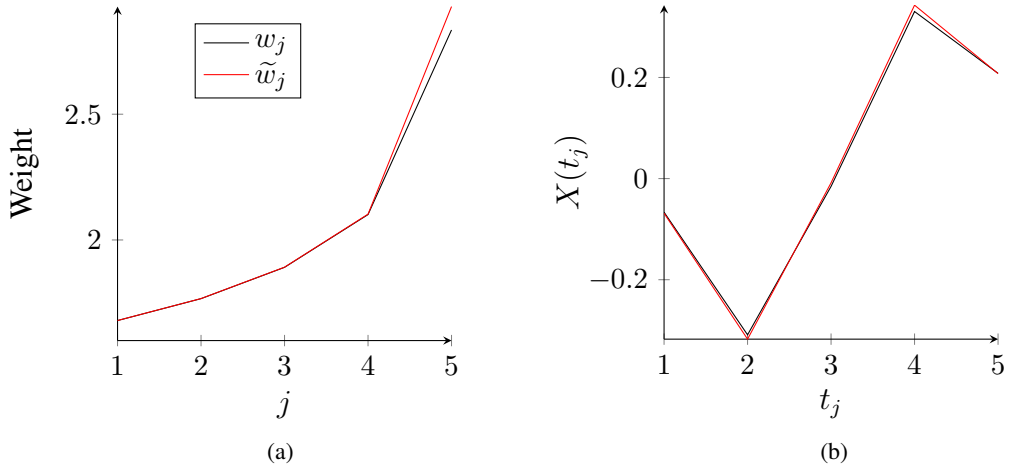


Figure 2.4: Comparison of weights for the Improved and Exact Discrete RL-fBm algorithms. (A) The newest Brownian noise component, ξ_N receives a higher weight for the Improved algorithm compared to the Exact algorithm. (B) The impact of the increased weight for trailing Brownian noise components in the Improved algorithmm is that the position process of the resulting path is slightly more extreme compared to the Discrete algorithm.

A comparison of the Exact and Improved weights over the length of a short path ($N = 5$) is shown in Fig. 2.4. The newest Brownian noise component, ξ_N, receives a higher weight for the Improved algorithm compared to the Exact algorithm. The two weighting functions are monotonic and convergent for earlier time points, i.e., as $N - i + 1$ increases. The increased weight on ξ_N in the Improved algorithm causes the position process of the resulting path to be slightly more extreme compared to the Discrete algorithm. A comparison of the weights for the Direct, Exact Discrete and Improved Discrete RL-fBm algorithms is shown in Fig. 2.5. There is a stronger correspondence between the Direct and Discrete versions of the algorithm weights compared with the corresponding

W-fBm algorithms. Within RL-fBm algorithms, the Discrete approaches have the advantage that they are straightforward to compute and do not require the evaluation of the Gaussian hypergeometric function, which is unwieldy, requiring different algorithms for different parameter values.

While RL-fBm is a promising candidate model for diffusion processes, solving the Starting Point Problem, and can be efficiently simulated via Discrete algorithms, it is not without limitations. In particular, RL-fBm in its current presentation is not well suited to model diffusion in heterogeneous fluids. To understand this, recall that the mobility of a tracer particle embedded in a fluid is directly related to the viscoelastic properties of the fluid via the GSER. In the case of a viscoelastic fluid the elasticity arises from a physical source. For mucus, the source is a gel network composed of interconnected glycosylated mucin proteins. The elasticity in the fluid engenders memory in the increment process of an embedded tracer particle, characterized by $H < 0.5$ (when no net force is applied to the particle), which may, in the case of mucus, vary with local disparities in mucin concentration or temporal changes in the chemical bonds forming the gel network itself. When we refer to a heterogeneous fluid, we are referring to a fluid that exhibits non-uniform elasticity–the case where elasticity is constant but viscosity is non-uniform is trivial–and in the context of an fBm model, the Hurst parameter H for such a fluid should naturally be variable. To generalize fractional Brownian motion for such situations, we turn to multi-fractional Brownian motion.

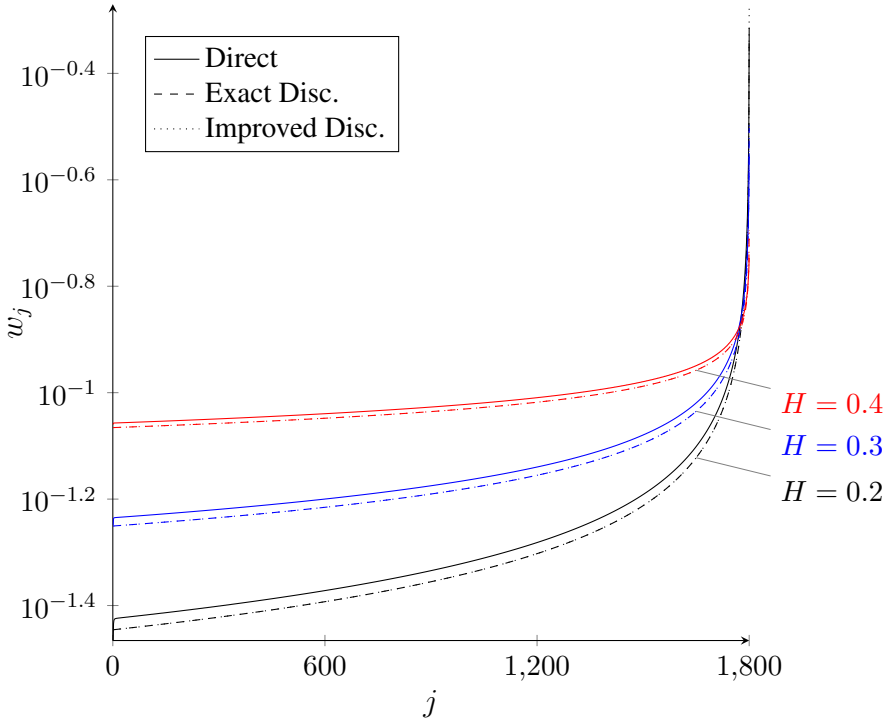


Figure 2.5: Comparison of weights for the Direct, Exact Discrete and Improved Discrete RL-fBm algorithms for $H = 0.2$ (black), $H = 0.3$ (blue) and $H = 0.4$ (red). The weights are for a path of temporal length $T = 30$ s, inter-observational time $\Delta t = 1/60$ and $N = 1,800$ steps. The difference between the Discrete Exact and Improved weights is indistinguishable except for at $j = 1800$.

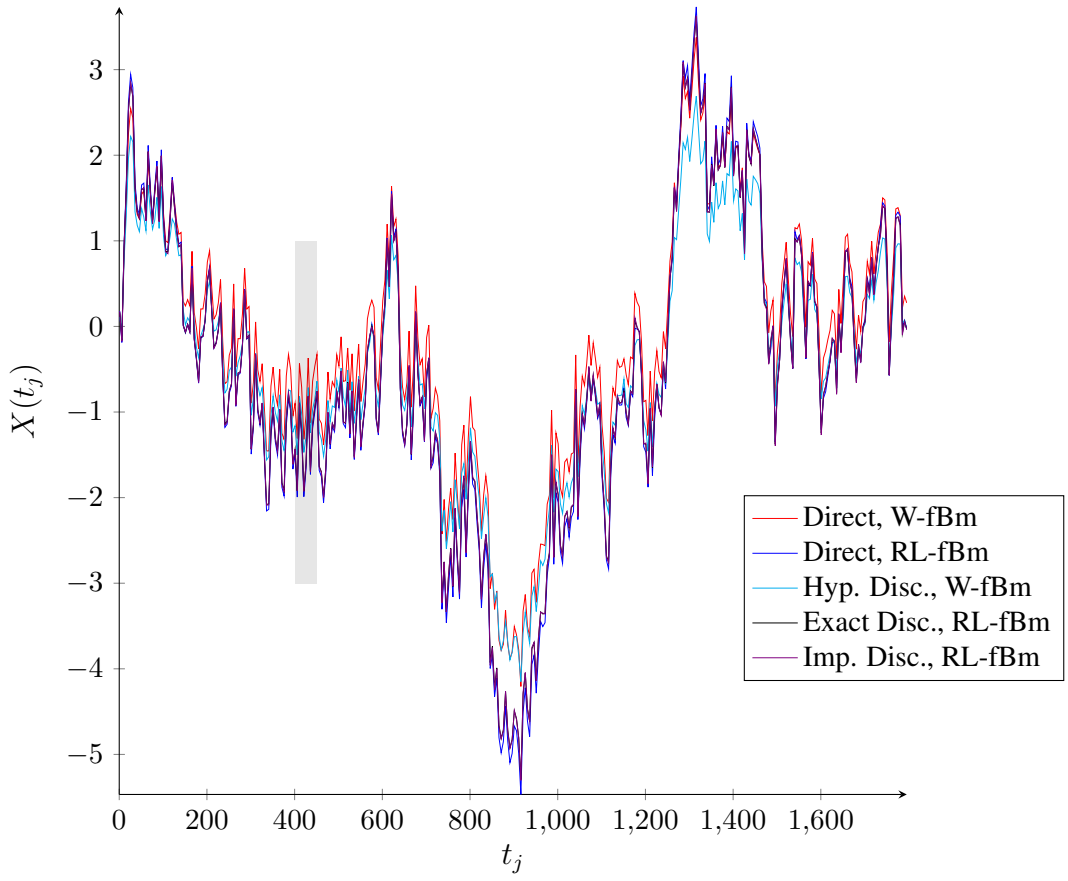


Figure 2.6: Comparison of position processes from W-fBm and RL-fBm algorithms. The paths have a temporal length $T = 30$ s with inter-observational time $\Delta t = 1/60$ and $N = 1,800$ steps. The gray region indicates data highlighted in Fig. 2.7.

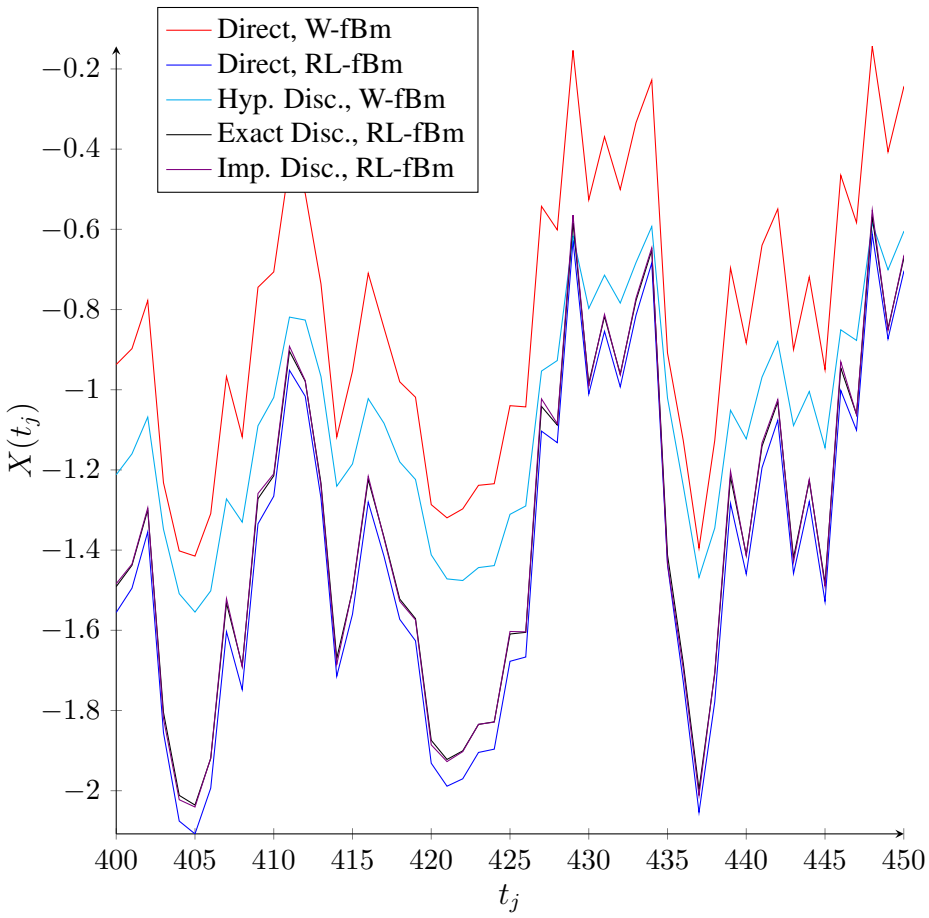


Figure 2.7: Comparison of position processes from W-fBm and RL-fBm algorithms. The paths have a temporal length $T = 30$ s with inter-observational time $\Delta t = 1/60$ and $N = 1,800$ steps. The data presented here correspond to the gray region indicated in Fig. 2.6.

2.5 Multi-fractional Brownian motion

2.5.1 Theory

Multi-fractional Brownian motion (mBm) is a generalization of fractional Brownian motion that allows for a time-varying Hurst parameter [58, 59]. Here, we focus on Riemann-Liouville mBm although various representations of mBm exist [60, 61]. The continuous time (RL) mBm process is defined as [55]

$$B_{H(t)}(t) = \frac{1}{\Gamma(H(t) + 1/2)} \int_0^t (t-u)^{H(t)-1/2} \xi(u) du, \qquad t \geq 0, \tag{2.30}$$

and has covariance,

$$\text{ACF}_B(t_1, t_2) = \frac{2 t_1^{H(t_1)+1/2} t_2^{H(t_2)-1/2}}{[2H(t_1)+1]\,\Gamma(H(t_1)+1/2)\Gamma(H(t_2)+1/2)} \times_2 F_1\left(\frac{1}{2} - H(t_2), 1, H(t_1) + \frac{3}{2}, \frac{t_1}{t_2}\right), \tag{2.31}$$

where $t_2 > t_1$. We observe that the key difference between mBm and RL-fBm is the substitution of the function $H(t)$ for the constant H.

2.5.2 Simulation techniques

Muniandy et al. [55] propose Exact and Improved Discrete algorithms for mBm using the same weighting functions described by (2.29) and (2.27), without modification. Their approach is as straight forward as it is computationally expensive. The position process is simulated as

$$X_j = B_{H(t_j)}(t_j), \qquad 0 \leq j \leq N. \tag{2.32}$$

For k discrete values of H, we recognize that (2.32) calls for the simulation of k distinct RL-fBm processes, whose elements are then selectively recombined according to $H(t_j)$. For example, given $H(t_j)$, we use the appropriate RL-fBm Discrete algorithm to generate $B_{H(t_i)}(t_i)$, $0 \leq i \leq j$, and store the final position $B(t_j)$ as $X(t_j)$. At the next time step, if $H(t_{j+1}) \neq H(t_j)$ we generate a new path $B_{H(t_i)}(t_i)$, $0 \leq i \leq (j+1)$ and store the final position $B(t_{j+1})$ as $X(t_{j+1})$. If

$H(t_{j+1}) = H(t_j)$, we simply extend the current path by one step and store the result in the same manner.

Let

$$H(t_j) = \begin{cases} 0.35 & 0 \leq j \leq 500 \\ 0.5 & 500 \leq j \leq 900 \\ 0.1 & 900 \leq j \leq 1800 \end{cases} \tag{2.33}$$

Figure 2.8 shows the position process resulting from this choice of $H(t)$.

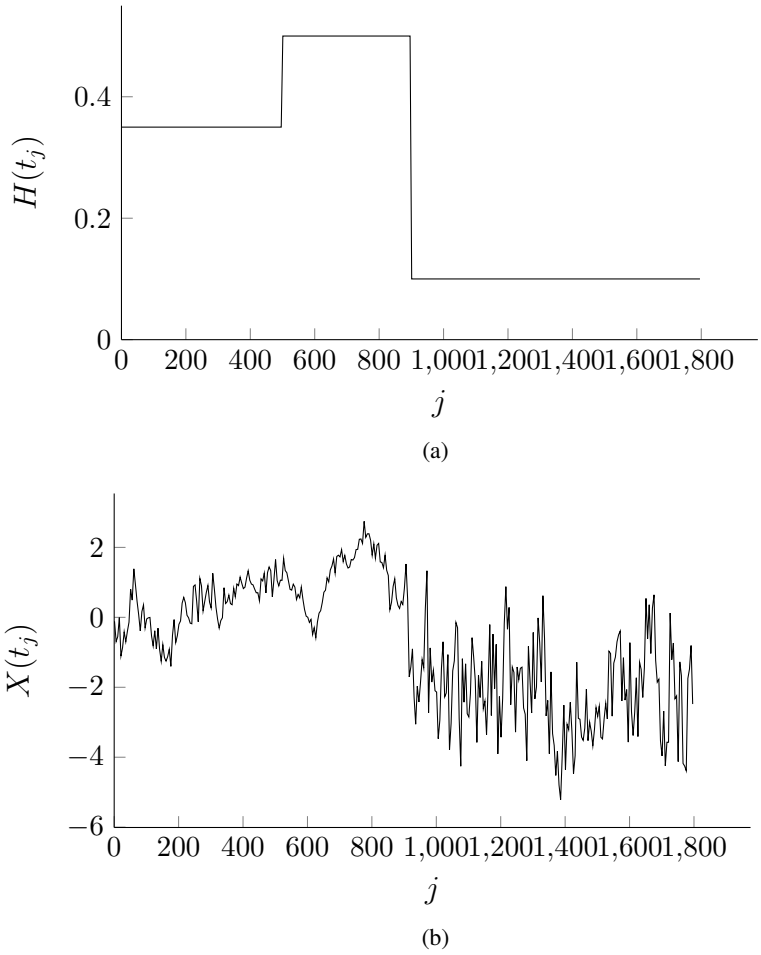


Figure 2.8: Example mBm path simulated using Discrete Improved weights. The paths have a temporal length $T = 30$ s with inter-observational time $\Delta t = 1/60$ and $N = 1,800$ steps. (A) The input function $H(t_j)$ is defined by (2.33). (B) The resulting possition process given $H(t_j)$.

The temporal dependence of H makes mBm a more suitable model for processes with variable memory, corresponding to variable elasticity in the fluid properties, compared to W-fBm or RL-fBm.

However, mBm is not without limitations. To understand these, consider a biphasic fluid, the left volume of which is a purely viscous fluid and the right volume of which is a viscoelastic fluid. *Within* each region, the fluid is homogeneous. Consider a particle that has been deposited in the purely viscous half of the volume at time $t = 0$. While the particle diffuses within the viscous fluid, we would apply standard Brownian motion to model its movements since there would be no correlation in the increment process. As soon as the particle transits to the viscoelastic half of the volume, we would instinctively apply RL-fBm to account for the correlation in the increment process created by the non-zero storage modulus of the fluid in that region. Within the viscoelastic region, we note that the diffusive movements of the embedded particle does not meet the formal definition of a long memory process. The formal definition of a long memory process is a process such that

$$\text{ACF}(t) \sim ct^{2d-1} \quad \text{as } t \to \infty, \tag{2.34}$$

where $c \neq 0$ and $d < 0.5$ [62]. In the case of RL-fBm (and W-fBm), the covariance scales as ct^{2H}, and $H \in [0, 0.5] \implies d \in [0.5, 1]$, thus, formally, there is no long-time memory. The weights however, *do decay* as ct^{2d-1} with $d < 1/2$. To see this, consider the formula for the Exact Discrete weights given in (2.27). Ignoring the leading gamma term, we have,

$$w_i = \{\sqrt{\Delta t}(H + 1/2)\}^{-1/2} \left[t_i^{H+1/2} - (t_i - \Delta t)^{H+1/2}\right] \tag{2.35}$$

$$= \{\sqrt{\Delta t}(H+1/2)\}^{-1/2} \left[t_i^{H+1/2} - t_i^{H+1/2} - \Delta t(H + 1/2)t_i^{H-1/2} + \frac{1}{2}\Delta t^2(H - 1/2)(H + 1/2)t_i^{H-3/2} + \mathcal{O}(\Delta t^3)\right] \tag{2.36}$$

$$= \Delta t^{1/2} t^{H-1/2} + \frac{1}{2}\Delta t^{3/2}(H - 1/2)t^{H-3/2} + \mathcal{O}(\Delta t^{5/2}) \tag{2.37}$$

As $t_i \to \infty$, the first two terms satisfy the scaling relation in (2.34) when $H < 1/2$. Thus, even though RL-fBm is not formally a long memory process, we would be remiss to neglect any of the noise terms in the calculation of X_N, even when $i \ll N$.

Since there is no point beyond which we may ignore the weights, RL-fBm depends on all previous behavior. This presents a contradiction when considering our hypothetical biphasic fluid– if the correlation in the increment process is attributable to the elastic nature of the fluid, and the correlation gives rise to a weighting function that does do not decay, then the diffusion of the

particle in the viscoelastic fluid is dependent upon its previous behavior in the purely viscous fluid. Furthermore, under the mBm model, the movement of the particle in the viscoelastic half of the fluid volume would depend on the stochastic movements of the particle in the purely viscous half *as if the purely viscous half had been viscoelastic.* We know this is inconceivable since there is no physical mechanism in the viscous fluid to store and transmit the particle's movements to the viscoelastic region for perpetuity, or otherwise. We refer to this as the "Agency Problem." This phenomena is manifested in mBm through the use of $H(t)$ as the exponent of the integrand in (2.30) and gives rise to discontinuity in the position process $X(t)$ when $H(t)$ is not a smooth function, and for many fluids, we have no reason to suspect this to be true. To address the Agency Problem, we introduce amnesiac Brownian motion.

2.6 Amnesiac Brownian motion

2.6.1 Theory

Continuous time amnesiac Brownian motion (aBm), presented here for the first time, is generalization of RL-fBmn which addresses the Agency Problem. It may be used to simulate processes with both constant and variable Hurst parameter, i.e., diffusion in heterogeneous and homogeneous, viscous and viscoelastic fluids. An aBm position process may be written as

$$B_{H(t)}(t) = \frac{1}{\Gamma(H_t(t) + 1/2)} \int_0^t (t-u)^{H_t(u)-1/2} \xi(u) du, \qquad t \geq 0 \tag{2.38}$$

The key difference between aBm and mBm is the use of $H(u)$ in the integrand, as opposed to $H(t)$.

To develop an intuitive understanding of aBm consider a volume of fluid with four regions, each region containing a different fluid with an storage modulus characterized by the appropriate corresponding Hurst parameter. From left to right, the respective values of H for each region are, $H = 0.4$, $H = 0.5$, $H = 0.1$ and $H = 0.2$. A particle, initially in the first region, is allowed to diffuse through the fluid and six observations of its location are made. This scenario is illustrated in Fig. 2.9.

For an mBm process, we would write the position $Y_n = Y(n\Delta t)$ as

$$Y_1 = \xi_1 w_{1,H=0.4} \tag{2.39}$$

$$Y_2 = \xi_1 w_{1,H=0.4} + \xi_2 w_{2,H=0.4} \tag{2.40}$$

$$Y_3 = \xi_1 w_{1,H=0.5} + \xi_2 w_{2,H=0.5} + \xi_2 w_{3,H=0.5}. \tag{2.41}$$

Note that the weights w are a function of H at the particle's current position. Under an aBm model, we may similarly write the first three positions as

$$X_1 = \xi_1 w_{1,H=0.4} \tag{2.42}$$

$$X_2 = \xi_1 w_{1,H=0.4} + \xi_2 w_{2,H=0.4} \tag{2.43}$$

$$X_3 = \xi_1 w_{1,H=0.4} + \xi_2 w_{2,H=0.4} + \xi_2 w_{3,H=0.5}. \tag{2.44}$$

Up to this point, the aBm and mBm processes for this scenario differ only by the third position. At this point, for the mBm process, Y_3 is generated as if the particle had only inhabited the $H = 0.5$ region, as evidenced by the use of the weights $w_{,H=0.5}$. According to the formulation of mBm (2.32) the fourth position should be written as

$$Y_4 = \xi_1 w_{1,H=0.1} + \xi_2 w_{2,H=0.1} + \xi_3 w_{3,H=0.1} + \xi_4 w_{4,H=0.1}. \tag{2.45}$$

Here, an even more extreme change in behavior has happened, the fourth position is calculated assuming that all previous behavior had happened in the $H = 0.1$ region. In contrast to mBm, the fourth position, under an aBm model, is simply

$$X_4 = \xi_1 w_{1,H=0.5} + \xi_2 w_{2,H=0.5} + \xi_3 w_{3,H=0.5} + \xi_4 w_{4,H=0.1}. \tag{2.46}$$

The remaining aBm positions are,

$$X_5 = \xi_1 w_{1,H=0.5} + \xi_2 w_{2,H=0.5} + \xi_3 w_{3,H=0.5} + \xi_4 w_{4,H=0.2} + \xi_5 w_{5,H=0.2} \tag{2.47}$$

$$X_6 = \xi_1 w_{1,H=0.5} + \xi_2 w_{2,H=0.5} + \xi_3 w_{3,H=0.5} + \xi_4 w_{4,H=0.2} + \xi_5 w_{5,H=0.2} + \xi_6 w_{6,H=0.1}. \tag{2.48}$$

The difference in the resulting paths is illustrated in Fig. 2.10 for a single realization of each the process using the same noise terms, ξ_i.

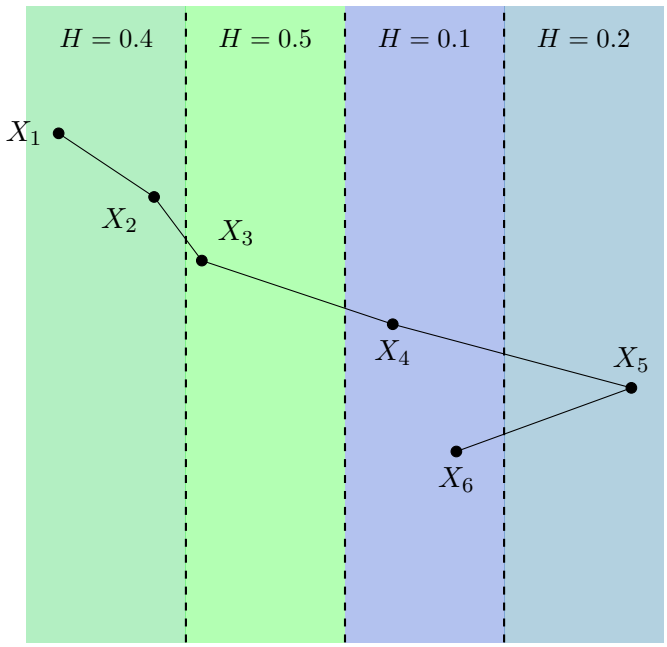


Figure 2.9: Illistration of diffusion through a spatially heterogeneous fluid. A particle, beginning in the first region, is allowed to diffuse through the volume of fluid. Each region exhibits a different level of memory characterized by their respective Hurst parameters.

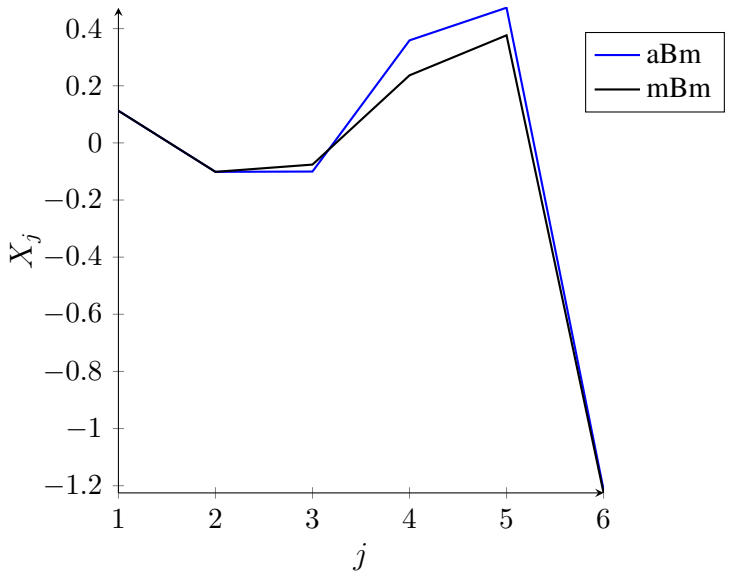


Figure 2.10: Example aBm and mBm paths corresponding to $H(t_j)$ presented in Fig. 2.9. Both paths incorporate the same nose terms. As expected, the position processes are identical for the first 2 steps.

2.6.2 Simulation Techniques

2.6.2.1 Path Splicing

The path splicing algorithm relies on the assumption that, in nature, the bombardment of the diffusing particle by the molecules of the surrounding fluid, represented here by the noise term ξ, cannot impact the position of the particle for an infinite amount of time, i.e., we assume there exists a ψ such that

$$\frac{\sum_{i=1}^{\psi} w_i}{\sum_{i=\psi+1}^{N} w_i} \approx 0. \tag{2.49}$$

This allows us to write the current state of an aBm process as a function of a finite number of previous states. Adjacent segments can then be "spliced" together by incorporating the noise terms used in the final ψ steps of the previous path segment into the initial ψ steps of the next path segment. To construct a position process of length N made of segments of length $a > \psi$, we need only compute the $a \times a$ correlation matrix. The first segment is given by,

$$\{X\}^1 = \sigma^{1/2} \begin{pmatrix} \xi_1^1 & \xi_2^1 & \cdots & \xi_a^1 \end{pmatrix} \mathbf{V}_{H(t_1)}, \tag{2.50}$$

where $\mathbf{V}_{H(t_1)}$ is the $a \times a$ weight matrix of the first segment and $\xi \sim \mathcal{N}(0,1)$. The elements of the weight matrix may be specified using any of the RL-fBm methods previously discussed. Here we use the Discrete Improved weights defined in (2.29). The superscripts on the ξ terms indicate that they are associated with the first segment. Each subsequent segment $b \neq 1$ is constructed as follows,

$$\{X\}^b = \sigma^{1/2} \begin{pmatrix} \xi_{a-q+1}^{b-1} & \xi_{a-q+2}^{b-1} & \cdots & \xi_a^{b-1} & \xi_{q+1}^b & \xi_{q+2}^b & \cdots & \xi_a^b \end{pmatrix} \mathbf{V}_H, \tag{2.51}$$

where $q \geq \psi$ is a measure of the overlap between adjacent segments. Note that for uniform H, the correlation matrix for each segment is the same, thus we only need to compute $\mathbf{V}$ once. To use this algorithm to simulate diffusion through heterogeneous media, we simply keep track of the Hurst parameter along the length of the segment and use this record to build the appropriate correlation matrix. At each time step where $H(t_j) \neq H(t_{j-1})$, only the a weights in the final row of $\mathbf{V}_{H(t_j)}$ need to be calculated.

Two adjacent segments, $\{X\}^{b-1}$ and $\{X\}^b$ are then concatenated as $(\{X\}^{b-1}, \widetilde{\{X\}}^b)$ where

$$\widetilde{\{X\}}^b = \{X_{q+1}^b \cdots X_a^b\} + \zeta \tag{2.52}$$

and

$$\zeta = X_a^{b-1} - X_q^b. \tag{2.53}$$

The shift of segment b by ζ ensures that the paths are continuous.

2.6.2.2 Modified Discrete Algorithms

Both the Improved Discrete and Exact Discrete algorithms may be modified to efficiently generate aBm process without resorting to path splicing if the assumption that the impact of noise increments exhibit a finite duration does not hold. To do so, we only need to keep track of the values of H associated with each previously time step. Figure 2.11 shows an example aBm path generated using a modified Improved Discrete algorithm where

$$H(t_j) = 0.4 \sin\left(\frac{j2\pi}{N}\right) + 0.5. \tag{2.54}$$

The path has $N = 1,800$ steps and inter-observational time $\Delta t = 1/60$ s.

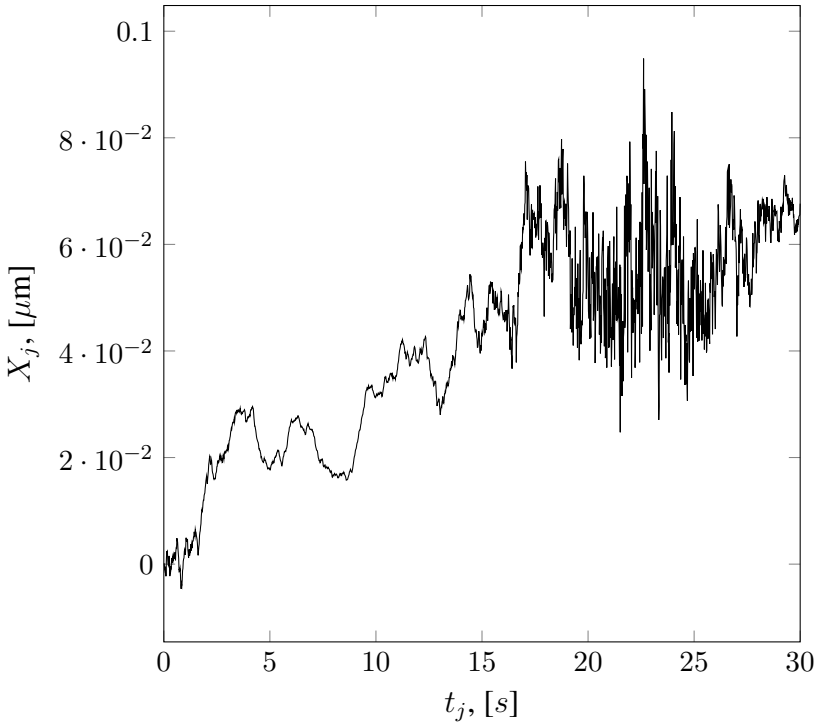


Figure 2.11: Example aBm path with sinosodial H specified by (2.54). The path was generated using a modified Improved Discrete algorithm and has $N = 1,800$ steps and inter-observational time $\Delta t = 1/60$ s, and $D = 0.005\ \mu\text{m}^2\text{s}^{-2H}$.

2.7 Local estimations of H

There are many alternative approaches to simulating exact and approximate fBm processes, such as the spectral, wavelet-based and Random Midpoint Displacement methods (for an excellent review, see [63, 64]). However, in this chapter, we have focused on a subset of methods that involve multiplication of a weight matrix by a vector of noise increments because they naturally give rise to efficient methods for local estimation of H and D.

Let $B_H(t)$ be fractional Gaussian noise with Hurst parameter H defined. The position process of a particle undergoing diffusion may be modeled as

$$X(t) = \sigma B_H(t), \tag{2.55}$$

with inter-observational time Δt, $X_n := X(n\Delta t)$ and $x_n := X_{n=1} - X_n$, $n = 1, 2, ...N + 1$. The prefactor σ is a scaling factor related to the diffusivity as $1/2D = \sigma^2$. The increments are normally

distributed and the corresponding log-likelihood function, given x and H is

$$\ell(\sigma|x,H) = -\frac{1}{2}\left\{\frac{x'V_H^{-1}x}{\sigma^2} + N\log(\sigma^2) + \log(|V_H|)\right\}, \tag{2.56}$$

where V_H is the correlation matrix of the increment process. The derivative of (2.56) with respect to σ is

$$\frac{d}{d\sigma}\ell(\sigma|x,H) = -\frac{1}{2}\left\{\frac{-2x'V_H^{-1}x}{\sigma^3} + \frac{2N}{\sigma}\right\} \tag{2.57}$$

$$= \frac{x'V_H^{-1}x}{\sigma^3} - \frac{N}{\sigma} \tag{2.58}$$

Setting (2.57) equal to zero and solving for σ^2, we get

$$\widehat{\sigma}^2 = \frac{x'V_H^{-1}x}{N} \tag{2.59}$$

Given this maximum likelihood estimations σ, the probability of observing the data is $p(x,H|\sigma) = \ell(\sigma|x,H)$. Substituting $\widehat{\sigma}$ into (2.56), we see

$$p(x,H|\widehat{\sigma}) = \ell(\widehat{\sigma}|x,H) = -\frac{1}{2}\left\{\frac{x'V_H^{-1}x}{\widehat{\sigma}^2} + N\log(\widehat{\sigma}^2) + \log(|V_H|)\right\} \tag{2.60}$$

$$= -\frac{1}{2}\left\{\frac{Nx'V_H^{-1}x}{x'V_H^{-1}x} + N\log(\widehat{\sigma}^2) + \log(|V_H|)\right\} \tag{2.61}$$

$$= -\frac{N}{2}\left\{1 + \log(\widehat{\sigma}^2) + \frac{1}{N}\log(|V_H|)\right\} \tag{2.62}$$

We have the Cholesky decomposition L of V_H such that $LL' = V_H$. We also know that

$$|V_H| = |L|^2 = \left[\prod L_{i,i}\right]^2 \tag{2.63}$$

If we assume that H is constant, then $L_{1,1} = L_{i,i}\ \forall\ i = 2,3,...N$ when using the Improved Discrete or Exact Discrete weighting scheme. Therefore

$$|V_H| = (L_{1,1})^{2N} \tag{2.64}$$

$$\implies \frac{1}{N}\log(|V_H|) = \frac{2N}{N}\log(L_{1,1}) = 2\log(L_{1,1}). \tag{2.65}$$

A local estimation of H and D can be computed by maximizing (2.60) for a subset of the path. Here, we have implemented a simple local estimator of H for overlapping neighborhoods of $n = 60$ steps in width. The algorithm was applied to the path shown in Figure 2.11 where $H(t_j)$ varies sinusoidally. A path with both $H(t_j)$ and $D(t_j)$ defined by step functions was also generated and the local estimates of H are shown in Figure 2.13 for three neighborhood sizes.

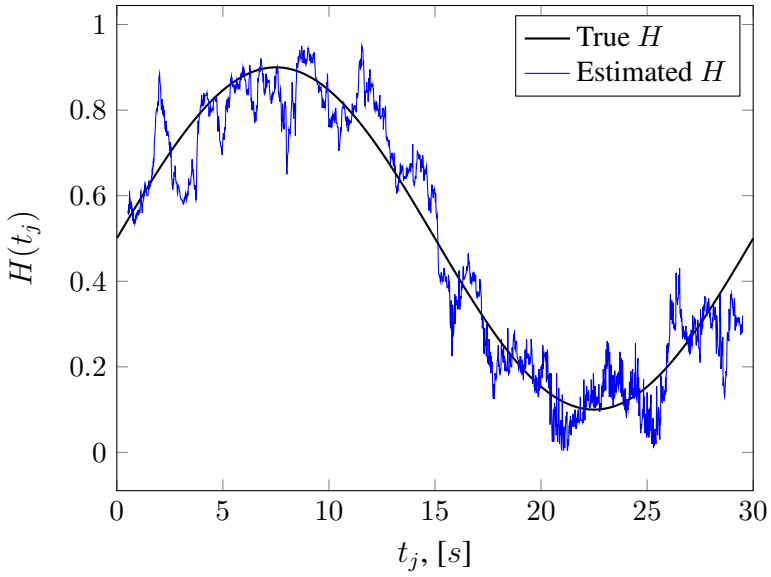
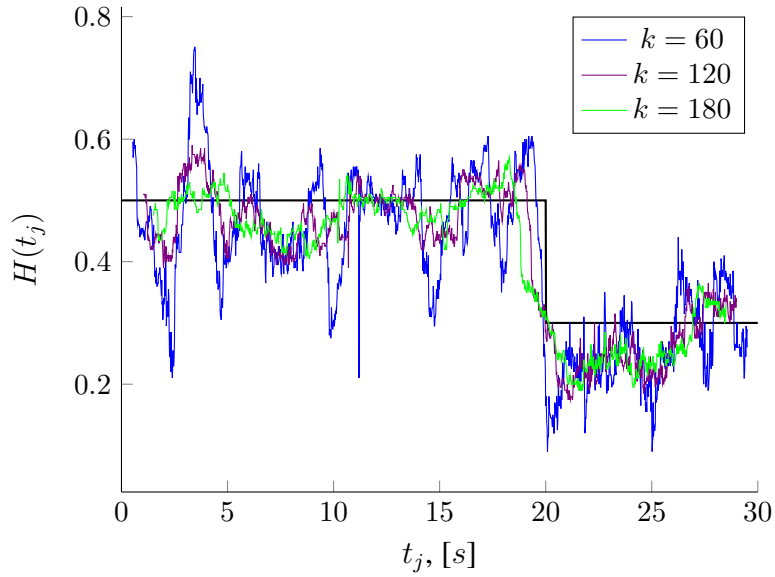
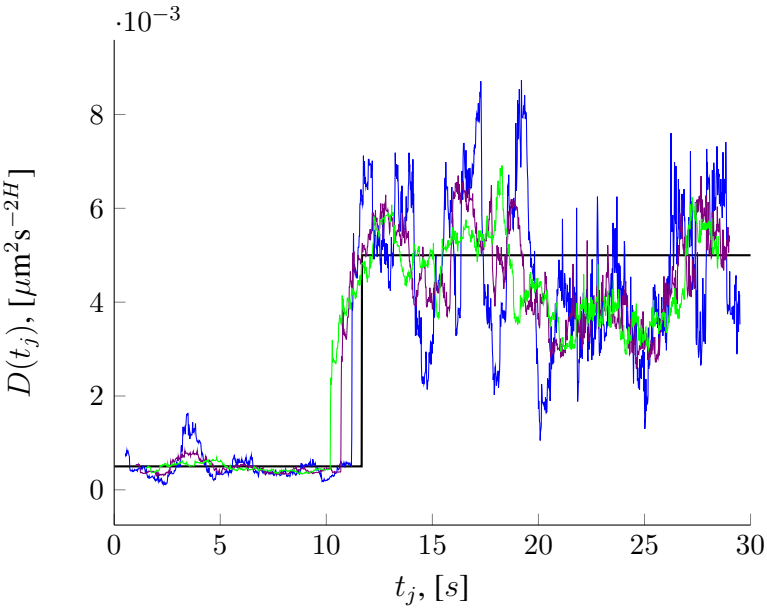


Figure 2.12: Local estimates of H for the path presented in Figure 2.11. The local estimates are computed for overlapping neighborhoods of 60 steps.



(a)



(b)

Figure 2.13: Local estimation of an aBm process with variable D and H for three neighborhood sizes, k. The true value of each parameter is indicated in black.

2.8 Conclusion

In this chapter, we presented five models for the stochastic behavior of a particle undergoing diffusion. The Weyl-fractional Brownian motion (W-fBm) was motivated by diffusion through

homogeneous viscoelastic fluids, a scenario that is poorly described by the standard Brownian model. Analysis of the weighting function for W-fBm showed that noise increments early in the time series received a disproportionate amount of weight compared to noise increments in the middle of the time series, a feature that likely has no physical basis for diffusion processes. W-fBm also exhibits what we termed the Starting Point Problem- the incorporation of memory into the initial increments when none should exist. To address the Starting Point Problem and the weighting of the noise increments, Riemann-Liouville fractional Brownian motion (RL-fBm) was introduced. The weighting function was shown to decrease monotonically, at the expense of non-stationary in the increment process due to the initial lack of memory.

To model the movements of a particle through a heterogeneous viscoelastic fluid, multifractional Brownian motion was introduced, however the model was seen to be ill-suited for physical data due to the discontinuities that arise when switching between regions characterized by different elastic moduli. We introduced the Agency Problem to highlight the way in which information is passed from regions with different properties. In an effort to address the Agency Problem in a manner motivated by the underlying physical processes at work, amnesiac fractional Brownian motion was proposed. Simulation methods, example paths, and an outline to parameter estimation for aBm processes were provided.

Accurately modeling the diffusive movement of particles through heterogeneous viscoelastic fluids is critical for the accurate simulation of particle movement and subsequent investigation of key parameters such as first passage times distributions discussed in chapter 5.

CHAPTER 3
Dealing with Drift

Whereas the previous chapter discussed simulation techniques for diffusive processes given model parameters, here we discuss approaches to estimating those parameters given a particle's diffusive position process. We advocate for a fully parametric description allowing for a direct estimate of the model parameters from the position process as opposed to the mean squared displacement of the position process, the current standard.

3.1 Introduction

Particle tracking microrheology focuses on the movement of tracer particles embedded in the fluid of interest, while "native" microrheology analyzes the motion of native fluid or cellular constituents *in situ* [66–70]. For a thorough review of cell-based microrheology applications, the reader is referred to the excellent reviews by Weihs et al. [35] and Gal et al. [71]. Particle tracking microrheology can be subdivided into passive [12, 23, 72–75] and active [76–78], depending on whether the particles are manipulated by an external force. Two-particle, also known as two-point microrheology [79–82], focuses on the auto- and cross-correlations of particle pairs as a means to screen uniform drift or to screen the modified fluid properties due to chemical interactions of the probe particle with the surrounding fluid. Faithful inference of the linear viscoelastic properties of a fluid via single particle passive microrheology requires that particles are sufficiently dispersed and have neutral chemical interactions. Passive microrheology is also used to infer the diffusive mobility of foreign particles of diverse size and surface chemistry, especially for biological applications such as diffusion of drug carrier particles through mucus barriers. Uniting all of these applications of passive microparticle microrheology is the analysis of the mean squared displacement (MSD) of a tracked particle in order to estimate either the particle's diffusive transport properties or the physical viscoelastic properties of the particle's local environment [1, 4–6, 71]. The analysis of the MSD

in this context stems from the seminal work of Einstein [37] for viscous fluids where the MSD scales linearly with time, and the pre-factor gives a direct inference of the fluid viscosity given the temperature and hydrodynamic diameter of the probe via the Stokes-Einstein relation.

The primary application that motivates this work is the determination of the viscoelastic properties (linear, dynamic storage and loss moduli) of biological soft matter, which directly involves the MSD of tracked particle paths using the Mason-Weitz protocol [1, 73] (discussed in Section 1.1.2). Video microscopy, in combination with passive microparticle tracking (MPT), has been used to explore the physical properties of a wide range of mucus biogels, including cervicovaginal [27, 83, 84], pulmonary [23] and gastrointestinal [85–87] mucus. Often in passive MPT experiments, an observed particle exhibits drift, a persistent, inadvertent, driven motion due to convection, movement of the optical stage [88, 89], or some other source, and is superimposed on the particle's diffusive increment process. In active biological fluids such as living cells where native DNA domains are fluoresced and tracked, the cells may translocate, such as with budding yeast. In viral trafficking within cells, the virus may hijack directed motion along microtubules. Since drift can significantly alter the MSD of a tracked particle, and thereby distort the inference of the viscous and elastic moduli of the particle's local environment [1, 90] as well as distort the inferred mobility, the question arises as to how drift should be accounted for in the analysis of MPT data.

In the case of optical stage drift, each particle in the field of view exhibits the same magnitude and direction of movement. Thus, if enough particles are present, the driven motion may be removed by estimating the ensemble average movement of the particles within the field of view and subtracting this from each particle path [90]. Other scenarios pose a more difficult challenge due to the potential for temporal and spatial heterogeneity in the driving force. In highly heterogeneous biological fluids such as mucus, differential heating caused by the microscope's lighting element may create small-scale convective flow within the fluid volume; regions of high elasticity, due to locally elevated protein concentrations, may cause some particles to appear immobile while neighboring particles are clearly undergoing net transport due to convective flow. In this scenario, if one were to subtract the ensemble-averaged movement of the particles in the field of view from each particle path, one would be subtracting directed motion from the more mobile particles and *adding* directed motion to the less mobile particles. To avoid this issue, it has been suggested that drift should be estimated and removed from each path individually [73].

We point out that the debate over the optimal way to remove drift assumes that directed motion needs to be removed in order to analyze the underlying diffusive process. Historically, this approach is natural because of the need to determine the scaling of the ensemble particle MSD due to purely diffusive dynamics [37, 71, 75, 91], which is often extremely sensitive to drift [92]. In this chapter, we take a different approach and show that deterministic *drift does not need to be removed from the particle path data to determine the MSD statistics* ***if*** *one posits and exploits a fully parametric statistical model for the underlying diffusive process*. This will clearly be the case for diffusion in a simple viscous fluid, but we further show, using numerical simulations, that this is also the case for sub-diffusive processes that have been demonstrated to be accurate models for diffusion in mucus gels [23, 25] and other biological soft matter [93, 94]. That is, we show that for simple diffusion and fractional Brownian motion with a sub-diffusive scaling typical of mucus gels, one can easily estimate the diffusive model parameters from MPT data with drift via a maximum likelihood approach that does not attempt to estimate the MSD directly. From these parameter estimates it is straightforward to generate the MSD of the purely diffusive dynamics and thereby deduce the dynamic viscoelastic moduli by the Mason-Weitz protocol. We illustrate the procedure with numerical simulations for a range of drift components relative to the diffusive mobility. We furthermore show the errors in dynamic moduli if one uses standard ordinary least squares fitting of the MSD with and without removal of the drift.

The structure of this chapter is as follows: first, we discuss drawbacks of a reliance on MSD-based analysis. We then introduce a simple model for particle diffusion accounting for drift and provide details on simulating particle paths in accordance with this model. Next, we review two MSD-based approaches to the recovery of diffusive parameters and present our increment process-based method. These three methods are then compared by applying them to simulated data sets where the drift has been tuned to span experimentally observed values.

3.2 Mean squared displacement

Given observations $X(0), X(\Delta t), X(2\Delta t), ..., X(M\Delta t)$ of a particle's position, the MSD statistic is calculated as,

$$\langle r_\tau \rangle^2 = \frac{1}{M-i+1} \sum_{j=0}^{M-i} [X(i\Delta t + j\Delta t) - X(j\Delta t)]^2 \tag{3.1}$$

where $\tau = i\Delta t$ is known as the lag time and Δt is the inter-observational time. For many diffusive processes, theory and observation suggest that the MSD of particles undergoing diffusion exhibits a power law scaling [4, 11, 23, 26, 95, 96]:

$$\langle r_\tau \rangle^2 = 2D\tau^\alpha \tag{3.2}$$

where the prefactor D is the diffusivity and α is a unitless real number on the interval $[0, 2]$. For a Brownian model, we have $\alpha = 2H$, thus for standard Brownian motion without drift, the power-law exponent is $\alpha = 1$ or, equivalently, $H = 0.5$. However, Weihs et al. [92] rigorously illustrated via simulated Brownian motion that linear (i.e., constant) drift causes a plot of the MSD versus lag time τ to tend toward a slope of 2 at large lag times. That is, as τ increases, $H \to 1$ and the larger the drift velocity, the smaller the lag time at which this transition occurs (Figure 3.1).

When one attempts to correct for directed motion by subtracting the mean displacement, one inadvertently changes the structure of the entire particle path by forcing it to begin and end at the same location in space. The resulting impact is more extreme at longer times, potentially altering one's understanding of the underlying stochastic process. Before showing why subtracting the mean displacement from a particle path has this effect, it is worth recalling that the displacements of a particle diffusing via Brownian motion are normally distributed. If $X_i = X(i\Delta t)$ is the location of such a particle at time $i\Delta t$ and $i = 1, 2, ..., M$, the displacements are given by $x_i = X_{i+1} - X_i$. The distribution of x_i is expected to have mean $\mu\Delta t$ and variance $2D\Delta t$ where μ is the drift velocity. When no drift is present, the mean of x_i, denoted $\bar{x}$, converges to zero as the number of particle positions increases, i.e., as $M \to \infty$. The fact that the distribution of x_i is symmetric with $\bar{x}$ converging to zero intuitively indicates that the particle is expected to make an equal number of steps to the left and right. This however, is *not to say* that a particle diffusing via Brownian motion never

travels a net distance. The net distance is $\bar{x}$ and when we subtract $\bar{x}$ from each displacement, x_i, we center the distribution of x_i at zero, inadvertently stipulating that the first and final positions of the particle are equivalent (Figure 3.2). To see this, consider $\bar{x}$, the mean of x_i,

$$\bar{x} = \frac{1}{M-1} \sum_{j=1}^{M-1} x_j. \tag{3.3}$$

$\bar{x}$ is subtracted from each increment to "remove drift," centering the distribution of displacements at zero. The resulting modified position process is computed by taking the cumulative sum of the shifted increments, denoted by $\widetilde{X}_i$,

$$\begin{aligned}
\widetilde{X}_1 &= X_1 \\
\widetilde{X}_2 &= X_1 + (x_1 - \bar{x}) \\
\widetilde{X}_3 &= X_1 + (x_1 - \bar{x}) + (x_2 - \bar{x}) \\
&\vdots \\
\widetilde{X}_M &= X_1 - (M-1)\bar{x} + \sum_{i=1}^{M-1} x_i.
\end{aligned}$$

The equation for the final position may be simplified further:

$$\widetilde{X}_M = X_1 - (M-1)\frac{1}{M-1} \sum_{i=1}^{M-1} x_i + \sum_{i=1}^{M-1} x_i \tag{3.4}$$

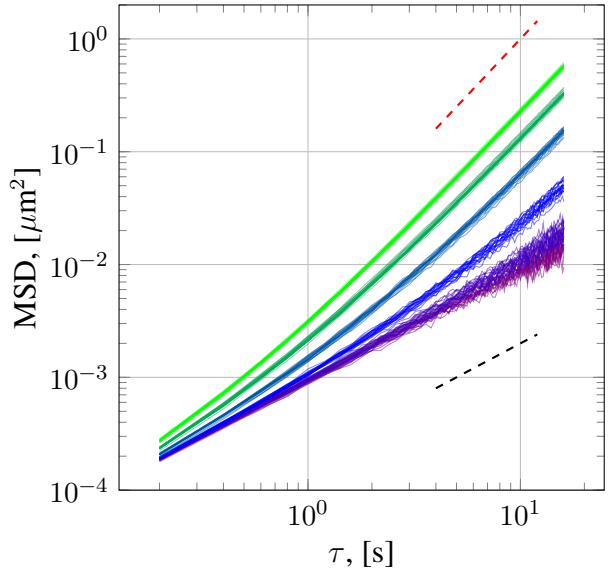
$$= X_1 - \sum_{i=1}^{M-1} x_i + \sum_{i=1}^{M-1} x_i \tag{3.5}$$

$$= X_1, \tag{3.6}$$

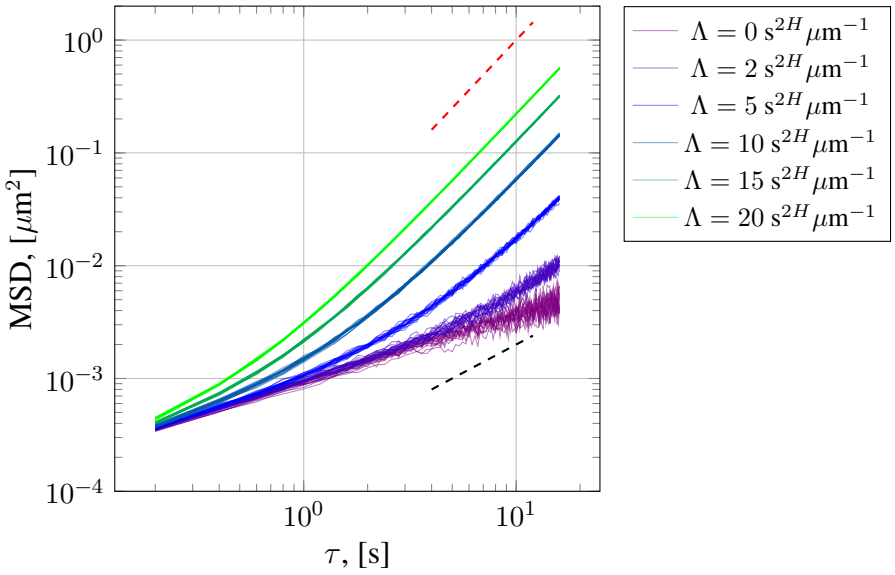
and thus we see the final position is equivalent to the initial position.

An additional drawback of a MSD-based approach to diffusive parameter estimation is the unreliability in the MSD at large lag times. As the lag time increases, the number of displacements included in the mean of the squared displacements decreases and thus becomes less stable. Weihs et al. [92] estimate that only the initial two-thirds of the MSD is statistically reliable. Due to experimental factors limiting the ability to collect data over long time scales, the uncertainty in

the MSD for large lag times can have a pronounced impact on the accurate recovery of diffusion parameters.

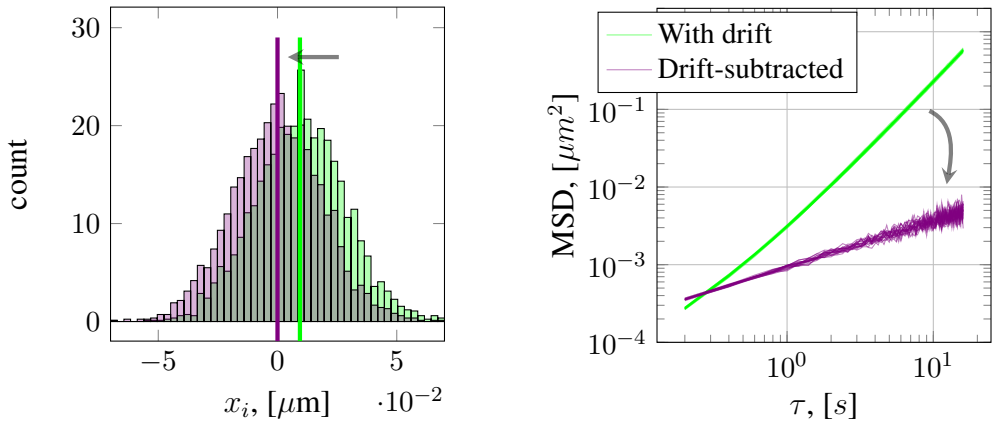


(a)



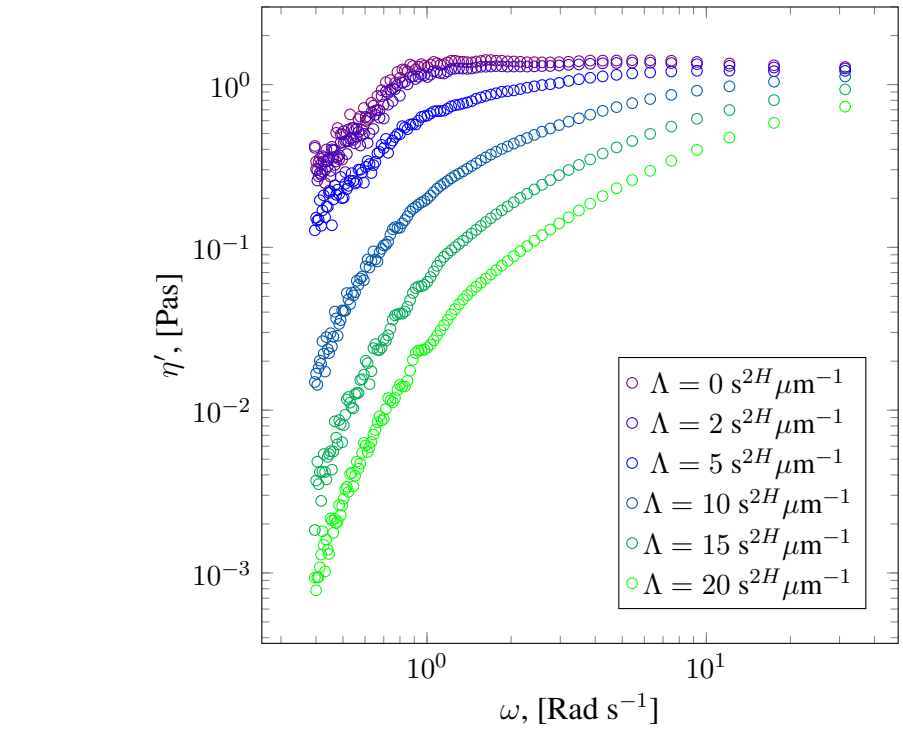
(b)

Figure 3.1: Pathwise MSD for simulated Brownian (a) and subdiffusive (b) particles. Drift is parameterized by the ratio of the drift velocity μ to diffusivity, and is denoted by Λ. Twenty representative MSD curves are shown for each value of Λ. The black and red dashed lines indicate slopes of 1 and 2, respectively. The subdiffusive paths are simulated with $H = 0.3$ and thus appear subdiffusive (slope < 1) for small Λ.

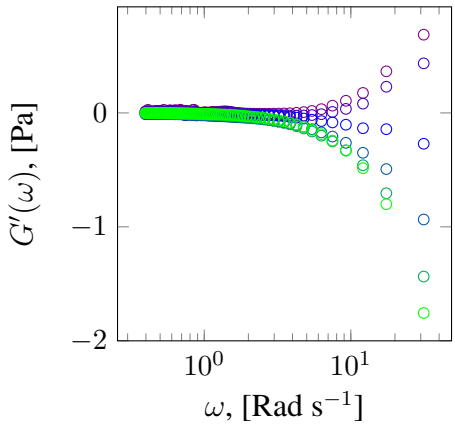


(a) (b)

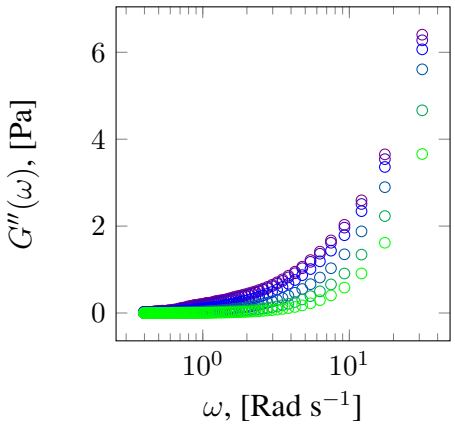
Figure 3.2: Impact of drift-subtraction on the distribution of displacements and MSD. The distribution of displacements (a) is shown at $\tau = 5$ s for a single particle path exhibiting drift of relative magnitude $\Lambda = 20\ s^{2H}\mu m^{-1}$ before (green) and after (violet) drift subtraction. Before drift subtraction, the mean of the distribution of displacements (solid green line) is 9.4×10^{-3} μm, 20 times the particle's diffusivity. Subtracting drift centers the distribution at zero (solid violet line). The MSD is shown for 20 representative paths (b) with drift of relative magnitude $\Lambda = 20\ s^{2H}\mu m^{-1}$ before and after dirft-subtraction. The ensemble average parameter values based on an ordinary least squares fit to the dift-subtracted MSD are $H = 0.299$ and $D = 9.37 \times 10^{-4}$ $\mu\text{m}^2\text{s}^{-2H}$. The true parameter values are $H = 0.3$ and $D = 4.67 \times 10^{-4}$ $\mu\text{m}^2\text{s}^{-2H}$.



(a)



(b)



(c)

Figure 3.3: Complex viscosity, η' (a) and the dynamic storage, G' (b), and loss, G'' (c), moduli for Brownian data for varying levels of linear drift parameterized by Λ.

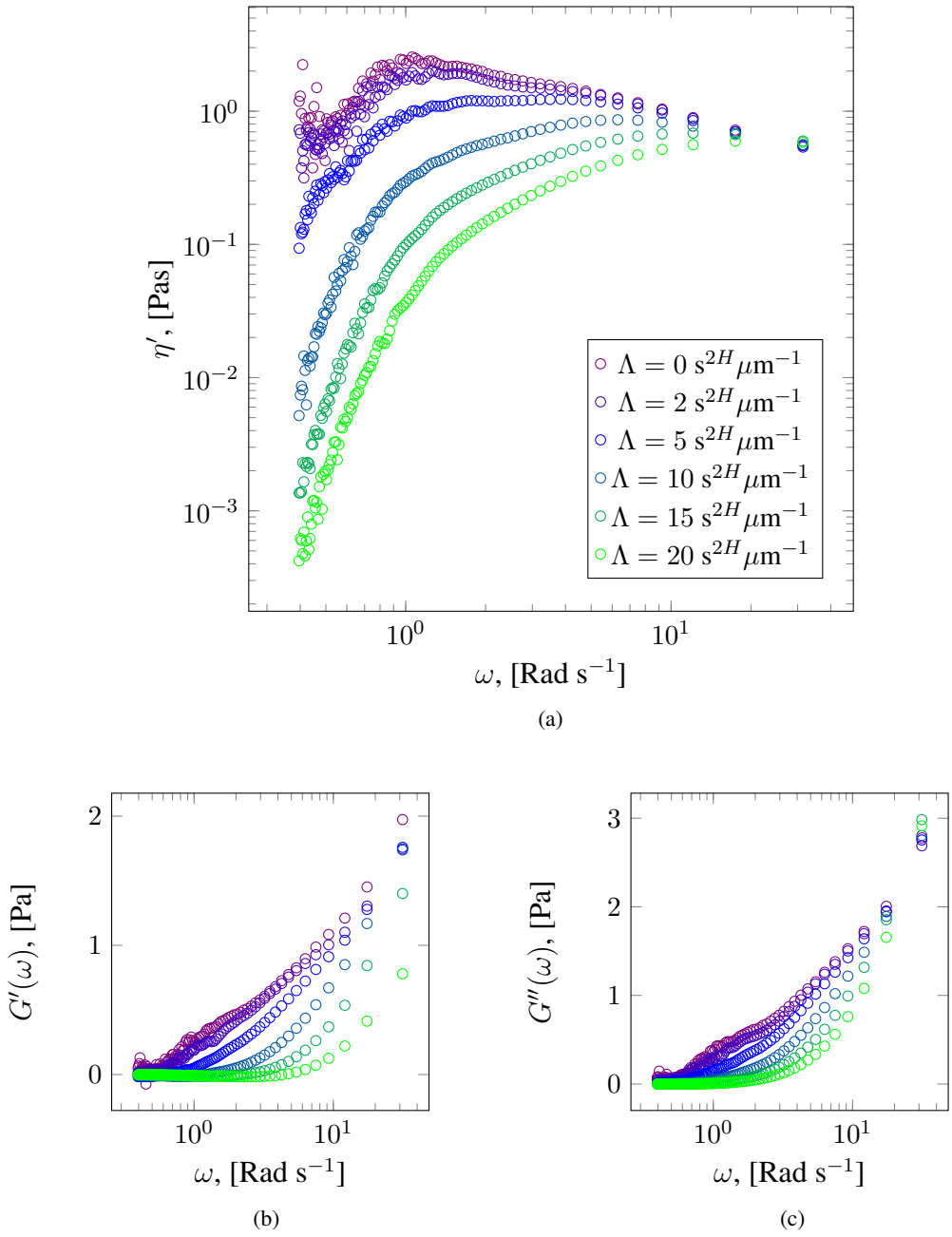


Figure 3.4: Complex viscosity, η' (a) and dynamic storage, G' (b), and loss, G'' (c), moduli for subdiffusive data for varying levels of linear drift parameterized by Λ.

3.3 Fractional Brownian motion and drift

Lysy et al. [25] employed Bayesian model comparison techniques to show that Weyl fractional Brownian motion (W-fBm) and a generalized Langevin equation with long memory spectrum are equally valid models for the movement of micron-scale particles in human bronchial epithelial (HBE) mucus, a known viscoelastic biological fluid [23]. For simplicity, we opt to use a fBm model. Under this model, a particle's increment process $x(t)$ can be written as a net motion μt plus a scaled noise term, $W_H(t)$,

$$x(t) = \mu t + 2DW_H(t). \tag{3.7}$$

The noise term, $W_H(t)$ represents a zero-mean Weyl fractional Gaussian process with initial condition $W_H(0) = 0$. Calculating the discrete displacements x_i provides a simple way to estimate the drift exhibited by a particle since $\mathbf{E}[x_i] = \mu\Delta t$, just as for a Brownian process. To generalize our analysis, we choose to focus not on μ, but on the ratio of the drift to the diffusivity of the particle,

$$\Lambda = \frac{\mu\Delta t}{D}. \tag{3.8}$$

Diffusivity is reported in units of $\mu\text{m}^2\text{s}^{-2H}$, therefore Λ has units $\text{s}^{2H}\mu\text{m}^{-1}$. $\Lambda = 1\ \text{s}^{2H}\mu\text{m}^{-1}$ corresponds to directed movement at each time step equal to the numeric value of the diffusivity.

3.4 Simulation design

To generate a particle path exhibiting linear drift and W-fBm dynamics, we first generate the increment process for W-fBm without drift using the direct algorithm described in Section 2.3.2.1, then add the desired drift to the path. Using this method, two sets of simulated data are generated. The first set is subdiffusive with Hurst parameter $H = 0.3$ and diffusivity $D = 4.67 \times 10^{-4}\mu\text{m}^2\text{s}^{-2H}$, mimicking the estimated parameter values based on experimental observations of 1 μm diameter particles in 4 weight percent (wt%) human bronchial epithelial (HBE) mucus [23]. See chapter 6 for further discussion of HBE mucus. The second data set exhibits standard Brownian motion with $H = 0.5$ and $D = 4.67 \times 10^{-4}\mu\text{m}^2\text{s}^{-1}$. Each simulated path is generated with a temporal resolution of $f = 5$ fps and a length of $M = 2992$ steps, mimicking experimental conditions. Linear drift with

relative magnitude Λ is added to the simulated paths by computing the displacements, adding $D\Lambda$ to each element, then taking the cumulative sum of the result. The fBm position process with linear drift is therefore given by,

$$X_j = \sum_{i=}^{j} x_i + D\Lambda. \tag{3.9}$$

To generate simulated paths with drift equivalent to experimentally observed values, 100 simulated paths are generated for Λ spanning the interval $[0, 20]$ in increments of 0.5 $\mathrm{s}^{2H}\mu\mathrm{m}^{-1}$, resulting in 4,100 simulated fBm paths ($H = 0.3$) and 4,100 simulated Brownian paths ($H = 0.5$). These data sets will be referred to as the fractional Brownian motion (FBM) and Brownian motion (BM) data sets. The MSD as a function of Λ is shown in Figure 3.1. The Generalized Stokes-Einstein relation (1.12) was applied to the observed scaling of the increments and used to calculate the path-wise complex viscosity and dynamic storage and loss moduli. The impact of drift on these parameters is shown in Figures 3.3 and 3.4.

3.5 Approaches to parameter estimation

We consider three approaches to diffusive parameter estimation,

Naïve LS The naïve ordinary least squares (NLS) approach estimates D and H from the scaling of the observed MSD given by (3.1), assuming no drift is present. The goal is to minimize the objective function,

$$\sum_{i=1}^{M} (y_i - c - 2Ht_i)^2, \tag{3.10}$$

for the model parameters c and H where,

$$y_i = \ln[\langle r_\tau \rangle_i^2], \qquad c = \ln[2D], \qquad t_i = \ln[\tau_i]. \tag{3.11}$$

Let $y = (y_1, y_2, ...y_M)$ and $t = (t_1, t_2, ...t_M)$. We assume that the values y_i are normally distributed with mean $f(t_i, \theta)$ and variance δ, where $\theta = (c, H)$. Therefore the probability of observing y_i given $f(t_i, \theta)$ and δ is expressed as

$$P(y_i | f(t_i, \theta), \delta) = \mathcal{N}(y_i | f(t_i, \theta), \delta). \tag{3.12}$$

The likelihood function of observing all the data points y, given all of the inputs t, is

$$\mathcal{L}(t, \theta, \delta|y) = P(y|t, \theta, \delta) = \prod_{i=1}^{M} \mathcal{N}(y_i|f(t_i, \theta), \delta), \tag{3.13}$$

with corresponding log-likelihood function,

$$\log[\mathcal{L}(t, \theta, \delta|y)] = \log[P(y|t, \theta, \delta)] = -\frac{1}{2\delta}\sum_{i=1}^{M}(f(t_i, \theta) - y_1)^2 + \frac{M}{2}\log(\delta^{-1}) - \frac{M}{2}\log(2\pi). \tag{3.14}$$

We recognize that maximizing the above log-likelihood function is equivalent to solving the standard least squares regression problem. If we compute the partial derivatives with respect to θ and set them equal to zero, we get

$$\frac{\partial \log[\mathcal{L}]}{\partial c} = \sum_{i=1}^{M}(c + 2Ht_i - y_i) = 0 \tag{3.15}$$

$$\frac{\partial \log[\mathcal{L}]}{\partial H} == \sum_{i=1}^{M}(c + 2Ht_i - y_i)t_i = 0, \tag{3.16}$$

which can be rearranged into the system of equations

$$\mathbf{T}\theta = \mathbf{Y} \tag{3.17}$$

where,

$$\mathbf{T} = \begin{bmatrix} M & \sum_{i=1}^{M} t_i \\ \sum_{i=1}^{M} t_i & \sum_{i=1}^{M} t_i^2 \end{bmatrix} \quad \text{and} \quad \mathbf{Y} = \begin{bmatrix} \sum_{i=1}^{M} y_i \\ \sum_{i=1}^{M} t_i y_i \end{bmatrix}. \tag{3.18}$$

The maximum likelihood parameter estimates are then given by

$$\widehat{\theta} = (\widehat{c}, \widehat{H}) = (\mathbf{T}'\mathbf{T})^{-1}\mathbf{T}'\mathbf{Y}. \tag{3.19}$$

Drift-subtracted LS The drift-subtracted ordinary least squares approach (DLS) subtracts the mean of the distribution of displacements $\bar{x}$ from each increment x_i, centering the distribution of displacements at zero, before applying the approach described above for naïve estimation.

Full model MLE This approach applies maximum likelihood estimation to (3.7) to estimate μ, D and H directly from the raw data without estimating the MSD statistic. (3.7) specifies that the increments x_i have a multivariate Gaussian distribution,

$$x_i \sim \mathcal{N}(\mu\Delta t, \sigma^2 V_H), \qquad \sigma^2[V_H]_{i,j} = \text{ACF}_x(|i-j|), \tag{3.20}$$

where ACF_x is given by (2.5) and $\sigma = 2D$. The likelihood function is thus $\mathcal{L}(\theta|x) = p(x|\theta)$ and the MLE of the model parameters $\theta = (\mu, \sigma, H)$ is $\widehat{\theta} = \text{argmax}_\theta \mathcal{L}(\theta|x)$. The three-dimensional optimization problem can be reduced to a one-dimensional problem by maximizing in (μ, σ) for fixed H. That is, let

$$y = [V_H]^{-1/2} x, \quad \text{and} \quad z = \Delta t [V_H]^{-1/2} 1_M, \tag{3.21}$$

where $1_M = (1, 1, 1, ...1)$. Moreover, the likelihood function $\mathcal{L}_H(\mu, \sigma|x)$ for fixed H is identical to the one for the linear regression model,

$$y_i = \mu z_i + \epsilon_i \qquad \epsilon_i \sim \mathcal{N}(0, \sigma^2). \tag{3.22}$$

Therefore, the values $(\widehat{\mu}_H, \widehat{\sigma}_H)$ that maximize $\mathcal{L}_H(\mu, \sigma|x)$ are

$$\widehat{\mu}_H = \frac{\sum_{i=1}^M z_i y_i}{\sum_{i=1}^M z_i^2}, \qquad \widehat{\sigma}_H = \left(\frac{\sum_{i=1}^M (y_i - \widehat{\mu}_H z_i)^2}{M} \right)^{1/2}. \tag{3.23}$$

The MLE of H for (3.7) is then obtained by maximizing the one-dimensional profile likelihood function

$$\mathcal{L}_{\text{prof}}(H|x) = \mathcal{L}((\widehat{\mu}_H, (\widehat{\sigma}_H, H|x). \tag{3.24}$$

Specifically, we have $\widehat{H} = \text{argmax}_H \ell_{\text{prof}}(H|x)$ where the profile log-likelihood function is

$$\ell_{\text{prof}}(H|x) = \log\left(\mathcal{L}_{\text{prof}}(H|x)\right) = -1/2 \left[M \log(\widehat{\sigma}_H^2) + \log(|V_H|) \right]. \tag{3.25}$$

We note that for arbitrary variance matrix V, the linear systems in (3.21) are obtained in $\mathcal{O}(M^3)$ operations. However, since V_H is a Toeplitz matrix [97], the systems can be solved in $\mathcal{O}(M^2)$ operations using the Durbin-Levinson algorithm [98, 99].

Much like the least squares (LS) approach involving the sample MSD, the maximum likelihood approach we have described hinges on a linear regression approach. However, whereas the LS approach estimates the drift only once, the MLE estimates the "optimal" drift and diffusvity for every value of H. That is, the LS estimate of the drift by $\bar{x}$ would be optimal if the increments were uncorrelated whereas the MLE approach estimates the drift by a weighted average of the increments $\widehat{\mu}_H$ that accounts for their correlation.

3.6 Results

3.6.1 Simulated data

For each simulated path, we compute the path-wise MSD given by (3.1). To estimate the viscous and elastic moduli, we follow the work of Mason and Weitz and transform the MSD using the generalized Stokes-Einstein relation given by (1.12). When comparing the moduli of the simulated data, we consider three scenarios: (1) application of the GSER to the observed MSD without accounting for drift, i.e., the Naïve approach, (2) application of the GSER to the observed MSD after centering the distribution of displacements, i.e., the Drift-subtraction approach, and (3) application of the GSER to the MSD calculated from the maximum likelihood estimates of D and H *assuming* (3.2) *holds*, i.e., the Parametric approach. This key assumption is afforded by the use of a fully parametric model for the diffusive dynamics.

Figures 3.5 and 3.6 show the estimates of the dynamic storage, G', and loss, G'', moduli for the BM and FBM data sets when $\Lambda = 20\ \mathrm{s}^{2H}\mu\mathrm{m}^{-1}$. For both the BM and FBM data sets, the Naïve approach, ignoring the presence of drift, results in highly erroneous predictions of G' and G''. When the drift is subtracted by centering the distribution of displacements, the predicted moduli more closely match the true moduli, however at low frequencies the estimations are corrupted by noise. Applying the GSER to the MSD predicted by (3.2), as opposed to the observed scaling of the MSD, produces a more consistent estimate of each moduli across the frequency spectrum.

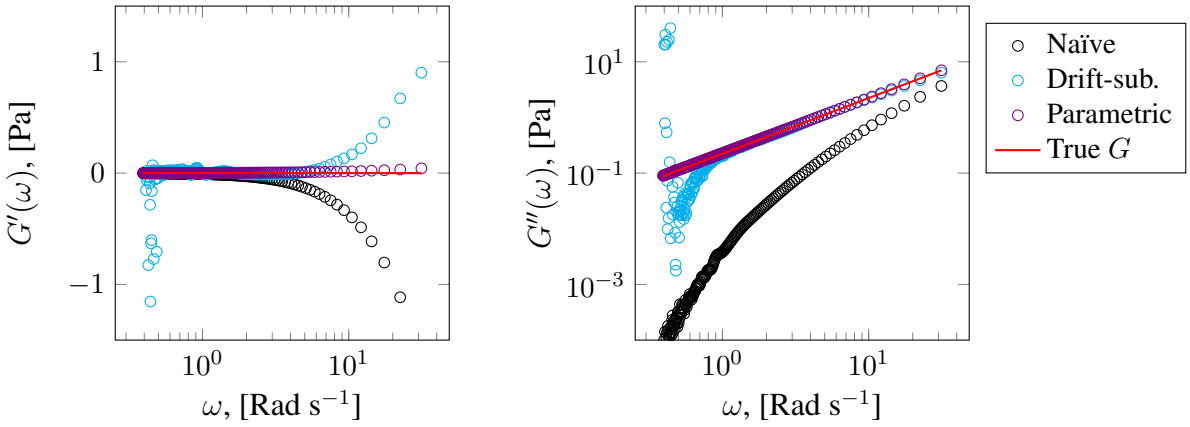


Figure 3.5: Dynamic storage, G' (a), and loss, G'' (b), moduli for the BM data set without drift subtraction (Naïve), with drift subtraction (Drift-sub.) and based on the parametric scaling of the MSD (Parametric). Ensemble averaged results over 100 paths are shown for $\Lambda = 20$ sμm^{-1}.

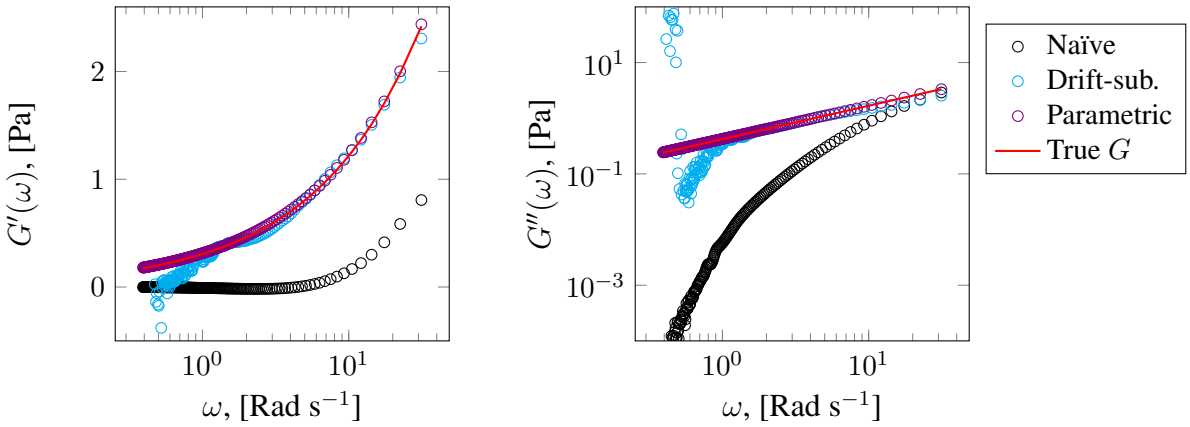


Figure 3.6: Dynamic storage, G' (a), and loss, G'' (b), moduli for the FBM data set without drift subtraction (Naïve), with drift subtraction (Drift-sub.) and based on parametric scaling of the MSD (Parametric). Ensemble averaged results over 100 paths are shown for $\Lambda = 20$ sμm^{-1}.

The errors in the estimation of G' and G'' for the BM data set are reported in Figures 3.7 and 3.8. For G'', the relative error is reported as a function of frequency ω, i.e., the difference between the estimated and true value of G'', divided by the true value of G'' at each frequency. For G', an absolute error is reported as a function of ω since, for Brownian motion, the expected value of the storage modulus is zero.

Figure 3.7 contains three panels with each panel corresponding to one of the three approaches to estimating G', Naïve, Drift-subtracted and Parametric. In each panel, all of the data from the three

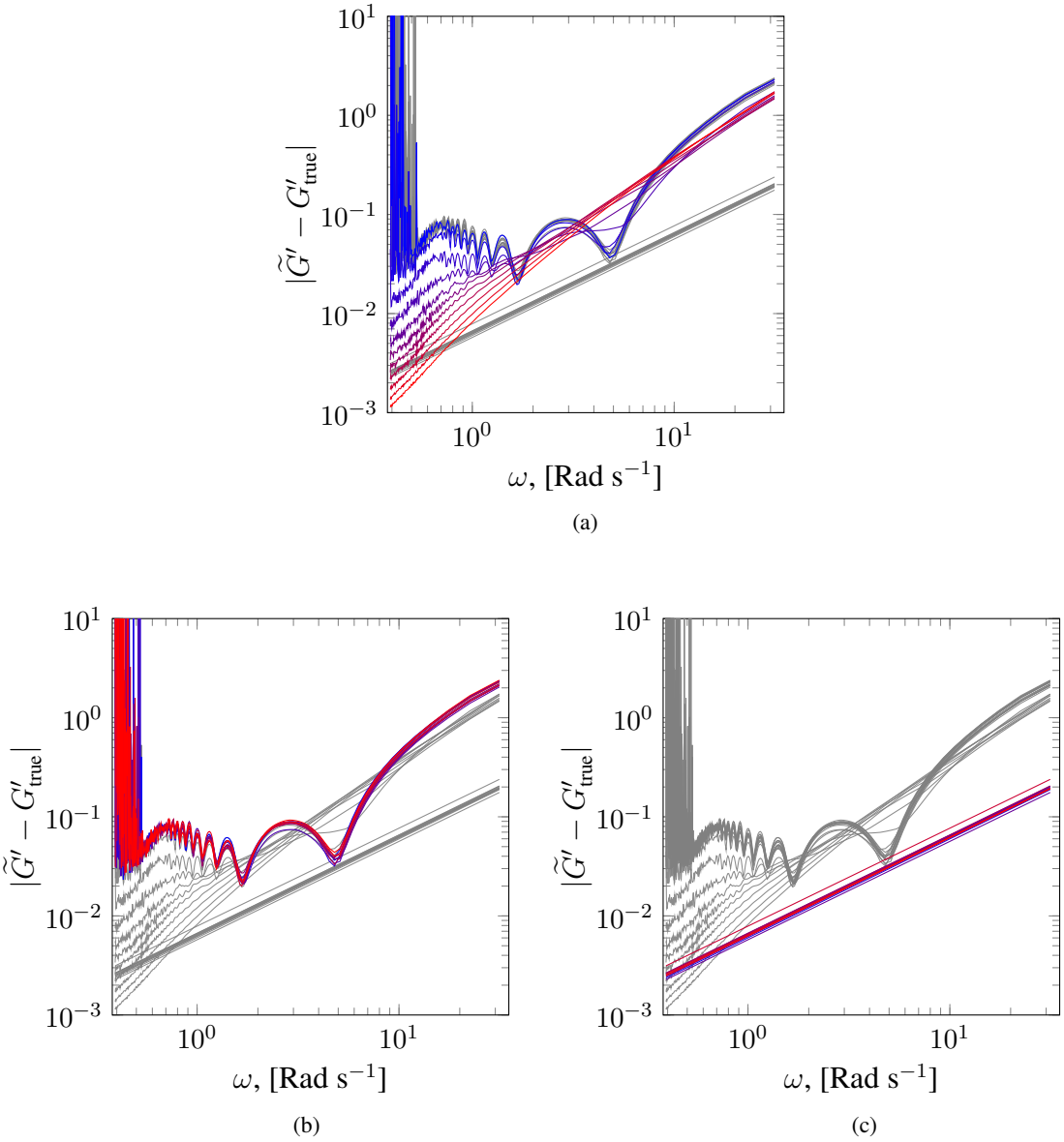


Figure 3.7: Absolute error in the estimation of G' from the BM data set using the Naïve (a), Drift-subtracted (b) and Parametric (c) approaches. The data are shadded from blue, representing no drift ($\Lambda = 0$ sμm^{-1}), to red, representing maximum drift ($\Lambda = 20$ sμm^{-1}).

estimation approaches are shown, however *only data from one of the three approaches is presented in color*. In Figure 3.7(a) for example, the error resulting from the Naïve approach is presented in color, while, to facilitate comparisons, the error from the Drift-subtracted and Parametric data are presented in gray. The color of the Naïve data ranges from blue, corresponding to no drift, i.e., $\Lambda = 0$ s$^{2H}\mu$m^{-1}, to red, corresponding to maximum drift, i.e., $\Lambda = 20$ s$^{2H}\mu$m^{-1}. It is apparent from Figure 3.7(a) that the error in the estimation of G' is a function of Λ. The data also indicate

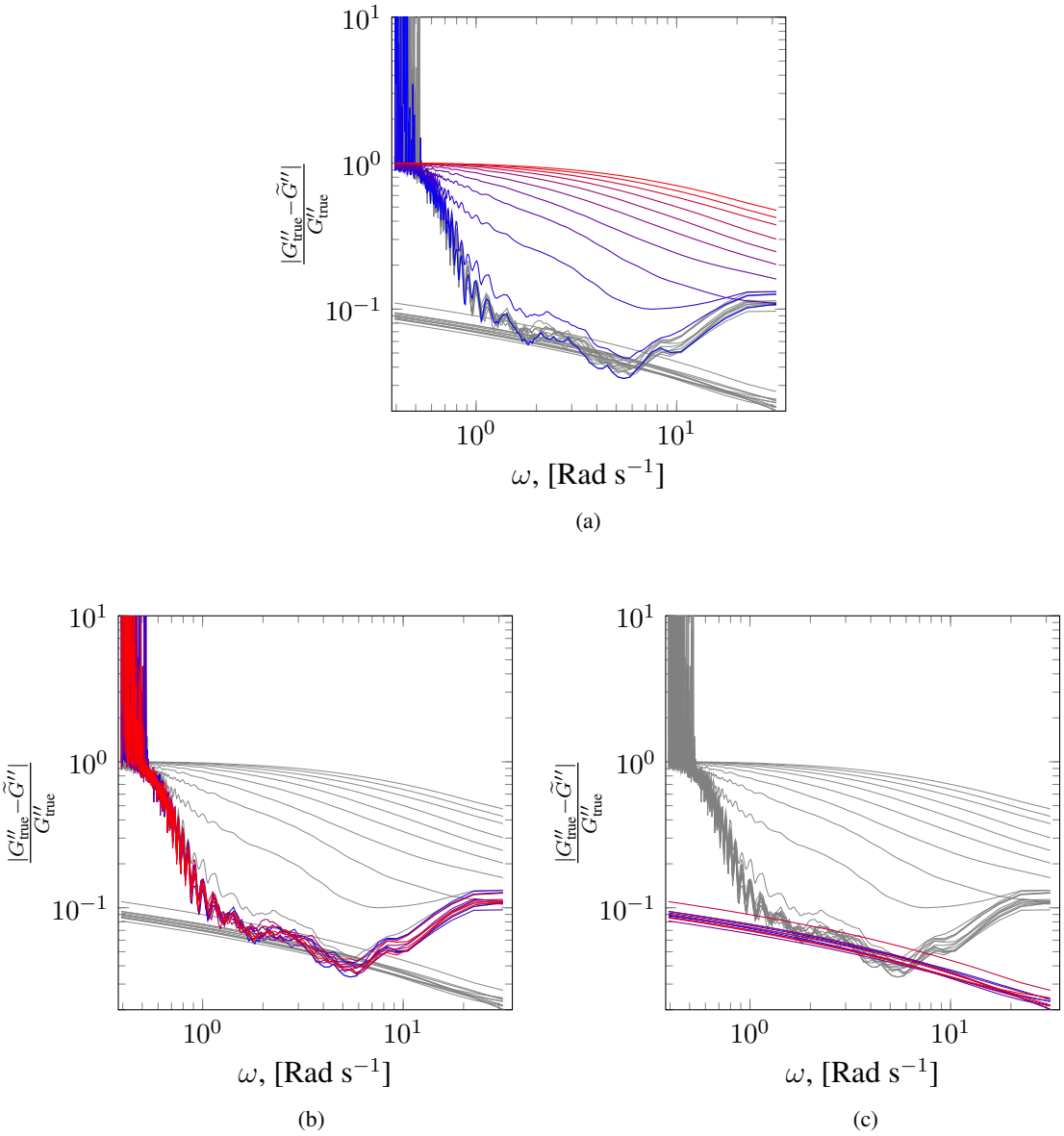


Figure 3.8: Relative error in the estimation of G'' from the BM data set using the Naïve (a), Drift-subtracted (b) and Parametric (c) approaches. The data are shadded from blue, representing no drift ($\Lambda = 0$ sμm^{-1}), to red, representing maximum drift ($\Lambda = 20$ sμm^{-1}).

that for small Λ, error at low frequencies can be significant. As the amount of drift increases, i.e., as Λ gets larger, signified by the transition in color from blue to red, the amount of error at low frequencies decreases. In general, the amount of error resulting from the Naïve estimation of G' is directly proportional to frequency and inversely proportional to drift.

Error in the Drift-subtracted and Parametric estimations of G' from the BM data set are highlighted (in color) in panels (b) and (c) of Figure 3.8, respectively. For both approaches, the distribution

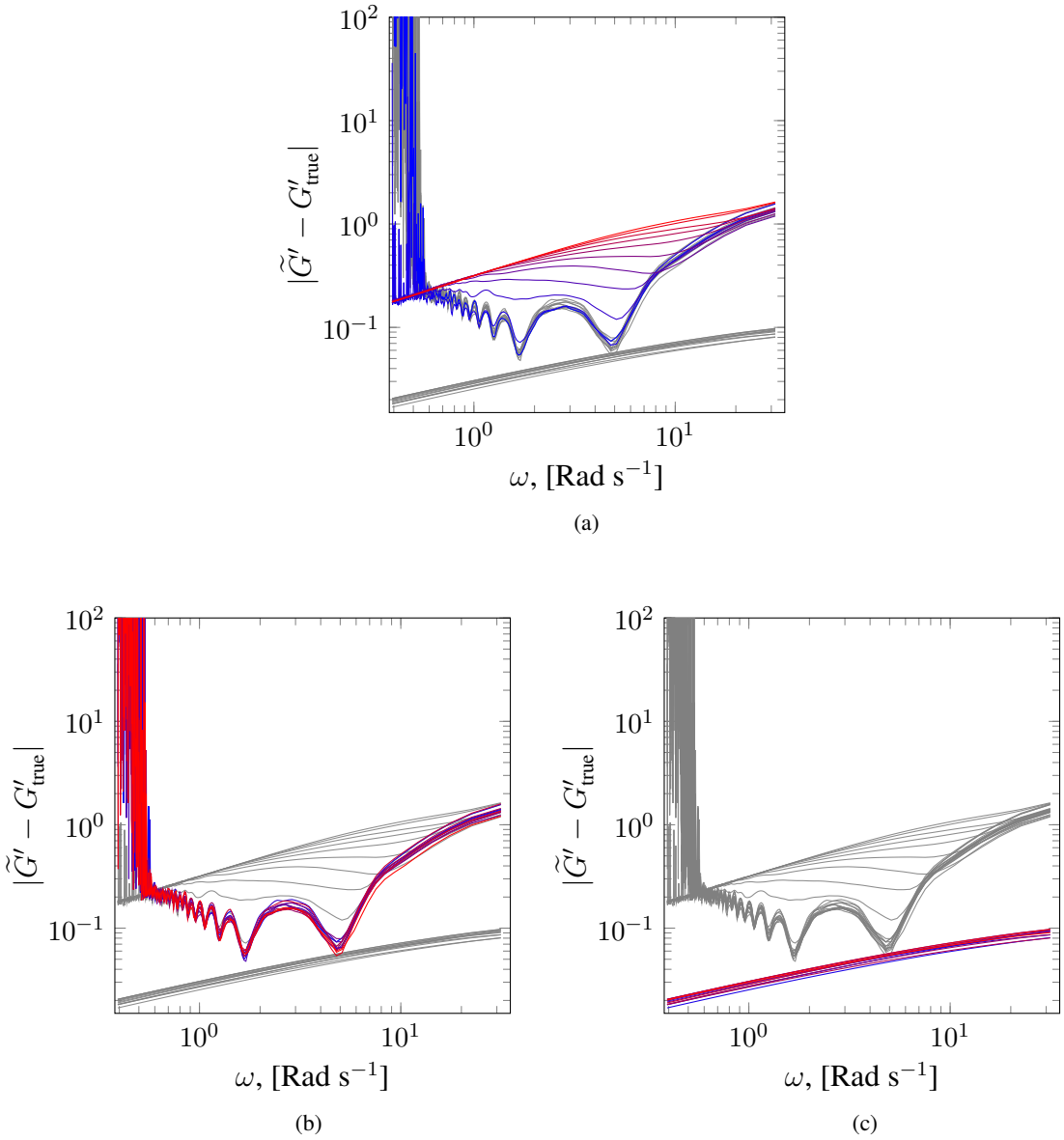


Figure 3.9: Absolute error in the estimation of G' from the FBM data set using the Naïve (a), Drift-subtracted (b) and Parametric (c) approaches. The data are shadded from blue, representing no drift ($\Lambda = 0$ s$^{2H}\mu$m^{-1}), to red, representing maximum drift ($\Lambda = 20$ s$^{2H}\mu$m^{-1}).

of blue and red data indicate that the error in G' is independent of Λ. Similar to the Naïve method, the Drift-subtracted method exhibits significant error at low frequencies and a complicated depend on ω over the observed frequency range. By contrast, the Parametric method offers a stable prediction of G' with error linearly related to frequency.

In Figure 3.8, the relative error in the estimation of G'' is reported for each method. Once again, each panel contains all of the data, but only data from one approach is presented in color. The error in

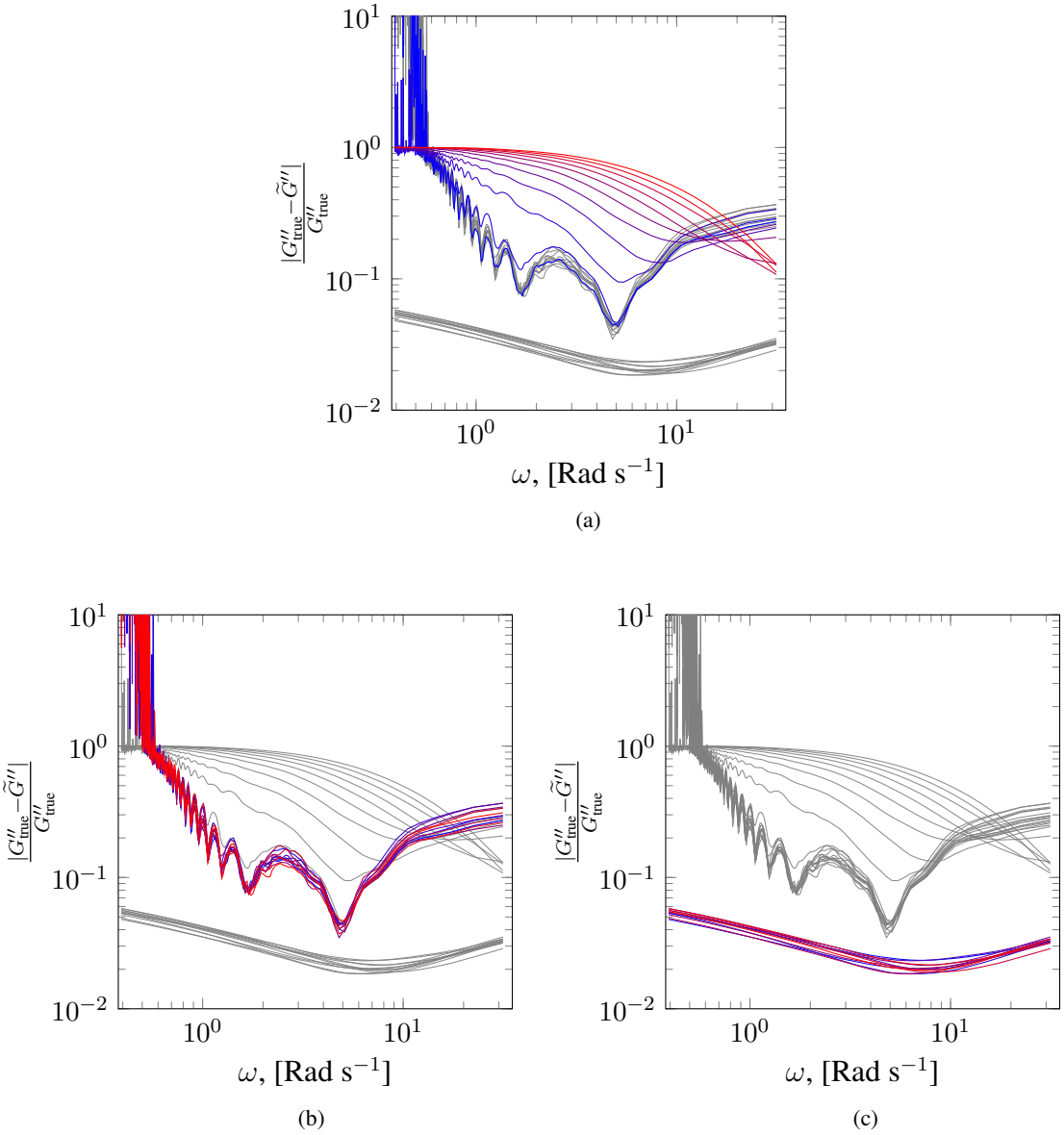


Figure 3.10: Relative error in the estimation of G'' from the FBM data set using the Naïve (a), Drift-subtracted (b) and Parametric (c) approaches. The data are shadded from blue, representing no drift ($\Lambda = 0$ s$^{2H}\mu$m^{-1}), to red, representing maximum drift ($\Lambda = 20$ s$^{2H}\mu$m^{-1}).

the Naïve estimates of G'' is a complex function of Λ and, similar to the error in G', is extreme at low frequencies when no drift is present. Increasing drift mitigates low frequency error at the expense of increased error over the rest of the frequency spectrum. Error in the middle of the frequency spectrum for the Drift-subtracted approach is approximately 10%, however more extreme values are observed at low frequencies. While the absolute error in G' is directly proportional to ω for the Drift-subtracted and Parametric approaches, an inverse relationship is generally observed for G''.

The absolute and relative error in G' and G'' based on the three estimation approaches is presented in Figures 3.9 and 3.10 for the FBM data set. As expected, only the error from the Naïve approach is impacted by drift.

The Naïve Least Squares (NLS), Drift-subtracted Least Squares (DLS) and full model maximum likelihood (MLE) parameter estimation approaches described in Section 3.5 were applied to each simulated data set. Estimations of the Hurst parameter H as a function of drift are shown in Figure 3.11 and the corresponding estimates of the diffusivity D are presented in Figure 3.12. The NLS estimation of H and D is strongly impacted by drift. As Λ increases, H converges to 1, as expected based on the simulation results presented in Figure 3.1. The NLS estimate of the diffusivity has a non-linear dependence on Λ, initially underestimating, and later overestimating D. In contrast, both the DLS and MLE estimates of D and H are independent of Λ and exhibit a similar level of accuracy, however MLE is the more precise estimator.

Using the Naïve, Drift-subtracted and Parametric approaches, the relative error in the recovery of the fluid's viscosity was calculated via the Stokes-Einstein relation (1.1) and Generalized Stokes-Einstein relation (1.12) (Figure 3.13 and 3.14). Although both the Drift-subtracted and Parametric approaches are independent of Λ, the Parametric approach is able to recover the fluid's viscosity with approximately one-third the error of the Drift-subtracted approach.

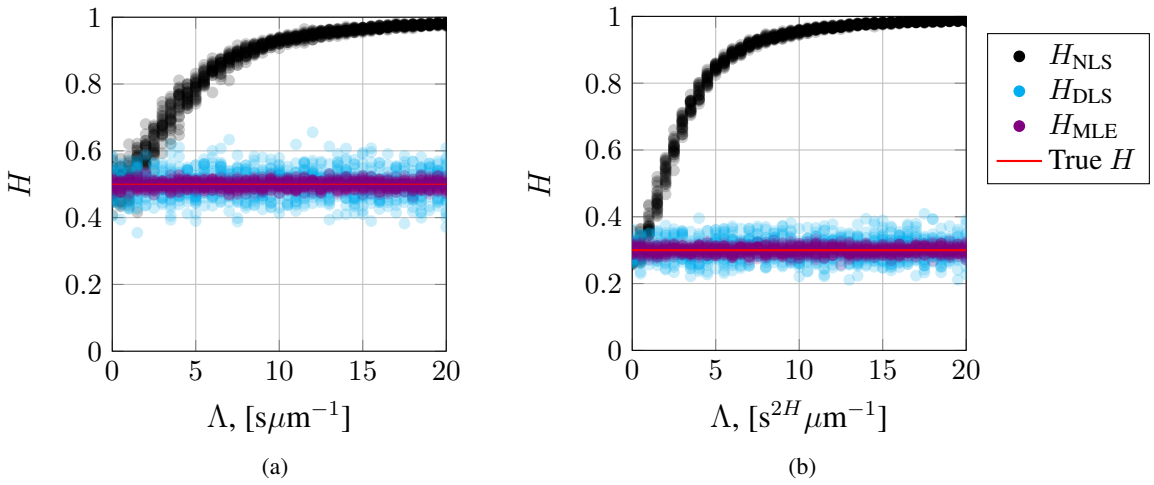


Figure 3.11: Estimated values of H using the Naïve Least Squares (H_{NLS}) and Drift-subtracted Least Squares (H_{DLS}) method applied to the MSD and the parametric MLE (H_{MLE}) method applied directly to the increment process. Estimated values are shown as a function of Λ for the BM (a) and FBM (b) data sets. The horizontal red line indicates the value of H used to generate the simulated paths in each data set.

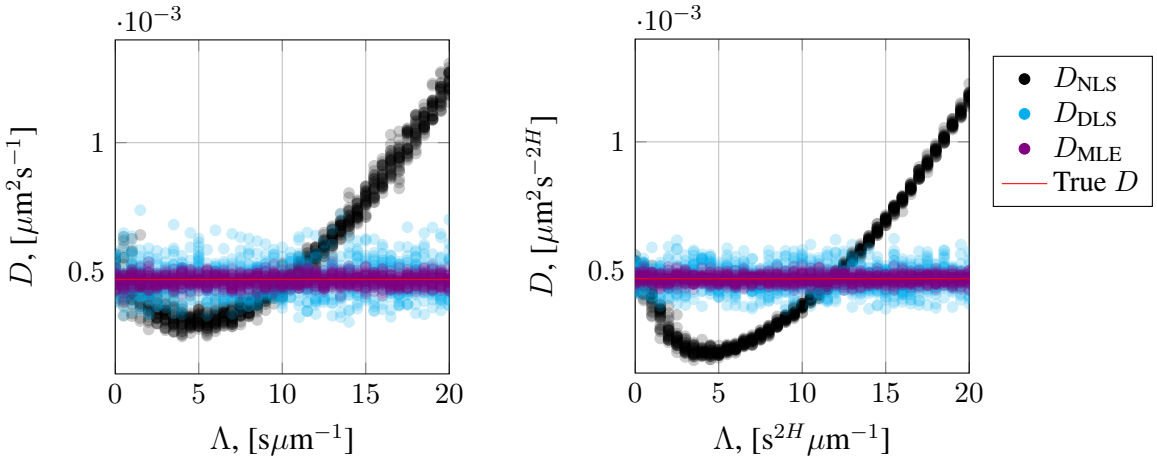


Figure 3.12: Estimated values of D using the Naïve Least Squares (D_{NLS}) and Drift-subtracted Least Squares (D_{DLS}) method applied to the MSD and the parametric MLE (D_{MLE}) method applied directly to the increment process. Estimated values are shown as a function of Λ for the BM (a) and FBM (b) data sets. The horizontal red line indicates the value of D used to generate the simulated paths in each data set.

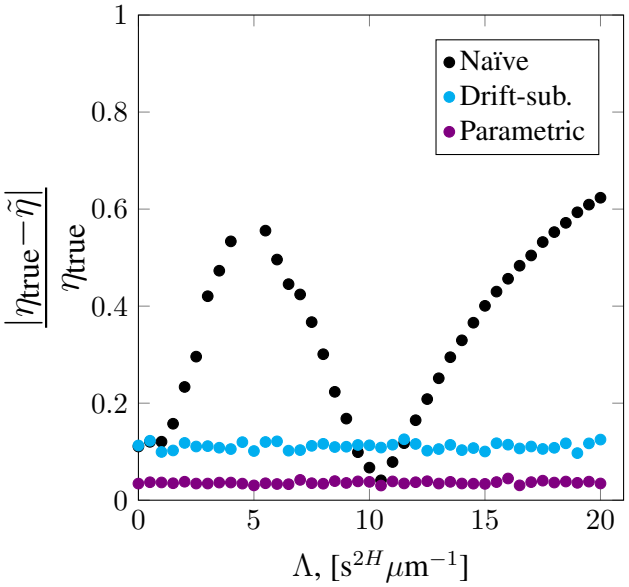


Figure 3.13: Relative error in estimates of η given by the Stokes-Einstein equation for the BM data set. The mean error for each approach is Naïve- 39.5%, Drift-subtracted- 11.1% and Parametric- 3.6%.

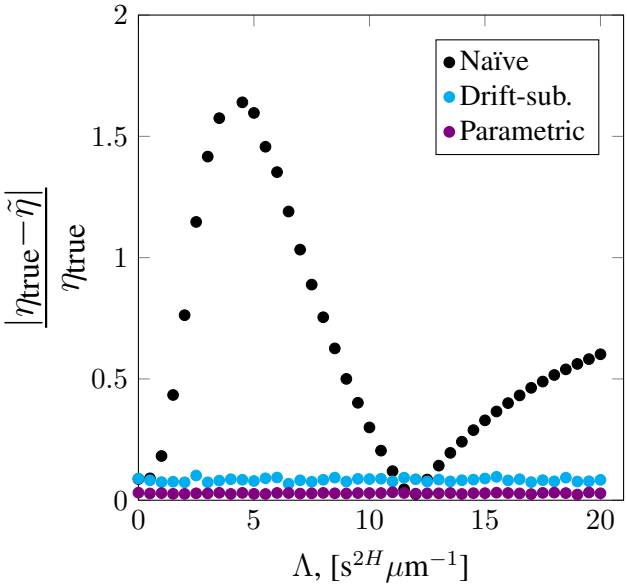
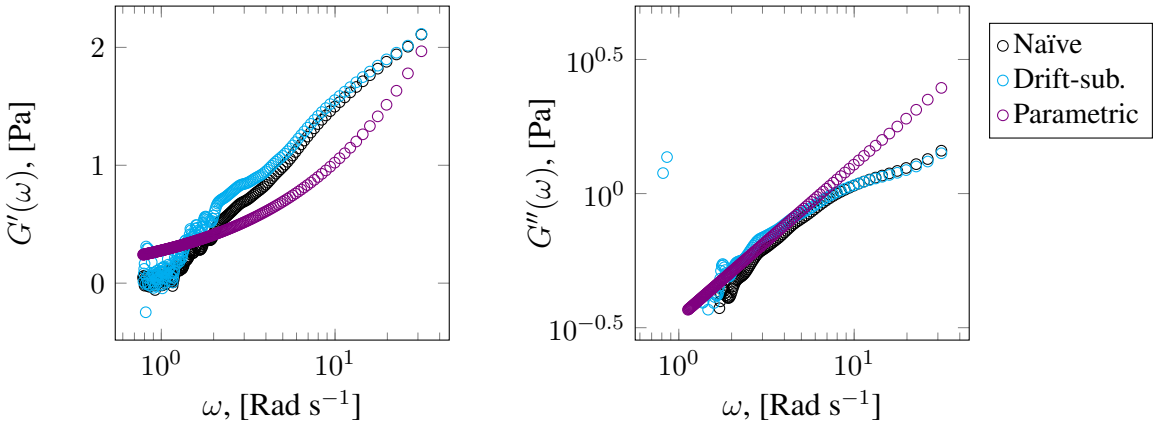


Figure 3.14: Relative error in estimates of η given by the Stokes-Einstein equation based on the three approaches for the FBM data set. The mean error for each approach is Naïve- 65.9%, Drift-subtracted- 8.34% and Parametric- 2.9%.

3.6.2 Experimental data

In this section, we analyze twenty-two representative 1 μm diameter particle paths in 4 wt% human bronchial epithelial (HBE) mucus, a fluid previously shown to exhibit viscoelastic properties [23]. The ensemble-averaged predicted storage and loss moduli based on the Naïve, Drift-subtracted and Parametric approaches all confirm the presence of viscoelastic characteristics (Figure 3.15). Relative to the Parametric estimate, the Naïve and Drift-subtracted approaches indicate elevated elasticity over the majority of the frequency spectrum and depressed viscosity at high and low frequencies.

We now turn to the estimation of H and D, and consider the ratio of the NLS and DLS predictions versus the MLE predictions for each parameter. In Figure 3.16, the estimated values of H are presented in panel (a). The MLE estimates of H are shown along the x-axis and the NLS and DLS estimates are shown on the y-axis. The domain of each axis is $[0, 0.5]$ and a dashed line indicates the diagonal. Each data point represents the predicted parameter value for one of the 22 paths. A data point on the diagonal indicates that the MLE and NLS (or DLS) estimates agree for that experimental particle path. A data point above the diagonal indicates that the NLS (or DLS) method *overestimated*



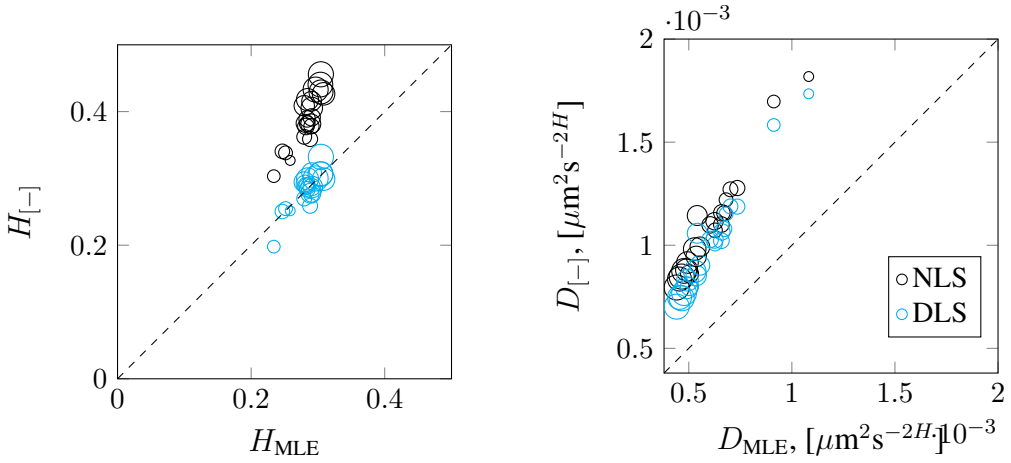
(a) (b)

Figure 3.15: Ensemble average estimates of the dynamic storage, G' (a), and loss, G'' (b), moduli for twenty-two 1 μm diameter particles in 4 wt% mucus.

H compared to the MLE approach. Conversely, a data point below the diagonal indicates the NLS (or DLS) method *underestimated* H compared to the MLE approach. The size of each data point is directly proportional to the drift exhibited by the corresponding particle.

In Figure 3.16(a) we see the DLS and MLE approaches exhibit strong agreement in their predictions of H, as evidenced by the distribution of the data points (cyan) along the diagonal dashed line. In contrast, the NLS estimates of H (black) fall above the diagonal, indicating an overestimation of H relative to the MLE estimates. The amount of overestimation is directly proportional to the amount of drift in the experimental data (larger markers are further from the diagonal than smaller markers).

Figure 3.16(b) presents the estimations of the diffusivity D. All data points fall above the diagonal, thus both the NLS and DLS approach estimate larger values of D compared with the MLE approach. Here, the amount of overestimation is inversely proportional to the amount of drift (larger markers are closer to the diagonal). We note that this is *not* a scenario observed in the simulated data. Returning to Figure 3.12, we see that the only time the NLS estimates of D are in increasing correspondence to the MLE estimates with increasing drift is when the NLS method *underestimates* D and Λ is approximately between 5 and 10 s$^{2H}\mu$m^{-1}. Furthermore, according to Figure 3.12, the DLS estimate of D relative to the MLE estimate should be independent of Λ. We hypothesize these incongruencies between the simulated and experimental data may be the result of non-linear drift.



(a) (b)

Figure 3.16: Ratio of the NLS and DLS predictions to the MLE prediction of the Hurst parameter H (a) and diffusivity D (b) for 1 μm diameter particles in 4wt% mucus. The size of the marker is proportional to the amount of drift exhibited by each particle.

3.7 Conclusion

Persistent linear drift over the course of a particle path is compounded at large lag times, resulting in a bending of the MSD toward a slope of 2, i.e., $H \to 1$. Directly estimating D and H from the scaling of the MSD using standard least squares (LS) estimation techniques without accounting for drift is highly erroneous. As the magnitude of the relative drift Λ increases, the slope of the path-wise MSD increases to 2 at increasingly smaller lag times, causing the LS estimate of H to converge to 1. By subtracting the mean displacement of each particle from the particle's path, thereby centering the distribution of displacements at zero, the LS estimate of each parameter is more stable, however the correlation in the increment process caused by the calculation of the displacements over increasingly large windows is not sufficiently addressed and error in the estimation of the diffusive parameters is on the order of 10%.

To address this issue, we advocate for a parametric maximum likelihood based approach that incorporates a parameter μ while fitting D and H. We demonstrate on numerically generated particle paths with physically relevant diffusive parameters that the use of the parametric maximum likelihood approach recovers H and D with approximately 1/3 of the error of the drift-subtracted least squares approach. Furthermore, this gain in accuracy has been demonstrated for both standard Brownian dynamics as well as fractional Brownian dynamics typical of many experimentally observed

subdiffusive processes. Finally, we applied each approach to experimental data obtained from 1 μm diameter particles in 4wt% HBE mucus and considered the relative magnitude of each approaches estimated parameter value. We observe a divergence between the maximum likelihood and naïve estimates of H with increasing drift, while the maximum likelihood and drift-subtracted estimates exhibit strong agreement independent of drift. A more complex relationship is observed in the relative estimates of D. Here, the experimental data suggest that the correspondence between the parameter estimation techniques is directly proportional to the amount of drift.

While we have only considered linear drift, the proposed approach may be easily adapted to non-linear drift by replacing the drift parameter μ with an appropriate functional form. Overall, the proposed method demonstrates promise for the more accurate recovery of diffusive parameter values in the presence of drift in both viscous and viscoelastic fluids.

CHAPTER 4

Methods for the Quantification of Heterogeneity

Many biological fluids are known to exhibit heterogeneous properties. In this chapter we present an algorithm adapted from the Machine Learning community for the analysis of path data from passive particle tracking microrheology that yields a quantitative characterization of diffusion in heterogeneous complex fluids. The utility of this algorithm is demonstrated on both simulated and experimental data[1].

4.1 Introduction

Passive particle tracking microrheology (PPTM), whereby the diffusive movements of tracer particles are used to infer the viscous and elastic moduli of the local environment, is an increasingly common method of fluid characterization [1]. Traditionally, the analysis of PPTM data relies on the ensemble averaging of individual path statistics derived from the observed increment process. For homogeneous fluids, this practice is admissible since, if all particles are probing the same environment, the increment process will follow a Gaussian distribution. The material parameters are then inferred from the scaling of the variance with lag time of this distribution. Spatial heterogeneity however, is often found in biological fluids across multiple length scales. When spatial heterogeneity is present, different particles probe different environments, and although the step-size distribution of individual particles follows a Gaussian distribution (if it was sampling a homogeneous region), the ensemble distribution of displacements for a collection of particles will be non-Gaussian. Heterogeneity can be measured by the deviation of the ensemble distribution of displacements from a Gaussian and several metrics exist within the literature based on this idea. In this chapter we present methods adapted

[1]This chapter is based on an article in the journal Soft Matter. The original citation is as follows: Mellnik J., Vasquez, P. A., McKinley, S. A., Witten, J., Hill, D. B. & Forest, M. G. Micro-heterogeneity metrics for diffusion in soft matter. *Soft Matter* **10**, 7781-7796 (2014) - Reproduced by permission of The Royal Society of Chemistry `http://pubs.rsc.org/en/content/articlelanding/2014/sm/c4sm00676c#!divCitation`.

from the statistics and machine learning communities that simultaneously detect heterogeneity and divide the path data into clusters based on the distribution of the statistic of choice.

We are also interested in predictive consequences of heterogeneity beyond the timescales of experimental observations, which requires a final model fitting step. For most biological soft matter, which, unlike simple viscous fluids, posses viscoelastic relaxation modes and thereby exhibit memory in the diffusive path data, there are few models for which a rigorous diffusive transport theory has been derived from first principles. The list shortens if one requires that the MSD scaling behavior and other statistical properties are exactly solvable. The rare model systems with these criteria include the Rouse model for dilute, monodisperse polymer melts, and the Zimm model that couples solvent hydrodynamic interactions to the Rouse model. Rubinstein and Colby [100] and Cai et al. [101] provide excellent discussions of scaling behavior associated with models for semi-dilute and entangled polymers. These first-principles models yield anomalous sub-diffusive, MSD scaling behavior with exponenets $1/2$ and $2/3$ on intermediate timescales, followed by convergence to simple diffusion and MSD power-law exponent of 1 at long time scales.

Complex biological fluids are typically composed of diverse molecular constituents, exhibiting active and repulsive interactions. Electrostatic interactions between tracer particles and soft matter can significantly alter particle diffusion [102]. This observation has been extensively explored in the context of particle-based drug delivery through mucus barriers in the lung [103]. For such biological soft matter systems, there is no rigorous theory to guide model selection beyond the ideal systems noted above, whereas PPTM data in biological fluids such as pulmonary mucus [23] yields MSD power-law exponents that span the entire interval $[0, 1]$.

Thus, until such time that a rigorous theory exists of diffusive properties of complex biological fluids and the effects of probe-fluid interactions, even for homogeneous complex fluids, the analysis of the particle path data must be performed by statistical methods with minimal assumptions of the underlying models to discern among different fluids and different particles in a given fluid. That is the perspective taken here in regard to the first two goals of heterogeneity detection among the ensemble of paths and clustering of the paths.

There are, nonetheless, *ad hoc* stochastic models that share several key features of the PPTM data in biological fluids. These include fractional Brownian motion [104] and generalized Langevin equations with special memory kernels [1, 105–107]. The proper statistical approach, given a

candidate list of potential models, is to rank the likelihood that the observed data arise from each model. Lysy et al. provides a rigorous protocol for model selection as applied to PPTM data [25].

In the next section we review the current best practices in PPTM at the level of detection of statistically significant heterogeneity without reference to a particular diffusive model. We emphasize that the techniques of data analysis discussed here are novel only in their application to PPTM data.

Many researcher teams have used PPTM data analysis to infer the degree of heterogeneity of soft biological materials [11, 12, 74, 108–114]. These efforts span two broad categories: one based on the "Gaussianity" of the distribution of particle displacements and the second on the statistics of the individual particle mean squared displacement (iMSD). The new protocol presented here combines standard machine learning techniques, such as the Expectation Maximization algorithm [115] and hierarchical clustering [116], to identify statistically distinct clusters based on the distribution of particle path statistics, without reference to the stochastic process that generated the paths. These techniques rely on two relatively weak assumptions: that each path has Gaussian increments and that the process generating each path is stationary. The resulting semi-parametric protocol is consistent both with a large number of stochastic processes and with current approaches to heterogeneity detection in the literature.

Once the particle paths have been assigned to statistically distinct clusters, we then consider the inverse problem of fitting the paths in each cluster to models for simple and anomalous diffusive processes. Unlike simple diffusion where the MSD grows linearly with lag time (τ), anomalous diffusion is described by the more general power law, MSD $\sim \tau^{2H}$ with $0 < H < 0.5$. Anomalous diffusion has been found in many biological contexts; diffusion of 1μm diameter particles in human bronchial epithelial mucus [23], diffusion of biopolymers inside cells [47], bacteria chromosomal loci [117], movement of lipids on model membranes [118], protein diffusion in organellar membranes [119] and in the nucleoplasm [120]. Model fitting of cluster members to candidate models for the underlying stochastic process and intra-cluster pooling of fitting results affords predictive power for elusive experimental properties such as passage times, as illustrated in Hill et al. [23], and discussed in Lysy et al. [25].

In the next section, we start by summarizing existing metrics for the detection and assessment of heterogeneity in PPTM data. In Section 4.4 we describe our metrics that have precedent in the statistics and machine learning communities and compare them with best practices on numerically

generated data. In Section 4.5 we apply our metrics to numerically generated and experimental data, beginning with systems where the heterogeneity is controlled in order to illustrate the precision of our tools. We close with application of these metrics to particle data in an agarose solution, an oft-used simulant for biological gels which is typically heterogeneous, and finally to particle data in human bronchial epithelial cell culture mucus. In the last two experiments, the degree of heterogeneity is not know *a priori* representing the typical scenario for application of these tools to experimental PPTM data.

4.2 Current metrics to detect heterogeneity in PPTM data

Several groups [11, 12, 74, 112, 113, 121] use the van Hove correlation function [122], $P(x(\tau))$, which is the probability distribution function constructed from the observed displacements, x, at lag time τ. Let

$$x_\tau = x(\tau) = X(t+\tau) - X(t) \tag{4.1}$$

where $X(t) = X_t$ is the position of a particle at time t.

For the majority of relevant stationary, stochastic increment processes used to model PPTM data, including normal diffusion, fractional Brownian motion, and generalized Langevin equations, the corresponding van Hove correlation function is Gaussian for each fixed set of model parameters. Paths generated from any of these classical stochastic processes can be considered homogeneous over the observed time scales if they arise from the same set of model parameters, or within some small neighborhood of a parameter set. The practical challenge for experimental path data is to develop a test that does not rely on *a priori* knowledge of the model that generated the data. In a heterogeneous environment, identical particles diffuse in regions with different local properties. One may also consider heterogeneity that arises from particles that are polydisperse in some aspect, e.g., diameter or surface chemistry.

When identical particles are embedded in a sufficiently heterogeneous sample, a single Gaussian fails to fit the ensemble-averaged van Hove correlation function. Heterogeneity can then be measured by the extent to which the van Hove correlation function deviates from a Gaussian. We refer to such metrics as "Stage 1 metrics" and note that they are useful for detection of heterogeneity but not for predictions beyond the observable data.

Whereas a Stage 1 metric implies the presence of statistically significant heterogeneity, one can proceed to probe further into the underlying heterogeneity by binning the paths into disjoint clusters, which we refer to as a "Stage 2 metric." We first survey Stage 1 metrics and then address existing Stage 2 metrics. Our approach is a Stage 2 metric that does not require a preliminary Stage 1 step.

4.2.1 Stage 1 metrics for detection of heterogeneity in PPTM data

- Rahman [123] proposed a non-Gaussianity parameter NG_τ, which measures the departure from an exact identity satisfied by the second and fourth moments of a Gaussian distribution. Namely, on takes these moments of the van Hove correlation function and constructs the metric NG_τ defined for each lag time τ by,

$$\text{NG}_\tau = \frac{\langle x_\tau^4 \rangle}{3\langle x_\tau^2 \rangle^2} - 1 \tag{4.2}$$

  If the increments are Gaussian, $\text{NG}_\tau = 0$ for every lag time τ, whereas non-zero values of NG_τ indicate a degree of heterogeneity. This parameter was applied to the analysis of colloidal systems by Kegel and van Blaaderen [113].

- Houghton et al. [74] used the excess kurtosis (ku) of the van Hove function, defined as

$$\text{ku} = \frac{\sum_{i=1}^{N}(x_i - \overline{x})^4}{(N-1)\sigma^4} - 3 \tag{4.3}$$

  to measure heterogeneity. Here $\overline{x}$ is the mean and σ is the standard deviation of the van Hove correlation function. For a Gaussian distribution ku$= 0$, and again non-zero values denote a degree of heterogeneity.

- Savin and Doyle [124] formulated estimators of the square of the ensemble mean squared displacement $M_1(\tau)$, and of its corresponding variance, $M_2(\tau)$. These estimators are derived from a weighted average of the iMSD where the weights are proportional to the length of the particle trajectory. This prevents the results from being biased by more mobile particles. Rick et al. [112] used these estimators to propose a heterogeneity ration (HR), defined as

$$\text{HR} = \frac{M_2(\tau)}{M_1(\tau)^2}. \tag{4.4}$$

Numerical simulations showed that the maximum value of HR for a bimodal fluid is 3 [112]. Smaller and larger values of HR are then used to indicate heterogeneity [12, 112].

- Tseng et al. [125] employed bin partitions of compliance values to determine the degree of heterogeneity. The compliance $\Gamma(\tau)$, is related to the MSD by [126],

$$\Gamma(\tau) = \frac{\pi a}{k_B T}\langle x_\tau^2 \rangle. \tag{4.5}$$

Bin partitions of the compliance distributions were obtained by comparing the relative contributions of the 10%, 25% and 50% highest values of the individual compliance to the ensemble mean compliance. The relative contributions of these values to the ensemble compliance should be close to 1 to a highly heterogeneous solution and close to 0.10, 0.25 and 0.50, respectively for a homogeneous solution.

- Another Stage 1 metric involves the calculation of the iMSD for a given particle. The iMSD is defined as

$$r_\tau^2 = \frac{1}{N-\tau}\sum_{i=1}^{N-\tau}\left[x_\tau^2 + y_\tau^2\right], \tag{4.6}$$

where x_τ is the displacement over time τ in the X direction and y_τ is the displacement over time τ in the Y direction. Duits et al. [110] constructed auto- and cross-correlation matrices of the amplitude of the iMSD, A_p, to detect both path-wise and temporal heterogeneities. Here the amplitude is found by fitting (4.6) to a power-law function,

$$r_\tau = A_p \tau^\alpha. \tag{4.7}$$

The authors used normalized variances, both with respect to time and space, to quantify the heterogeneity in the distribution of A_p. We note that this strategy mixes pure path analysis with a presumed model for the scaling of the iMSDs with lag time τ.

We highlight one feature of the iMSD strategy that we adopt in our approach, namely that it is based on the cross-correlations among all particle paths, removing any reliance on comparison of the ensemble with one representative path. However, we seek ta clustering strategy that does not rely as strongly on a model for the underlying particle increment process. We therefore choose to defer any fitting to parametric models after decomposing particle paths into clusters, using only statistics of the raw data to cluster the ensemble. After clustering is complete, we then entertain best-fit models and parameter estimation for each cluster.

4.2.2 Stage 2 metrics for decomposition of paths into clusters

Stage 2 metrics aim to assign particle paths to statistically distinct clusters.

- Valentine et al. [11] compared the standard deviation of individual particle step size distributions relative to one chosen particle in the ensemble using the F-statistic,

$$f_{l,k} = \frac{\sigma_k^2/n_k}{\sigma_l^2/n_l} \tag{4.8}$$

  where σ_k^2 and n_k are, respectively, the variance and the number of statistically independent time steps in the van Hove function (degrees of freedom) of particle k, and σ_l^2 and n_l are the corresponding statistics for an arbitrarily chosen reference particle l. Using a 95% certainty of difference for N particle paths, the $F-$test is applied to all $N(N-1)/2$ pair-wise combinations of particle paths. Clusters are then formed by merging statistically indistinguishable paths based on the results of the $F-$test.

When designing our algorithm, we drew inspiration from the two complementary methods proposed by Duits et al. [110] and Valentine et al. [11]. The former incorporates the cross-correlation among all particles, making it robust to any individual outlier or small perturbations, but it also requires a model for the underlying increment process before heterogeneity can be quantified. In contrast, the Valentine et al. method does not require a model to investigate heterogeneity and separates particles into clusters, however it does not uniquely cluster the data. Without a well-defined way to determine the reference particle used at each iteration, applying the Valentine algorithm to the same data multiple times can produce different results. See Section 4.4.5 for further discussion.

Based on their work, we sought to construct a robust and consistent semi-parametric method to assign particles to statistically distinct clusters; for this, we turn to techniques from the field of Machine Learning.

It is common to assume that each particle path is best described by a stationary stochastic process, i.e. the dynamics are non-transient. While analysis of particle paths that violate this assumption pose an additional mathematical challenge, the results can provide insight into temporal or spatial dependencies in a particle's dynamics. Transient behavior has been observed in a wide range of biological settings, including the movement of secretory vesicles [127], viruses [109] and membrane proteins [128, 129], and multiple approaches exist for he identification and characterization of non-stationary behavior. For a discussion of analytical approaches when the stationarity assumption is relaxed, refer to Chapter 2. In this chapter we focus on the analysis of paths exhibiting stationary dynamics. That is, we assume that either each particle's behavior is stationary over the length of the path or a path segmentation algorithm has already been applied to the data to segment the paths into stationary intervals.

4.3 Materials and methods

4.3.1 Materials

A 2 molar sucrose solution was prepared by dissolving sucrose (Sigma) in deionized, distilled (DI) water. We use this sucrose solution as our experimental model for a Newtonian material. Hyaluronic acid solutions (HA), with concentrations of 8 and 10 mg mL^{-1}, were prepared from hyaluronic acid socium salt from *Streptococcus equi* (Sigma), dissolved in DI water and allowed to mix at room temperature for 2 days while rotating at 20 rpm. 10 mg mL^{-1} HA solution is our experimental model for a homogeneous viscoelastic solution. HA is monodisperse in molecular weight, therefore we expect the dynamics of embedded uniform particles to be monodisperse as well, as shown in the work of De Smedt et al. [130]. Low melting point agarose (Fischer) samples were prepared at 0.2% by weight (wt%) agarose mixed in PBS at 45° C for 24 hours. Human Bronchial Epithelial (HBE) cell culture 2.5 wt% mucus samples were prepared as described in Button and Hill [131] and Hill et al. [23] (see Section 6.2.1 for further description). One and two micron diameter carboxylated fluorescent particles (Life Technologies) were used in sucrose, HA

and agarose experiments, and 500 nm particles were used in mucus experiments. The particles in all experiments were added while the solution was at 45 $^\circ$ C and mixed for an additional 24 hours. Samples were then allowed to cool to room temperature. All particles are added to stock solutions at a 0.001 volume fraction and allowed to mix on a 20 rpm rotator for 12 hours prior to use to insure thorough mixing.

4.3.2 Particle tracking

A Nikon Eclipse TE2000-U at $40\times$ magnification and standard video microscopy techniques were used to collect video of particles undergoing thermal diffusion. For all experimental data, the total length of each video was $T = 30$ s and the camera frame rate was $\Delta t^{-1} = 60$ fps. The number of frames or time steps in each particle path is then given by $M = T\Delta t^{-1}$. Video spot tracking software (`http://cismm.cs.unc.edu/`) extracts the positions of each particle of interest in the field of view as a function of time. In this chapter, only particles with recorded positions at each of the 1800 time steps are analyzed. While this has the potential to bias our results toward slower moving particles that are more likely to remain in the field of view during video acquisition [124], the diffusivity of the particles is such that very few particles could not be tracked over the entire length of the video.

4.4 Mathematical protocol

Our Stage 2 analysis is based on the standard deviations of the individual van Hove correlation functions. We do not draw any inference at this stage, i.e., we skip the analog of Stage 1 metrics described in Section 4.2.1. Hierarchical agglomerative clustering [116] is used in our Stage 2 approach, primarily because the resulting dendrogram (defined below) shows the hierarchical "relatedness" between each path based on the statistic of choice [132]. The issue of partitioning the dendrogram to create a disjoint clustering of the data is solved by employing the gap statistic [133]. By comparing the data to multiple null reference distributions, we are able to consistently and uniquely assign particles to clusters. Finally, a model of the underlying process is proposed for each cluster and the relevant parameters are determined.

4.4.1 Calculation of displacements and standard deviations of individual step size distributions

Given N particle paths of length M, we will denote the two-dimensional position process as $\{X^j(i), Y^j(i)\}$, $i = 1, 2...M$, $j = 1, 2, ...N$. We calculate the van Hove correlation functions for a specific lag h corresponding to lag time $\tau = h\Delta t$, where Δt is the inter-observational time, i.e. the time between successive camera frames. In our analysis, we consider $h = 1, 2, ...16$. The displacements are given by $x^j = X^j(1 + ih) - X^j(1 + (i - 1)h)$ and $y^j = Y^j(1 + ih) - Y^j(1 + (i - 1)h)$. The use of non-overlapping displacements severely limits the maximum value of h (h_{max}) due to small sample size, but they are necessary in order to treat the displacements as independent measurements of the displacement process. Fitting each column to a Gaussian distribution produces two $N \times h_{max}$ matrix of standard deviations of the particle displacements, denoted $\mathbf{s}_x$ and $\mathbf{s}_y$ for the X and Y directions. The matrix $\mathbf{s}_{xy}$ is produced by concatenating $\mathbf{s}_x$ and $\mathbf{s}_y$ resulting in a $N \times 2h_{max}$ matrix. Principal component analysis (PCA) is then applied to $\mathbf{s}_{xy}$ and the projection of $\mathbf{s}_{xy}$ onto the first and second principal components, denoted $\{\text{PC}_1(j), \text{PC}_2(j)\}$, constitute the data to which our clustering protocol are applied. PCA allows for information across multiple time scales to be pooled in a manner that maximizes the information stored in the initial principal components and removes the arbitrariness of applying the clustering algorithm to one of many possible projections of the data into $\mathbf{s}_x$-$\mathbf{s}_y$ space.

4.4.2 Determining the number of clusters

The goal is to partition the projection of the standard deviations onto the first two principal components into statistically distinct clusters. We choose not to use standard clustering algorithms such as K-means [134] or K-medoids [135] because these methods require prior knowledge of the number of clusters in the data. Instead, we use agglomerative hierarchical clustering [136, 137] with the average linkage function and the standard Euclidean distance metric; for details see Hastie et al. [132].

4.4.2.1 Hierarchical cluster

The pair-wise distances between all scalar pairs $\{\mathrm{PC}_1(j), \mathrm{PC}_2(j)\}$ are calculated using the Euclidean distance metric and the distance between clusters is determined by computing the average distance between all points in both clusters, a metric known as the average linkage function. In agglomerative hierarchical clustering, each data point is initially its own cluster. The two closest clusters based on the Euclidean distance in PC_1-PC_2 space (points 1 and 2 in Fig. 4.1) are then merged to form a new (violet) cluster. This process is repeated (blue cluster containing points 1, 2 and 3, Fig. 4.1B) until all of the data points have been merged into a single cluster (green cluster containing all points, Fig. 4.1B). Recording the order in which clusters are merged allows one to construct a dendrographic representation of the data showing the hierarchical similarity between clusters (Fig. 4.1C)[116]. The height of each connection in the dendrogram is equal to the average distance between the connected clusters, encoding a hierarchical metric of cluster similarity based on their van Hove correlation functions.

After all the distances are calculated (Fig. 4.1C), the number of clusters, K, is determined by a cutoff value ζ that partitions the dendrogram at resolution ζ. For instance, if we choose any $\zeta < 1$ in Fig. 4.1, all data points remain in their own cluster, and there are 4 clusters at this resolution. For any $1 < \zeta < 2.12$, the two data points making up the violet cluster are now indistinguishable. Thus we declare 3 clusters for this range of ζ. Next, for $2.12 < \zeta < 4.75$, there are only 2 clusters, the blue cluster and data point 4. Finally, for $\zeta > 4.75$, the green cluster is the only cluster and contains all data points. In this way, the parameter ζ solely determines the partitioning of the data, and as ζ varies from the smallest to the largest values, the number of clusters K_τ ranges from 1 to N. The next critical step is to select the appropriate degree of resolution, i.e. the value of ζ, and thus determine the number of clusters K that best delineates the ensemble of data points.

4.4.2.2 Optimal number of clusters and the gap statistic

To find the optimal number of clusters, K^*, we use a gap statistic [133]. We start by defining the parameter W_K as [138]

$$W_K = \sum_{c=1}^{K} \frac{1}{2n_c} D_c, \tag{4.9}$$

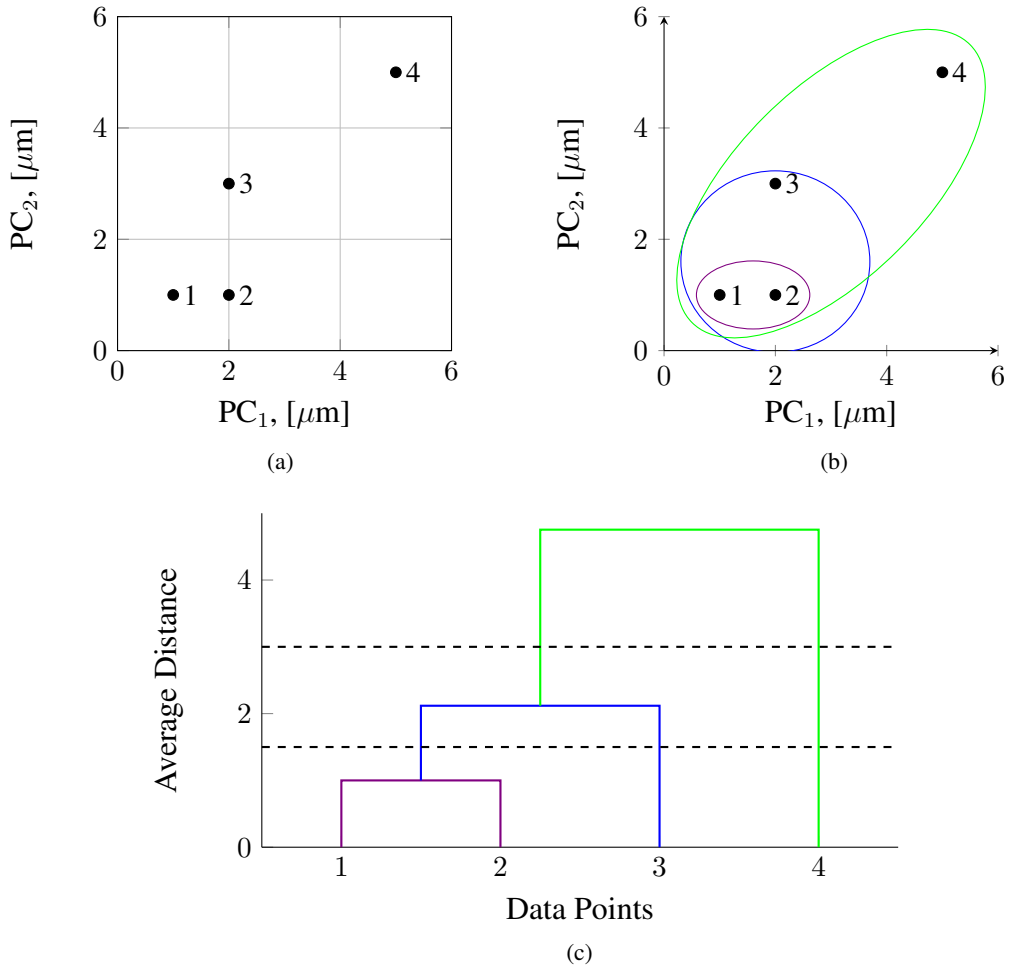


Figure 4.1: Example of hierarchical clustering. (A) The distribution of data points to be clustered. Each data point is initially assigned to a cluster containing only itself. The pair-wise distances between all clusters are calculated and the closest two clusters are merged to form a new cluster. (B) This process is repeated until a single cluster contains all data points. (C) A dendrogram showing the distance between each cluster and the order in which they were merged. The dashed lines at 3 and 1.5 show cutoff values that produce two and three clusters, respectively

where n_c is the number of data points in cluster c and D_c is the sum of the pair-wise squared distances between all the elements of cluster c. As ζ decreases, the number of clusters, K increases, which in turn results in a decrease in W_K due to the increasing mean intra-cluster density.

Next, we use these values of W_K to compare the projection of standard deviations of the van Hove functions into the PC_1-PC_2 space, which may or may not contain statistically distinct clusters, to a null reference data set containing only one cluster and with uniform density. In order to ensure that the null reference data set only contains a single cluster with uniform density, these data are

generated from a uniform distribution. To match the input data as closely as possible (apart from the number of clusters present), the reference data set is created such that its cardinality and domain are the same as the input data, i.e. the distribution of $\{\mathrm{PC}_1(j), \mathrm{PC}_2(j)\}$. To remove the variability associated with comparison of the input data to a single reference data set, it is common practice to compare the input data to multiple reference data sets. We have determined that 40 reference data sets suffice to consistently partition the data.

To illustrate this procedure, we numerically generate paths of 150 1 μm diameter spherical particles diffusing via Brownian motion in a heterogeneous medium with diffusion coefficients: 1.67×10^{-2} (50 paths), 1.40×10^{-2} (50 paths), 1.00×10^{-4} (49 paths), and 1.00×10^{-3} $\mu\mathrm{m}^2\mathrm{s}^{-1}$ (1 path). This data set will be referred to as the "Numerically Generated Heterogeneous Newtonian" (NGHN) data set. First we fit the van Hove correlation function of each particle path to a Gaussian, doing so separately for each coordinate. For particle diffusion in viscous fluids, the van Hove correlation function in any direction is expected to have mean 0 and variance $s(\tau)$ where $s(\tau) = \sqrt{2D\tau}$, and the diffusion coefficient is given by the Stokes-Einstein relation,

$$D = \frac{k_B T}{6\pi\eta a}, \tag{4.10}$$

where a is the particle radius and η is the fluid viscosity. In our example, the projection of the distribution of standard deviations onto the first and second principal components is shown in Figure 4.2A. We next calculate W_K for the path data and W_{ref}, which is the mean of the W_K's calculated using (4.9) applied to each of the reference data sets described previously. The results are plotted in Fig. 4.2B as a function of the number of clusters, K. A measure of the variability introduced by the use of a finite number of reference data sets and the number of particles being clustered has the form $s_k = \mathrm{sd}(K)\sqrt{(1 + 1/B)(1 + 1/N)}$, where sd is the standard deviation of the reference data sets and B is the number of reference data sets and N is the number of data points [133].

The optimal number of clusters in the distribution of standard deviations of van Hove functions, for a given lag time, is estimated as

$$K^* = \underset{K}{\operatorname{argmin}} \left\{ K \,|G(K) - s_K > G(K-1) + s_{K-1} \right\}, \tag{4.11}$$

where argmin returns the value of the input argument that minimizes the input function. This equation chooses K^* to be the smallest number of clusters such that the lower bound of the gap statistic at K clusters is greater than the upper bound of the gap statistic when $K-1$ clusters are present. In our example, at this stage of the algorithm, two clusters are identified ($K^* = 2$), as shown in Fig. 4.2C.

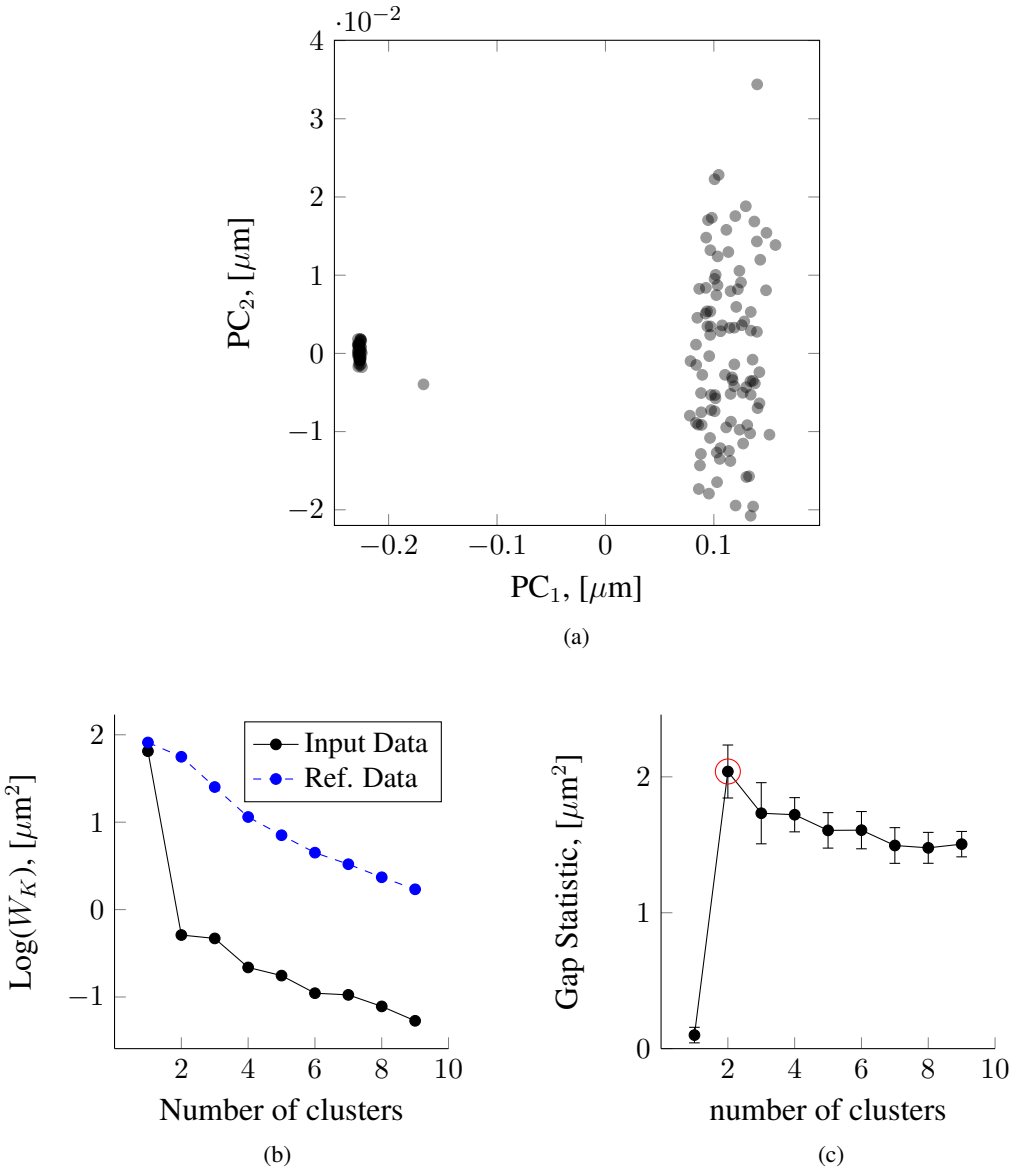


Figure 4.2: Example clustering applied to the Numerically Generated Heterogeneous Nestonian (NGHN) data set. (A) Standard deviations of van Hove functions for particles moving in a Newtonian, heterogenous fluid. The heterogeneity of the fluid is characterized by four different diffusion coefficients. (B) Value of W_K is given by (4.9). The gap statistic is calculated based on the differences between the reference and input data (C) as described in Section 4.4.2.2. The algorithm indicates that at this stage, there are two clusters present (red circle).

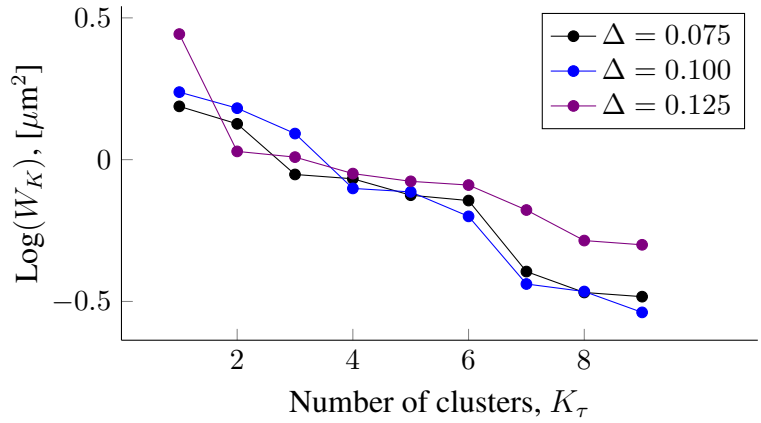
A question of interest is, "what is the minimal separation in variances of the step size distribution between two clusters so that they appear distinct at this stage of the algorithm." Recall that in our example NGHN data set, this is the same as asking what minimum difference in diffusivities is distinguishable. To investigate this, we generated three heterogeneous Newtonian data sets of size $N = 150$. Each data set contains two clusters: particles belonging to the first cluster in each data set have diffusivity $D_1 = 1.61 \times 10^{-2} \mu\text{m}^2\text{s}^{-1}$, while particles in the second cluster have diffusivity $D_2 = D1(1 + \Delta)$. The value D_1 is the diffusion coefficient of a one-micron particle diffusing in a fluid with viscosity 27 mPA s. We choose three values of Δ: 0.075, 0.100 and 0.125. Our algorithm correctly identifies the two clusters when $\Delta \geq 0.100$. We note that the NG_τ metric (4.2), the heterogeneity ratio (4.4) and the percent contribution of the bin partitions described by Tseng et al. [125] steadily increases as Δ increases, as expected for increasing heterogeneity. The Stage 2 metric (4.8) of Valentine et al. [11] identifies two to three clusters in each data set. These results are given in Table 4.1.

Metric	$\Delta = 0.050$	$\Delta = 0.075$	$\Delta = 0.10$
NG, [113]	$9.11E^{-2}$	$3.25E^{-1}$	$3.57E^{-1}$
ku, [74]	$2.73E^{-1}$	$1.08E^{-1}$	$1.19E^{-1}$
HR, [124]	$8.44E^{1}$	$9.88E^{1}$	$1.14E^{2}$
Relative Compliance, 10%, [125]	$1.12E^{1}$	$1.13E^{1}$	$1.15E^{1}$
Relative Comp., 25%	$2.72E^{1}$	$2.74E^{1}$	$2.77E^{1}$
Relative Comp., 50%	$5.28E^{1}$	$5.29E^{1}$	$5.34E^{1}$
F-test, [11]	2-3	2-3	2-3

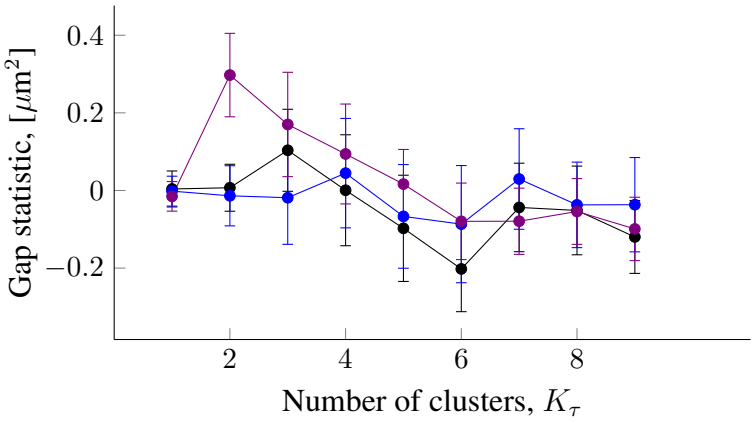
Table 4.1: Sensitivity test of existing heterogeneity metrics: numerical data with controlled degrees of heterogeneity in the diffusion coefficients. Three heterogeneous Newtonian data sets are generated, where each data set consists of 150 paths of particles of diameter 1 μm. For each data set, the first 75 paths have diffusion coefficients $D_1 = 1.61 \times 10^{-2}$ $\mu\text{m}^2\text{s}^{-1}$ while the next 75 paths have diffusion coefficient $D_2 = D_1(1 + \Delta)$ for $\Delta = 7.5\%$, 10% and 12.5%.

Given these results from other methods in the literature, we now apply our method. Figure 4.3A shows values of $\log(W_K)$ vs. K for each of the three data sets. As Δ increases, the 'bend' in the plot at $K = 2$ becomes more pronounced. Figure 4.3B shows the gap statistic as a function of K. We see that for a Newtonian fluid in this range of diffusvities, the distribution of the principal component projection of the standard deviations of the van Hove correlation functions of two data sets with diffusivities that vary by only 7.5% are indistinguishable. However as the difference in

the diffusivities increases beyond 10%, the distributions become distinguishable by our metrics and the correct number of clusters is successfully recovered. It is important to point out that this 10% threshold may not hold for different data sets and its value depends on, among other variables, the total number of clusters, presence of outliers, distribution of data points within each cluster and experimental error.



(a)



(b)

Figure 4.3: Test of the gap statistic: numerical data with controlled degrees of heterogeneity in the diffusion coefficients. Three heterogeneous Newtonian data sets are generated, where each data set consists of 150 paths of particles of diameter 1 μm. For each data set, the first 75 paths have diffusion coefficients $D_1 = 1.61 \times 10^{-2}$ μm^{2}s^{-1} while the next 75 paths have diffusion coefficient $D_2 = D_1(1 + \Delta)$ for $\Delta = 7.5\%$, 10% and 12.5%. (A) As Δ increases, the 'bend' in the $\log(W_K)$ vs. K plot at $K = 2$ becomes more pronounced. (B) The gap statistic correctly indicates two clusters for $\Delta \geq 0.125$. The number of clusters selected by the gap statistic is indicated by a red circle.

4.4.2.3 Cluster refining

After the main clusters are identified, we repeat the hierarchical clustering and gap statistic steps for each cluster $c = 1, 2...K^*$. The first clustering steps (Sections 4.4.2.1-4.4.2.2) serve to identify well-separated clusters while the second round of clustering, introduced here, inspects each previously identified cluster for the presence of sub-clusters. The final number of clusters, K^{final}, is the total number of clusters found after applying the clustering algorithm to the previously identified K^* clusters. This two-pass clustering approach is robust to outliers that normally cause single-pass clustering to fail. At this stage, each of the previously identified clusters in the NGHN data set is split into two clusters, i.e. $K^{final} = 4$ (Fig. 4.4).

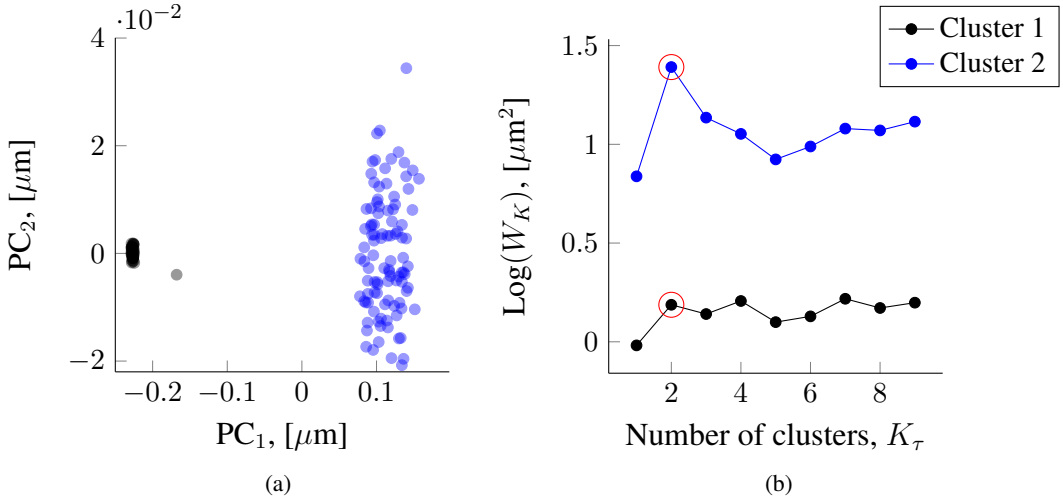


Figure 4.4: Cluster refining on the NGHN data set (Section 4.4.2.2). (A) Results of the initial cluster step. (B) The clustering algorithm is applied to each individual cluster to identify sub clusters. The original Cluster 1 and 2 are found to each contain two sub clusters (red circles).

4.4.3 Cluster distribution fitting

Once the data have been fully partitioned, i.e. we have K^{final}, we may apply model selection and fitting procedures on a path-by-path basis, then, if appropriate, pool data from particles within the same cluster in order to increase the accuracy of the estimates of the corresponding diffusive parameters. Path-wise fitting is carried out using the maximum likelihood approach described in Section 3.5. Here, we use fractional Brownian motion (fBm) based on its ability to describe the

autocovariance observed in the displacements of particles undergoing passive thermal diffusion in a wide range of both simple and complex fluids. Figure 4.5 shows the final clustering results in PC_1-PC_2 space and the maximum likelihood fitting results.

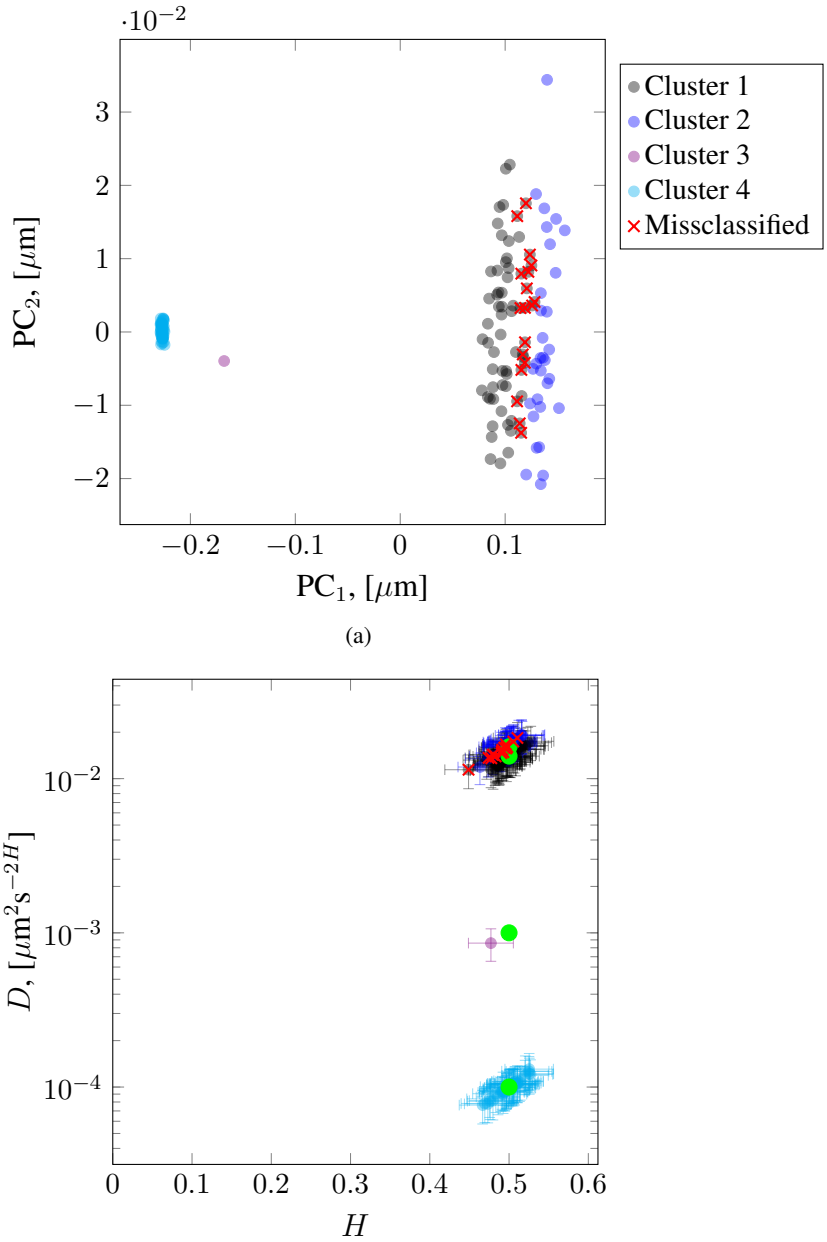


Figure 4.5: Final clustering the NGHN data set. (A) Clusters projected onto the first and second principal components. (B) Path-wise maximum likelihood estimation of Hurst parameter (H) and diffusivity (D). The green marker indicates the expected value of each parameter. Eighteen of the 150 particles (12%) were missclassified. Error bars indicate 95% confidence intervals.

4.4.4 Algorithm to simulate numerical data

For the purpose of validating the protocol described in Section 4.4.2, we perform simulations of particles moving by both standard Brownian motion (sBm) and fractional Brownian motion (fBm). Simulated data were generated using the Direct method and Cholesky decomposition described in Section 2.3.2.1. For all simulated data, the interobservational time is $\Delta t = {}^{1}/_{60}$ and path length $M = 1800$, matching the experimental data.

4.4.5 Metric comparison

When our algorithm is applied to the numerically generated heterogeneous Newtonian (NGHN) data set (Fig. 4.5), we find three main clusters corresponding to the three clusters generated with mean diffusivities $1.67{\times}10^{-2}$, $1.40{\times}10^{-2}$ and $1.00{\times}10^{-4}$ μm^{2}s^{-1}. The outlying point, generated with $D = 1.00 \times 10^{-3}$ μm^{2}s^{-1}, was also correctly identified. Assuming fBm as the underlying process, we fit D and H on a per-path basis for the three main clusters and pool estimates within a given cluster. Intra-cluster estimates of D and H are shown in Table 4.2. To compare the performance of our algorithm with the metrics described in Sections 4.2.1 and 4.2.2, we applied those metrics to this same set of data.

	Count (%)	H (SD) $\times10^{-1}$[-]		D (SD) $\times10^{-2}$ [μm^{2}s^{-2H}]	
		expected	*predicted*	*expected*	*predicted*
Cluster 1	69 (46)	5.00	4.95 (0.14)	1.40	1.42 (0.16)
Cluster 2	31 (20.7)	5.00	5.00 (0.11)	1.67	1.68 (0.16)
Cluster 3	1(0.7)	5.00	4.77 (0.00)	0.10	0.081 (0.00)
Cluster 4	49 (32.7)	5.00	5.00 (0.14)	0.01	0.010 (0.001)

Table 4.2: Cluster results for the NGHN data set.

All Stage 1 metrics presented in Section 4.2.1 correctly indicated that the simulated data set was heterogeneous. The non-Gaussianity parameter (4.2), excess kurtosis (4.3), and heterogeneity ratio (4.4), are 0.121, 0.362 and 1,115, respectively. The relative contributions of the 10%, 25% and 50% highest values of the individual compliance to the ensemble mean compliance were 18.0%, 39.9%

and 71.0%, respectively. Finally, the mean spatial relative standard deviation in the iMSD amplitudes was 1.02 μm^2.

The Stage 2 metric of Valentine et al. [11] described in Section 4.2.2 was applied to the simulated data set multiple times. Clusters were formed by randomly selecting "representative" particle paths of the particles not yet clustered and assigning all particle paths to a cluster based on the results of an F-test. In each instance, the data was correctly determined to be heterogeneous while the number of statistically distinct clusters within the data predicted by the algorithm varied between 5 and 6, demonstrating sensitivity to the choice of the representative particle path.

4.5 Results and discussion

We set out to test our methods on a variety of simulated and experimental data sets exhibiting various degrees of heterogeneity. In each instance, the simulated data were generated with parameter values comparable to the measured values for the corresponding experimental data set. This provides a way to distinguish the error inherent in our algorithm from experimental error [139].

4.5.1 Homogeneous data: simulated and experimental

4.5.1.1 Newtonian paths and data analysis

(i) *Simulated.* 100 particle paths were generated with $H = 0.5$ and $D = 0.0161$ μm^{2}s^{-2H}. These parameter values were chosen to match the expected values for the experimental homogeneous sucrose data. See Table 4.3 for intra-cluster estimates of H and D.

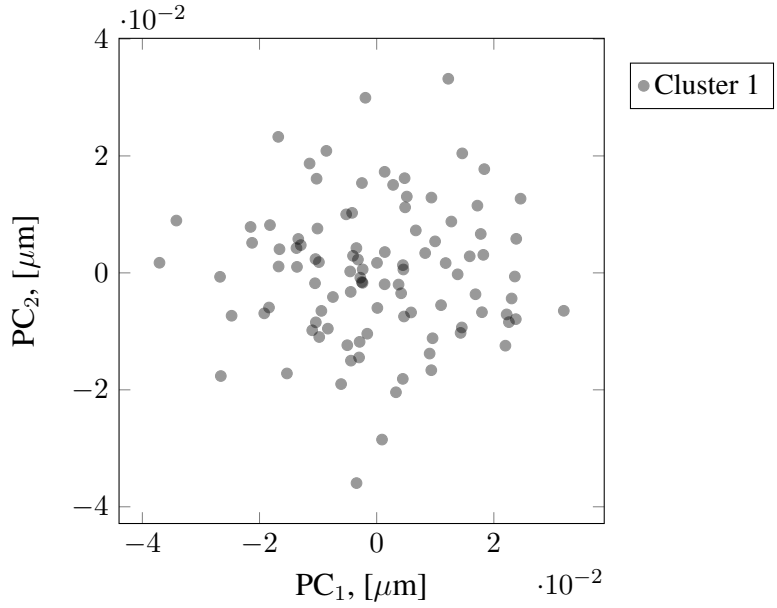
(ii) *Experimental.* Position time series were collected for 100 1 μm diameter particles undergoing passive thermal diffusion in a 2 molar sucrose solution. The viscosity of the 2 molar sucrose solution was calculated to be 0.027 Pa s based on the MSD of embedded tracer particles. See Table 4.4 for intra-cluster estimates of H and D.

	Count (%)	H (SD) $\times 10^{-1}$[-]		D (SD) $\times 10^{-2}$ [μm^{2}s^{-2H}]	
		expected	*predicted*	*expected*	*predicted*
Cluster 1	100 (100)	5.00	4.98 (0.14)	1.61	1.59 (0.19)

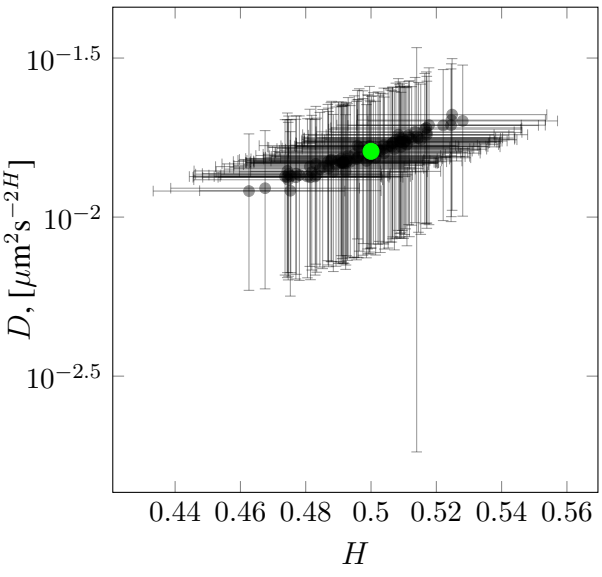
Table 4.3: Cluster results for simulated homogeneous Newtonian data.

	Count (%)	H (SD) $\times 10^{-1}$[-]		D (SD) $\times 10^{-2}$ [μm^{2}s^{-2H}]	
		expected	*predicted*	*expected*	*predicted*
Cluster 1	100 (100)	5.00	5.12 (.37)	1.61	1.56 (.30)

Table 4.4: Cluster results for experimental homogeneous viscoelastic data.

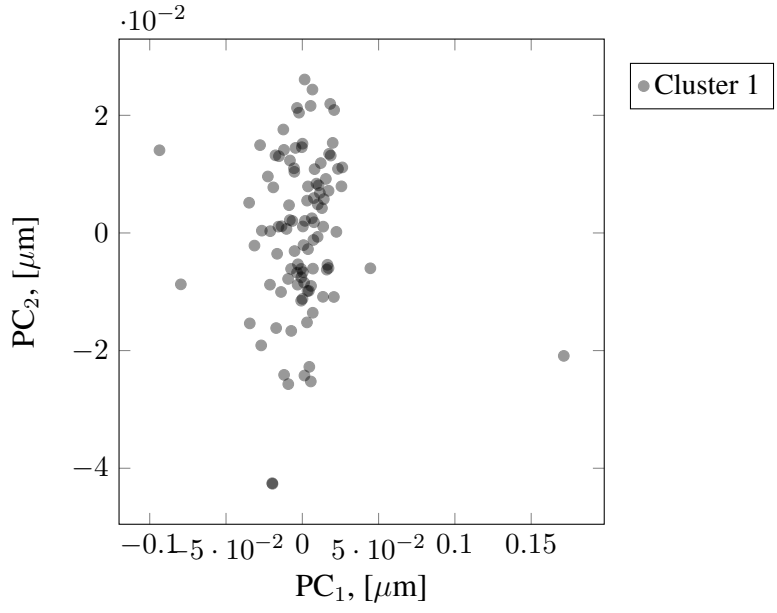


(a)

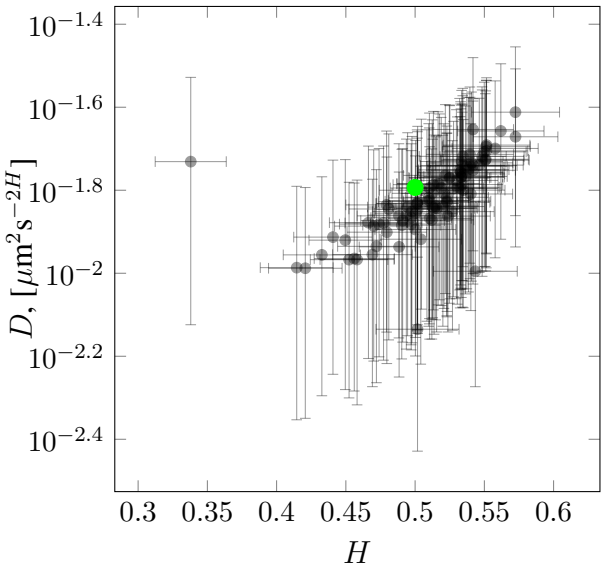


(b)

Figure 4.6: Simulated homogeneous Newtonian cluster results. (A) Clusters projected onto the first and second principal components. (B) Path-wise maximum likelihood estimation of Hurst parameter (H) and diffusivity (D). The green marker indicates the expected value of each parameter. Error bars indicate 95% confidence intervals.



(a)



(b)

Figure 4.7: Experimental (sucrose) homogeneous Newtonian cluster results. (A) Clusters projected onto the first and second principal components. (B) Path-wise maximum likelihood estimation of Hurst parameter (H) and diffusivity (D). The green marker indicates the expected value of each parameter. Error bars indicate 95% confidence intervals.

4.5.1.2 Viscoelastic paths and data analysis

(i) *Simulated.* 175 particle paths were generated with $H = 0.288$ and $D = 9.3 \times 10^{-5}$ $\mu\text{m}^2\text{s}^{-2H}$. These parameter values were chosen based on the inferred experimental values of D and H for the homogeneous HA path data. See Table 4.5 for intra-cluster estimates of H and D.

(ii) Experimental. Position time series were collected for 175 1 μm diameter particles undergoing passive thermal diffusion in a 8 mg mL^{-1} solution. See Table 4.6 for intra-cluster estimates of H and D.

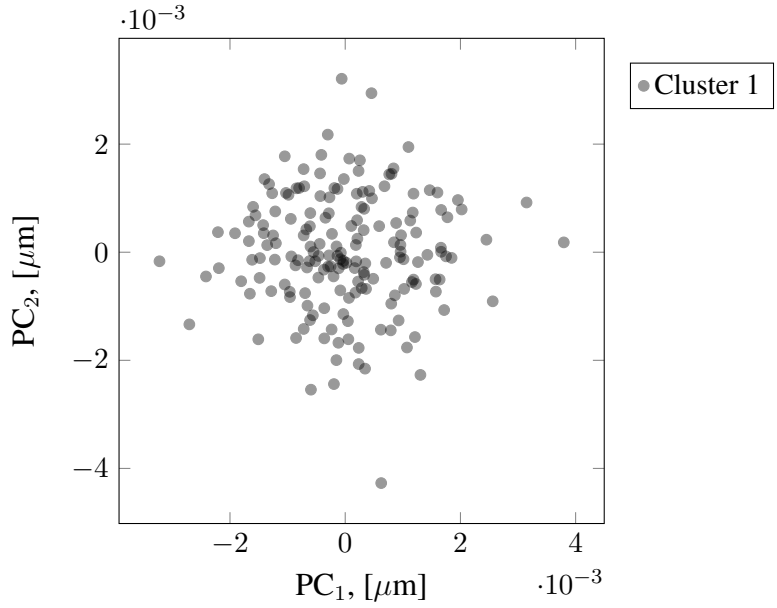
The results for the experimental homogeneous viscoelastic data (Fig. 4.9) indicate the presence of two clusters, one of which contains only a single point.

	Count (%)	H (SD) $\times 10^{-1}$[-]		D (SD) $\times 10^{-5}$ [$\mu\text{m}^2\text{s}^{-2H}$]	
		expected	*predicted*	*expected*	*predicted*
Cluster 1	175 (100)	2.88	2.87 (0.13)	9.30	9.34 (.934)

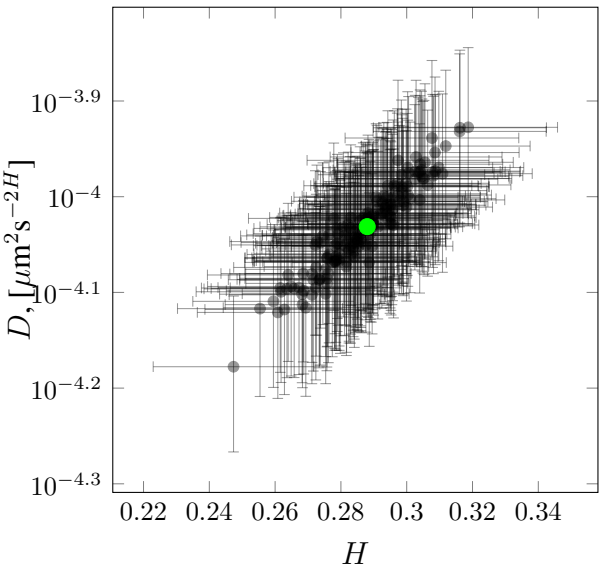
Table 4.5: Cluster results for simulated homogeneous viscoelastic data.

	Count (%)	H (SD) $\times 10^{-1}$[-]		D (SD) $\times 10^{-4}$ [$\mu\text{m}^2\text{s}^{-2H}$]	
		expected	*predicted*	*expected*	*predicted*
Cluster 1	1 (0.5)	-	4.25 (0)	-	41.0 (0)
Cluster 2	206 (99.5)	-	3.01 (0.56)	-	3.44 (2.65)

Table 4.6: Cluster results for experimental homogeneous viscoelastic data.

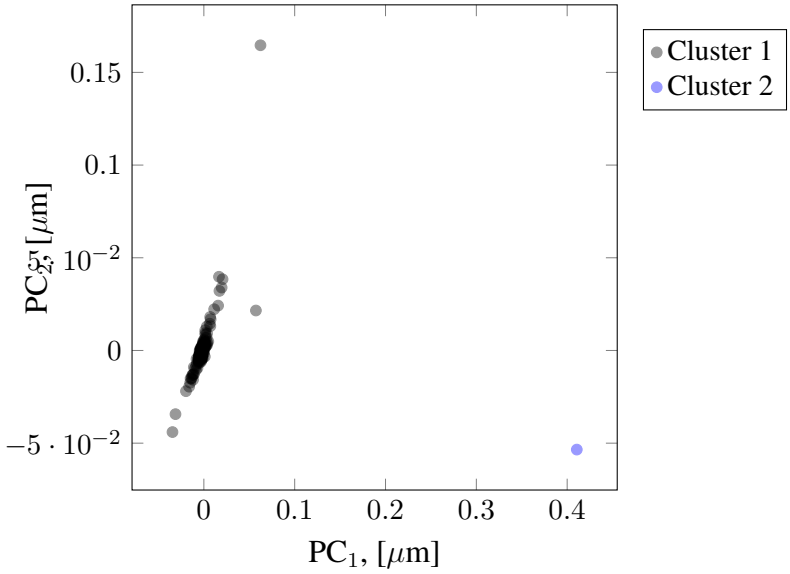


(a)

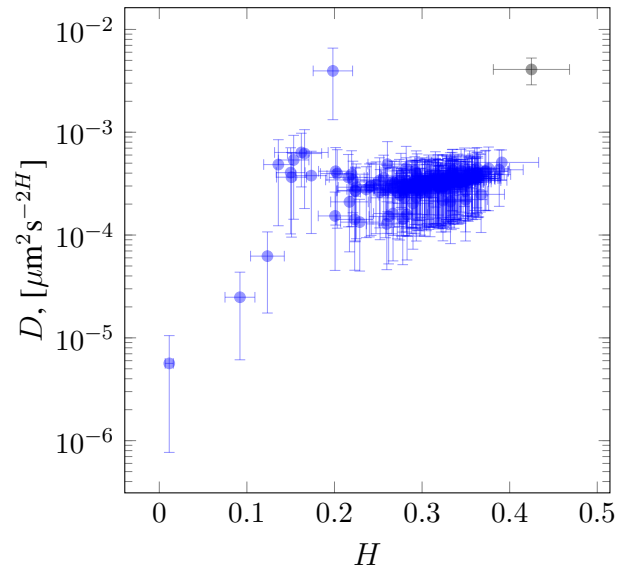


(b)

Figure 4.8: Simulated homogeneous viscoelastic cluster results. (A) Clusters projected onto the first and second principal components. (B) Path-wise maximum likelihood estimation of Hurst parameter (H) and diffusivity (D). The green marker indicates the expected value of each parameter. Error bars indicate 95% confidence intervals.



(a)



(b)

Figure 4.9: Experimental (HA) homogeneous viscoelastic cluster results. (A) Clusters projected onto the first and second principal components. (B) Path-wise maximum likelihood estimation of Hurst parameter (H) and diffusivity (D). One outlier was identified (Cluster 2). Error bars indicate 95% confidence intervals.

4.5.2 Heterogeneous data: simulated and experimental

4.5.2.1 Newtonian paths and data analysis

(i) *Simulated.* 90 particle paths were generated with $H = 0.5$ and $D = 8.05 \times 10^{-3}$ μm^{2}s^{-2H} and combined with 100 particles generated with $H = 0.5$ and $D = 0.0161$ μm^{2}s^{-2H}. These parameter values were chosen to match the expected experimental values of D and H for the heterogeneous experimental sucrose data containing 1 μm and 2 μm diameter particles. See Table 4.7 for intra-cluster estimates of H and D.

(ii) *Experimental.* Position time series were collected for 90 2 μm diameter particles undergoing passive thermal diffusion in a 2 molar sucrose solution and combined with the 1 μm experimental data presented in Section 4.5.1.1. See Table 4.8 for intra-cluster estimates of H and D.

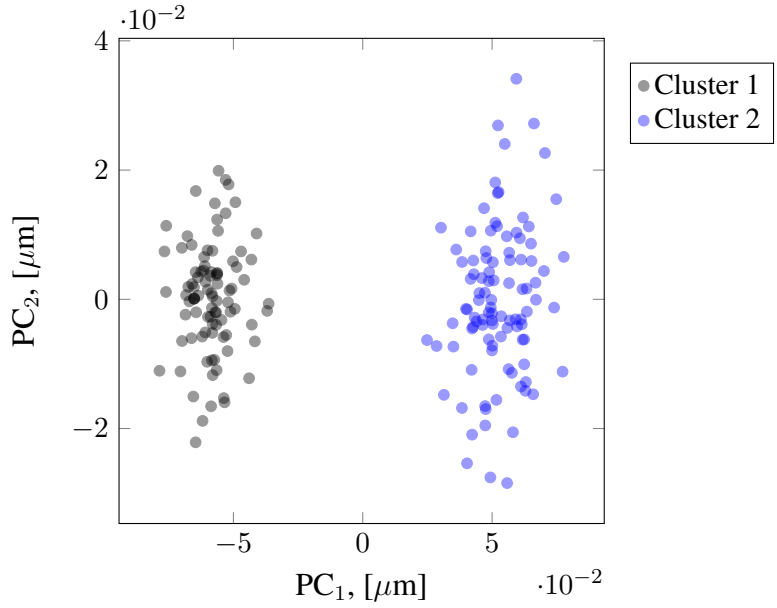
For the simulated data, the correct number of clusters was identified while an additional cluster, containing a single particle, was identified in the experimental data. By cross-referencing the cluster assignments with the file the data came from, we were able to identify 5 out of the 200 particles (2.5%) which were assigned to the wrong cluster (Fig. 4.11).

	Count (%)	H (SD) $\times 10^{-1}$[-]		D (SD) $\times 10^{-2}$ [μm^{2}s^{-2H}]	
		expected	*predicted*	*expected*	*predicted*
Cluster 1	90 (47.4)	5.00	5.02 (0.14)	0.81	0.82 (0.001)
Cluster 2	100 (52.6)	5.00	5.00 (0.14)	1.61	1.60 (0.20)

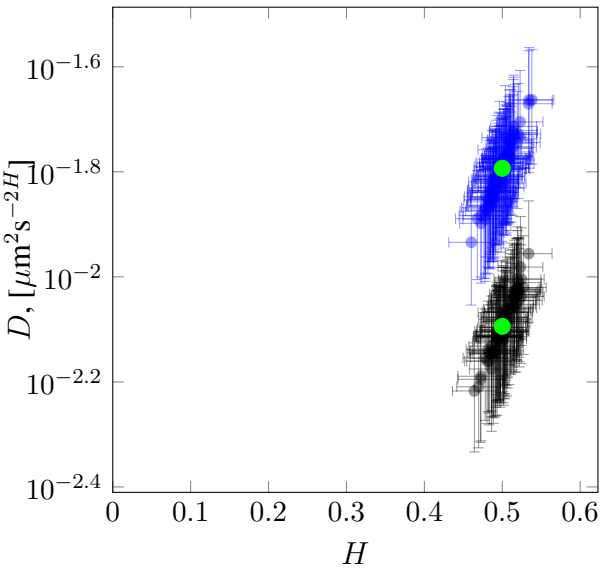
Table 4.7: Cluster results for simulated heterogeneous Newtonian data.

	Count (%)	H (SD) $\times 10^{-1}$[-]		D (SD) $\times 10^{-2}$ [μm^{2}s^{-2H}]	
		expected	*predicted*	*expected*	*predicted*
Cluster 1	100 (50)	5.00	5.13 (0.03)	1.61	1.56 (0.29)
Cluster 2	1 (1)	5.00	4.91 (0.01)	-	6.4 (3.8)
Cluster 3	1(0.5)	5.00	3.38 (0.00)	-	1.86 (0.00)
Cluster 4	97 (48.5)	5.00	5.04 (0.0458)	0.81	1.02 (0.77)

Table 4.8: Cluster results for experimental heterogeneous Newtonian data.

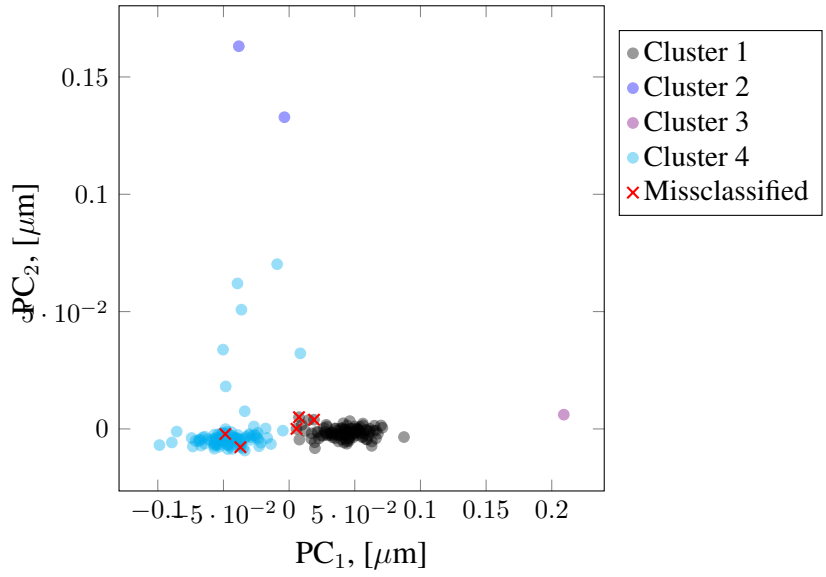


(a)

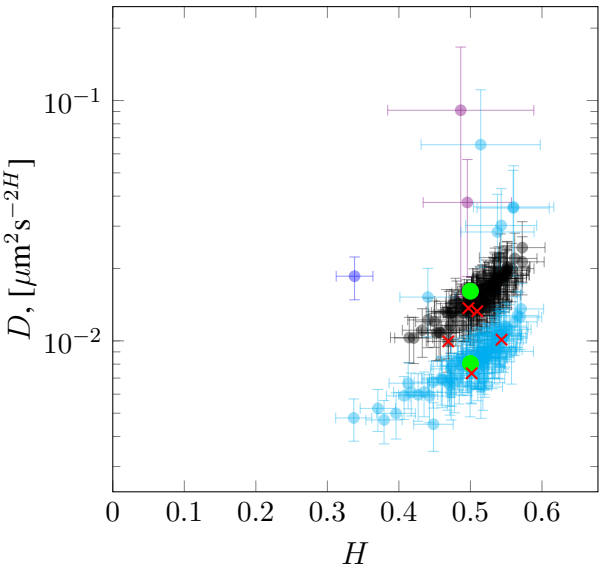


(b)

Figure 4.10: Simulated heterogeneous Newtonian cluster results. (A) Clusters projected onto the first and second principal components. (B) Path-wise maximum likelihood estimation of Hurst parameter (H) and diffusivity (D). The green marker indicates the expected value of each parameter. Error bars indicate 95% confidence intervals.



(a)



(b)

Figure 4.11: Experimental (sucrose) heterogeneous Newtonian cluster results. (A) Clusters projected onto the first and second principal components. (B) Path-wise maximum likelihood estimation of Hurst parameter (H) and diffusivity (D). The green marker indicates the expected value of each parameter. Five of the 200 particles (2.5%) were misclassified. Three outliers were identified (Cluster 2 and 3). Error bars indicate 95% confidence intervals.

4.5.2.2 Viscoelastic paths and data analysis

(i) *Simulated.* 180 particle paths were generated with $H = 0.4$ and $D = 4.4 * 10^{-4}$ μm^{2}s^{-2H} and combined with 175 particles generated with $H = 0.32$ and $D = 9.3 \times 10^{-5}$ μm^{2}s^{-2H}. These parameter values were chosen based on the inferred experimental values of D and H for the heterogeneous experimental HA data containing 1 μm and 2 μm diameter particles. See Table 4.9 for intra-cluster estimates of H and D.

(ii) *Experimental HA.* The 175 1 μm diameter particles in 8 mg mL^{-1} HA solution presented in Section 4.5.1.2 were combined with 208 1 μm particles undergoing passive thermal diffusion in a 10 mg mL^{-1} HA solution. See Table 4.10 for intra-cluster estimates of H and D.

The algorithm correctly identified two clusters in the simulated data set. Seven additional clusters, four of which containing a single particle, were identified in the experimental data (Fig. 4.13). The two largest clusters contain 322 of the 355 particles (90.1%). Cluster 2 is composed of 164 particles, 98.8% of which are from the 8 mg mL^{-1} data subset. The other large cluster, Cluster 8 is composed of 158 particles, 98.7% of which are from the 10 mg mL^{-1} data subset. The distribution of outlying clusters suggests imperfect mixing of the HA solutions during preparation.

We now apply the clustering algorithm to path data from two putatively heterogeneous complex fluids with unknown heterogeneity. To our knowledge, there is no guidance in the literature for a quantitative heterogeneous characterization of agarose solutions or HBE cell culture mucus.

(iii) *Experimental 0.2wt% agarose.* Position time series were collected for 198 1 μm particles undergoing passive thermal diffusion in 0.2wt% agarose solution

The results from a 0.2wt% agarose solution are shown in Fig. 4.14. It is clear from Fig. 4.14B that the ensemble of particles exhibit a range of diffusive behavior from relatively mobile (large H and D) to nearly immobile (small H and D). These disparities in diffusive behavior are resolved with out clustering methods into two main clusters containing 142 and 42 particles, and four smaller clusters containing 7, 4, 2 and 1 particles. The path data for each cluster are then fit to fBm, with the results shown in Table 4.11. We note that no particles appear well fit by standard Brownian motion(sBm) due to the small Hurst parameter values.

(iv) *Experimental 2.5wt% HBE cell culture mucus.* Position time series were collected for 355 0.5 μm diameter particles undergoing passive thermal diffusion in 2.5wt% HBE mucus. See Table 4.12 for the resulting best fit values of H and D per cluster. The resulting clusters are shown in Fig. 4.15.

	Count (%)	H (SD) $\times 10^{-1}$[-]		D (SD) $\times 10^{-4}$ [μm^{2}s^{-2H}]	
		expected	*predicted*	*expected*	*predicted*
Cluster 1	175 (49.3)	3.20	3.19 (0.14)	0.93	0.90 (0.01)
Cluster 2	180 (50.7)	4.00	3.98 (0.13)	4.40	4.35 (0.46)

Table 4.9: Cluster results for simulated heterogeneous viscoelastic data.

	Count (%)	H (SD) $\times 10^{-1}$[-]		D (SD) $\times 10^{-4}$ [μm^{2}s^{-2H}]	
		expected	*predicted*	*expected*	*predicted*
Cluster 1	17 (4.4)	-	2.56 (0.34)	-	3.75 (0.51)
Cluster 2	164 (42.8)	-	0.32 (0.41)	-	3.32 (0.48)
Cluster 3	1 (0.3)	-	1.74 (0)	-	3.78 (0)
Cluster 4	1 (0.3)	-	1.98 (0)	-	40.0 (0)
Cluster 5	1 (0.3)	-	4.25 (0)	-	41.0 (0)
Cluster 6	33 (8.6)	-	2.30 (0.83)	-	1.51 (0.62)
Cluster 7	1 (0.3)	-	0.11 (0)	-	0.06 (0)
Cluster 8	158 (41.3)	-	2.18 (0.45)	-	0.70 (0.15)
Cluster 9	7 (1.8)	-	1.65 (0.32)	-	6.40 (3.62)

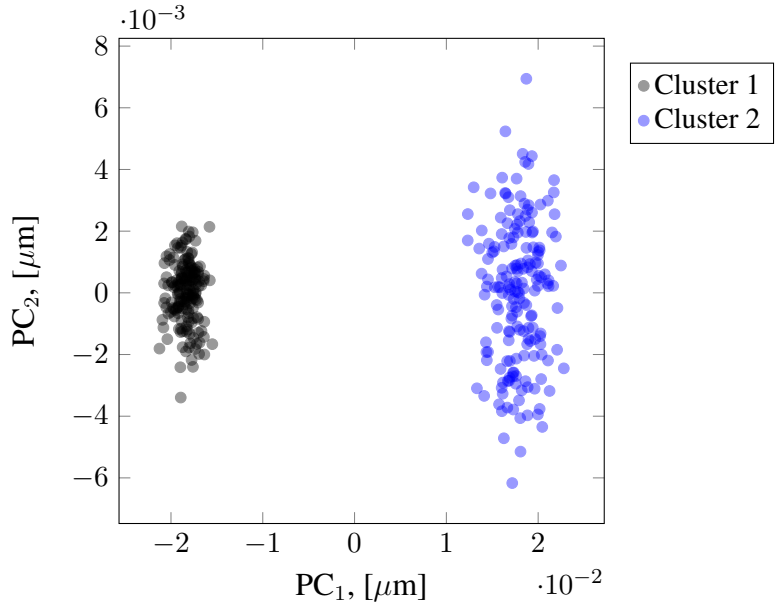
Table 4.10: Cluster results for experimental heterogeneous viscoelastic data.

	Count (%)	H (SD) $\times 10^{-1}$[-]		D (SD) $\times 10^{-4}$ [μm^{2}s^{-2H}]	
		expected	*predicted*	*expected*	*predicted*
Cluster 1	73 (36.9)	-	0.94 (0.05)	-	1.07 (0.56)
Cluster 2	68 (34.3)	-	0.71 (0.32)	-	2.17 (1.01)
Cluster 3	2 (1.0)	-	2.11 (0.08)	-	48.1 (17.1)
Cluster 4	1 (0.5)	-	3.61 (0)	-	167 (0)
Cluster 5	44 (22.2)	-	0.84 (0.34)	-	6.18 (4.50)
Cluster 6	10 (5.1)	-	1.35 (0.61)	-	17.1 (14.5)

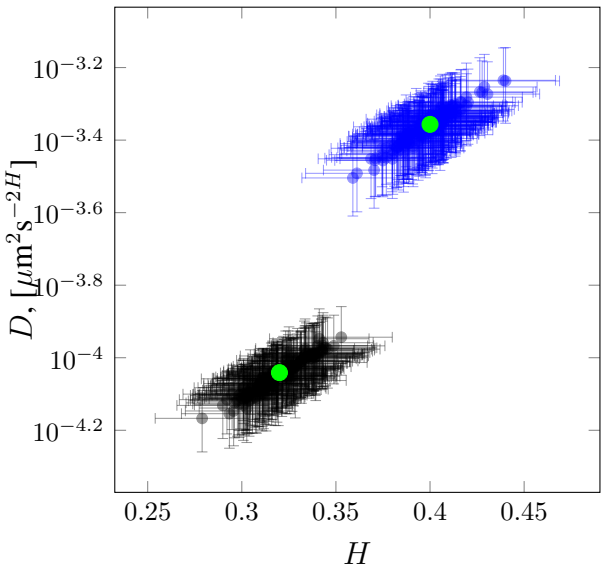
Table 4.11: Cluster results for experimental 0.2wt% agarose data.

	Count (%)	H (SD) $\times 10^{-1}$[-]		D (SD) $\times 10^{-2}$ [μm^{2}s^{-2H}]	
		expected	*predicted*	*expected*	*predicted*
Cluster 1	31 (8.7)	-	2.79 (0.67)	-	12.1 (5.72)
Cluster 2	285 (80.3)	-	1.27 (0.74)	-	0.57 (1.35)
Cluster 3	23 (6.5)	-	3.70 (0.63)	-	61.0 (33.4)
Cluster 4	16 (4.5)	-	4.60 (0.59)	-	226 (118)

Table 4.12: Cluster results for experimental 2.5wt% mucus data.

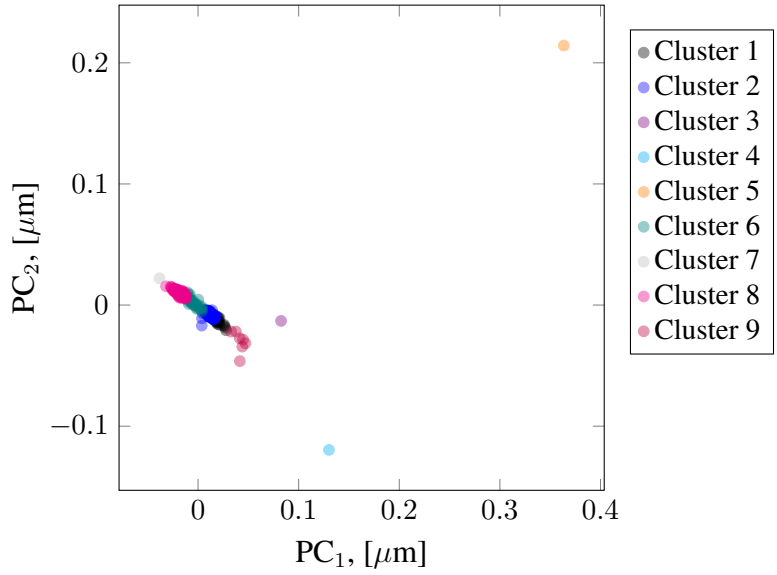


(a)

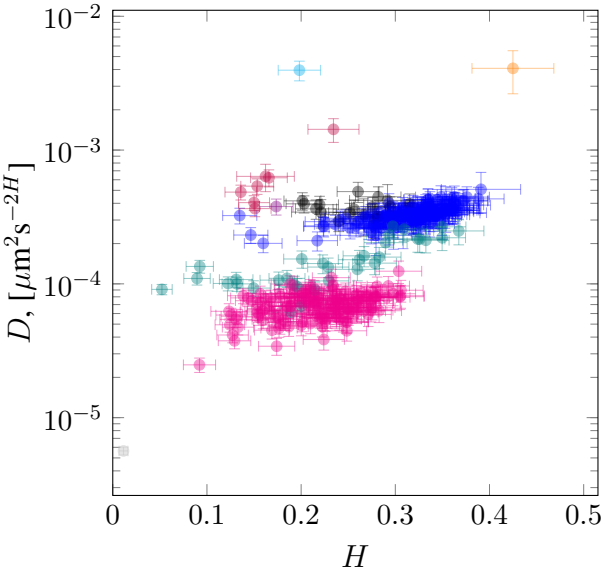


(b)

Figure 4.12: Simulated heterogeneous viscoelastic cluster results. (A) Clusters projected onto the first and second principal components. (B) Path-wise maximum likelihood estimation of Hurst parameter (H) and diffusivity (D). The green marker indicates the expected value of each parameter. Error bars indicate 95% confidence intervals.

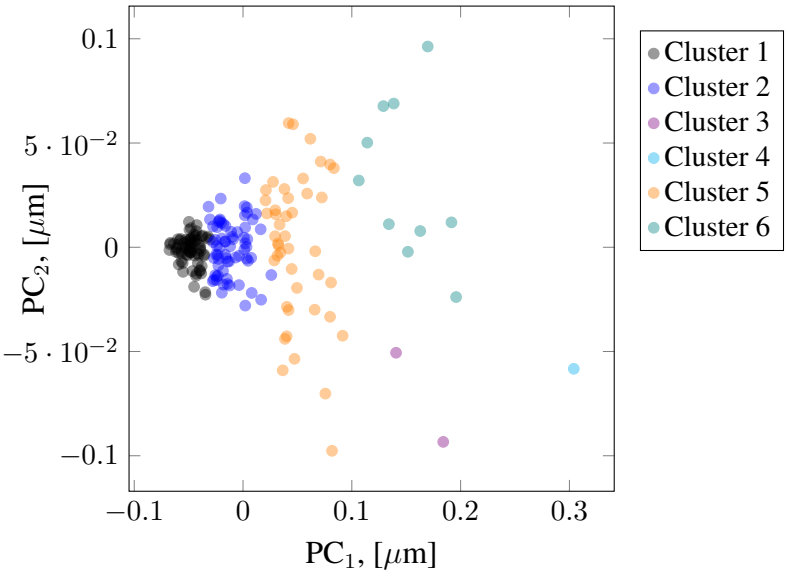


(a)

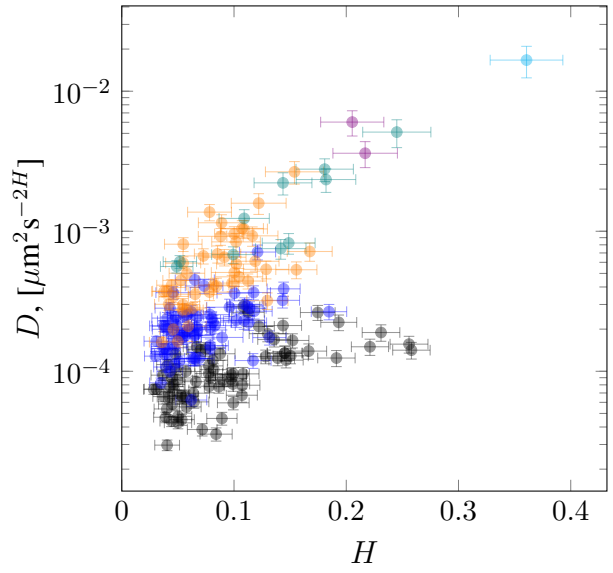


(b)

Figure 4.13: Experimental (HA) heterogeneous viscoelastic cluster results. (A) Clusters projected onto the first and second principal components. (B) Path-wise maximum likelihood estimation of Hurst parameter (H) and diffusivity (D). Nine clusters were identified when two were expected. Two clusters (Clusters 2 and 8) contain 322 of the 355 particles (90.1%). Error bars indicate 95% confidence intervals.

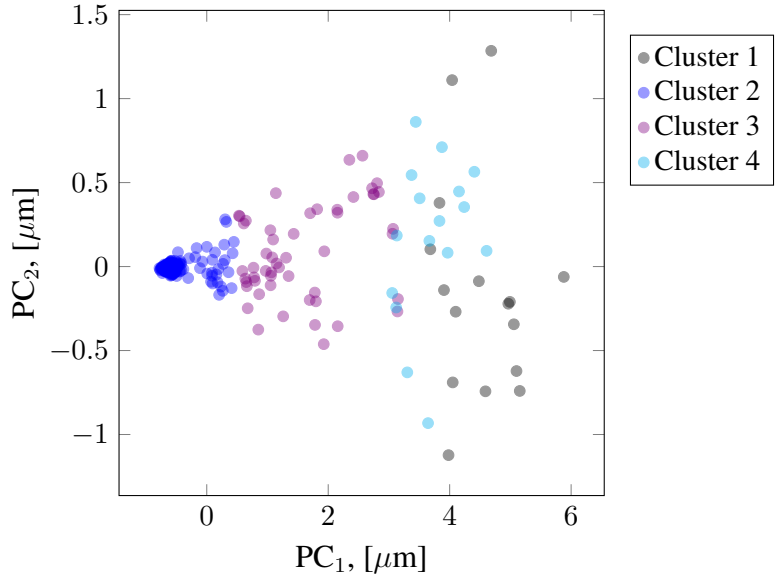


(a)

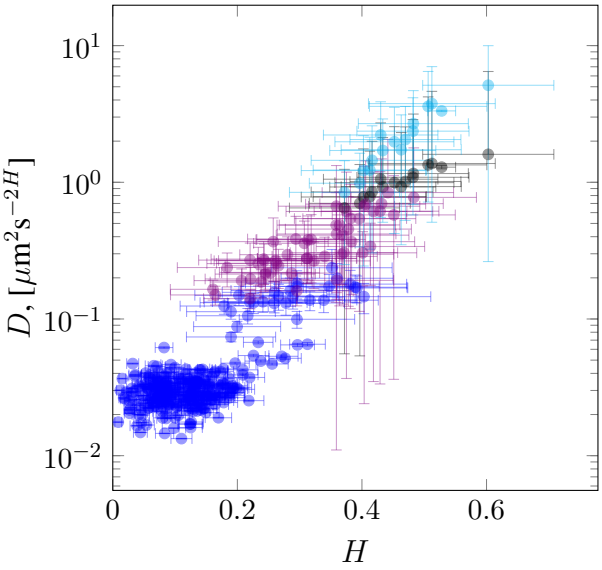


(b)

Figure 4.14: Clustering of experimental agarose data. (A) Clusters projected onto the first and second principal componenets. (B) Path-wise maximum likelihood estimation of Hurst parameter (H) and diffusivity (D). Error bars indicate 95% confidence intervals.



(a)



(b)

Figure 4.15: Clustering of experimental 2.5wt% HBE cell culture mucus. (A) Clusters projected onto the first and second principal componenets. (B) Path-wise maximum likelihood estimation of Hurst parameter (H) and diffusivity (D). Error bars indicate 95% confidence intervals.

4.6 Conclusions

A protocol for analysis of path data from passive particle tracking microrheology is presented that yields a quantitative characterization of diffusive heterogeneity in complex fluids. This protocol

is based on methods adapted from the statistics and machine learning literature. The first goal is to design an algorithm to quantify the observed heterogeneity based on the primitive path data, without reliance on a presumed model of the underlying stochastic process, beyond the minimal assumption that the increments of single paths are stationary and Gaussian. The second goal is to have a technique that yields unique, reproducible clustering of the given ensemble of paths. Similar to other approaches discussed in Section 4.2, our algorithm is applied to the position time series of passive particles in simple or complex fluids. Specifically, we partition the paths into clusters based on the principal component analysis of the scaling of the path-wise van Hove correlation functions. These clusters may arise either due to differences in particle characteristics or complex fluid characteristics, or both. Using the standard deviation of the van Hove correlation function as our metric of interest and two-pass hierarchical clustering with the gap statistic to partition the data, our algorithm yields a robust and consistent method for the detection and quantification of heterogeneity in complex fluids. The method to this point is weakly parametric, only relying on the assumption that each path is stationary and the increments are Gaussian. After the clustering step is complete, our protocol fits the parameters of a proposed model on a per-path basis, then pools the parameter values within a cluster. This approach was illustrated for standard Brownian motion and fractional Brownian motion, on both numerical and experimental data.

To benchmark our algorithm, we created data sets containing known, discrete levels of heterogeneity. We analyzed experimental data with "artificial" heterogeneity using two methods. For analysis of heterogeneity in Newtonian fluids, we embedded particles of two different diameters in a homogeneous solution (sec. 4.5.2.1). For analysis of heterogeneity in viscoelastic fluids, identical particles were embedded in two hyaluronic acid solutions of different concentrations and then the path data were combined into one dataset (sec. 4.5.2.2). For Newtonian fluids, doubling particle diameter is a proxy for doubling viscosity, or equivalently halving the diffusion coefficient. In addition to controlling the degree of heterogeneity in the paths, combining dissimilar data sets provides a way to test the accuracy of our particle-clustering assignments. Finally, we applied our protocol to monodisperse particles in two putative heterogeneous complex fluids, an agarose gel and mucus derived from human bronchial epithelial cell cultures. The data analysis reveals that both fluids are heterogeneous, and indicates a quantitative variability in sub-diffusive behavior that would have strong implications for passage times through mucus barriers.

The accuracy of our method, the small necessary volume of fluid, and the short collection times required to quantify the heterogeneous composition of viscous and viscoelastic samples, combine to make our methods promising for a wide range of applications in passive particle tracking microrheology.

CHAPTER 5
First Passage Times

In this chapter we investigate a functional form for the first passage time distribution of fractional Brownian motion from an arbitrary interval and with arbitrary scaling. Such a function may be used to predict the transit times of compounds through a viscoelastic fluid layer over time-scales that cannot be experimentally observed. We identify an interesting phenomena exhibited by fractional Brownian motion that is contrary to the common interpretation of the terms "subdiffusion" and "superdiffusion."

5.1 Introduction

First passage time is the time required for an object to exit a region for the first time. In biology, first passage times play a fundamental role in dictating infection rates, drug onset times and other important processes. Transmucosal drug delivery techniques often require a drug to transit the body's natural mucus layer before reaching the target site [20–22, 140]. The inhalation-based delivery of dry powder insulin for the treatment of diabetes is an example of such a scenario [141]. If the drug never transits the mucosal layer, it will not reach the target site and would thus fail to benefit the patient (apart from a placebo response). Similarly, HIV and other pathogens must physically transit the mucosal layer before they can infect the host and the exact mechanism by which mucus and its constituents retard transit is an active area of research [27, 142–144].

The human body has a host of natural macro-scale defenses against foreign microbes that work in tandem with the mucosal layer to prevent infection. In the lungs, mucociliary clearance and cough transport foreign matter embedded in pulmonary mucus to the back of the throat where it is swallowed, neutralized in the acidity of the digestive system, and subsequently removed from the body. Mucus drainage in the female reproductive track and the blinking of the eye similarly transport threats away from vulnerable epithelial surfaces. Techniques to retard the movement of pathogens,

should be judged, not by their ability to completely arrest the movement of pathogens, but by their ability to slow movement to the point where the body's natural clearance mechanisms may take over. Indeed, a reduction in infection rates may be achievable by altering the movement of a pathogen such that its expected first passage time is comparable to the timescale of the relevant clearance process.

Whether to decrease the onset time of an inhaled medication or reduce viral infection rates, it is important to understand how modifications of underlying diffusive processes, either by altering the diffusing object or the physical properties of the surrounding fluid, impact first passage time distributions.

For standard Brownian motion (sBm), explicit formulas exist describing the scaling of the mean first passage time (FPT) distribution with the interval width. Recent work by das Neves et al. [26] and others [24, 142, 145, 146] have applied transit time estimates based on sBm dynamics to particles that clearly exhibit subdiffusive, non-Brownian behavior. We strongly advise against this and demonstrate that the results of this type of approximation are highly erroneous. O'Malley et al. [147] recently proposed a scaling relationship for the FPT for fractional Brownian motion (fBm) as a function of the baseline mean FPT from the unit interval. We are unaware of any result for this baseline FPT expectation for fBm. In Section 5.2 of this chapter, we present the scaling relation proposed by O'Malley et al. In Section 5.3 we analyze the discrete exact algorithm for simulation of Riemann-Liouville fBm (RL-fBm) to motivate a functional form of the baseline mean FPT from the unit interval on which the O'malley conjecture is based. This function found to agree well with simulation results and may be used to estimate the mean first passage time of particles exhibiting subdiffusive fractional Brownian behavior through a fluid layer of a given thickness as a function of Hurst parameter H and and diffusivity D.

5.2 Theory

As noted in chapter 2.1, Weyl-fBm $\{B_H(t)\}_{t\geq 0}$, and RL-fBm under certain conditions, exhibit the self-similarity scaling relation

$$B_H(ct) \sim c^H B_H(t). \tag{5.1}$$

We are interested in motion that does not necessarily begin at the origin and denote such processes as

$$X_H(t) := x_o + B_H(t). \tag{5.2}$$

The exit time of this process from an interval is defined by

$$\tau_H(x_0, [a, b]) := \inf\{t > 0 : X_H(t) \notin [a, b]\}. \tag{5.3}$$

Given the initial position x_0, we introduce the function $T(x_0)$ that expresses the mean FPT from the unit interval $[0, 1]$

$$T_H(x_0) := \mathbf{E}[\tau_H(x_0, [0, 1])]. \tag{5.4}$$

For Brownian motion this function takes the form $T_{1/2}(x_0) = x_0(1 - x_0)$. O'Malley et al. [147] recently noted that the scaling relationship in (5.1) implies similar rescaling in the mean first passage time. To complete the argument started there, we observe that

$$\mathbf{E}\left[\tau_H(x_0, [a, b])\right] = (b - a)^{1/H} T_H\left(\frac{x_0 - a}{b - a}\right). \tag{5.5}$$

Let $\hat{\tau}_H(L)$ be shorthand for $\tau_H(0, [-L, L])$. For any L, the cumulative distribution function for the random variable $\hat{\tau}_H(L)$ can be expressed in terms of the cdf for $\hat{\tau}_H(1)$. Indeed

$$\begin{aligned}
P\left\{\hat{\tau}_H(L) > t\right\} &= P\left\{\sup_{s\in[0,t]} |B_H(s)| < L\right\} && (5.6)\\
&= P\left\{\sup_{s\in[0,t]} \frac{1}{L}|B_H(s)| < 1\right\} && (5.7)\\
&= P\left\{\sup_{s\in[0,t]} |B_H(L^{-1/H}s)| < 1\right\} && (5.8)\\
&= P\left\{\sup_{s\in[0,L^{-1/H}t]} |B_H(s)| < 1\right\} && (5.9)\\
&= P\left\{\hat{\tau}_H(1) > L^{-1/H}t\right\} && (5.10)
\end{aligned}$$

Using the tail formula for the expectation, we have

$$\begin{aligned}\mathbf{E}[\hat{\tau}_H(L)] &= \int_0^\infty P\left\{\hat{\tau}_H(L) > t\right\} dt \\ &= \int_0^\infty P\left\{\hat{\tau}_H(1) > L^{-1/H}t\right\} dt \\ &= L^{1/H}\int_0^\infty P\left\{\hat{\tau}_H(1) > u\right\} du \\ &= L^{1/H}\mathbf{E}[\hat{\tau}_H(1)].\end{aligned}$$

where the third line follows from the substitution $u = L^{-1/H}t$.

O'Malley et al. further conjectured that the exit time baseline function will generalize as

$$T_H(x_0) = T_H\left(0.5\right) 2^{1/H} x_0^{1/2H} (1 - x_0)^{1/2H}. \tag{5.11}$$

We confirm this conjecture (Figure 5.4). To our knowledge, there is no known form or conjecture for the leading coefficient and baseline mean FPT $T_H(0.5)$. In the next section, we analyze discrete exact simulation method for Riemann-Liouville fBm to motivate a functional form for $T_H(0.5)$.

5.3 Simulations and analysis

To investigate the relationship between H and the mean FPT, we generate paths using the Exact Discrete simulation method discussed in Section 2.4.2.2. Recall the following representation of Riemann-Liouville fractional Brownian motion (RL-fBm) from Section 2.4:

$$B_H(t) = \frac{1}{\Gamma(H + 1/2)} \int_0^t (t-u)^{H-1/2} \xi(u) du, \qquad t \geq 0. \tag{5.12}$$

Furthermore, recall that several groups have used RL-fBm simulation algorithms based on the discrete approximation to (5.12), such as Muniandy et al. [55] and Rambaldi et al. [57]. The discrete approximation is

$$B_H(t_j) = \sum_{i=1}^{j} \int_{(i-1)\Delta t}^{i\Delta t} (t_j - u)^{H-1/2} \xi_i du, \tag{5.13}$$

where $\xi \sim \mathcal{N}(0, 1)$. Exact evaluation of the integral results in

$$B_H(t_j) = \Delta t^{1/2} \sum_{i=1}^{j} \xi_i w_{j-i+1} \tag{5.14}$$

with

$$w_i = \frac{1}{\Gamma(H+1/2)} \left[\frac{t_i^{H+1/2} - (t_i - \Delta t)^{H+1/2}}{(H+1/2)} \right]. \tag{5.15}$$

We can Taylor expand the $(t_i - \Delta t)^{H+1/2}$ term at $\Delta t = 0$:

$$w_{i,H} = \Gamma(H+1/2)^{-1}(H+1/2)^{-1} \Big[t_i^{H+1/2} - t_i^{H+1/2} + \Delta t (H+1/2) t_i^{H-1/2} - \frac{1}{2}\Delta t^2 (H-1/2)(H+1/2) t_i^{H-3/2} + \mathcal{O}(\Delta t^3) \Big] \tag{5.16}$$

$$= \Gamma(H+1/2)^{-1} \left[\Delta t \; t^{H-1/2} - \frac{1}{2}\Delta t^{3/2} (H-1/2) t^{H-3/2} + \mathcal{O}(\Delta t^{5/2}) \right]. \tag{5.17}$$

Let the leading term be denoted as $g(t, H)$, i.e.,

$$g(t, H) := \frac{\Delta t \; t^{H-1/2}}{\Gamma(H+1/2)}. \tag{5.18}$$

We may have also arrived at the same form for $g(t, H)$ by recognizing that (5.15) represents the discrete derivative of $t_i^{H+1/2}$ as Δt approaches zero. That is,

$$\lim_{\Delta t \to 0} w_i = \lim_{\Delta t \to 0} \frac{1}{\Gamma(H+1/2)} \left[\frac{t_i^{H+1/2} - (t_i - \Delta t)^{H+1/2}}{(H+1/2)} \right] \tag{5.19}$$

$$= \lim_{\Delta t \to 0} \frac{\Delta t}{(H+1/2)\Gamma(H+1/2)} \left[\frac{t_i^{H+1/2} - (t_i - \Delta t)^{H+1/2}}{\Delta t} \right] \tag{5.20}$$

$$= \frac{\Delta t}{(H+1/2)\Gamma(H+1/2)} \frac{d}{dt_i} t_i^{H+1/2} \tag{5.21}$$

$$= \frac{\Delta t \; t^{H-1/2}}{\Gamma(H+1/2)}. \tag{5.22}$$

Figure 5.1 shows $g(t, H)$ for three different values of H. The results are striking in that they imply RL-fBm paths with smaller values of H take *less* time to exit short intervals than paths with larger values of H. Generally, we associate $H < 0.5$ with sub-diffusion and $H > 0.5$ with

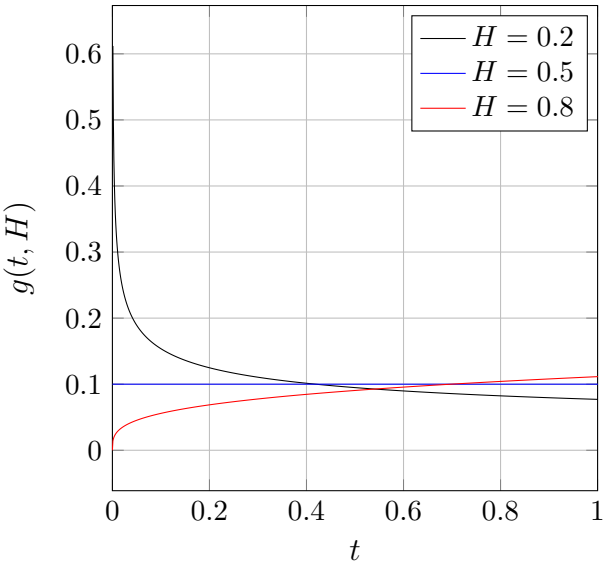


Figure 5.1: Approximate scaling of the leading weighting term $g(t, H)$ for the Exact Discrete approximation to Riemann-Liouville fBm. $g(t, H)$ is shown for three values of H and $t \in [0, 1.5]$.

super-diffusion, therefore we would think $B_{H<0.5}(t)$ requires *longer* to exit any interval compared to $B_{H>0.5}(t)$. We note that a similar derivation for Weyl-fBm shows that this qualitative behavior over small intervals is *not unique to RL-fBm*. Independent of Δt, the point t^*, at which $g(t, H < 0.5)$ becomes subordinate to $g(t, H = 0.5)$, or the point that $g(t, H = 0.5)$ becomes subordinate to a $g(t, H < 0.5)$, can be computed as

$$t^* = \Gamma(H + 1/2)^{\frac{2}{2H-1}}. \tag{5.23}$$

The two cross-over points t^* seen in Figure 5.1 are $t^* = 0.419$ and $t^* = 0.697$. To better understand the connection between $g(t, H)$ and exit times, recall the discrete approximation to (5.12) given by (5.13). For the first time step, the position is,

$$B_H(t_1) = \xi_i w_{1,H}. \tag{5.24}$$

For two processes defined by $\{H_1, H_2 \mid 1 > H_2 \geq .5 > H_1 > 0\}$, the ratio of the distances traveled during the first time step, denoted r_1, is

$$r_1 := \frac{B_{H_1}(t_1)}{B_{H_2}(t_1)} = \frac{\sqrt{\Delta t}\xi_1 w_{1,H_1}}{\sqrt{\Delta t}\xi_1 w_{1,H_2}} = \frac{w_{1,H_1}}{w_{1,H_2}} \approx \frac{\Gamma(H_2 + 1/2)\Delta t^{H_1 - 1/2}}{\Gamma(H_1 + 1/2)\Delta t^{H_2 - 1/2}} = \frac{\Gamma(H_2 + 1/2)\Delta t^{H_1 - H_2}}{\Gamma(H_1 + 1/2)} \tag{5.25}$$

where $H_1 - H_2 \in (-1, 0)$. Figure 5.2 shows r_1 for fixed Δt as a function of H. The relationship given by (5.25) indicates that if $B_{H_2}(\Delta t)$ travels a distance L over the course of a simulation composed of a single time step, then $B_{H_1}(\Delta t)$ travels r_1 times farther *in the same amount of time*. $B_{H_2}(\Delta t)$ and $B_{H_1}(\Delta t)$ therefore exited the intervals $(-L, L)$ and $(-r_1 L, r_1 L)$ in Δt time, respectively. We consider the ratio of the mean FPT of $B_{H_1}(t)$ and $B_{H_2}(t)$:

$$\mathbf{E}\left[\tau_H\left(0, [-r_1 L, r_1 L]\right)\right] = \mathbf{E}\left[\tau_{H_2}\left(0, [-L, L]\right)\right] \tag{5.26}$$

$$\implies (2r_1 L)^{1/H_1} T_{H_1}(0.5) = (2L)^{1/H_2} T_{H_2}(0.5) \tag{5.27}$$

$$\implies \frac{T_{H_1}(0.5)}{T_{H_2}(0.5)} = \frac{(2L)^{1/H_2}}{(2r_1 L)^{1/H_1}} \tag{5.28}$$

$$= \frac{\Gamma(H_1 + 1/2)^{1/H_1} 2^{1/H_2} L^{1/H_2}}{L^{1/H_1} 2^{1/H_1} \Gamma(H_2 + 1/2)^{1/H_1} \Delta t^{1 - H_2/H_1}} \tag{5.29}$$

$$= \Gamma(H_2 + 1/2)^{-1/H_1} \Gamma(H_1 + 1/2)^{1/H_1} 2^{1/H_2 - 1/H_1} L^{1/H_2 - 1/H_1} \Delta t^{H_2/H_1 - 1} \tag{5.30}$$

Let $H_2 = 0.5$, then we have

$$\frac{T_{H_1}(0.5)}{T_{0.5}(0.5)} = \Gamma(H_1 + 1/2)^{1/H_1} 2^{2 - 1/H_1} L^{2 - 1/H_1} \Delta t^{1/2H_1 - 1}. \tag{5.31}$$

We know $T_{0.5}(0.5) = 0.25$, therefore if $L = 0.25$, we can solve for $T_{H_1}(0.5)$:

$$T_{H_1}(0.5) = 1/4 \Gamma(H_1 + 1/2)^{1/H_1} 0.5^{2 - 1/H_1} \delta^{1/2H_1 - 1}, \tag{5.32}$$

where we have replaced Δt with the variable δ. For a fBm process that is parameterized by the diffusivity D as well as H, the mean FPT through an arbitrary interval of width L is thus given by,

$$T_H(L) = \frac{1}{4}\left(\frac{L^2}{D}\right)^{1/2H} \Gamma(H_1 + 1/2)^{1/H_1} 0.5^{2 - 1/H_1} \delta^{1/2H_1 - 1}. \tag{5.33}$$

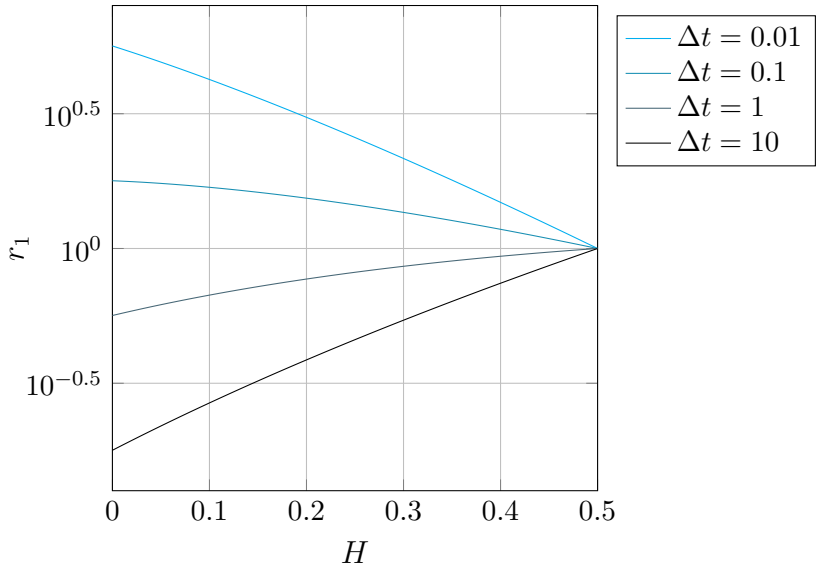


Figure 5.2: Ratio of the initial step size r_1 for the Exact Discrete approximation to Riemann-Liouville fBm. r_1 is shown for four values of Δt as a function of H.

5.4 Results

Figure 5.3 shows $T_H(0.5)$ as a function of H. It is not clear how the term δ should be interpreted since in our motivation of (5.32) we assumed that the simulation was only one step in length. Clearly, for certain choices of H and Δt, a fBm process cannot exit the interval $[-0.5, 0.5]$ in a single time step. We thus choose to disassociate this parameter from the temporal resolution of the simulation (hence the change in symbol) and treat it as a free parameter to be inferred from simulations. In Figure 5.3, $T_H(1/2)$ was fit to the simulation results with δ as the sole free parameter. The best fit value of δ was 0.01271 with 95% confidence bounds (0.009524, 0.0159). Because all simulation results represent over-estimations of the true mean FPT, the best fit value of δ is likely an over-estimation of the true value.

To investigate the "bioavailability" of our simulated paths at the interval boundary, let $\Omega_H(p)$ be the time required for proportion p of the particles to escape from the interval $[0, 1]$ given $x_0 = 0.5$, i.e.,

$$\Omega_H(p) := \{t \mid P\{\hat{\tau}_H(1) < t\} = p\}. \tag{5.34}$$

Figure 5.5 presents simulation-based estimations of $\Omega_H(p)$ for various values of H.

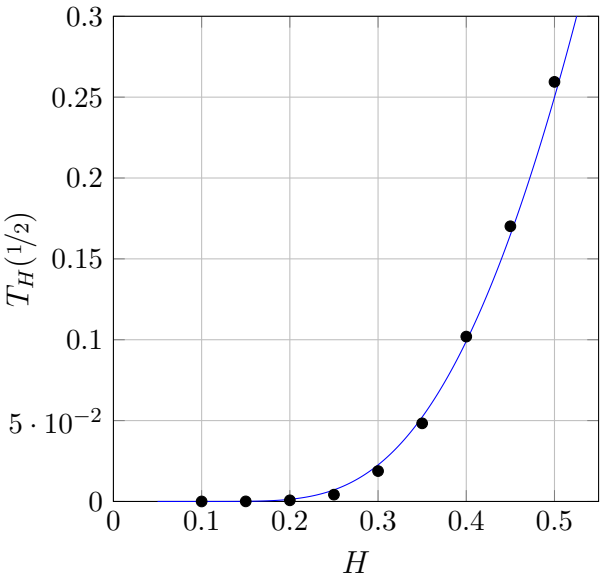


Figure 5.3: Scaling of the mean FPT from the interval $[0, 1]$ starting from $x_0 = 0.5$ as a function of H for Riemann-Liouville fBm based on (5.32) (blue line) and simulation results (black dots).

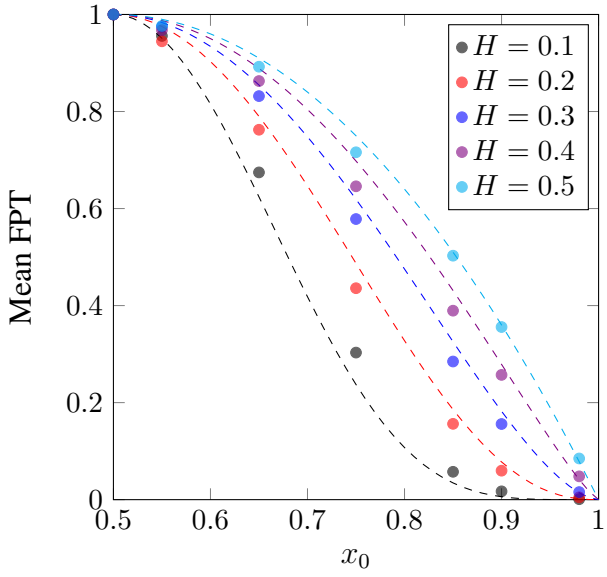


Figure 5.4: Scaling of mean FPT from the interval $[0, 1]$ with initial location x_0, for different values of H. The circles indicate the simulation results and dashed lines are the fit to the theoretical scaling. Results confirm the conjecture by O'Malley et al [147], (5.11)).

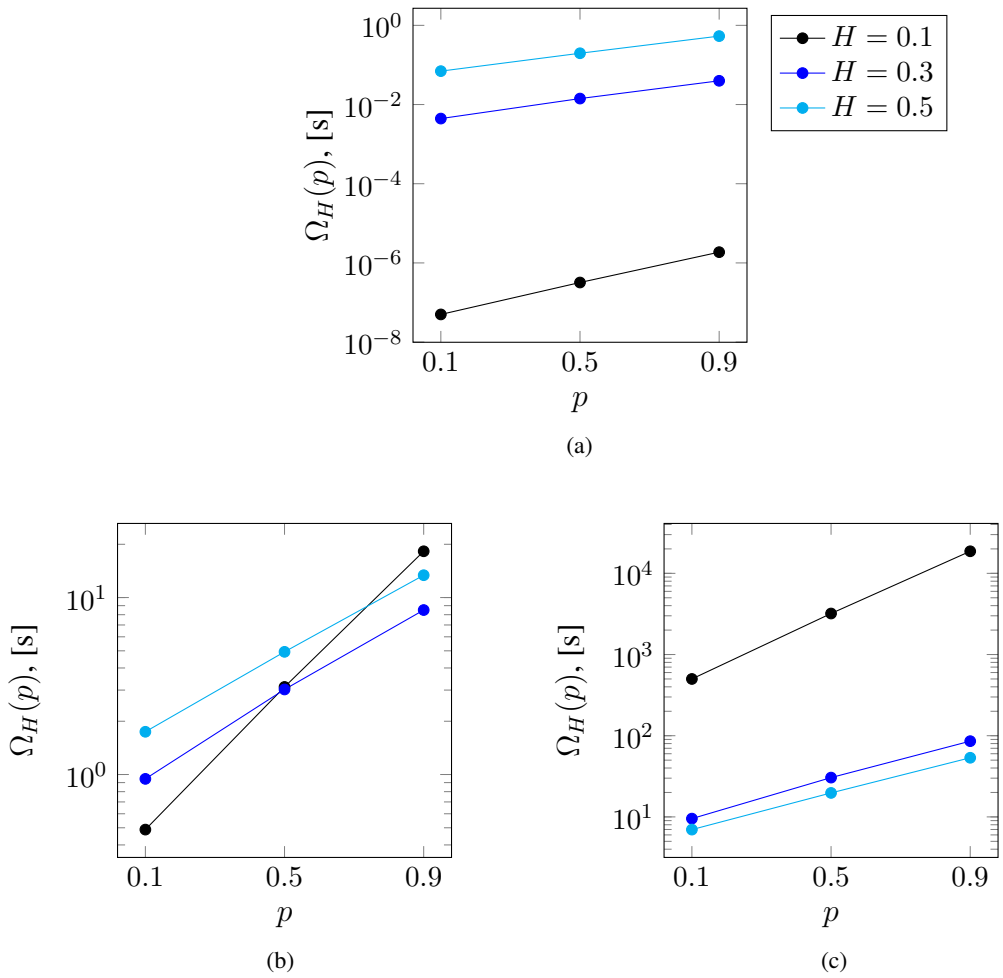


Figure 5.5: Time required to achieve a specified bioavailability at the interval boundary for three values of H for three intervals. Simulations were conducted on the unit interval (a), then the scaling relation (5.1) was applied to estimate $\Omega_H(p)$ for intervals of width $L = 5$ and $L = 10$. Initially, for $H = 1$ and an interval of width $L = 1$, the time required to achieve 10%, 50% and 90% bioavailability is a small proportion of the time required for less subdiffusive processes. As the width of the interval increases, this trend is reversed and the Brownian process with $H = 0.5$ is able to transit the interval most efficiently.

5.5 Conclusion

A functional form for the baseline mean first passage time (FPT) from the unit interval is presented as a function of the Hurst parameter H. This function was motivated by analysis of the Exact Discrete simulation method for Riemann-Liouville fractional Brownian motion (RL-fBm) presented in Section 2.4.2.2. For RL-fBm processes exiting small intervals of width $L \in [0, \sim 1]$, the mean first passage time is directly related to H, resulting in simulated particle paths with low H exiting faster than particles with large H. This behavior is contrary to the usual association of subdiffusive behavior with retarded movement. We emphasize that this behavior is not unique to RL-fBm and is also present in Weyl-fBm. When the generalized mean FPT function is used to scale the baseline mean FPT function to investigate exit times from intervals larger than the unit interval, the expected relationship between mean FPT and Hurst parameter is recovered, i.e., particles with low H take longer to exit than particles with larger H. This extreme change in the qualitative behavior of the mean FPT function emphasizes the importance of accurately estimating the diffusive parameter values describing a particle's motion.

CHAPTER 6

Analysis of Pediatric Bronchoalveolar Lavage Samples

In this chapter we analyze the microrheological properties of bronchaolveolar lavage samples from a pediatric cystic fibrosis cohort utilizing the tools discussed in the previous chapters. We address the primary issue associated with such studies– the differentiation of the fluid used to perform the bronchaolveolar lavage from the airway surface liquid. Strong agreement is observed between the diffusive parameters of particles in the bronchaolveolar lavage samples and particles in mucus harvested from human bronchial epithelial cell cultures, emphasizing the importance of such systems as a model for pulmonary pathogenesis.

6.1 Introduction

Cystic fibrosis (CF) affects between 70,000 to 100,000 people worldwide and is caused by mutations in CFTR, the CF transmembrane conductance regulator gene. CFTR malfunction disrupts cellular sodium and chloride balance across the epithelial membrane and impacts organs throughout the body. The pancreas, in particular, often exhibits pronounced damage, even at birth [148]. In contrast, the lungs of newborns with CF often appear normal, however *pulmonary infections*, specifically B. cepacia and S. matophilia, are strongly correlated with adult morbidity and mortality [149–152]. Several theories have been proposed for the pulmonary manifestations of CF disease progression, such as hindered secretion of microbial factors and bicarbonate [153, 154], ceramide and salt accumulation [155, 156], and disrupted phagocytosis [157], however recent research has focused on the role of airway surface liquid hydration [158–160].

Airway surface liquid is a general term used to refer to the mixture of surfactants; mucin molecules and other proteins; DNA from dead cells; immunological agents such as macrophages, pathogens, and inhaled particulates; and bacteria coating the pulmonary epithelium. This fluid layer may be roughly conceptualized as a viscoelastic gel bilayer. The mucus half of the bilayer, at the

air-liquid interface, is the most prominent, between 0.5-60 μm thick [160–164], and composed of a mixture of proteins, including the major gel forming glycoproteins MUC5B and MUC5AC [165, 166]. Under the mucus layer, the periciliary layer (PCL) makes up the second half of the gel bilayer and also exhibits viscoelastic properties, although a different combination of proteins leads to properties distinct from those of the mucus layer. The PCL is approximately 7 μm thick, and in direct contact with the epithelium where it simultaneously facilitates cilia movement and hinders the diffusion of foreign particulates that were not trapped by the mucus layer above [162, 167]. Dehydration of the PCL and mucus layers is believed to lead to a collapse of the PCL layer, decreasing in the effectiveness of mucociliary and cough clearance of mucus, and resulting in mucus stasis. As mucus accumulates, the airways become occluded by "mucus plugs," creating an ideal environment for bacterial colonization [159].

The Australian Respiratory Early Surveillance Team for CF (AREST CF) study is an ongoing effort in Australia to understand the early progression of CF lung disease. Computed tomography (CT) chest scans of children with CF collected through the AREST CF study reveal a spatially heterogeneous manifestation of disease symptoms throughout the lungs [168]. In order to quantify factors driving early, heterogeneous manifestations of CF, the AREST CF study is conducting bronchoalveolar lavages (BAL) and CT scans, and collecting exhaled breath condensate from its enrolled pediatric CF cohort. The invasive BAL results are analyzed for biochemical and biophysical measures of mucus hydration, as well as indications of inflammation and bacterial colonization. The goal is to correlated these metrics with alternative biomarkers derived from the less invasive CT scans and exhaled breath condensate.

Recent work by Hill et al. [23] demonstrated that mucus weight percent solids (wt%), the percent of a fluid sample's weight that is not due to water, is a robust biomarker of pulmonary pathogenesis. Hill et al. collected sputum samples (mucus and saliva) from normal (1.7 ± 0.56 wt%), COPD (3.7 ± 2.3 wt%) and CF (7.0 ± 2.3 wt%) patients by either expectoration or induction using hypertonic saline. The wt% solids of each sample was measured by comparing the sample's weight before and after dehydration. This metric found to scale robustly with disease state. Mucus was then harvest from primary human bronchial epithelial (HBE) cell cultures and reconstituted to 2.5, 3, 4 and 5 wt%. Systematic characterization of the HBE mucus was conducted using passive microparticle rheology and microparticle tracking (MPT) techniques to investigate the scaling of

the linear, viscoelastic properties of mucus with wt% as a proxy for disease state. Weyl-fBm fits to the MPT path data revealed the Hurst parameter H exhibits a negative linear relationship with wt% while the diffusivity D exhibits a negative exponential relationship with wt%, both results confirm qualitative observations that subdiffive characteristics become more pronounced with increasing wt%. A prominent increase in the spread of the distribution of the path-wise Hurst parameter was observed at 4wt%, corresponding to the hypothesized sol-gel transition, the point at which the viscous modulus becomes subordinate to the elastic modulus. The change in the distribution of H at this wt% suggests structural change in the mucus gel network that may manifest as increased heterogeneity in the spatial distribution of the fluid's viscoelastic properties.

In this chapter, we utilize the analysis techniques presented in the previous chapters to compare recently collected, unpublished MPT data from the AREST CF study and MPT data from the HBE mucus samples presented by Hill et al. [23]. Our analysis suggests that mucus derived from HBE cell cultures is representative of the viscoelastic properties of mucus contained in clinical BAL samples collected from children with CF. Furthermore, the data suggest that early CF is characterized by spatial heterogeneity in the viscoelastic properties of pulmonary mucus, which may be a driving factor behind the observed heterogeneous bacterial colonization and mucus plugging.

6.2 Materials and methods

6.2.1 Sample collection

6.2.1.1 HBE Mucus

Mucus was harvest from HBE cell cultures as described by Hill et al. [23]. Briefly, non-pathological human bronchial epithelial (HBE) cells were cultured on a polycarbonate membrane scaffold coated with collagen [169–171]. Over the course of 6 weeks, confluent cultures developed cilia, generated and established a PCL and mucus layer. The cilia exhibited coordinated beating, qualitatively mimicking *in vivo* mucociliary clearance, resulting in mucus transport within the culture dish. Washings from $>$100 cultures were aggregated and concentrated to the desired wt%. Concentrated mucus was dialyzed against PBS to insure isotonicity as described elsewhere [131, 169]. Concentrations of 2.5 wt%, 3 wt%, 4 wt% and 5 wt% were utilized.

6.2.1.2 AREST CF

The AREST CF study plans to conduct bronchoalveolar lavages on 50 pediatric CF subjects. For each subject, the first BAL is scheduled to occur at age 3 months and is repeated yearly from age 1 to age 4. During a BAL, a small volume of 0.9% NaCl is deposited on the airway surface, then removed so as to entrain a sample of native airway surface liquid. Upon collection, BAL fluid samples (mucus and NaCl) are divided into 1 ml aliquots and frozen at -80 °C, then transported from Australia to UNC-Chapel Hill for microrheological analysis. To date, BAL samples have been collected from 23 children 0.5 to 5.5 years old. No replicates are currently available.

6.2.2 Particle tracking

AREST CF and HBE cell culture mucus samples were combined with 1 μm diameter carboxylated fluorescent particles diluted in light scattering buffer (LSB) and allowed to mix for 24 hours prior to particle tracking. Particle tracking for both the HBE and AREST CF samples was conducted as described in section 4.3. Briefly, a Nikon Eclipse TE2000-U at $40\times$ magnification and standard video microscopy techniques were used to collect video of particles undergoing thermal diffusion. Each video is 30 seconds in length with a temporal resolution of 60 fps. Video spot tracking software (`http://cismm.cs.unc.edu/`) extracts the positions of each particle of interest in the field of view as a function of time. These position time series form the basis of our analysis.

6.2.3 Initial path filtering

The microparticle path data from the AREST CF samples underwent four screens to remove "corrupt" data. Here we describe the first three screens while the fourth is addressed in Section 6.2.4.

The video spot tracking software used to extract the positions of each particle is increasingly inaccurate the closer the particle is to the edge of the field of view, and this error has yet to be systematically investigated. To mitigate this, particles whose initial positions were within 9 μm of the edge of the field of view were excluded from further analysis (Figure 6.1). The width of the boundary region was heuristically chosen based on a manual inspection of representative position time series. 11.27% ($n = 3415$) of all particles fell within this boundary region and excluded from the final data analysis.

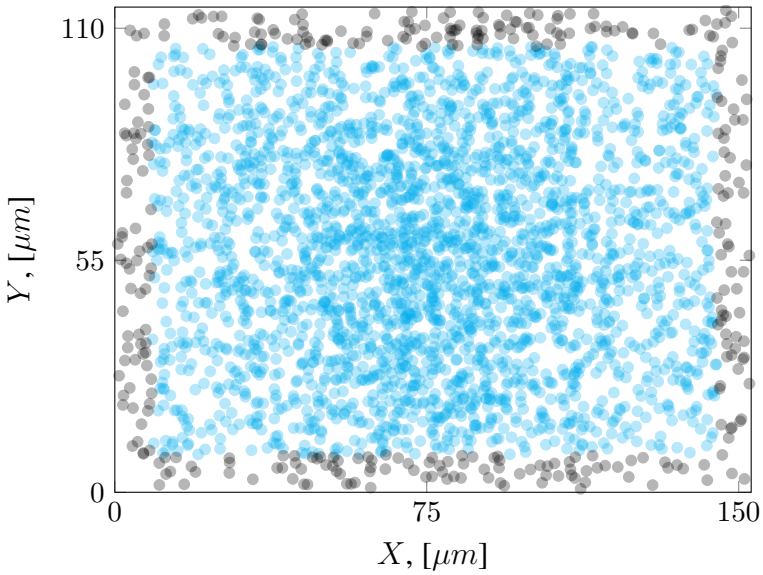


Figure 6.1: Overlay of the initial positions of a representative sample of particle paths from the AREST CF data set. The field of view of the microscope is approximately 150 μm by 110 μm. Particles in the exclusion region near the edge of the field of view are shown in black.

The position time series reported by the video spot tracking software is not representative of the actual movements of a particle if that particle exits the field of view (laterally) or the focal plane (vertically) during the video. In these circumstances, the position time series will be distorted and indicate a prolonged period without movement in one or more directions (Figure 6.2). Paths with more than 200 time points where the displacement was reported to be exactly zero were identified and removed. A sequence of 200 consecutive time points with zero displacement results in an approximate 5% error in the estimation of D and a 1% error in the estimation of H. 1.07% ($n = 330$) of all particles, including those previously screened, met this threshold and withheld from subsequent analysis.

Particle paths with position time series containing fewer than 100 data points were removed to ensure enough data were present for reasonably accurate fitting. 2.11% ($n = 654$) of all particles, including those previously screened, met this threshold and withheld from subsequent analysis.

After applying the three filters described in this section, 88.2% ($n = 26725$) of the data points remain. Figure 6.3 shows a representative sampling of the data points remaining and the data points that were removed.

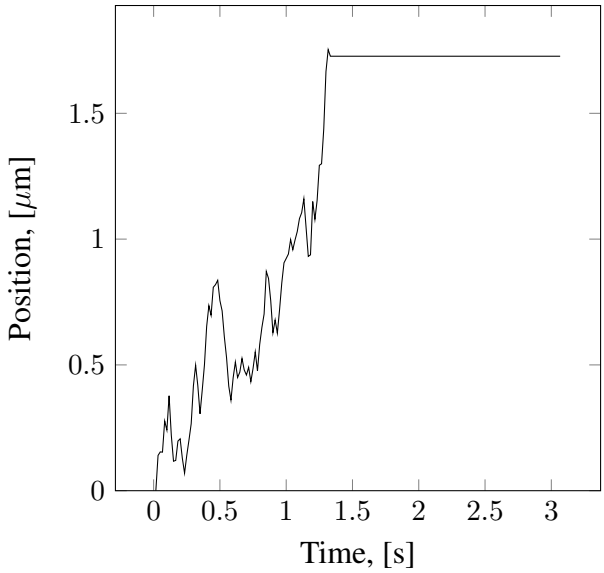
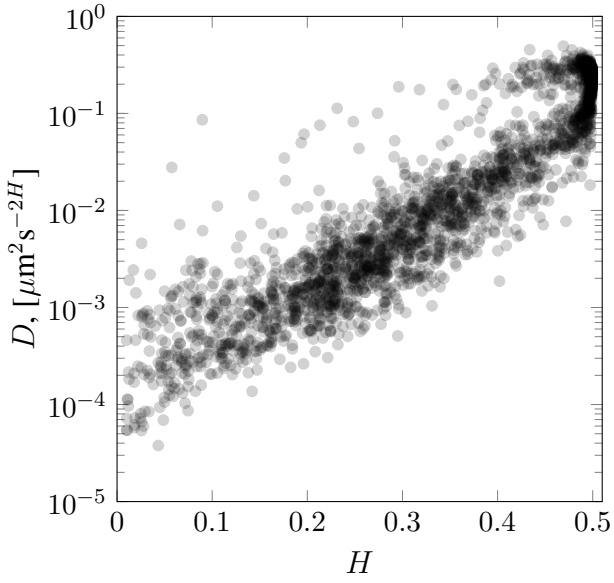


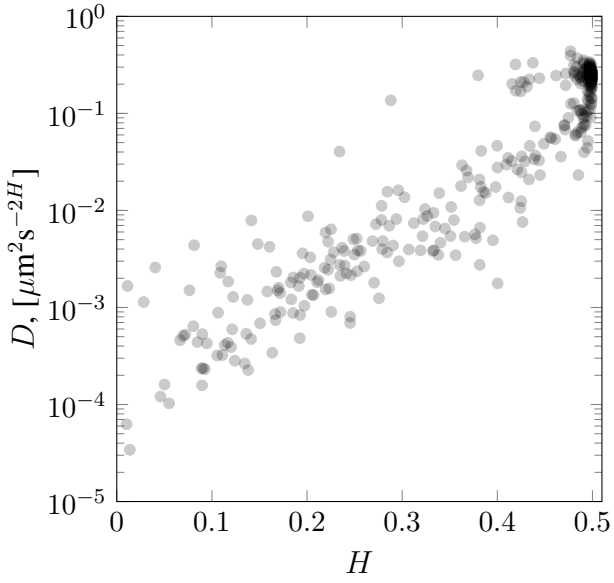
Figure 6.2: Example particle path exhibiting no movement over multiple time steps. The particle likely diffused vertially out of the focal plane since at 1.5 seconds, it was not at the edge of the field of view.

6.2.4 The Background Fluid Problem

A pernicious impediment to the rheological investigation of BAL samples is the difficulty in distinguishing the properties of the airway surface liquid from the properties of the fluid used to carry out the BAL–in this case, 0.9% NaCl [172, 173]. We refer to this as the "Background Fluid Problem." Additionally, limited fluid volume makes rheometer-based characterization difficult. Due to these factors, most analysis of BAL samples focuses on the enumeration of cellular constituents and their characteristics, and the culturing of native bacteria [172]. Microrheology is ideally suited to address the constraints posed by limited BAL volumes. It is not initially clear however, how the Background Fluid Problem may be solved via microrheological techniques given that additional dilution of the BAL sample occurs upon the addition of tracer particles, which have been diluted in light scattering buffer (LSB)—as reported in Section 6.2.1. We are unaware of any microrheological investigations of pediatric BAL samples to date. To address the Background Fluid Problem, we employ the model fitting techniques presented in Chapter 3 in conjunction with the heterogeneity metrics introduced in Chapter 4.



(a)



(b)

Figure 6.3: Representative samples of data points remaining (a), and data points removed (b), after accounting for the path length, initial position and abnormal displacements, as described in Section 6.2.3. Data points are presented in terms of their maximum likelihood diffusive parameter values.

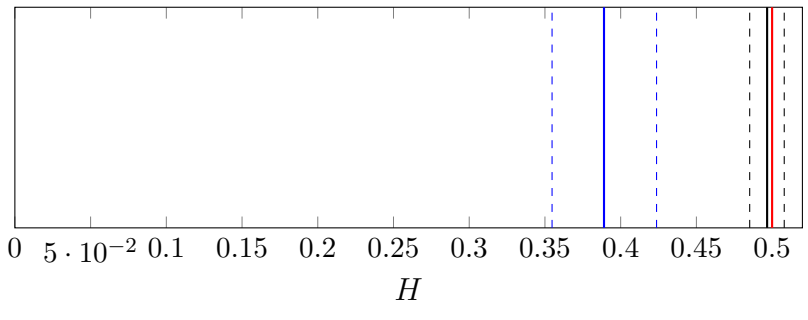
Upon injection of the BAL sample with tracer particles and the tracking of their movements, the goal is to fit the increment process of each particle path (as opposed to the MSD) to a fully parametric model, such as that discussed in Section 3.3. Under the proposed model, maximum

likelihood estimates of each parameter and their confidence intervals can be calculated. Microparticle tracking data was collected for 1 μm diameter particles in a 0.9% NaCl, LSB solution as a control (the "background solution"). This data set contained 254 particles that underwent a similar filtering process to that described in Section 6.2.3. Based on the movements of these particles, we calculate the expected value of the diffusivity, D_{bf} and Hurst parameter, H_{bf}. These values are $D_{bf} = 0.38$ μm^2 s^{-2H} and $H_{bf} = 0.49$ Particles from the AREST CF data set with 95% confidence intervals about their MLE diffusive parameter values that include the expected diffusive parameter values in the background fluid are noted, and we refer to these data points as Background Fluid Points. Table 6.1 provides a summary of the comparative results and Figure 6.4 illustrates this process for two particles from the AREST CF data set. In this figure, D_{bf} and H_{bf} are shown in red. The solid blue and black lines represent the MLE estimates for two particles and the dotted lines represent the 95% confidence interval. The 95% confidence interval about the MLE of the diffusive parameters for the particle shown in blue does not include D_{bf} and H_{bf}. The data therefore suggest that this particle is not probing an unadulterated solution of NaCl and LSB. In contrast, the 95% confidence interval about H for the particle shown in black includes H_{bf}, we therefore declare this point to be a Background Fluid Point. Background Fluid Points make up 36.6% ($n = 9789$) of the data points, after applying the filters described in the previous section.

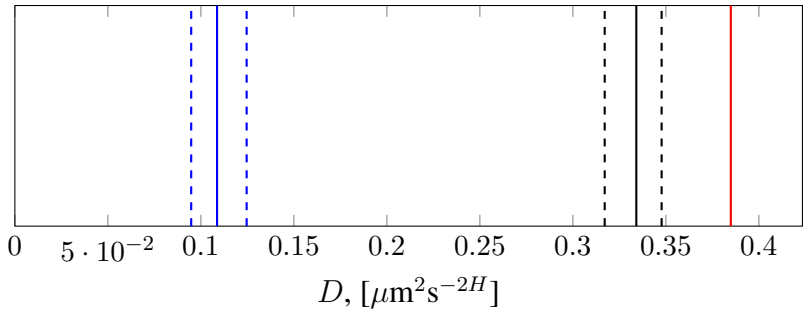
n, (%)	$H \approx H_{bf}$	$H \not\approx H_{bf}$
$D \approx D_{bf}$	495 (1.9)	132 (0.5)
$D \not\approx D_{bf}$	9162 (34.3)	16936 (63.4)

Table 6.1: Breakdown of data points based on whether or not the 95% confidence interval about each parameter value includes the expected parameter values for the background solution. 1.9% of particles have 95% confidence intervals about both H and D that include H_{bf} and D_{bf}. 63.4% of particles have 95% confidence intervals about both parameter that do not include H_{bf} and D_{bf}.

Since this work represents the first microrheological characterization of airway surface liquid from a pediatric CF cohort we hesitate to declare that non-Background Fluid Points are sampling the airway surface liquid in a biologically relevant context. It is not known how swelling of the mucins proteins might impact particle movement or if local aggregation of mucin proteins within the airway surface liquid results in sub-regions with decreased diffusivity but minimal elasticity.



(a)



(b)

Figure 6.4: Example maximum likelihood estimates of H and D for two particles (blue and black) from the AREST CF data set. The dashed lines indicate the 95% confidence interval about the MLE parameter value (solid line). The red lines indicate the expected parameter values D_{bf} and H_{bf}, for a 1 μm diameter particle in the background fluid, NaCl+LSB. We declare the data point represented in black to be a Background Fluid point because its 95% confidence interval about H includes H_{bf}.

The constituents of the the airway surface liquid layer may diffuse into the surrounding background fluid, altering both the rheological properties of the background fluid, e.g., through a increase in protein concentrations, and the airway surface liquid, through a corresponding decrease in protein concentrations. It is therefore possible that particles exhibiting values of H and D distinct from H_{bf} and D_{bf} may not be sampling a fluid representative of the *in vivo* airway surface liquid layer.

To investigate this question further, we apply the clustering algorithm from Chapter 4 to the data with and without the Background Fluid Points, after applying the filters discussed in Sec. 6.2.3. Figure 6.6 shows the clustering of the data with the Background Fluid points in PC_1-PC_2 and D-H space. The dashed red lines in panel (b) indicate D_{bf} and H_{bf}. The mean values of D and H for each cluster, and the cluster sizes are reported in Table 6.2. Figure 6.7 and Table 6.3 present the same

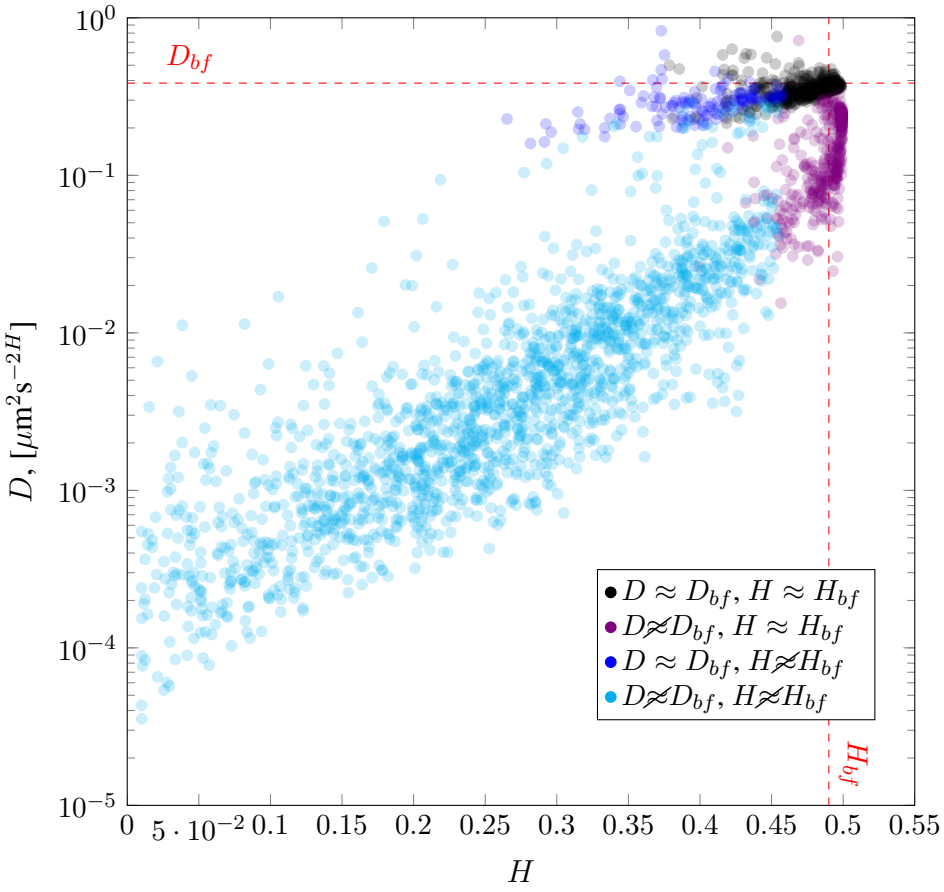
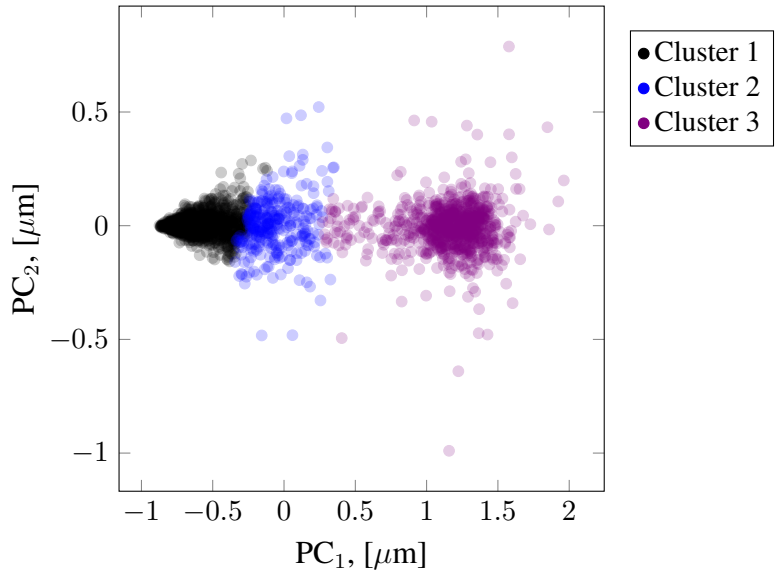


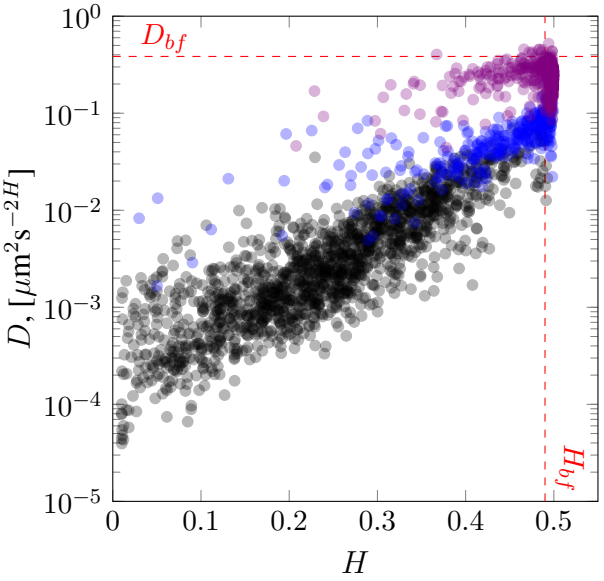
Figure 6.5: Representative distribution of the AREST CF data points, grouped based on their distinguishability from the background fluid. The dashed red lines indicate H_{bf} and D_{bf}, the expected parameter values in the background solution.

data after the Background Fluid Points have been removed. Importantly, the clustering results of both data sets indicate the presence of three clusters in approximately the same locations.

As observed in Table 6.1, the majority of the data points characterized as Background Fluid Points received that designation solely based on the maximum likelihood estimate of H. This suggests that proteins and other cellular constituents of the airway surface liquid are diffusing into the NaCl+LSB solution and increasing its apparent viscosity, thus decreasing the diffusivity of the embedded probe particles. By comparing the measured diffusivity of particles in the NaCl control solution with the mean diffusivity of Cluster 3 in Table 6.2, we note a decrease of approximately 36%. We refer to this as the "modified background solution." Our goal is to identify data points with



(a)



(b)

Figure 6.6: Clustering of pediatric bronchoalveolar lavage samples from the AREST CF study with Background Fluid Points. (a) Clusters of representative data projected onto the first and second principal componenets. Cluster 1 contains 57.0% ($n = 15239$) of the data points while Cluster 2 and Cluster 3 contain 11.92% ($n = 3186$) and 31% ($n = 8286$) of the data points, respectively. (b) Path-wise maxiimum likelihood estimation of the Hurst parameter (H) and diffusivity (D) for the representative data. The dashed red lines indicated the expected values of D and H for a 1 μm diameter particle in the background fluid.

a high likelihood of inhabiting the modified background solution that have not been classified as Background Solution Points.

In Figure 6.6, Cluster 1 and Cluster 3 seem to correspond to discrete sets of fluid properties, with Cluster 2 representing an intermediate transition region. Upon removal of the Background Fluid Points, the number of data points in Cluster 1 decreased by 2.9% while the number of data points in Cluster 2 and Cluster 3 decrease by 48.1% and 94.2%, respectively. We may associate the mean diffusivity of Cluster 3 with the mean diffusivity of the modified background solution since they are in strongest agreement. The shape of Cluster 3 in D-H space indicates a relatively constant diffusivity over a range of H values decreasing from $H = 0.5$. One explanation for this is that the modified background fluid has permeated the mucin mesh and displaced the native intra-mesh fluid. An alternative explanation is that the gel forming mucins in the airway surface liquid have reassembled to formed a new mesh network in the background fluid. Regardless of the explanation, the continued existence of points in Cluster 2 and Cluster 3 after the removal of the Background Fluid Points suggests that this phenomena has a statistically significant impact on the viscoelastic properties of the fluid, whether it be an increase in the local diffusivity of the airway surface liquid or an increase in the local elasticity of the background fluid. Furthermore, it suggests that there are in fact particles inhabiting a modified background solution that simultaneously exhibit D and H distinct from D_{bf} and H_{bf}, but who are also unlikely to be sampling the airway surface liquid. For subsequent analysis, we choose to focus on the viscoelastic properties of airway surface liquid that has had minimal interaction with the NaCl+LSB solution. The cluster results suggest that Cluster 1 contains the desired data points.

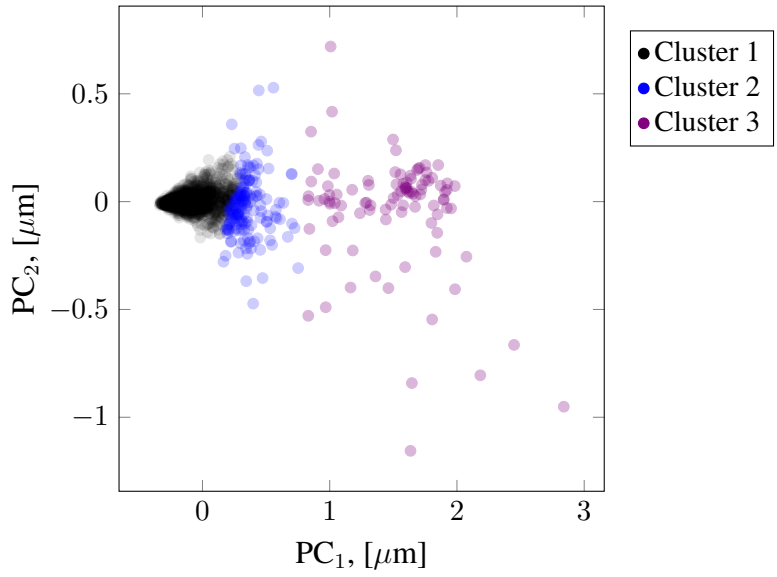
	Count (%)	H (SD) $\times 10^{-1}$[-]	D (SD) $\times 10^{-2}$ [μm^{2}s^{-2H}]
Cluster 1	15239 (57.0)	2.46 (1.09)	0.64 (0.95)
Cluster 2	3186 (11.9)	4.28 (0.79)	5.86 (3.33)
Cluster 3	8286 (31.0)	4.87 (0.32)	24.6 (21.7)

Table 6.2: Cluster results for pediatric bronchoalveolar lavage samples from the AREST CF study with Background Fluid Points.

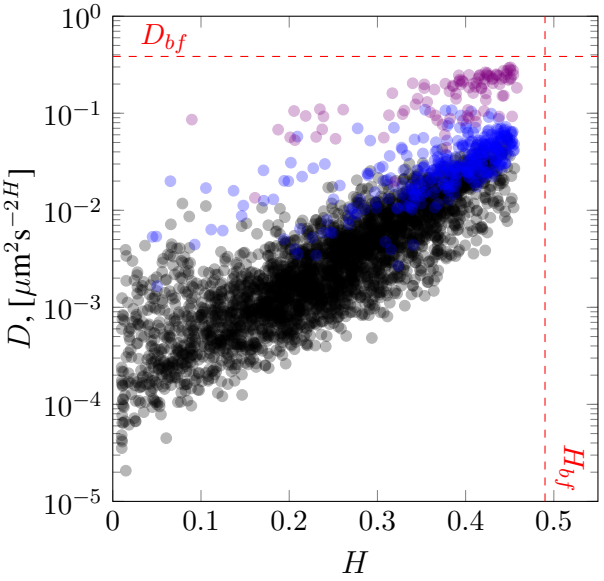
	Count (%)	H (SD) $\times 10^{-1}$[-]	D (SD) $\times 10^{-2}$ [μm^{2}s^{-2H}]
Cluster 1	14797 (87.4)	2.40 (1.05)	0.55 (0.74)
Cluster 2	1655 (9.8)	3.71 (0.79)	3.37 (1.93)
Cluster 3	483 (2.9)	3.86 (0.64)	16.8 (7.05)

Table 6.3: Cluster results for bronchoalveolar lavage samples from the AREST CF study without Background Fluid Points

After applying all filters to the data and removing the Background Solution Points, Cluster 1 contains 14,797 particles. Within this cluster, the maximum likelihood estimates of the Hurst parameter range from $H = 0.01$ to $H = 0.46$, and in each case is distinct from $H = 0.49$. The maximum likelihood estimates of the diffusivity range from $D = 1.9 \times 10^{-5}$ μm^{2}s^{-2H} to $D = 5.6 \times 10^{-2}$ μm^{2}s^{-2H}.



(a)



(b)

Figure 6.7: Clustering of pediatric bronchoalveolar lavage samples from the AREST CF study without Background Fluid Points. (a) Clusters of representative data projected onto the first and second principal componenets. Cluster 1 contains 87.4% ($n = 14797$) of the data points while Cluster 2 and Cluster 3 contain 9.8% ($n = 1655$) and 2.9% ($n = 483$) of the data points, respectively. (b) Path-wise maximum likelihood estimation of Hurst parameter (H) and diffusivity (D) for representative data. The dashed red lines indicate the expected values of D and H for a 1 μm diameter particle in the background fluid.

6.3 Results

6.3.1 HBE data

The path-wise maximum likelihood estimates of D and H were calculated as described in Chapter 3 and the clustering algorithm presented in Chapter 4 was applied to the HBE microparticle data. In each of the 2.5 wt%, 3 wt%, and 5 wt% data sets, one cluster was identified. In the 4 wt% data set, 6 clusters were identified, one of which contained only a single data point. A summary of the results at each wt% are presented in Tables 6.4-6.7.

	Count (%)	H (SD) $\times 10^{-1}$[-]	D (SD) $\times 10^{-2}$ [μm^{2}s^{-2H}]
Cluster 1	76 (100)	3.57 (0.53)	2.94 (1.53)

Table 6.4: Cluster results for 1 μm diameter particles in 2.5 wt% HBE mucus

	Count (%)	H (SD) $\times 10^{-1}$[-]	D (SD) $\times 10^{-2}$ [μm^{2}s^{-2H}]
Cluster 1	99 (100)	3.21 (0.51)	1.63 (1.49)

Table 6.5: Cluster results for 1 μm diameter particles in 3 wt% HBE mucus

	Count (%)	H (SD) $\times 10^{-1}$[-]	D (SD) $\times 10^{-3}$ [μm^{2}s^{-2H}]
Cluster 1	53 (29.4)	1.90 (0.43)	1.74 (0.67)
Cluster 2	51 (28.3)	2.21 (0.45)	3.30 (1.02)
Cluster 3	10 (5.6)	4.25 (0.18)	35.0 (7.76)
Cluster 4	39 (21.7)	2.81 (0.47)	6.89 (2.21)
Cluster 5	26 (14.4)	3.24 (0.70)	13.65 (4.04)

Table 6.6: Cluster results for 1 μm diameter particles in 4 wt% HBE mucus

	Count (%)	H (SD) $\times 10^{-1}$[-]	D (SD) $\times 10^{-4}$ [μm^{2}s^{-2H}]
Cluster 1	178 (100)	1.10 (0.35)	3.90 (2.45)

Table 6.7: Cluster results for 1 μm diameter particles in 5 wt% HBE mucus

6.3.2 AREST CF data

The path-wise maximum likelihood estimates of D and H with 95% confidence intervals are shown in Figure 6.8. The data are grouped by their cluster assignment after the removal of the Background Fluid Points. In [23], Hill et al. report the following heuristic equations relating H and D to mucus weight percent solids (wt%):

$$D \approx 1.6e^{-1.5\times \text{wt\%}}, \qquad 1.5 \leq \text{wt\%} \leq 5, \tag{6.1}$$

$$2H \approx -0.17 \times \text{wt\%} + 1.1, \qquad 1.5 \leq \text{wt\%} \leq 5. \tag{6.2}$$

Based on the mean MLE values of each parameter for Cluster 1–thought to most closely represent the airway surface liquid–the predicted weight percents given by (6.1) and (6.2) are 3.6 wt% and 3.5 wt%, respectively. The predicted weight percents for Cluster 2 are 1.68 wt% and 2.14 wt% from (6.1) and (6.2), respectively. The predicted weight percents for Cluster 3 are 1.46 wt% and 1.25 wt% from (6.1) and (6.2), respectively.

In Figure 6.9, the cluster results from the HBE data (shown in color) are overlaid on the path-wise estimations of D and H for the BAL data from Cluster 1 (shown in black). For the HBE data, the size of the marker is directly proportional to the percentage of the data points at the given wt% in each cluster. For 2.5 wt%, 3 wt%, and 5 wt%, one cluster was identified. For 4 wt%, five clusters (and one outlier, not shown) were identified. Of note, the data from Cluster 1 *alone* spans the distribution of parameter values observed in the HBE data, suggesting that we did not "over filter" the data in Sections 6.2.3 and 6.2.4. The BAL data exhibit a slightly depressed diffusivity compared to the HBE data near $H = 0.5$, suggesting that at low mucin concentrations, other factors, such as cellular debris or inhaled particulates that are not present in the HBE culture mucus, play a more important role in

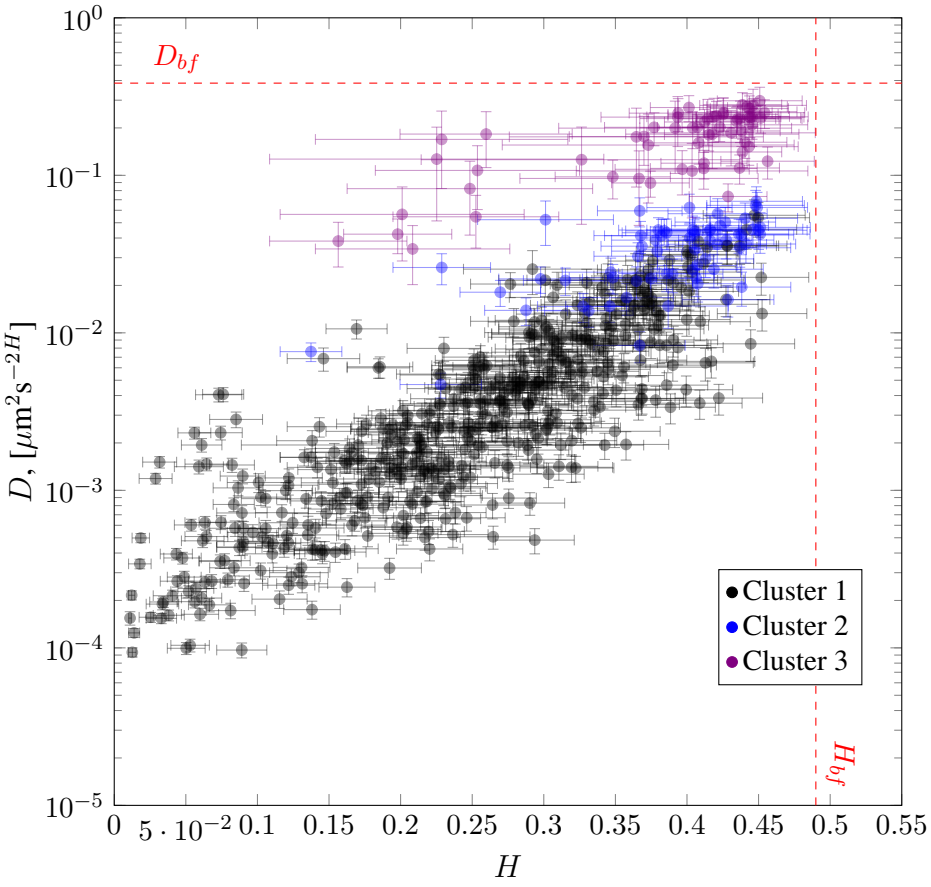


Figure 6.8: Maximum likelihood estimates of D and H for a representative subset of the AREST CF data. The error bars represent 95% confidence intervals. All filters from Section 6.2.3 have been applied and the Background Points have been removed.

defining the viscoelastic properties of the airway surface liquid. As disease progresses, gel-forming mucins begin to have a dominate affect or airway surface liquid properties, evidence be the strong agreement between the 5wt% HBE data and the BAL data at low values of H.

Path data were matched to subject-level demographic descriptors, such as age and sex, and sample-level markers of inflammation, such as IL-8 and neutrophil elastase. These "meta data" were available for 52 subjects. Of the 23 BAL samples available for analysis, only 17 had corresponding meta data. These 17 subjects correspond to $n = 21687$ data points, or 71.6% of the original data set. In Figure 6.10, the mean values of D and H within Cluster 1 are shown for each of the 17 subjects. The size of the marker is proportional to the number of particles available for that subject. The color of the marker is proportional to the total cell count (tcc) for each patient. It is important to note that different BAL samples, from different parts of the lung, were used to determine the

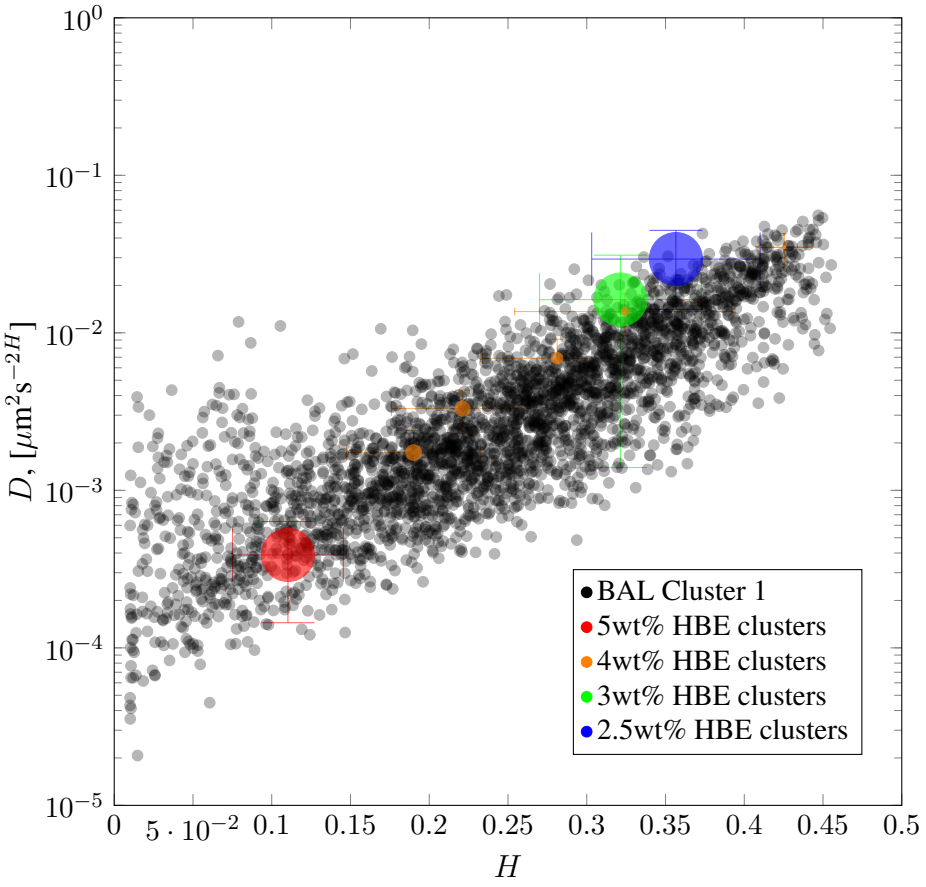


Figure 6.9: Comparison of bronchoalveolar lavage sample results from the AREST CF study to clustering of HBE data. Error bars indicate one standard deviation about the cluster mean. The size of the HBE markers is commiserate with the proportion of particles in the respective data set within each cluster. For 2.5 wt%, 3 wt%, and 5 wt%, one cluster was identified. For 4 wt%, five clusters were identified.

microrheological properties and the total cell count. No statistically distinct correlation is observed between the sample-wise ensemble average of H and D and tcc, or any other subject-level meta data. Additionally, the sample-wise ensemble average of H and D *within Cluster 1* was not correlated with any of the meta data variables in a statistically distinct manner. This may be due to the limited number of subjects associated with both meta data and microparticle data ($n = 17$), or it may indicate that changes in the rheological properties of the airway surface liquid are "upstream" of existing metrics for CF lung disease.

An additional limitation observed in Figure 6.10 is that a compression of the distributions of D and H to a subject-level mean and standard deviation masks a significant amount of potentially

biologically-relevant variability in the data. The presentation of the subject-level data in this fashion does not adequately represent the richness of the data, which follows a highly non-Gaussian distribution. As an alternative, Figure 6.11 shows distributions of particles in D-H space for six subjects. These data highlight the inter-subject diversity present in the BAL data.

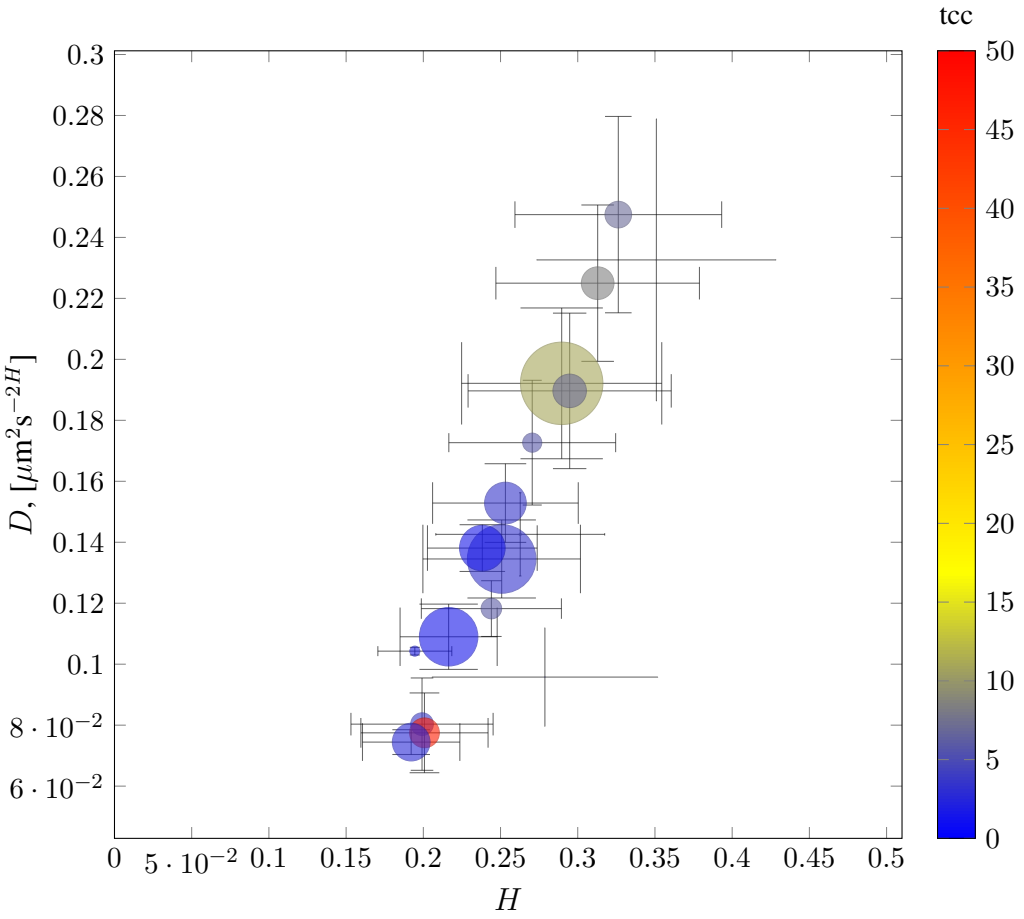


Figure 6.10: Inter-subject comparison of total cell count in D-H space. Each circle represents one subject. The size of the marker is proporitonal to the number of data points for that subject. Circles are colorcoded by the subject's total cell count (tcc) and error bars represent standard deviations about the mean. No obvious relationship between H and D, and total cell count is apparent.

The Improved Discrete Riemann-Liouville simulation algorithm (Section 2.4.2.3) was used to simulate trajectories for 1 μm diameter particles in HBE mucus. The simulations were run for a total of 18 minutes with an inter-observational time of $1/60$ seconds. The number of particles able to transit a distance L in the given amount of time may be thought of as the bioavailability of the particles at the pulmonary epithelium after transiting a mucus layer of thickness L. Figure 6.13

shows the predicted bioavailability of 1 μm particles through the given wt% mucus as a function of mucus layer thickness. Intuitively, as the thickness of the mucus layer decreases, all particles are expected to make a successful transit. As the mucus layer thickness increases, drastic differences are

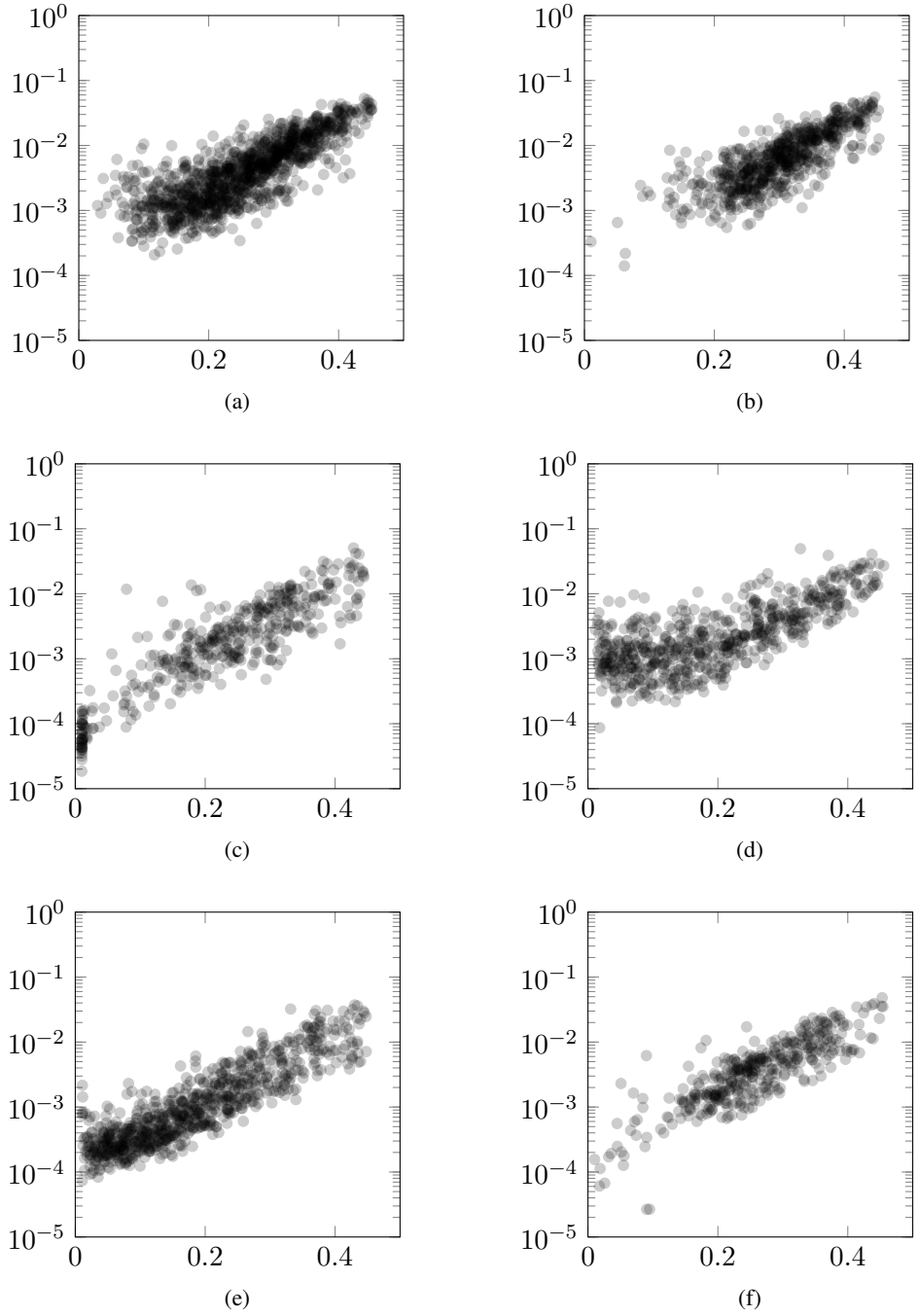
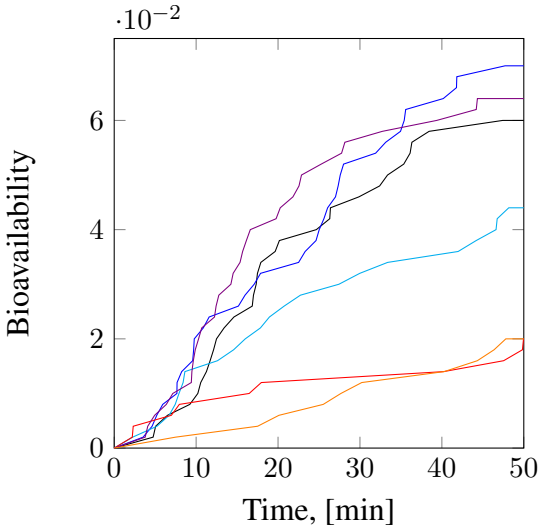
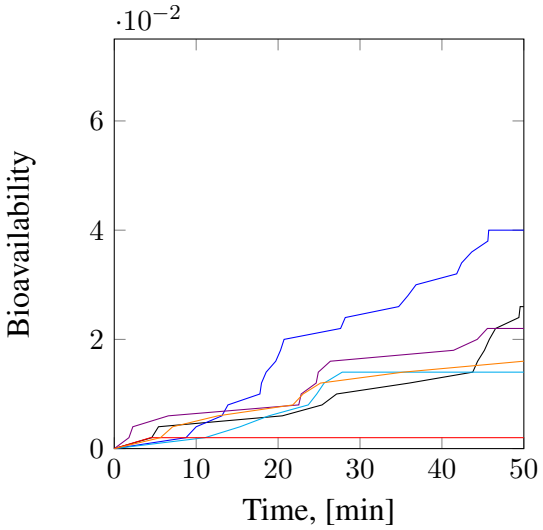


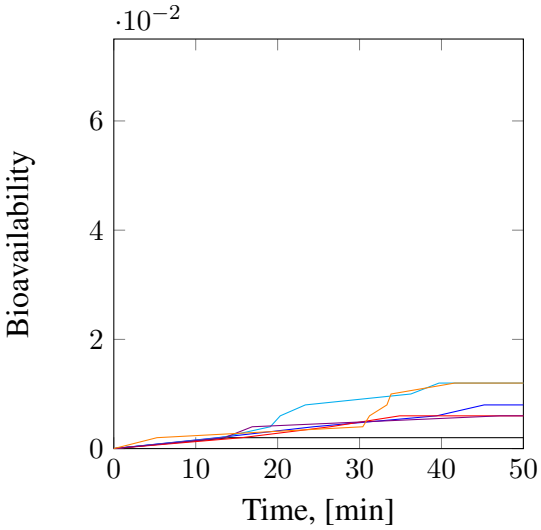
Figure 6.11: Patient-level distributions of D (y axis) and H (x axis) from the AREST CF data set for six subjects. The x- and y-labels match those of the previous figures.



(a)



(b)



(c)

Figure 6.12: Subject-level predictions of the bioavailability of a 1 μm diameter particle through an airway surface liquid layer with thickness 10 μm (a), 15 μm (b), and 20 μm (c) for the six subjects represented in Figure 6.11.

apparent between 2.5 wt% mucus, serving as a proxy for healthy mucus, and 5 wt% mucus, serving as a proxy for adult CF mucus. The dashed black line indicates the typical mucus layer thickness in the upper respiratory track. At this mucus layer thickness, the data suggest a significant disparity bioavailability between healthy and early-stage CF subjects with mucus concentrations near 3.6 wt%.

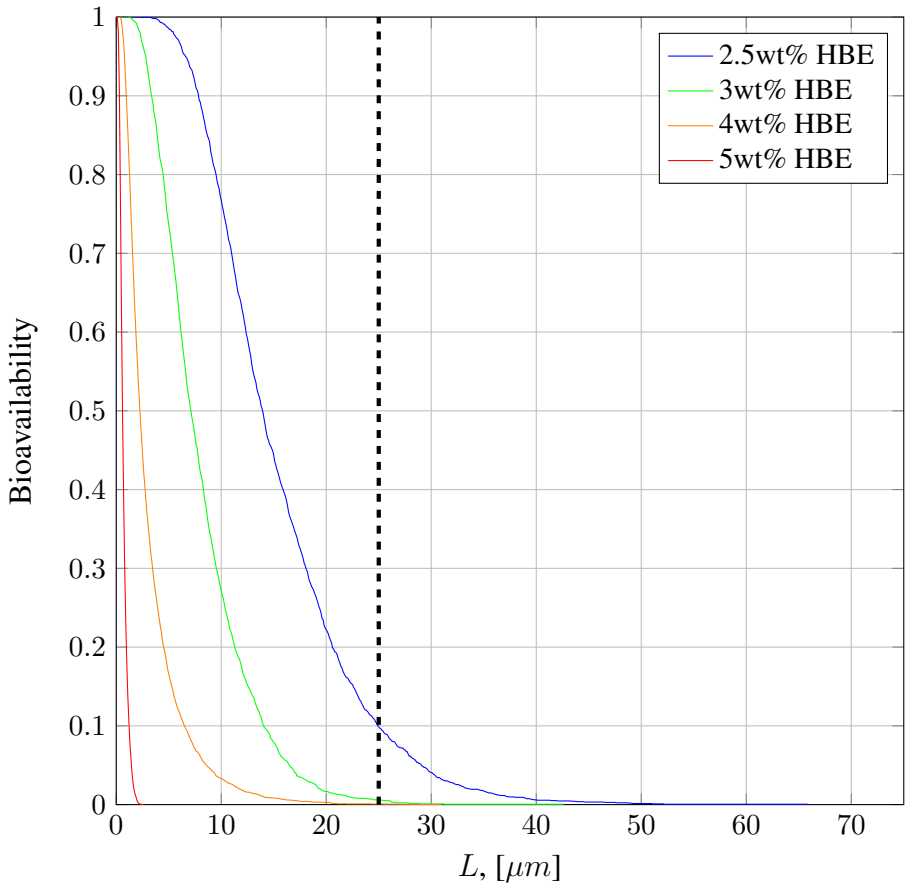


Figure 6.13: Predicted bioavailability of a 1 μm diameter carboxylated particle through a mucosal layer of thickness L in 18 minutes for four different mucus concentrations. The vertical black line (dashed) shows the typical thickness of the mucosal layer in the upper airways.

6.4 Conclusion

Changes in the viscoelastic properties of the airway surface liquid have the potential to be an early indicator of CF-related pathogenesis, presaging bacterial colonization, macrophage recruitment, mucus stasis and even collapse of the periciliary layer. Despite this promise, practical limitations regarding the volume of airway surface liquid available following a BAL have hindered the widespread macro-scale characterization of those properties. While not as limited by volumetric constraints, microrheology has suffered from the Background Fluid Problem–differentiation of the physical properties of the airway surface liquid from the physical properties of the background fluid, i.e., the fluid used to perform the lavage.

Initial analysis of microparticle tracking data from BAL fluids suggests a statistically significant interaction between the airway surface liquid and the background fluid, either due to displacement of the native intra-gel fluid by the background fluid, the formation of a gel structure in the background fluid itself, or some combination thereof. In an effort to quantify this complex behavior and characterize the airway surface liquid within BAL fluid, we have presented a protocol that can be used to identify, and discard data from embedded tracer particles with a high likelihood of having sampled the background fluid and not the airway surface liquid layer.

Subject-level distributions of the particle-wise diffusive parameter values reveals a potentially rich source of information that will be more thoroughly explored over the course of the AREST CF study. No statistically significant correlation was observed between the diffusive parameters of the tracer particles and the subject-level meta data. This may either due to a lack of statistical power, a true lack of correlation, or it may indicate that changes in the viscoelastic properties of the airway surface liquid are upstream, early indicators of pulmonary pathogenesis before more commonplace indicators of disease have manifested, and have therefore not yet been observed. Importantly, simulations suggest that the observed changes in viscoelastic properties of the airway surface liquid are nonetheless biologically relevant, even before mucus stasis and plugging occur.

Understanding the time course of disease will be critical to the future development of therapeutics, both to gauge the effectiveness of compounds designed to treat the underlying genetic cause of CF, as well as compounds designed to mitigate the downstream effects. To this end, these data reinforce the utility of human bronchial epithelial cell cultures as a relevant model for early CF pathogenesis on which new therapies may be tested. The use of human bronchial epithelial cell cultures facilitates the personalization of therapies and we have demonstrated how data derived from such cell culture models may be used to predict bioavailability of a compound through the airway surface liquid layer for a subpopulation.

REFERENCES

1. Mason, T. G. & Weitz, D. A. Optical measurements of frequency-dependent linear viscoelastic moduli of complex fluids. *Physical Review Letters* **74,** 1250 (1995).

2. Solomon, M. J. & Lu, Q. Rheology and dynamics of particles in viscoelastic media. *Current Opinion in Colloid and Interface Science* **6,** 430–437. ISSN: 13590294 (2001).

3. Levine, A. J. & Lubensky, T. C. Response function of a sphere in a viscoelastic two-fluid medium. *Physical review. E, Statistical, nonlinear, and soft matter physics* **63,** 041510. ISSN: 1063-651X (2001).

4. Mason, T. G. Estimating the viscoelastic moduli of complex fuids using the generalized Stokes-Einstein equation. *Rheologica Acta* **39,** 371–378 (2000).

5. Mason, T., Ganesan, K., van Zanten, J., Wirtz, D. & Kuo, S. Particle Tracking Microrheology of Complex Fluids. *Physical Review Letters* **79,** 3282–3285. ISSN: 0031-9007 (Oct. 1997).

6. Squires, T. M. & Mason, T. G. Fluid Mechanics of Microrheology. *Annual Review of Fluid Mechanics* **42,** 413–438 (2010).

7. Hess, W. & Klein, R. Generalized hydrodynamics of systems of Brownian particles. *Advances in Physics* **32,** 173–283. ISSN: 0001-8732 (1983).

8. Chaikin, P. M. & Lubensky, T. *Principles of Condensed Matter Physics* (Cambrdige University Press, Cambridge, 1994).

9. Fletcher, D. A. & Geissler, P. L. Active biological materials. *Annual review of physical chemistry* **60,** 469–486. ISSN: 0066-426X (2009).

10. Mackintosh, F. C. & Levine, A. J. Nonequilibrium mechanics and dynamics of motor-activated gels. *Physical Review Letters* **100,** 1–4. ISSN: 00319007 (2008).

11. Valentine, M. T. *et al.* Investigating the microenvironments of inhomogeneous soft materials with multiple particle tracking. *Physical Review E Stat Nonlin Soft Matter Phys* **64,** 061506 (2001).

12. Aufderhorst-Roberts, A., Frith, W. J. & Donald, A. M. Micro-scale kinetics and heterogeneity of a pH triggered hydrogel. *Soft Matter* **8,** 5940–5946 (2012).

13. Aufderhorst-Roberts, A., Frith, W. J., Kirkland, M. & Donald, A. M. Microrheology and microstructure of Fmoc-derivative hydrogels. *Langmuir* **30,** 4483–4492. ISSN: 15205827 (2014).

14. Aufderhorst-Roberts, A., Frith, W. J. & Donald, A. M. A microrheological study of hydrogel kinetics and micro-heterogeneity. *The European physical journal. E, Soft matter* **37,** 44. ISSN: 1292-895X (May 2014).

15. Tuteja, A., Mackay, M. E., Narayanan, S., Asokan, S. & Wong, M. S. Breakdown of the continuum Stokes-Einstein relation for nanoparticle diffusion. *Nano Letters* **7,** 1276–1281. ISSN: 15306984 (2007).

16. Round, A. N. *et al.* Heterogeneity and persistence length in human ocular mucins. *Biophysical journal* **83,** 1661–70. ISSN: 0006-3495 (Sept. 2002).

17. Mellnik, J. *et al.* Micro-heterogeneity metrics for diffusion in soft matter. *Soft matter* **10,** 7781–96. ISSN: 1744-6848 (Oct. 2014).

18. Dawson, M., Wirtz, D. & Hanes, J. Enhanced viscoelasticity of human cystic fibrotic sputum correlates with increasing microheterogeneity in particle transport. *The Journal of biological chemistry* **278,** 50393–401. ISSN: 0021-9258 (Dec. 2003).

19. Kirch, J. *et al.* Optical tweezers reveal relationship between microstructure and nanoparticle penetration of pulmonary mucus. *Proceedings of the National Academy of Sciences of the United States of America* **109,** 18355–60. ISSN: 1091-6490 (Nov. 2012).

20. Ensign, L. M., Schneider, C, Suk, J. S., Cone, R & Hanes, J. Mucus penetrating nanoparticles: biophysical tool and method of drug and gene delivery. eng. *Adv Mater* **24,** 3887–3894 (2012).

21. Rabanel, J. M., Aoun, V, Elkin, I, Mokhtar, M & Hildgen, P. Drug-loaded nanocarriers: passive targeting and crossing of biological barriers. *Current medicinal chemistry* **19,** 3070–102. ISSN: 1875-533X (2012).

22. Fröhlich, E. & Roblegg, E. in *Intracellular Delivery II* (eds Prokop, A., Iwasaki, Y. & Harada, A.) 139–163 (Springer Netherlands, Graz, Austria, 2014). ISBN: 978-94-017-8895-3. doi:10.1007/978-94-017-8896-0.

23. Hill, D. B. *et al.* A biophysical basis for mucus solids concentration as a candidate biomarker for airways disease. *PloS one* **9,** e87681. ISSN: 1932-6203 (2014).

24. Suk, J. S. *et al.* The penetration of fresh undiluted sputum expectorated by cysitic fibrosis patients by non-adhesive polymer nanoparticles. *Engineering* **30,** 2591–2597 (2009).

25. Lysy, M. *et al.* Model comparison for single particle tracking in complex fluids, arXiv:1407.5962v1 [stat.AP] (Submitted to the Journal of the American Statistical Association).

26. Das Neves, J. *et al.* Interactions of microbicide nanoparticles with a simulated vaginal fluid. *Molecular pharmaceutics* **9,** 3347–56. ISSN: 1543-8392 (Nov. 2012).

27. Lai, S. K., Wang, Y.-Y., Cone, R., Wirtz, D. & Hanes, J. Altering mucus rheology to "solidify" human mucus at the nanoscale. *PloS one* **4,** e4294. ISSN: 1932-6203 (Jan. 2009).

28. Schuster, B. S., Suk, J. S., Woodworth, G. F. & Hanes, J. Nanoparticle diffusion in respiratory mucus from humans without lung disease. *Biomaterials Biomaterials* (2013).

29. Rubin, B. K. Mucus structure and properties in cystic fibrosis. *Paediatric respiratory reviews* **8,** 4–7. ISSN: 1526-0542 (2007).

30. Yoshida, T. & Tuder, R. M. Pathobiology of cigarette smoke-induced chronic obstructive pulmonary disease. *Physiological reviews* **87,** 1047–82. ISSN: 0031-9333 (2007).

31. Fernandes, C. A. & Vanbever, R. Preclinical models for pulmonary drug delivery. *Expert Opin Drug Deliv* **6,** 1231–45 (2009).

32. Eixarch, H., Haltner-Ukomadu, E., Beisswenger, C. & Bock, U. Drug Delivery to the Lung : Permeability and Physicochemical Character- istics of Drugs as the Basis for a Pulmonary Biopharmaceutical Classifica- tion System (pBCS). *Journal of Epithelial Biology & Pharmacology* **3,** 1–14 (2010).

33. Moschakis, T. Microrheology and particle tracking in food gels and emulsions. *Current Opinion in Colloid and Interface Science* **18,** 311–323. ISSN: 13590294 (2013).

34. Wilking, J. N. & Mason, T. G. Optically driven nonlinear microrheology of gelatin. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics* **77,** 1–4. ISSN: 15393755 (2008).

35. Weihs, D, Mason, T. G. & Teitell, M. A. Bio-microrheology: a frontier in microrheology. *Biophys J* **91,** 4296–4305 (2006).

36. Hoffman, B. D., Massiera, G., Van Citters, K. M. & Crocker, J. C. The consensus mechanics of cultured mammalian cells. *Proceedings of the National Academy of Sciences of the United States of America* **103,** 10259–10264. ISSN: 0027-8424 (2006).

37. Einstein, A. On the movement of small particles suspended in stationary liquids required by the molecular-kinetic theory of heat. *Annalen der Physik* **17,** 549–560 (1905).

38. Kolmogorov, A. N. The Wiener helix and other interesting curves in the Hilbert space. *Dokl. Acad. Sci. USSR* **26** (1940).

39. Mandelbrot, B. B. & Van Ness, J. W. Fractional Brownian Motions, Fractional Noises and Applications. *SIAM Review* **10,** 422–437 (1968).

40. Lim, S. C. Fractional Brownian motion and multifractional Brownian motion of Riemann-Liouville type. *Journal of Physics A: Mathematical and General* **34,** 1301–1310 (2001).

41. Phatarfod, R. M. Riverflow and Reservoir storage models. *Mathematical and Computer Modeling* **12,** 1057–1077 (1989).

42. Hosking, J. R. M. Modeling Persistence In Hydrological Time Series Using Fractional Differencing. *Water Resources Research* **20,** 1898–1908 (1984).

43. Hurst, H. E. Long term storage capacity of reservoirs. *Transactions of the American Society of Civil Engineers* **116,** 770–799 (1951).

44. Hurst, H. E. Methods of using long term storage in reservoirs. *Proc. Instn civ. Engrs* **1,** 519–543 (1956).

45. Willinger, W., Taqqu, M. S., Leland, W. E. & Wilson, D. V. Self-similarity in high-speed packet traffic: analysis and modeling of ethernet traffic measurements. *Statistical Science* **10,** 67–85 (1995).

46. Leland, W. E., Taqqu, M. S., Willinger, W. & Wilson, D. V. On the Self-Similar Nature of Ethernet Traffic. *ACM SIGCOMM Computer Communication Review* **23,** 183–193 (1993).

47. Guigas, G. & Weiss, M. Sampling the cell with anomalous diffusion - The discovery of slowness. *Biophys J* **94,** 90–94 (2008).

48. Davis, M. W. Production of Conditional Simulations via the LU Triangular Decomposition of the Covariance Matrix. *Mathematical Geology* **19,** 91–98 (1987).

49. Dietrich, C. R. & Newsam, G. N. Fast and exact simulation of stationary Gaussian processes through circulant embedding of the covariance matrix. *SIAM Journal of Scientific Computing* **18,** 1088–1107 (1997).

50. Krishnamoorthy, A. & Menon, D. Matrix Inversion Using Cholesky Decomposition. *CoRR* **abs/1111.4.** arXiv: arXiv:1111.4144v2[cs.MS] (2011).

51. Decreusefond, L. Stochastic analysis of the fractional Brownian motion. *Potential Analysis* **10,** 177–214 (1999).

52. Carmona, P., Coutin, L. & Montseny, G. Stochastic integration with respect to fractional Brownian motion. *Annales de l'institut Henri Poincare (B) Probability and Statistics* **39,** 27–68. ISSN: 02460203 (2003).

53. Lim, S. & Sithi, V. Asymptotic properties of the fractional Brownian motion of Riemann-Liouville type. *Physics Letters A* **206,** 311–317. ISSN: 03759601 (1995).

54. Lim, S. & Muniandy, S. Self-similar Gaussian processes for modeling anomalous diffusion. *Physical Review E* **66,** 021114. ISSN: 1063-651X (2002).

55. Muniandy, S. & Lim, S. Modeling of locally self-similar processes using multifractional Brownian motion of Riemann-Liouville type. *Physical Review E* **63,** 046104. ISSN: 1063-651X (2001).

56. Barnes, J. A. & Allan, D. W. A Statistical Model of Flicker Noise. *Proceedings of the IEEE* **54,** 176–178 (1966).

57. Rambaldi, S. & Pinazza, O. An accurate fractional Brownian motion generator. *Physica A* **208,** 21–30 (1994).

58. Peltier, Romain Francois and Lévy-Véhel, J. *Multifractional Brownian motion: definition and preliminary results* 1995.

59. Benassi, A., Jaffard, S. & Roux, D. Elliptic gaussian random processes. *Revista Matemática Iberoamericana* **13,** 19–90 (1997).

60. Lim, S. C. & Muniandy, S. V. On some possible generalizations of fractional Brownian motion. *Physics Letters, Section A: General, Atomic and Solid State Physics* **266,** 140–145 (2000).

61. Cohen, S. in *Fractals: Theory and Applications in Engineering* (eds Dekking, M., Lévy Véhel, J., Lutton, E. & Tricot, C.) 3–16 (Springer London, 1999). ISBN: 978-1-4471-0873-3. doi:10.1007/978-1-4471-0873-3_1.

62. Brockwell, P. J. & Davis, R. *Time series: theory and methods* (Springer-Verlag, New York [etc.], 1991).

63. Dieker, A. B. & Mandjes, M. On spectral simulation of fractional Brownian motion. *Probability in the Engineering and Informational Sciences* **17,** 417–434 (2003).

64. Dieker, T. *Simulation of fractioinal Brownian motion* PhD thesis (2004).

65. Bishop, C. M. *Pattern recognition and machine learning* (Springer, New York, 2006).

66. Verdaasdonk, J. S. *et al.* Centromere tethering confines chromosome domains. *Molecular Cell* **52,** 819–831. ISSN: 10972765 (2013).

67. Dieterich, P., Klages, R., Preuss, R. & Schwab, A. Anomalous dynamics of cell migration. *PNAS* **105,** 459–63. ISSN: 1091-6490 (Jan. 2008).

68. Kenwright, D. A., Harrison, A. W., Waigh, T. A., Woodman, P. G. & Allan, V. J. First-passage-probability analysis of active transport in live cells. *Physical Review E* **86,** 031910. ISSN: 1539-3755 (Sept. 2012).

69. Bronstein, I. *et al.* Transient Anomalous Diffusion of Telomeres in the Nucleus of Mammalian Cells. *Physical Review Letters* **103,** 018102. ISSN: 0031-9007 (July 2009).

70. Mak, M., Kamm, R. D. & Zaman, M. H. Impact of Dimensionality and Network Disruption on Microrheology of Cancer Cells in 3D Environments. *PLoS Computational Biology* **10,** e1003959. ISSN: 1553-7358 (2014).

71. Gal, N., Lechtman-Goldstein, D. & Weihs, D. Particle tracking in living cells: a review of the mean square displacement method and beyond. *Rheologica Acta* **52,** 425–443. ISSN: 0035-4511 (2013).

72. Gardel, M. L., Valentine, M. T., Crocker, J. C., Bausch, A. R. & Weitz, D. A. Microrheology of entangled F-actin solutions. *Phys Rev Lett* **91,** 158302 (2003).

73. Fong, E. J. *et al.* Decoupling directed and passive motion in dynamic systems: particle tracking microrheology of sputum. *Annals of Biomedical Engineering* **41,** 837–846 (2013).

74. Houghton, H. A., Hasnain, I. A. & Donald, A. M. Particle tracking to reveal gelation of hectorite dispersions. *European Physical Journal E* **25,** 119–127 (2008).

75. Qian, H., Sheetz, M. P. & Elson, E. L. Single particle tracking. Analysis of diffusion and flow in two-dimensional systems. *Biophys J* **60,** 910–21 (1991).

76. Bertseva, E. *et al.* Optical trapping microrheology in cultured human cells. *Eur Phys J E Soft Matter* **35,** 63 (2012).

77. Crocker, J. C. & Grier, D. G. Microscopic measurement of the pair interaction potential of charge-stabilized colloid. *Physical Review Letters* **73,** 352–355. ISSN: 00319007 (1994).

78. Tassieri, M., Evans, R., Warren, R., Bailey, N. & Cooper, J. Microrheology with optical tweezers: data analysis. *New Journal of Physics* **115032,** 115032. ISSN: 1367-2630 (2012).

79. Hohenegger, C. & Forest, M. G. Two-bead microrheology: modeling protocols. eng. *Phys Rev E Stat Nonlin Soft Matter Phys* **78,** 31501 (2008).

80. Crocker, J. C. & Hoffman, B. D. Multiple-particle tracking and two-point microrheology in cells. *Methods Cell Biol* **83,** 141–178 (2007).

81. Levine, A. J. & Lubensky, T. C. Two-point microrheology and the electrostatic analogy. *Phys Rev E Stat Nonlin Soft Matter Phys* **65,** 11501 (2002).

82. Crocker, J. C. *et al.* Two-point microrheology of inhomogeneous soft materials. *Phys Rev Lett* **85,** 888–891 (2000).

83. Wang, Y.-Y. *et al.* The microstructure and bulk rheology of human cervicovaginal mucus are remarkably resistant to changes in pH. *Biomacromolecules* **14,** 4429–35 (2013).

84. Das Neves, J., Nunes, R., Machado, A. & Sarmento, B. Polymer-based nanocarriers for vaginal drug delivery. *Advanced Drug Delivery Reviews.* ISSN: 0169409X (2014).

85. Georgiades, P., Pudney, P. D. A., Thornton, D. J. & Waigh, T. A. Particle tracking microrheology of purified gastrointestinal mucins. *Biopolymers* **101,** 366–77. ISSN: 0006-3525 (2014).

86. Georgiades, P., Pudney, P. D. A., Rogers, S., Thornton, D. J. & Waigh, T. A. Tea derived galloylated polyphenols cross-link purified gastrointestinal mucins. *PloS one* **9,** e105302. ISSN: 1932-6203 (2014).

87. Macierzanka, A. *et al.* Transport of Particles in Intestinal Mucus under Simulated Infant and Adult Physiological Conditions: Impact of Mucus Structure and Extracellular DNA. *PloS one* **9,** e95274. ISSN: 1932-6203 (2014).

88. Dangaria, J. H. & Butler, P. J. Macrorheology and adaptive microrheology of endothelial cells subjected to fluid shear stress. eng. *Am J Physiol Cell Physiol* **293,** C1568–75. ISSN: 15378276 (2007).

89. Dangaria, J. H., Yang, S. & Butler, P. J. Improved nanometer-scale particle tracking in optical microscopy using microfabricated fiduciary posts. *Biotechniques* **42,** 437–440 (2007).

90. Hasnain, I. A. & Donald, A. M. Microrheological characterization of anisotropic materials. *Phys Rev E Stat Nonlin Soft Matter Phys* **73,** 31901 (2006).

91. Michalet, X. Mean square displacement analysis of single-particle trajectories with localization error: Brownian motion in an isotropic medium. *Phys Rev E Stat Nonlin Soft Matter Phys* **82,** 41914 (2010).

92. Weihs, D., Teitell, M. A. & Mason, T. G. Simulations of complex particle transport in heterogeneous active liquids. *Microfluid and Nanofluid* **3,** 227–237. ISSN: 1613-4982 (Nov. 2007).

93. Kou, S. C., Xie, X. S. & Liu, J. S. Markov chain Monte Carlo in the analysis of single-molecule experimental data. *AIP Conference Proceedings* **690,** 123–133. ISSN: 0094-243X (2003).

94. Kou, S. C., Olding, B. P., Lysy, M. & Liu, J. S. A Multiresolution Method for Parameter Estimation of Diffusion Processes. *Journal of the American Statistical Association* **107,** 1558–1574. ISSN: 0162-1459 (2012).

95. Seisenberger, G. *et al.* Real-Time Single-Molecule Imaging of the Infection Pathway of an Adeno-Associated Virus. *Science* **294,** 1929–1933 (2001).

96. Caspi, A., Granek, R. & Elbaum, M. Enhanced Diffusion in Active Intracellular Transport. *Physical Review Letters* **85,** 5655–5658. ISSN: 0031-9007 (Dec. 2000).

97. Bareiss, E. H. Numerical solution of linear equations with Toeplitz and vector Toeplitz matrices. *Numerische Mathematik* **13,** 404–424. ISSN: 0945-3245 (1969).

98. Ljung, L. *System Identification: Theory for the User* 278–280 (Prentice-Hall, 1987).

99. Durbin, J. The fitting of time-series models. *Review of the International Statistical Institute* **28,** 233–244 (1960).

100. Rubinstein, M. & Colby, R. H. *Polymer Physics* (Oxford, 2003).

101. Cai, L.-H., Panyukov, S. & Rubinstein, M. Mobility of Nonsticky Nanoparticles in Polymer Liquids. *Macromolecules* **44,** 7853–7863 (2011).

102. MacKintosh, F. C. Microrheology in soft materials. *Abstracts of Papers of the American Chemical Society* **216,** U661–U661 (1998).

103. Lai, S. K., Wang, Y.-Y., Wirtz, D. & Hanes, J. Micro-and macrorheology of mucus. *Adv Drug Delivery Rev* **61,** 86–100 (2009).

104. Min, W., Luo, G., Cherayil, B. J., Kou, S. C. & Xie, X. S. Observation of a power-law memory kernel for fluctuations within a single protein molecule. *Phys Rev Lett* **94,** 198302 (2005).

105. Meyer, A., Marshall, A., Bush, B. G. & Furst, E. M. Laser tweezer microrheology of a colloidal suspension. *J Rheol* **50,** 77–92 (2006).

106. Lele, P. P., Swan, J. W., Brady, J. F., Wagner, N. J. & Furst, E. M. Colloidal diffusion and hydrodynamic screening near boundaries. *Soft Matter* **7,** 6844–6852 (2011).

107. McKinley, S. A., Yao, L. & Forest, M. G. Transient anomalous diffusion of tracer particles in soft matter. *J Rheol* **53,** 1487–1506 (2009).

108. Montiel, D., Cang, H. & Yang, H. Quantitative characterization of changes in dynamical behavior for single-particle tracking studies. *J Phys Chem B* **110,** 19763–19770 (2006).

109. Helmuth, J. A., Burckhardt, C. J., Koumoutsakos, P., Greber, U. F. & Sbalzarini, I. F. A novel supervised trajectory segmentation algorithm identifies distinct types of human adenovirus motion in host cells. *J Struct Biol* **159,** 347–58 (2007).

110. Duits, M. H., Li, Y., Vanapalli, S. A. & Mugele, F. Mapping of spatiotemporal heterogeneous particle dynamics in living cells. *Physical Review E Stat Nonlin Soft Matter Phys* **79,** 051910 (2009).

111. De Bruyn, J. R. & Oppong, F. K. Microrheology and dynamics of an associative polymer. *Eur Phys J E Soft Matter* **31,** 25–35 (2010).

112. Rich, J. P., McKinley, G. H. & Doyle, P. S. Size dependence of microprobe dynamics during gelation of a discotic colloidal clay. *Journal of Rheology* **55,** 273–299 (2011).

113. Kegel, W. K. & van Blaaderen, A. Direct observation of dynamical heterogeneities in colloidal hard-sphere suspensions. *Science* **287,** 290–293 (2000).

114. Penaloza, D. P., Hori, K., Shundo, A. & Tanaka, K. Spatial heterogeneity in a lyotropic liquid crystal with hexagonal phase. *Physical Chemistry Chemical Physics* **14,** 5247–5250 (2012).

115. Dempster, A. P., Laird, N. M. & Rubin, D. B. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **39,** 1–38 (1977).

116. Jain, A. K., Murty, M. N. & Flynn, P. J. Data Clustering: A Review. *ACM Comput. Surv.* **31,** 264–323. ISSN: 0360-0300 (Sept. 1999).

117. Weber, S. C., Spakowitz, A. J. & Theriot, J. A. Bacterial Chromosomal Loci Move Subdiffusively through a Viscoelastic Cytoplasm. *Physical Review Letters* **104,** 238102 (2010).

118. Schütz, G. J., Schindler, H. & Schmidt, T. Single-molecule microscopy on model membranes reveals anomalous diffusion. *Biophys J* **73,** 1073–1080 (1997).

119. Weiss, M., Hashimoto, H. & Nilsson, T. Anomalous protein diffusion in living cells as seen by fluorescence correlation spectroscopy. *Biophys J* **84,** 4043–4052 (2003).

120. Wachsmuth, M., Waldeck, W. & Langowski, J. Anomalous diffusion of fluorescent probes inside living cell nuclei investigated by spatially-resolved fluorescence correlation spectroscopy. *J Mol Biol* **298,** 677–689 (2000).

121. Oelschlaeger, P. Outsmarting metallo-beta-lactamases by mimicking their natural evolution. *J Inorg Biochem* **102.** Oelschlaeger, Peter J Inorg Biochem. 2008 Dec;102(12):2043-51. doi: 10.1016/j.jinorgbio.2008.05.007. Epub 2008 May 28., 2043–2051 (2008).

122. Van Hove, L. Correlations in Space and Time and Born Approximation Scattering in Systems of Interacting Particles. *Phys Rev* **95,** 249–262 (1954).

123. Rahman, A. Correlations in the motion of atoms in liquid argon. *Phys Rev* **136,** A405 (1964).

124. Savin, T. & Doyle, P. S. Statistical and sampling issues when using multiple particle tracking. *Physical Review E* **76.** PRE, 021501 (2007).

125. Tseng, Y., Kole, T. & Wirtz, D. Micromechanical mapping of live cells by multiple-particle-tracking microrheology. *Biophys J* **83,** 3162–3176 (2002).

126. Xu, J., Viasnoff, V. & Wirtz, D. Compliance of actin filament networks measured by particle-tracking microrheology and diffusing wave spectroscopy. *Rheologica Acta.* **37,** 387–398 (1998).

127. Huet, S. *et al.* Analysis of transient behavior in complex trajectories: application to secretory vesicle dynamics. *Biophys J* **91,** 3542–3559 (2006).

128. Meilhac, N., Le Guyader, L., Salome, L. & Destainville, N. Detection of confinement and jumps in single-molecule membrane trajectories. *Physical Review E* **73,** 011915 (2006).

129. Pinaud, F. *et al.* Dynamic partitioning of a glycosyl-phosphatidylinositol-anchored protein in glycosphingolipid-rich microdomains imaged by single-quantum dot tracking. *Traffic* **10,** 691–712 (2009).

130. De Smedt, S. C. *et al.* Structural information on hyaluronic acid solutions as studied by probe diffusion experiments. *Macromolecules* **27,** 141–146 (1994).

131. Hill, D. B. & Button, B. in *Mucins* 245–258 (Springer, 2012).

132. Hastie, T., Tibshirani, R. & Friedman, J. *The Elements of Statistical Learning* 2nd, 745 (Springer, New York, 2009).

133. Tibshirani, R., Walther, G. & Hastie, T. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **63,** 411–423 (2001).

134. MacQueen, J. Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley symposium on mathematical statistics and probability* **I,** 281–297 (1967).

135. Velmurugan, T. & Santhanam, T. Computational complexity between K-means and K-medoids clustering algorithms for normal and uniform distributions of data points. *J. Comput. Sci. Journal of Computer Science* **6,** 363–368 (2010).

136. Ward Jr, J. H. Hierarchical grouping to optimize an objective function. *J Am Statist Assoc* **58,** 236–244 (1963).

137. Johnson, S. C. Hierarchical clustering schemes. *Psychometrika* **32,** 241–254 (1967).

138. Mohajer, M., Englmeier, K.-H. & Schmid, V. J. A comparison of Gap statistic definitions with and with-out logarithm function. *Department of Statistics: Technical Reports* **96,** arXiv:1103.4767.

139. Savin, T. & Doyle, P. S. Static and dynamic errors in particle tracking microrheology. *Biophys J* **88,** 623–38 (2005).

140. Khanvilkar, K, Donovan, M. D. & Flanagan, D. R. Drug transfer through mucus. *Advanced drug delivery reviews* **48,** 173–93. ISSN: 0169-409X (June 2001).

141. Klonoff, D. C. Afrezza Inhaled Insulin : The Fastest-Acting FDA-Approved Insulin on the Market Has Favorable Properties. *Journal of Diabetes Science and Technology* **8,** 2014–2016 (2015).

142. Shukair, S. A. *et al.* Human cervicovaginal mucus contains an activity that hinders HIV-1 movement. *Mucosal Immunology* **6,** 427–434. ISSN: 1933-0219 (2012).

143. McKinley, S. A. *et al.* Modeling neutralization kinetics of HIV by broadly neutralizing monoclonal antibodies in genital secretions coating the cervicovaginal mucosa. *PLoS ONE* **9.** ISSN: 19326203. doi:10.1371/journal.pone.0100598 (2014).

144. Chen, A. *et al.* Transient Antibody-Mucin Interactions Produce a Dynamic Molecular Shield against Viral Invasion. *Biophysical journal* **106,** 2028–36. ISSN: 1542-0086 (May 2014).

145. Suk, J. S. *et al.* Rapid transport of muco-inert nanoparticles in cystic fibrosis sputum treated with N-acetyl cysteine. *Nanomedicine (London, England)* **6,** 365–375. ISSN: 1748-6963 (2011).

146. Wang, Y. Y. *et al.* Mucoadhesive nanoparticles may disrupt the protective human mucus barrier by altering its microstructure. *PLoS ONE* **6,** 1–7. ISSN: 19326203 (2011).

147. O'Malley, D., Cushman, J. H. & Johnson, G. Scaling laws for fractional Brownian motion with power-law clock. *Journal of Statistical Mechanics: Theory and Experiment,* L01001 (2011).

148. Davis, P. B. Cystic fibrosis since 1938. *American Journal of Respiratory and Critical Care Medicine* **173,** 475–482. ISSN: 1073449X (2006).

149. Tomaiuolo, G. *et al.* A new method to improve the clinical evaluation of cystic fibrosis patients by mucus viscoelastic properties. *PloS one* **9,** e82297. ISSN: 1932-6203 (Jan. 2014).

150. Waters, V. *et al.* Stenotrophomonas maltophilia in cystic fibrosis: Serologic response and effect on lung disease. *American Journal of Respiratory and Critical Care Medicine* **183,** 635–640. ISSN: 1073449X (2011).

151. Harrison, F. Microbial ecology of the cystic fibrosis lung. *Microbiology* **153,** 917–923. ISSN: 13500872 (2007).

152. Gilligan, P. H. Microbiology of airway disease in patients with cystic fibrosis. *Clinical microbiology reviews* **4,** 35–51. ISSN: 0893-8512 (1991).

153. Joo, N. S., Cho, H. J., Khansaheb, M. & Wine, J. J. Hyposecretion of fluid from tracheal submucosal glands of CFTR-deficient pigs. *Journal of Clinical Investigation* **120,** 3161–3166. ISSN: 00219738 (2010).

154. Quinton, P. M. Cystic fibrosis: impaired bicarbonate secretion and mucoviscidosis. *The Lancet* **372,** 415–417. ISSN: 01406736 (2008).

155. Teichgräber, V. *et al.* Ceramide accumulation mediates inflammation, cell death and infection susceptibility in cystic fibrosis. *Nature medicine* **14,** 382–391. ISSN: 1078-8956 (2008).

156. Smith, J. J., Travis, S. M., Greenberg, E. P. & Welsh, M. J. Cystic fibrosis airway epithelia fail to kill bacteria because of abnormal airway surface fluid. *Cell* **85,** 229–236. ISSN: 00928674 (1996).

157. Haggie, P. M. & Verkman, A. S. Defective organellar acidification as a cause of cystic fibrosis lung disease: reexamination of a recurring hypothesis. *American journal of physiology. Lung cellular and molecular physiology* **296,** L859–L867. ISSN: 1040-0605 (2009).

158. Tarran, R. *et al.* Normal and Cystic Fibrosis Airway Surface Liquid Homeostasis. *Journal of Biological Chemistry* **280,** 35751–35759 (2005).

159. Boucher, R. C. New concepts of the pathogenesis of cystic fibrosis lung disease. *European Respiratory Journal* **23,** 146–158. ISSN: 09031936 (2004).

160. Matsui, H. *et al.* Evidence for Periciliary Liquid Layer Depletion, Not Abnormal Ion Composition, in the Pathogenesis of Cystic Fibrosis Airways Disease. *Cell* **95,** 1005–1015 (1998).

161. Verkman, A., Song, Y. & Thiagarajah, J. Role of airway surface liquid and submucosal glands in cystic fibrosis lung disease. *Am J Physiol Cell Physiol* **284,** C2–C15 (2003).

162. Fahy, J. V. & Dickey, B. F. Airway mucus function and dysfunction. *The New England journal of medicine* **363,** 2233–47. ISSN: 1533-4406 (Dec. 2010).

163. Tarran, R, Grubb, B. R., Gatzy, J. T., Davis, C. W. & Boucher, R. C. The relative roles of passive surface forces and active ion transport in the modulation of airway surface liquid volume and composition. *The Journal of general physiology* **118,** 223–236. ISSN: 00221295 (2001).

164. Clunes, M. T. & Boucher, R. C. Cystic fibrosis: the mechanisms of pathogenesis of an inherited lung disorder. *Drug Discov Today Dis Mech* **4,** 63–72. ISSN: 15378276 (2007).

165. Davies, J. R., Kirkham, S., Svitacheva, N., Thornton, D. J. & Carlstedt, I. MUC16 is produced in tracheal surface epithelium and submucosal glands and is present in secretions from normal human airway and cultured bronchial epithelial cells. *The international journal of biochemistry & cell biology* **39,** 1943–54. ISSN: 1357-2725 (Jan. 2007).

166. Kirkham, S., Sheehan, J. K., Knight, D., Richardson, P. S. & Thornton, D. J. Heterogeneity of airways mucus: variations in the amounts and glycoforms of the major oligomeric mucins MUC5AC and MUC5B. *Biochem, J.* **361,** 537–546 (2002).

167. Button, B. & Boucher, R. C. Role of mechanical stress in regulating airway surface hydration and mucus clearance rates. *Respiratory Physiology and Neurobiology* **163,** 189–201. ISSN: 15699048 (2008).

168. Mott, L. S. *et al.* Progression of early structural lung disease in young children with cystic fibrosis assessed using CT. *Thorax* **67,** 509–516. ISSN: 0040-6376 (2012).

169. Matsui, H. *et al.* A physical linkage between cystic fibrosis airway surface dehydration and Pseudomonas aeruginosa biofilms. *Proceedings of the National Academy of Sciences of the United States of America* **103,** 18131–6. ISSN: 0027-8424 (Nov. 2006).

170. Matsui, H. *et al.* Reduced Three-Dimensional Motility in Dehydrated Airway Mucus Prevents Neutrophil Capture and Killing Bacteria on Airway Epithelial Surfaces. *The Journal of Immunology* **175,** 1090–1099. ISSN: 0022-1767 (July 2005).

171. Randell, S. H. and Fulcher, M. L. in *Methods in molecular biology* 2nd, 469 (Humana Press; Springer, New York, 2012).

172. *Flexible Bronchoscopy* 3rd ed. (eds Wang, K.-P., Mehta, A. C. & Turner, J. F.) ISBN: 9781444346428. doi:10.1002/9781444346428 (Blackwell Science, Ann Arbor, MI, 2004).

173. Babb, J. *et al.* British Thoracic Society guidelines on diagnostic flexible bronchoscopy. *Thorax* **56,** i1–i21. ISSN: 00406376 (2001).